Efficient Inversion of Multi-frequency and Multi-source Electromagnetic Data

Final report
15 August, 2007– 14 February, 2011

Gary D. Egbert

College of Earth Oceanic and Atmospheric Sciences
Oregon State University
CEOAS Adm in Bldg 104
Corvallis, OR 97331-5503

Prepared for the
U.S. Department of Energy
Office of Basic Energy Sciences
Under grant DE-FG02-06ER15819

Summary

BES grant DE-FG02-06ER15819 supported efforts at Oregon State University (OSU) to develop improved inversion methods for 3D subsurface electromagnetic (EM) imaging. Three interrelated activities have been supported by this grant, and its predecessor (DE-FG02-06ER15818): (1) collaboration with a former student of the PI, Dr. Weerachai Siripunvaraporn, who is now Professor at Mahidol University in Bangkok, Thailand (Siripunvaraporn and Egbert, 2007; 2009). (2) Development at OSU of a new modular system of computer codes for EM inversion (Egbert and Kelbert, 2012; Egbert et al., 2013), and initial testing and application of this inversion on several large field data sets (Patro and Egbert, 2008; 2011; Kelbert et al., 2012; Meqbel et al., 2013). (3) Research on more efficient approaches to EM inverse problems, exploiting special features of the multi-transmitter problems that are common in EM imaging applications (Egbert, 2012). The last of these activities was the main motivation for this research project. The first two activities were important enabling steps, and produced useful products and results in their own right. In the following we provide brief summaries of these three activities, and results from each; further technical details are contained in the cited references, which are attached.

The project provided partial support for three post-doctoral scholars, who worked with the PI on various aspects of EM inversion, either in terms of development of code or theory, or in applications and testing: Dr. Prasanta Patro (2007-2008); Dr. Anna Kelbert (2008-2011); and Dr. Naser Meqbel (2010-2011). Project funding also helped the PI to support and interact with several visitors (who brought most or all of their own funding). These include Dr. Aihua Wang, an assistant Professor from Jilin University in China, who visited for one year (2009-2010); extended visits from two PhD students from Thailand (both now graduated: Dr. Weerachai Sarakorn, and Dr. Chatchai Vachiratienchai), as well as shorter (~6 week) visits from PhD students working with Dr. Oliver Ritter at GFZ-Potsdam in Germany (Dr. Kristina Tietze, and Dr. Xiao-Ming Chen).

(1) Collaboration with W. Siripunvaraporn

This collaborative activity supported our initial efforts on 3D inversion, resulting in a total of six publications over this project and its predecessor. During the first project (DE-FG02-06ER15818) collaboration with Dr. Siripunvaraporn resulted in development, and release to the academic EM community, of the first freely available 3D inversion code for magnetotelluric (MT) data, WSINV3DMT (http://mucc.mahidol.ac.th/~scwsp/wsinv3dmt/). During the subsequent project period (grant DE-FG02-06ER15819, covered by this report) we continued this collaboration, completing our joint work on development of new approaches to EM inversion (i.e., first steps towards activity 3; Siripunvaraporn and Egbert, 2007) and adding new capabilities to WSINV3DMT (parallelization, inversion for vertical field TFs; (Siripunvaraporn and Egbert, 2009). The OSU PI also hosted two PhD students from Mahidol University, for work on projects related to their dissertations. Mr. Weerachai Sarakorn visited for approximately one year (2008-9), working on 3D finite element modeling for EM geophysics. Mr. Chatchai Vachiratienchai spent 7 months at OSU (Dec. 2010-June 2011), working on controlled source EM inversion, using the ModEM system described below. Both of these students have subsequently completed the PhD degree program, and are working in Thailand.

(2) Development of ModEM

Our ultimate goal of exploring more efficient search algorithms for multi-transmitter EM inverse problems (activity 3) motivated our development of a fully modular system of computer codes for EM inversion, which we call ModEM (Egbert and Kelbert, 2012). We focused first on a simple (two-dimensional magnetotelluric; MT) problem as a specific example, but developed the code using an object oriented approach, independent of details of this specific problem. The top level of modules implements gradient calculations, and allows straightforward implementation of a range of specific inversion algorithms, including standard Gauss-Newton and mathematical optimization (e.g., conjugate gradients; quasi-Newton) schemes which have been widely applied in this field, as well as more novel schemes, as discussed in the next section. These calculations are implemented in an abstract way, to simplify generalization to treat a wide range of EM inverse problems (e.g., with different sources/receivers or data types; different model parameterization or regularization; different modeling schemes; different search algorithms). Our next focus was to develop capabilities for a general class of 3D EM inverse problems, based on a finite element forward solver. Again, the initial application focus was on MT. We next parallelized ModEM (over forward problems, using MPI). This effort was begun in collaboration with Naser Meqbel, then a PhD student at GFZ-Potsdam, and then completed with support of this project as a post-doc at OSU. The parallelization scheme was again developed in a generic manner, independent of details of the inversion algorithm, or the specific EM technique, using the Message Passing Interface (MPI) library. Extension of ModEM to frequency domain controlled source EM (CSEM) methods for land (with visiting Prof. Aihua Weng), and for marine CSEM (Dr. Chatchai Vachiratienchai) was then pursued, although these capabilities have not been applied to real datasets yet.

ModEM is now a mature parallel 3D inversion code for MT data, which we are distributing freely for academic use, and licensing for commercial applications. Funding has recently been obtained from NSF to help us support maintenance and further development of the 3D MT code for academic use. We are also continuing collaboration with Dr. N. Meqbel, now back at GFZ-Potsdam, on development of ModEM for controlled source, and joint CSEM-MT-DC inversion methods, with applications to real datasets now. One paper describing the theory underlying ModEM has been published (Egbert and Kelbert, 2012), and a second describing implementation details will be submitted soon (Egbert et al., 2013).

As part of our development and testing effort we have worked extensively throughout this project with real MT datasets. Papers describing this work, and citing support from this grant, include (Patro and Egbert (2008; 2011); Kelbert (2012) and Meqbel et al. (2013).

(3) Progress on new inversion approaches for multi-transmitter EM geophysical data

This activity was the main project focus, and is more novel, so we provide a more detailed description of key ideas here. A full technical development is provided in Egbert (2012). Our focus has been on regularized inversion of EM data, which is accomplished by minimizing a penalty functional such as

$$\mathcal{P}(\mathbf{m},\mathbf{d}) = \left\|\mathbf{f}(\mathbf{m}) - \mathbf{d}\right\|^2 + \lambda \left\|\mathbf{m} - \mathbf{m}_0\right\|^2 \tag{1}$$

where the first term represents data misfit, and the second a regularization term enforcing smoothness, proximity to a "prior" solution, etc. The tradeoff parameter λ is generally required to adjust the relative weighting of the data misfit and model regularity terms. All practical 3D EM inversion involves linearizing the non-linear data mapping $\mathbf{f}(\mathbf{m})$. Two general schemes based on this linearization have been applied: Gauss-Newton (GN), which is based on a second order Taylor series approximation to the penalty functional $\mathcal{P}$, and non-linear optimization algorithms based on generating a series of conjugate search directions, and then minimizing $\mathcal{P}$ with a line search along each successive direction. Algorithms of the second class (e.g., non-linear conjugate gradients (NLCG), quasi-Newton) have become the standard approach for large scale 3D EM inversion (e.g., Comer and Newman, 2009, and many previous works). This is largely because the simplest application of GN requires (i) solving the equivalent of one forward problem for each observation to obtain complete information about the linearized model parameter-data mapping (the Jacobian), and (ii) forming, and then solving, a very large dense system of normal equations ($M \times M$, where M is the number of model parameters). The first task (solving a 3D partial differential equation (PDE) thousands of times for each linearization) is computationally challenging, and the second is virtually impossible for realistic 3D model parameterizations, where M may easily exceed 10^6. Both of these complications are avoided by NLCG and related direct optimization methods.

However, it has long been appreciated that there are variants on GN that are at least feasible. In our own previous work (Siripunvaraporn et al., 2005) a data-space variant on GN was shown to be practical, particularly if run in parallel with both computations and storage distributed over a small cluster (Siripunvaraporn and Egbert, 2009). This effort illustrates one of the advantages of GN over NLCG: using the so called "Occam" scheme an optimal value for the tradeoff parameter (λ) can be found at very little cost as part of the GN inversion process; in NLCG the entire optimization process must be repeated a number of times to optimize the relative weighting of data and model norm. Our work also showed that NLCG and GN require a similar number of forward model solutions (Siripunvaraporn and Egbert, 2007). A key observation is that each NLCG step (and each forward model solution in the line search) requires solving the forward problem once for each frequency/transmitter—so with 20 or so frequencies, 50 minimization steps, and 4-5 forward solves per line search, NLCG also requires many thousands of forward solutions, comparable to GN.

Another approach to GN has also long been known: the $M \times M$ system of normal equations can be solved iteratively using conjugate gradients (CG) without actually computing all of the data sensitivities, and forming the normal equation matrices. This

GN variant requires about the same number of total forward solves as a more direct approach (Siripunvaraporn and Egbert, 2007), but eliminates storing and solving the large dense matrix. Unfortunately, with the simple application of CG that has generally been used in the past, the Occam scheme, which optimizes λ, cannot be used. The first (rather simple) innovation discussed in Egbert (2012) is a hybrid scheme that allows the Occam approach to be used with the CG variant on GN. The basic idea is very simple: CG effectively solves the normal equations by building up a low-dimensional approximation to the full sensitivity matrix, factored as a product of an orthogonal matrix and a bi-diagonal matrix. The standard implementation of CG is super-memory efficient: rather than store these matrices, an approximate solution to the normal equations (for a fixed value of λ) is computed “on the fly”. By saving the matrix factors (which have been computed at great cost, involving as many as a thousand forward solver calls) the normal equations can be solved (again approximately) for *any* value of the damping parameter, allowing application of Occam schemes, and efficient optimization of this parameter. The insight that CG methods effectively generate the most important parts of the sensitivity matrix suggests further possible computational efficiencies, as we discuss below.

The more significant innovation discussed in Egbert (2012) carries the hybrid CG/Occam idea one step further. For EM problems with multiple frequencies (e.g., MT) or multiple transmitters (e.g., controlled source EM, with a towed transmitter—as used for marine applications) each CG step requires solving the forward problem separately for each of the *K* frequencies/transmitters. Each of these calculations results in an independent sensitivity for a linear combination of the data components recorded for one transmitter. The standard CG scheme simply sums these sensitivity components, throwing away valuable information about the individual components, and hence the Jacobian. By retaining the separate components for each frequency/transmitter an accurate approximation to the Jacobian can be built up more rapidly than with the simpler hybrid algorithm based on standard CG. Although we have been pursing this general idea over the past several years, we have discovered only recently how to make this scheme work reliably. Very briefly, (see Egbert (2012) for further detail) the key is to actually do a full solution to the *linearized* inverse problem after *each* multi-component gradient calculation. This entails fitting the data (to within the linearized approximation) using all sensitivity components (saved from all previous iterations to that point, separately for each transmitter) to fit the projected data. The resulting model is then multiplied by the Jacobian to generate the next set of residuals, and the process continued. Each of the inverse solutions is computed in a relatively low-dimensional subspace, so these extra computations are actually not too onerous. Tests (so far limited to two-dimensional MT inverse problems) show that this scheme reduces the required number of forward solutions by a factor of 2-3, relative to all of the standard inversion approaches (GN/NLCG), and yet produces results identical to an Occam solution based on a complete calculation of the Jacobian.

We believe that the results of our research on methods will ultimately have a very significant impact on 3D EM inversion. Initial tests of the multi-transmitter hybrid scheme already show a factor of 2-3 in computational speedup. We still need to implement and test these ideas on other problems, but our intuition is that the advantage of the approach will be significantly greater for 3D MT (where some additional advantage can be gained from the fact that there are two polarizations, in addition to multiple frequencies), and in marine CSEM (where the number of transmitters and receivers is very large). There are also extensions that we have not yet considered. For example, the Occam inversion has an outer loop, which usually has to be executed 3-4 times. The approximated Jacobian computed in the hybrid scheme can effectively be used as a pre-conditioner for the next iteration, likely resulting in further speedup. The schemes we have developed are likely to be especially useful for joint inversion (e.g., MT and CSEM; EM and seismic). In the first place, multiple data types require running multiple forward models, and this can also be exploited within the multiple transmitter framework. Furthermore, experience inverting multiple data types demonstrates that multiple tradeoff parameters are required to allow for differential weighting of data types. And, one approach to joint inversion is to enforce structural similarity between two disparate physical parameters (e.g., conductivity and seismic velocity). This constraint is enforced by introducing another term into the penalty functional (1), with yet another adjustable weight. With the NLCG approach the full inversion must be run many times to choose optimal weights in a joint inversion. Efficient schemes for choosing these weights, as offered by the hybrid schemes we have developed, are thus likely to prove very valuable for joint inversion.

One might argue that even a factor of 4 decrease in inversion run time is dwarfed by the impact of Moore's law—the increase in computational power over the current project period has certainly exceeded a factor of 4. However, any advantages obtained with algorithmic efficiency can be multiplied by speedups obtained through developments in computing hardware. As long as the available computational resources fall short of requirements—and for 3D EM inversion they certainly do now, and will for the foreseeable future—speedups of the sort we are finding are certainly of practical value, and, we would argue, worth developing further.

Peer Reviewed Publications citing support from this Grant

Siripunvaraporn, W., and Egbert, G.D., 2007, Data space conjugate gradient inversion for 2-D magnetotelluric Data, *Geophys., J. Int*, 170, 986-994.

Siripunvaraporn, W., and G. Egbert, 2009, WSINV3DMT: Vertical magnetic field transfer function inversion and parallel implementation, *Phys. Earth. Planet. Inter.*, 317-329.

Patro P. K., G. D. Egbert, 2008. Regional conductivity structure of Cascadia: Preliminary results from 3D inversion of USArray transportable array magnetotelluric data, *Geophys. Res. Lett.,* 35, L20311, doi:10.1029/2008GL035326.

Patro, P. K., and G. D. Egbert, 2011. Application of 3D Inversion to Magnetotelluric profile data from the Deccan volcanic province of western India, P*hys, Earth, Planet. Int., 33-46* DOI: 10.1016/j.pepi.2011.04.005as.

Egbert, G. D. and A. Kelbert, 2012. Computational Recipes for Electromagnetic Inverse Problems, *Geophys. J. Int*. 189: 251–267. doi: 10.1111/j.1365-246X.2011.05347.x

Egbert, G. D. 2012, Hybrid conjugate gradient-Occam algorithms for inversion of multifrequency and multitransmitter EM data. *Geophys J. Int.* 190, 255-266, doi: 10.1111/j.1365-246X.2012.05523.x

Kelbert, A., G. D. Egbert and C. deGroot-Hedlin, 2012. Crust and upper mantle electrical conductivity beneath the Yellowstone Hotspot Track, *Geology,* 40, 447-450, doi: 10.1130/G32655.1.

Papers in preparation citing support from this grant

Egbert, G. D., A. Kelbert and N. M. Meqbel, 2013. Computational Recipes for EM Inverse Problems: I. Modular Implementation. To be submitted to *Comput. Geosci*.

Meqbel, N., Egbert G. D., Wannamaker, P. Kelbert, A., and A. Schultz, 2013. Deep electrical resistivity structure of the Pacific Northwestern U. S. derived from 3-D inversion of USArray Magnetotelluric data. To be submitted to *Earth Planet. Sci. Lett.*

doi: 10.1111/j.1365-246X.2007.03478.x

Data space conjugate gradient inversion for 2-D magnetotelluric data

Weerachai Siripunvaraporn[1] and Gary Egbert[2]

[1]*Department of Physics, Faculty of Science, Mahidol University, Rama VI Rd., Rachatawee, Bangkok* 10400*, Thailand. E-mail: scwsp@mahidol.ac.th*
[2]*College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis,* OR 97331*, USA*

Accepted 2007 April 25. Received 2007 April 25; in original form 2006 December 18

SUMMARY

A data space approach to magnetotelluric (MT) inversion reduces the size of the system of equations that must be solved from $M \times M$, as required for a model space approach, to only $N \times N$, where M is the number of model parameter and N is the number of data. This reduction makes 3-D MT inversion on a personal computer possible for modest values of M and N. However, the need to store the $N \times M$ sensitivity matrix **J** remains a serious limitation. Here, we consider application of conjugate gradient (CG) methods to solve the system of data space Gauss–Newton equations. With this approach **J** is not explicitly formed and stored, but instead the product of **J** with an arbitrary vector is computed by solving one forward problem. As a test of this data space conjugate gradient (DCG) algorithm, we consider the 2-D MT inverse problem. Computational efficiency is assessed and compared to the data space Occam's (DASOCC) inversion by counting the number of forward modelling calls. Experiments with synthetic data show that although DCG requires significantly less memory, it generally requires more forward problem solutions than a scheme such as DASOCC, which is based on a full computation of **J**.

Key words: data space method, inversion, magnetotellurics.

INTRODUCTION

Three-dimensional (3-D) magnetotelluric (MT) inversion can reveal the 3-D resistivity structure beneath the Earth's surface, and can be applied to 3-D data sets (e.g. Tuncer *et al.* 2006), as well as to 2-D profile data (Siripunvaraporn *et al.* 2005b). In recent years a number of 3-D MT inversion algorithms have been developed (e.g. Mackie & Madden 1993; Mackie, personal communication 2002; Newman & Alumbaugh 2000; Zhdanov *et al.* 2000; Sasaki 2001; Siripunvaraporn *et al.* 2004, 2005a). There are many similarities in the formulation of the inverse problem used by all of these authors—in all cases a data misfit/model roughness penalty functional is minimized—but a number of different computational approaches have been pursued. All approaches have pros and cons, as discussed in Siripunvaraporn *et al.* (2005a).

Newman & Alumbaugh (2000) and Mackie (personal communication 2002) used the non-linear conjugate gradient method to minimize a data misfit/model roughness penalty functional. Sasaki (2001) and Mackie & Madden (1993) both used a Gauss–Newton (GN) method, however in the latter case the system of normal equations was solved by the conjugate gradient method. Siripunvaraporn *et al.* (2004, 2005a) developed a 3-D inversion algorithm based on the Occam inversion of Constable *et al.* (1987), another variant of the GN method. In this work, the data space approach previously used for 2-D MT (Siripunvaraporn & Egbert 2000) was extended to the 3-D case. This transformation to the data space significantly reduced memory requirements, and making it possible to run 3-D MT inverse problems of modest size on a desktop PC. However, memory required to store the sensitivity matrix is still quite substantial, and this limits the size of both data sets and model parametrization. Here, we consider another possible approach, the 'data space conjugate gradient' (DCG) inversion. This is again a GN variant, formulated in the data space as in Siripunvaraporn & Egbert (2000), but without forming and storing the sensitivity matrix as in Mackie & Madden (1993).

We begin the paper by reviewing the Occam inversion, comparing model and data space approaches. We then introduce the DCG method, and test this using synthetic 2-D MT data set. In these tests we compare computational efficiency of DCG and previously described, proven MT inverse methods (Siripunvaraporn & Egbert 2000).

REVIEW OF OCCAM'S INVERSION

The data space Occam's (DASOCC) inversion has been successfully applied to 2-D (Siripunvaraporn & Egbert 2000) and 3-D (Siripunvaraporn *et al.* 2004, 2005a) magnetotelluric (MT) inversion. DASOCC follows the general Occam approach of Constable *et al.* (1987) to seek the 'minimum structure' model subject to an appropriate fit to the data. Mathematically, an unconstrained functional $\mathrm{U}(\mathbf{m}, \lambda)$ is varied :

$$\mathrm{U}(\mathbf{m}, \lambda) = (\mathbf{m} - \mathbf{m}_0)^{\mathrm{T}}\mathbf{C}_{\mathbf{m}}^{-1}(\mathbf{m} - \mathbf{m}_0) + \lambda^{-1}\left\{(\mathbf{d} - \mathbf{F}[\mathbf{m}])^{\mathrm{T}}\mathbf{C}_{\mathbf{d}}^{-1}(\mathbf{d} - \mathbf{F}[\mathbf{m}]) - X^{*2}\right\}, \qquad (1)$$

to minimize the model norm subject to the condition that the normalized squared total misfit is equal to $X^{*2.}$ Here $\mathbf{m}$ is the resistivity model of dimension M, $\mathbf{m_0}$ the prior model, $\mathbf{C_m}$ the model covariance matrix which defines the model norm, $\mathbf{d}$ the observed data with dimension N, $\mathbf{F}[\mathbf{m}]$ the forward model response, $\mathbf{C_d}$ the data covariance matrix, X^* the target misfit, and λ^{-1} a Lagrange multiplier.

The Occam scheme of Constable *et al.* (1987) is based on linearizing the forward response to obtain the following iterative sequence of linear equations (see Constable 1987; Siripunvaraporn & Egbert 2000),

$$\mathbf{m}_{k+1} - \mathbf{m}_0 = \left[\lambda \mathbf{C}_\mathbf{m}^{-1} + \mathbf{J}_k^{\mathrm{T}} \mathbf{C}_\mathbf{d}^{-1} \mathbf{J}_k\right]^{-1} \mathbf{J}_k^{\mathrm{T}} \mathbf{C}_d^{-1} \mathbf{đ}_k, \tag{2}$$

where the subscript k denotes iteration number, $\mathbf{J}_k = (\partial \mathbf{F}/\partial \mathbf{m})_k$ is the $N \times M$ sensitivity matrix calculated at $\mathbf{m}_\mathrm{k}$, and $\mathbf{đ}_k = \mathbf{d} - \mathbf{F}[\mathbf{m}_k] + \mathbf{J}_k(\mathbf{m}_k - \mathbf{m}_0)$. In (2) the dimension of the inverted matrix is $M \times M$, controlled by the size of the model space. For realistic 3-D problems M is usually very large, making application of this model space approach impractical.

To reach the ultimate goal of finding a stationary point of (1), in each iteration (2) is solved with a series of trial values of λ. In early iterations (Phase I), the Occam algorithm searches over λ for the model that minimizes misfit. The process continues until the target X^{*2} is attained. Once the misfit reaches the desired level, the next stage (Phase II) begins by keeping the misfit at the desired level, varying λ to seek the model of smallest norm achieving the target misfit. One advantage of Occam's inversion is that only a small number of iterations are required to converge to the solution.

Siripunvaraporn & Egbert (2000) transformed the Occam scheme for the 2-D MT problem from the model space to the data space, developing a variant of Occam in which the size of the inversion depends on the number of data N, instead of the number of model parameters M. See Parker (1994), Bennett *et al.* (1996) and Egbert (1997) for data space approaches to other inversion problems. In the data space approach, the series of iterative approximate solutions is obtained as

$$\mathbf{m}_{k+1} - \mathbf{m}_0 = \mathbf{C}_\mathbf{m} \mathrm{J}_k \left[\lambda \mathbf{C}_d + \mathbf{J}_k \mathbf{C}_\mathbf{m} \mathbf{J}_k^{\mathrm{T}}\right]^{-1} \mathbf{đ}_k^{\mathrm{T}}, \tag{3}$$

see Siripunvaraporn & Egbert (2000) and Siripunvaraporn *et al.* (2005) for details. The system of equation as given in (3) shows that the system of equations that must be solved for the inversion is in the data space, and thus of size $N \times N$. As in the model space Occam scheme, (3) is solved for a series of trial values of λ to search for the minimal misfit (Phase I) and then to minimize the model norm while keeping the misfit constant (Phase II). We refer to this 'data space' variant on Occam as DASOCC. Provided N is much less than M, DASOCC will be considerably more efficient than the original model space Occam. In particular, DASOCC allows an Occam type scheme to be used for 3-D inversion of MT data on a personal computer or workstation, as shown in Siripunvaraporn *et al.* (2004; 2005a). Pseudo-code for the DASOCC algorithm is given in Fig. 1.

Though the size of the system of equations that must be solved in the inversion can be significantly reduced with a data space approach, very significant computer memory is still required to store the $N \times M$ sensitivity matrix $\mathbf{J}_k$ for realistic values of N and M, particularly for 3-D MT problems. Furthermore, computation of the sensitivity matrix requires many forward model solutions. Here, we present an alternative approach that avoids storing the large matrix $\mathbf{J}_k$. Instead of forming and factoring the matrix ($\lambda\ \mathbf{C}_d + \mathbf{J}_k \mathbf{C}_\mathbf{m} \mathbf{J}_k^{\mathrm{T}}$) as in Siripunvaraporn & Egbert (2000) and Siripunvaraporn *et al.* (2004, 2005a), we apply a conjugate gradient (CG) technique to solve (3). With the CG method, there is no need to explicitly form the full $N \times M$ sensitivity matrix. Rather, only multiplication of the sensitivity matrix or its transpose with a given vector ($\mathbf{p}$ or $\mathbf{q}$) to form $\mathbf{J}_k\mathbf{p}$ or $\mathbf{J}_k^{\mathrm{T}}\mathbf{q}$ is required. Each of these matrix vector products in turn requires one forward model solution per period. A very similar approach has been used before in the model space EM inversion algorithms developed by Mackie & Madden (1993), Newman & Alumbaugh (1996), Rodi & Mackie (2001) and Haber *et al.* (2000) among

```
d = observed data, C_d = data error, m_0 = initial model, C_m = model covariance

Solve forward problem and compute misfit from model $\mathbf{m_0}$
Start DASOCC outer loop iteration $k$:
    For $i$ = 1 to $N_s*N_m*N_p$
        Call forward solver to form sensitivity for data $i$
    End
    Compute $\mathbf{đ}_\mathrm{k} = \mathbf{d} - \mathbf{F}[\mathbf{m}_\mathrm{k}] + \mathbf{J}_\mathrm{k}(\mathbf{m}_\mathrm{k} - \mathbf{m}_0)$
    Compute $\mathbf{\Gamma}_\mathrm{k} = \mathbf{C_d}^{-½}\mathbf{J}_\mathrm{k}\mathbf{C_m}\mathbf{J}_\mathrm{k}^{\mathrm{T}}\mathbf{C_d}^{-½}$
    For various values of λs
        Use $\mathbf{J}_\mathrm{k}$ to compute representer matrix $\mathbf{R}_\mathrm{k} = [\lambda\ \mathbf{I} + \Gamma_\mathrm{k}]$
        Use Cholesky decomposition to solve $\mathbf{m}_{\mathrm{k}+1}$ - $\mathbf{m_0} = \mathbf{C_m}\mathbf{J}_\mathrm{k}\mathbf{C_d}^{-½}\mathbf{R}_\mathrm{k}^{-1}\mathbf{C_d}^{-½}\mathbf{đ}_\mathrm{k}$
        Solve forward problem and Compute misfit from model $\mathbf{m}_{\mathrm{k}+1}$
        Phase I : Compare misfit from different  λs to seek for minimum misfit
        Phase II: Compare norm from different λs to seek minimum norm
    End
    Exit when misfit less than desired level with minimum norm
End DASOCC outer loop iteration
```

Figure 1. Pseudo-code for DASOCC.

others. Here, we describe and test DCG, a data space variant on this algorithm. Although the primary rationale for developing this limited memory scheme is to increase practicality of 3-D inversion, we report here initial tests and comparisons on synthetic 2-D MT. A key goal here is to compare computational efficiency of DCG and DASOCC, and these simpler tests are already instructive.

Note that an alternative approach to improving computational efficiency is the Reduced Basis Occam (REBOCC) approach of Siripunvaraporn & Egbert (2000). REBOCC is based on the observation that the updated inverse solution $\mathbf{m}_{k+1}$ of (3) is a linear combination of the N columns of $\mathrm{C}_\mathbf{m}\mathbf{J}_k$. In REBOCC sensitivites for a subset of K data are calculated (e.g. skipping every other frequency or every other site in a profile) and an approximate solution is sought as a linear combination of the corresponding K columns of $\mathrm{C}_\mathbf{m}\mathbf{J}_k$. The full data set is still fit, using this reduced set of basis functions. This scheme is more efficient than DASSOCC, particularly for MT data sets that are highly redundant, either in spatial or frequency sampling, To simplify our comparisons here we only consider the DASOCC and DCG schemes, and we restrict our comparisons to test data sets which are not heavily oversampled, for which only modest gains in efficiency would be achieved with REBOCC. Indeed, it is not obvious how, or even if, REBOCC might be usefully extended to make use of a subset of sites for general 3-D problems. Furthermore, as we shall see, DASOCC is generally already more efficient in terms of computational time than DCG, so there is little point to direct comparison of efficiency of DCG and REBOCC.

DATA SPACE CONJUGATE GRADIENT (DCG) METHOD

With the DASOCC approach, eq. (3) is solved for a series of values of λ using Cholesky decomposition. In the data space, each such solution is very fast, compared to the time required for forming the Jacobian. Such an Occam approach is not so well suited to using CG as the solver, since in the latter case **J** is not explicitly calculated and stored. To literally apply the Occam approach, the CG method would have to be applied to solve (3) for each λ, requiring a very large number of forward solutions.

We therefore, take a more traditional regularized optimization approach, taking λ as a fixed damping parameter. Thus, instead of solving the constrained optimization problem implied by (1), we minimize the penalty functional $\mathrm{W}_\lambda(\mathbf{m})$,

$$W_\lambda(\mathbf{m}) = (\mathbf{m}-\mathbf{m}_0)^\mathrm{T}\mathbf{C}_\mathrm{m}^{-1}(\mathbf{m}-\mathbf{m}_0) + \lambda^{-1}\left\{(\mathbf{d}-\mathbf{F}[\mathbf{m}])^\mathrm{T}\mathbf{C}_d^{-1}(\mathbf{d}-\mathbf{F}[\mathbf{m}])\right\}, \tag{4}$$

with λ fixed. Linearizing $\mathbf{F}[\mathbf{m}]$, we obtain the same system of data space eq. (3). With the data normalized with diagonal matrix $\mathbf{C_b}^{-1/2}$, this can be written

$$\mathbf{m}_{k+1}-\mathbf{m}_0 = \mathbf{C}_\mathrm{m}\mathbf{J}_k\mathbf{C}_\mathrm{d}^{-1/2}\left[\lambda\mathbf{I}+\mathbf{C}_d^{-1/2}\mathbf{J}_k\mathbf{C}_\mathrm{m}\mathbf{J}_k^\mathrm{T}\mathbf{C}_d^{-1/2}\right]^{-1}\mathbf{C}_d^{-1/2}\mathbf{đ}_k, \tag{5}$$

where I is the identity matrix. This simple transformation results in a better conditioned system, with the term λ **I** acting to stabilize the inversion. This simple transformation is analogous to the preconditioning of the model space equations by approximate solution of Poisson's equation, used by Haber & Ascher (2001) and Rodi & Mackie (2001).

CG is a relaxation method for solving the symmetric system of equations $\mathbf{Rx} = \mathbf{b}$ by iteratively minimizing the quadratic form $Q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Rx} - \mathbf{x}^T\mathbf{b}$. The CG algorithm and its details can be found in various publications (e.g. Press *et al.* 1992; Barret *et al.* 1994). In our application **R** is $[\lambda\,\mathbf{I}+\mathbf{C_d}^{-1/2}\mathbf{J}_\mathrm{k}\mathbf{C_m}\mathbf{J}_\mathrm{k}^\mathrm{T}\mathbf{C_d}^{-1/2}]$, **b** is $\mathbf{C_d}^{-1/2}\mathbf{đ}_\mathrm{k}$ and **x** is the unknown, which must be multiplied with $\mathbf{C_d}^{-1/2}$ to obtain the model $\mathbf{m}_{\mathrm{k}+1}$ as given in eq. (5). Implementation of CG requires only code to form the matrix–vector product **Rp** for arbitrary data space vectors **p,** rather than actually forming the matrix **R**. Thus we can also avoid forming and storing $\mathbf{J}_k$, provided we have routines for multiplication of model space vectors by $\mathbf{J}_k$ and data space vectors by $\mathbf{J}_k^\mathrm{T}$. Both of these matrix–vector products can be computed by solving one forward problem, as shown in Mackie & Madden (1993). Pseudo-code for the DCG algorithm is given in Fig. 2. Since $\mathbf{J}_k$ is never explicitly computed, one clear advantage of this approach is that storage of the large dense matrix $\mathbf{J}_k$ is not needed, as it is with DASOCC.

To compare the computational efficiency of DASOCC and DCG, we consider the total number of forward modelling steps required.

```
$\mathbf{d}$ = observed data, $\mathbf{C_d}$ = data error, $\mathbf{m_0}$ = initial model, $\mathbf{C_m}$ = model covariance

Solve forward problem and compute misfit from model $\mathbf{m_0}$
Select λ
Start DCG outer loop iteration $k$:
    Solve forward problem and Compute $\mathbf{đ}_\mathrm{k} = \mathbf{d} - \mathbf{F}[\mathbf{m}_\mathrm{k}] + \mathbf{J}_\mathrm{k}(\mathbf{m}_\mathrm{k} - \mathbf{m}_0)$
    Start CG iteration icg
        Solve forward problem twice to find $\mathbf{m}_{\mathrm{k}+1} - \mathbf{m_0} = \mathbf{C_m}\mathbf{J}_\mathrm{k}\mathbf{C_d}^{-1/2}\mathbf{R}_\mathrm{k}^{-1}\mathbf{C_d}^{-1/2}\mathbf{đ}_\mathrm{k}$,
            where $\mathbf{R}_\mathrm{k} = [\lambda\,\mathbf{I} + \mathbf{C_d}^{-1/2}\mathbf{J}_\mathrm{k}\mathbf{C_m}\mathbf{J}_\mathrm{k}^\mathrm{T}\mathbf{C_d}^{-1/2}]$
        Stop CG iteration if $\mathrm{r}_\mathrm{stop}$ less than desired level
    End icg
    Solve forward problem and Compute misfit from model $\mathbf{m}_{\mathrm{k}+1}$
    Exit when misfit less than desired level
End DCG outer loop iteration
```

Figure 2. Pseudo-code for DCG.

In DASOCC, where the full sensitivity is formed, the number of forward solver calls required to form all of **J** is $N_m N_s N_p$ using the reciprocity technique (Rodi 1976), where N_m is the number of modes (1 or 2 for MT), N_s is the number of sites and N_p is the number of periods. Then for each iteration, a further N_p forward solutions per mode are required for each λ in order to compute the actual data misfit. Thus the total number of forward solutions required per outer loop DASOCC iteration is about $N_m N_s N_p + N_\lambda N_p N_m$ where N_λ is the typical number of values of λ tried in each iteration. Since N_λ is typically 4–5, $N_\lambda N_p N_m$ is negligible compared to $N_m N_s N_p$, and will thus be ignored in the following comparisons.

With the DCG approach, the number of forward problems to be solved depends on the number of CG iterations in each step in the outer loop. For each (inner loop) CG iteration the number of forward solver calls required is $2N_p N_m$: one for computing $\mathbf{J}_{\boldsymbol{k}}\mathbf{p}$ and a second for computing $\mathbf{J}_k^T\mathbf{q}$, for each mode and for each period. At the end of one outer loop iteration of DCG, $N_m N_p$ forward modelling calls are required to form the background solution required for the next iteration, and to determine the misfit. Thus, the number of forward solver calls per outer loop DCG iteration is $2N_p N_m N_{cg} + N_m N_p$, where N_{cg} is number of CG iterations. Similar to the DASOCC case, we ignore $N_m N_p$ here because it is a small fraction of $2N_p N_m N_{cg}$. Thus, we can see that the DCG method will be more efficient than DASOCC only if the total number of CG iteration (N_{cg}) is less than $N_s/2$, and if the number of outer loop iterations remains the same

NUMERICAL EXAMPLES

Two 2-D synthetic data examples are used to test the relative efficiency of DCG and DASOCC. For this comparison we consider only the numbers of forward modelling calls used in each method, ignoring other computational overhead, such as solving the system of data space eq. (5) with Cholesky decomposition, as these represent only a small part of the total computational burden.

Synthetic Example I

First, we test DCG on the simple synthetic example illustrated in Fig. 3(a). The model is discritized into 100 × 31 blocks. The impedance Z_{xy} (TM mode) and Z_{yx} (TE mode) are generated from this model with 36 stations distributed uniformly from −40 to 40 km with a site spacing of 2.5 km. At each site, nine periods distributed uniformly in logarithmic period in the range from 0.01 to 100 s were computed. Random errors with a relative magnitude of 5 per cent were added to the real and imaginary part of the impedance data before inversion. The initial model for all inversion tests is a 50 Ohm-m half-space.

Data space Occam's inversion (DASOCC)

Convergence statistics from using DASOCC to invert TM mode, TE mode and TM + TE mode data are summarized in Table 1. For these three cases the inversion required 2, 3 and 3 outer loop iterations, respectively, to reach the desired target misfit of 1. This corresponds to Phase I of the Occam algorithm. For comparison with DCG we omit the additional 1–2 Phase II iterations, which fine tune the regularization parameter, and generally modify the solution only slightly. The result from joint inversion of the TM and TE data, fitting to an RMS misfit of one, is shown in Fig. 3(b). For DASOCC, the number of forward solver calls is fixed ($= N_s N_p N_m$), since the sensitivity matrix $\mathbf{J}_{\boldsymbol{k}}$ is explicitly formed. Thus, in each iteration of DASOCC, the number of forward solutions required for this example is 324 (36 × 9) for TM and TE single mode inversions, and 648 (36 × 9 × 2) for the joint TM + TE inversion. The total number of forward solutions required to reach the target misfit is thus 648 (324 × 2) for TM, 972 (324 × 3) for TE and 1944 (648 × 3) for TM + TE. These numbers provide a standard for evaluating the computational efficiency of the DCG algorithm.

Data space conjugate gradient method (DCG)

When solving (5) with CG, some stopping criteria must be defined. Rodi & Mackie (2001) terminate the CG process at three iterations per GN step in their MM method. Here, instead of fixing the number of iterations, we terminate when the relative error in the system of equations $||\mathbf{Ax} - \mathbf{b}||/||\mathbf{b}||$ reaches a specified tolerance r_{stop}. Initially we fix λ = 1, and compare the overall computational efficiency of the DCG scheme with different values of r_{stop}, such as 10^{-6}, 10^{-4}, 10^{-2} and 10^{-1}. The outer loop is terminated when the inversion reaches (or drops below) the desired RMS misfit of 1. Results for the TM mode are given in Table 2(a).

When r_{stop} is small, the number of iterations required is high, but when the actual data misfit is computed, the RMS is not reduced relative to the case $r_{stop} = 10^{-2}$. Clearly it is not necessary (or useful) to use a very stringent stopping criterion for inner loop DCG iterations. When r_{stop} is reduced further to 10^{-1}, the number of inner loop iterations is reduced, but the outer loop does not converge to the desired misfit in this case. Furthermore, even when the outer loop does converge, the number of outer loop iterations may be greater, resulting in a larger total number of forward modelling calls with this reduced value of r_{stop}. This example suggests that terminating the CG solver at a fixed small number of iterations, as in Rodi & Mackie (2001), will not always allow convergence to the target level. Indeed, in their tests examples Rodi & Mackie (2001) found that the CG scheme stalled in the later iterations, unable to achieve reduction in the objective function to levels achieved by GN, and non-linear CG (NLCG) approaches. At the same time it is worth

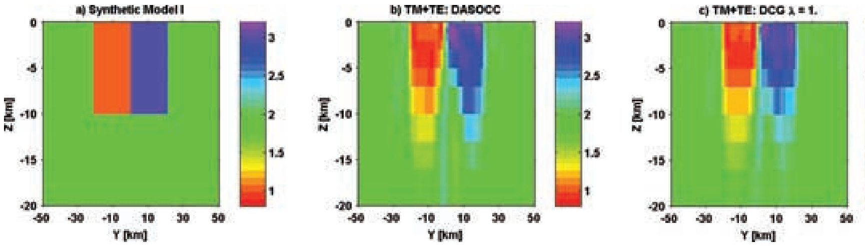


Figure 3. (a) Model I used to generate synthetic data, $\mathbf{Z}_{xy}$ and $\mathbf{Z}_{yx}$ for TM and TE modes. (b) Inverse model recovered from joint inversion of TM and TE modes using DASOCC inversion. (c) Same as (b) but using DCG with λ = 1. The 36 stations are distributed uniformly from −40 to 40 km.

Table 1. Number of iteration for DASOCC inversion to reach desired level of misfit for TM, TE and joint TM + TE inversions for synthetic test case I.

Outer loop DASOCC Iter no.	TM		TE		TM + TE	
	RMS	# of FWD to form $\mathbf{J}_k$	RMS	# of FWD to form $\mathbf{J}_k$	RMS	# of FWD to form $\mathbf{J}_k$
0	12.49	–	8.60	–	10.73	–
1	3.24	324	2.89	324	3.82	648
2	0.97	324	1.29	324	1.36	648
3			0.99	324	1.00	648
Total FWD		648		972		1944

Note: The number of forward modelling calls (FWD) required for each iteration, and the total over all iterations are also given.

Table 2. Number of iterations for the DCG inversion with different r_{stop} levels for TM (a), TE (b) and TM + TE (c) inversions for test case I.

Outer loop DCG Iter. no.	Relative error (r_{stop}) for stopping CG iterative process ($\lambda = 1.0$)							
	r_{stop} = 1.E-06		r_{stop} = 1.E-04		r_{stop} = 1.E-02		r_{stop} = 1.E-01	
	RMS	No. of CG Iter	RMS	No. of CG Iter	RMS	No. of CG Iter	RMS	No. of CG Iter
(a) TM: Initial RMS = 12.49								
1	3.24	58	3.24	39	3.22	23	3.70	11
2	1.44	45	1.44	32	1.44	16	1.57	7
3	1.01	43	1.01	29	1.01	14	1.37	6
4							1.37	6
5							1.33	6
6							1.33	6
Total CG Iter.		146		100		53		–
Total FWD		2628		1800		954		–
(b) TE: Initial RMS = 8.60								
1	2.89	46	2.89	31	2.89	17	2.96	10
2	1.94	37	1.94	26	1.94	15	2.04	8
3	1.19	34	1.19	24	1.20	13	1.30	7
4	1.04	34	1.04	24	1.04	12	1.25	6
5							1.21	6
6							1.21	6
Total CG Iter.		151		105		57		–
Total FWD		2718		1890		1026		–
(c) TM + TE: Initial RMS = 10.73								
1	4.38	77	4.38	49	4.38	28	4.38	16
2	2.60	61	2.60	41	2.61	22	2.98	12
3	1.17	50	1.17	36	1.19	20	1.75	10
4	0.95	50	0.95	36	0.96	19	1.18	9
5							1.08	7
6							1.08	7
Total CG Iter.		238		162		89		–
Total FWD		8568		5832		3204		–

Note: For r_{stop} = 1.E-01, the inversion cannot reach the desired level of misfit. Essentially the same RMS misfit is attained for all values of r_{stop} = 1.E-02 or less. Note that for each CG iteration 2 forward model solutions are required for each period.

noting that in the early outer loop steps similar misfit values are achieved with many fewer iterations when a larger value of r_{stop} is used. A more complex stopping criteria, with r_{stop} becoming smaller as the inversion converges may be worth considering.

Similar results are obtained for the TE and joint TM + TE inversions, as shown in Table 2(b) and (c), respectively. The inverse model obtained from the TM + TE inversion is shown in Fig. 3(c). Another general observation from these tables is that as the outer loop converges the number of CG iterations is reduced, even though it becomes necessary to use a more stringent stopping criteria to continue to make progress.

Next, we apply DCG using various values of λ, but with r_{stop} now fixed at 10^{-2}. Results for these experiments are given in Table 3(a) for TM, Table 3(b) for TE and Table 3(c) for joint TM + TE inversions. From these tables, we conclude that with smaller values of λ a larger number of CG iterations is required. This is because the system of equations becomes much stiffer. Higher values of λ on the other hand result in a well-conditioned system which converges in a smaller number of iterations. However, in this case, it may be impossible to reach the target misfit.

In all three inversion tests (TE, TM and TE + TM), optimal convergence occurs when $\lambda = 1$, and r_{stop} is 10^{-2}. For each of the outer loop iterations, the number of CG steps is roughly half the number of stations. However, the total number of outer loop iterations is slightly greater than what is required by DASOCC, that is, three for TM, four for TE and four for TM + TE inversions. The

Table 3. Number of iterations for the DCG inversion with different values of λ for TM (a), TE (b) and TM + TE (c) inversions of synthetic test case I.

Outer loop DCG Iter. no.	Different values of λ (r_{stop} = 1.0E-02)					
	λ = 0.1		λ = 1		λ = 10	
	RMS	No. of CG Iter	RMS	No. of CG Iter	RMS	No. of CG Iter
(a) TM: Initial RMS = 12.49						
1	4.98	48	3.22	23	4.83	11
2	3.03	44	1.44	16	3.64	7
3	1.81	45	1.01	14	3.54	6
4	0.76	37			3.47	6
5					3.46	6
Total CG Iter.		174		53		–
Total FWD		3132		954		–
(b) TE: Initial RMS = 8.60						
1	3.33	40	2.89	17	3.80	8
2	3.65	40	1.94	15	3.33	6
3	4.36	35	1.20	13	3.23	6
4	4.14	38	1.04	12	3.22	6
5	2.95	38			3.21	6
6	4.23	38			3.21	6
Total CG Iter.				57		–
Total FWD		–		1026		–
(c) TM + TE: Initial RMS = 10.73						
1	5.48	69	4.38	28	4.06	13
2	4.49	64	2.61	22	2.69	8
3	3.40	60	1.19	20	2.50	7
4	2.00	50	0.96	19	2.45	7
5	1.43	50			2.44	7
6	0.64	45			2.43	7
Total CG Iter.		338		89		–
Total FWD		12 168		3204		–

Note: For λ = 10, the inversion cannot reach the desired level of misfit, and for λ less than 0.1, the inversion diverges. For this test case λ = 1 is at least approximately optimal.

total number of CG steps required to reach the target misfit are thus 53, 57 and 89 for TM, TE and joint TM + TE inversions, respectively (Table 3). Each CG step requires two forward solutions for TM and TE, and four forward solutions for joint TM + TE inversions, per period. Thus, the total number of forward solver calls required are 954 (53 × 2 × 9), 1026 (57 × 2 × 9) and 3204 (89 × 4 × 9) for TM, TE and TM + TE inversions, respectively.

These numbers are higher than were required by the DASOCC method, by factors of roughly 1–1.6 times: 954 to 648 for TM, 1026 to 972 for TE and 3204–1944 for TM + TE inversions. Thus, for this example the computational efficiency of the DCG method is not superior to DASOCC in terms of CPU time. Numerous experiments with other synthetic examples support the general validity of this conclusion. Another issue for DCG is that we may need to try several different values of λ, particularly for real data sets. Values of λ that are too large may result in failure of the inversion to converge, while values that are too small will require a high number of CG iterations to converge, or may not result in convergence. However, DCG does have a very significant advantage with regard to memory, since storage of the sensitivity matrix is not required. Thus, there is a trade-off between computational efficiency and memory.

Synthetic Example: Case II

We next compare DCG and DASOCC on the more complicated structure shown in Fig. 4(a). This synthetic model may not look geologically realistic, but it provides a more challenging test of the inversions, and demonstrates that the relative performance of DCG and DASOCC will depend on the data set. As in the first example impedances Z_{xy} (TM mode) and Z_{yx} (TE mode) are generated for

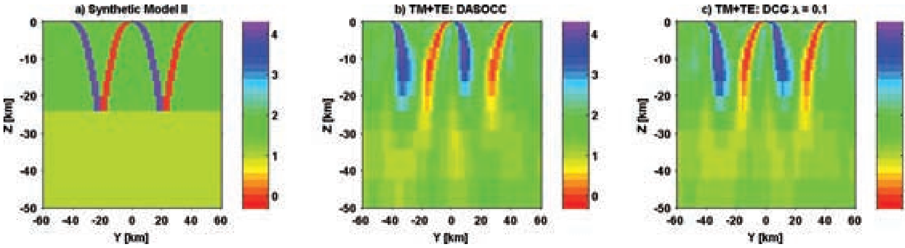


Figure 4. (a) Model II used to generate synthetic data, $\mathbf{Z}_{xy}$ and $\mathbf{Z}_{yx}$ for TM and TE modes. (b) Inverse model recovered from joint inversion of TM and TE modes using DASOCC inversion. (c) Same as (b) but using DCG with λ = 0.1. The 36 stations are distributed uniformly from −40 to 40 km.

Table 4. Number of iteration for DASOCC inversion to reach desired level of misfit for TM, TE and joint TM + TE inversions for synthetic test case II.

Outer loop DASOCC Iter no.	TM		TE		TM + TE	
	RMS	# of FWD to form $\mathbf{J}_k$	RMS	# of FWD to form $\mathbf{J}_k$	RMS	# of FWD to form $\mathbf{J}_k$
0	26.34	–	23.19		24.82	
1	8.38	324	6.71	324	9.27	648
2	3.96	324	2.82	324	3.76	648
3	3.93	324	1.54	324	2.13	648
4	4.35	324	1.11	324	1.35	648
5	3.62	324	0.97	324	1.24	648
6	1.53	324			1.00	648
7	0.97	324				
Total FWD		2268		1620		3888

Note: The number of forward modelling calls (FWD) required for each iteration, and the total over all iterations are also given.

36 stations and nine periods from 0.01 to 100 s with 5 per cent random errors. The model discretization is again 100 × 31 blocks, and the initial model for all inversion tests is a 50 Ohm-m half-space.

Data space occam's inversion (DASOCC)

Convergence of the DASOCC inversion for the TM, TE and TM + TE modes are shown in Table 4. The result from the joint TM + TE inversion at RMS misfit one is shown in Fig. 4(b) along with results from the comparable DCG inversion. Because the model is more complicated than the first case, the number of main loop iterations is higher: seven, five and six iterations are required to reach the desired target misfit of 1.0 for TM, TE and joint TM + TE inversion, respectively. This results in 2268 (324 × 7), 1620 (324 × 5) and 3888 (648 × 6) forward solutions for the three cases, as listed in Table 4.

Data space conjugate gradient method (DCG)

Next, we apply the DCG method to the same synthetic data sets. Results are summarized in Tables 5(a)–(c). Here, in all case r_{stop} was set at 10^{-2}. For the TE mode, with $\lambda = 0.1$, the inversion converges to below the target misfit in three iterations. Although this is less than what was required by DASOCC, the number of CG steps per outer loop iteration is about 1.5 times the number of stations, and the total number of CG iterations is 119. Thus the total number of forward solutions ($2142 = 119 \times 2 \times 9$) still exceeds that required by DASOCC (1620). The joint inversion requires 14 outer loop iterations, for a total of 806 CG steps, or 29 016 (806 × 4 × 9) forward solver calls. These numbers are huge compared to those required for DASOCC (Table 4). Tests with other values for λ did not yield better results; for λ lower than 0.1 there was generally no convergence. For the TM mode case no value of λ resulted in convergence of the DCG inversion. This example thus illustrates two potential shortcomings of the DCG approach. First, convergence can sometimes be very slow, and DCG may even fail to converge, even in cases where a DASOCC scheme works perfectly well. Second, DCG can be sensitive to the choice of regularization parameter λ, and the optimal choice is seldom known a priori. In the first synthetic example $\lambda = 1$ was optimal, but in the second example DCG worked considerably better with $\lambda = 0.1$. With real data sets one should plan on running the inversion for a range of values of this damping parameter.

DISCUSSION AND CONCLUSION

We have developed and tested a data space variant on the CG scheme (DCG) for 2-D MT data. The proposed scheme is essentially a GN scheme reformulated in the data space. Solution of the data space equivalent of the standard GN equations is then accomplished with CG, instead of computing the sensitivity matrix, forming the dense data space cross-product matrix, and solving the normal equations using Cholesky decomposition.

A widely perceived advantage of such CG approaches is that because they avoid explicit calculation of the sensitivity matrix, they are faster and computationally more efficient. However, in our numerical tests for 2-D MT data we find that a CG approach generally requires as many or more forward solver calls than an algorithm (DASOCC) which computes the full sensitivity. This is similar to results reported by Rodi & Mackie (2001). In their computational experiments, the numbers of forward solutions used in both their CG based MM and preconditioned NLCG methods were greater than those required for a more conventional GN method. However, they used a model space formulation, and the additional computational time required to form and solve the very large $M \times M$ system of normal eqs (2), made the GN approach slow, especially for large problems. This additional computational time is very significantly reduced when the problem is formulated in the data space, as with the DASOCC approach used here.

One disadvantage of a data space formulation is that there is no analogue of NLCG, which Rodi & Mackie (2001) found to be somewhat more efficient than CG in the late stages of convergence. However, these authors did not find substantial overall performance differences (in terms of forward solver calls required) between NLCG and the model space CG scheme they tested. Thus, it is far from clear that NLCG would require fewer forward calls than a GN approach such as DASOCC. It is possible that as the number of sites N_{s} is increases DCG may achieve convergence in many fewer than $N_{\mathrm{s}}/2$ iterations, and hence be faster than DASSOC, although this remains to be demonstrated. Furthermore, as the number of sites increases a reduced basis approach such as REBOCC (Siripunvaraporn & Egbert 2000) will also become more favourable.

One advantage of Occam in general, and DASOCC in particular, is that once $\mathbf{J}_k$ is computed and stored, this system of equations can

Table 5. Number of iterations for the DCG inversion with different values of λ for TM (a), TE (b) and TM + TE (c) inversions, synthetic test case II.

	Different values of λ (r_{stop} = 1.0E-02)					
	λ = 0.1		λ = 1		λ = 10	
Outer loop DCG Iter. No.	RMS	No. of CG Iter	RMS	No. of CG Iter	RMS	No. of CG Iter
(a) TM: Initial RMS = 26.34						
1	8.39	80	8.96	35	9.47	13
2	4.40	55	3.88	20	5.74	9
3	3.02	52	2.88	17	5.34	7
4	3.51	46	2.53	17	5.38	7
5	2.84	49	2.30	17	5.34	7
6	3.69	48	2.24	17	5.35	7
7	2.81	50	2.32	17	5.34	7
8	3.70	45	2.96	17	5.34	7
9	2.92	48	3.59	18	5.34	7
10	3.59	45	5.65	20	5.34	7
Total CG Iter.		–		–		–
Total FWD		–		–		–
(b) TE: Initial RMS = 23.19						
1	6.92	48	7.12	23	7.84	11
2	2.50	39	2.53	15	4.08	7
3	0.99	32	1.50	13	3.77	6
4			1.44	13	3.75	6
5			1.42	13	3.73	6
6			1.42	13	3.73	6
Total CG Iter.		119		–		–
Total FWD		2142		–		–
(c) TM + TE: Initial RMS = 24.80						
1	9.53	99	9.32	39	9.31	16
2	4.08	70	3.88	25	4.86	10
3	2.55	59	2.56	20	4.23	8
4	2.10	58	2.23	20	4.21	8
5	2.13	54	2.05	20	4.19	8
6	2.83	53	1.97	20	4.19	8
7	2.18	57	1.94	20	4.19	8
8	1.83	53	1.94	20	4.19	8
9	1.47	52	1.94	21	4.19	8
10	1.42	52	1.94	21	4.19	8
11	1.19	52	1.94	21	4.19	8
12	1.18	49	1.94	21	4.19	8
13	1.08	49	1.94	21	4.19	8
14	1.01	49	1.94	21	4.19	8
Total CG Iter.		806		–		–
Total FWD		29 016		–		–

Note: For TM, none of the values of λ tested allow the target level of misfit to be reached.

be solved repeatedly for different values of λ. Thus, the Lagrange multiplier λ^{-1} can be used both for damping and for step length control (Parker 1994). This guarantees, at least in theory, convergence to a local minimum of the model norm, subject to the data misfit achieved (Parker 1994). This property cannot be guaranteed for more standard GN-CG or NLCG methods, where λ is independently chosen and left fixed during penalty functional optimization. Because $\mathbf{J}_{\boldsymbol{k}}$ is not explicitly formed and stored in the DCG scheme, we also cannot directly use an Occam style approach. The optimal prior choice of λ is not obvious, and, as shown in our numerical tests, performance of the CG inversion can be greatly influenced by this parameter. Possible approaches to picking λ are given, for example, in Haber *et al.* (2000). Another idea, which deserves further exploration but is beyond the scope of this paper, would be to use Lanczos tridiagonalization (the basis for CG; Gloub & Van Loan 1989). At the cost of increased memory (required to store all search directions) the system (3) could then be efficiently solved for a range of values of λ.

For realistic 3-D problems, both model and data sizes become significantly larger. Therefore, DASOCC for 3-D MT inversion (Siripunvaraporn *et al.* 2004; 2005a) requires huge amounts of RAM. For example, in the EXTECH data set (Tuncer *et al.* 2006), $N = 16 \times 131 \times 4 = 8{,}384$ and $M = 56 \times 56 \times 33 = 103\,488$, requiring about $8NM \approx 7$ Gbyte to store just the sensitivity matrix. This data set thus requires running the 3-D inversion on a workstation or even a supercomputer. Applying DCG to 3-D MT inversion is straightforward, and would allow running large problems such as this on a common desktop PC. However, our 2-D numerical tests suggest that the number of forward modelling calls are actually likely to be larger for DCG, resulting in even longer run times. Clearly there is a trade-off between memory used and CPU run time, and the choice between DASOCC and DCG will depend on the application.

ACKNOWLEDGMENTS

This research has been supported by a grant from the Thailand Research Fund (RSA4780021) to WS, and by DOE-FG0302ER15318 to GE. The authors would like to thank Oliver Ritter (Editor) and two anonymous reviewer for their comments to help improving the manuscript.

REFERENCES

Barret, R. *et al.*, 1994. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, Soc. Ind. Appl. Math., Philadelphia, PA.

Bennett, A.F., Chua, B.S. & Leslie, L.M., 1996. Generalized inversion of a global numerical weather prediction model, *Meteorol. Atmos. Phys.,* **60,** 165–178.

Constable, C.S., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics,* **52,** 289–300.

Egbert, G.D., 1997. Tidal data inversion: interpolation and inference, *Prog. Oceanog.,* **40,** 53–80.

Golub, G. & Van Loan, C., 1989. *Matrix Computations,* 2nd ed., The Johns Hopkins University Press.

Haber, E., Ascher, U.M., Aruliah, D.A. & Oldenburg, D., 2000. On optimization techniques for solving nonlinear inverse problems, *Inverse Problems,* **16,** 1263–1280.

Haber, E. & Ascher, U., 2001. Preconditioned all-at-once methods for large, sparse parameter estimation problems, *Inverse Problems,* **17,** 1847–1864.

Mackie, R.L. & Madden, T.R., 1993. Three-dimensional magnetotelluric inversion using conjugate gradients, *Geophys. J. Int.,* **115,** 215–229.

Newman, G.A. & Alumbaugh, D.L., 1996. Three-dimensional massively parallel electromagnetic inversion-I. Theory, *Geophys. J. Int.,* **128,** 340–354.

Newman, G.A. & Alumbaugh, D.L., 2000. Three-dimensional magnetotelluric inversion using non-linear conjugate gradients, *Geophys. J. Int.,* **140,** 410–424.

Parker, R.L.,1994. *Geophysical Inverse Theory,* Princeton University Press, Princeton, NJ.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P., 1992. *Numerical Recipes in FORTRAN: The Art of Scientific Computing,* 2nd ed., Cambridge Univ. Press, Cambridge, UK.

Rodi, W.L., 1976. A technique for improving the accuracy of finite element solutions for Magnetotelluric data, *Geophys. J. Roy. Astr. Soc.,* **44,** 483–506.

Rodi, W.L. & Mackie, R.L., 2001. Nonlinear conjugate gradients algorithm for 2-D magnetotelluric inversion, *Geophys.,* **66,** 174–187.

Sasaki, Y., 2001. Full 3D inversion of electromagnetic data on PC., *J. Appl. Geophys.,* **46,** 45–54.

Siripunvaraporn, W. & Egbert, G., 2000. An efficient data-subspace inversion method for 2-D magnetotelluric data, *Geophysics,* **65,** 791–803.

Siripunvaraporn, W., Uyeshima, M. & Egbert, G., 2004. Three-dimensional inversion for Network-Magnetotelluric data, *Earth Planets Space,* **56,** 893–902.

Siripunvaraporn, W., Egbert, G., Lenbury, Y. & Uyeshima, M., 2005a. Three-dimensional magnetotelluric inversion: data space method, *Phys. Earth Planet. Interior.,* **150,** 3–14.

Siripunvaraporn, W., Egbert, G. & Uyeshima, M., 2005b. Interpretation of two-dimensional magnetotelluric profile data with three-dimensional inversion: synthetic examples, *Geophys. J. Int.,* **160,** 804–814.

Zhdanov, M.S., Fang, S., Hursan, G., 2000. Electromagnetic inversion using quasi-linear approximation, *Geophysics,* **65,** 1501–1513.

Tuncer, V., Unsworth, M.J., Siripunvaraporn, W. & Craven, J.A., 2006, Exploration for unconformity-type uranium deposits with audiomagnetotelluric data: a case study from the McArthur River mine, Saskatchewan, Canada, *Geophysics,* **71,** B201–B209.

Physics of the Earth and Planetary Interiors 173 (2009) 317–329

Contents lists available at ScienceDirect

Physics of the Earth and Planetary Interiors

journal homepage: www.elsevier.com/locate/pepi

WSINV3DMT: Vertical magnetic field transfer function inversion and parallel implementation

Weerachai Siripunvaraporn[a,*], Gary Egbert[b]

[a] Department of Physics, Faculty of Science, Mahidol University, Rama VI Rd., Rachatawee, Bangkok 10400, Thailand
[b] College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331, USA

ARTICLE INFO

Article history:
Received 15 September 2008
Received in revised form 21 January 2009
Accepted 25 January 2009

Keywords:
Magnetotellurics
Vertical magnetic transfer function
Data space method
3-D inversion
Occam's inversion

ABSTRACT

We describe two extensions to the three-dimensional magnetotelluric inversion program WSINV3DMT (Siripunvaraporn, W., Egbert, G., Lenbury, Y., Uyeshima, M., 2005, Three-dimensional magnetotelluric inversion: data-space method. Phys. Earth Planet. Interiors 150, 3–14), including modifications to allow inversion of the vertical magnetic transfer functions (VTFs), and parallelization of the code. The parallel implementation, which is most appropriate for small clusters, uses MPI to distribute forward solutions for different frequencies, as well as some linear algebraic computations, over multiple processors. In addition to reducing run times, the parallelization reduces memory requirements by distributing storage of the sensitivity matrix. Both new features are tested on synthetic and real datasets, revealing nearly linear speedup for a small number of processors (up to 8). Experiments on synthetic examples show that the horizontal position and lateral conductivity contrasts of anomalies can be recovered by inverting VTFs alone. However, vertical positions and absolute amplitudes are not well constrained unless an accurate host resistivity is imposed *a priori*. On very simple synthetic models including VTFs in a joint inversion had little impact on the inverse solution computed with impedances alone. However, in experiments with real data, inverse solutions obtained from joint inversion of VTF and impedances, and from impedances alone, differed in important ways, suggesting that for structures with more realistic levels of complexity the VTFs will in general provide useful additional constraints.

1. Introduction

WSINV3DMT (Siripunvaraporn et al., 2005) has been developed to invert Magnetotelluric (MT) impedance tensor components for three-dimensional (3-D) Earth conductivity. It was made freely available to the MT research community in 2006 and has since become one of the standard tools for 3-D inversion and interpretation (e.g., Tuncer et al., 2006; Heise et al., 2008; among others). The inversion algorithm used closely follows the two-dimensional (2-D) data space Occam's inversion of Siripunvaraporn and Egbert (2000) which has also been widely used for 2-D interpretation (e.g., Pous et al., 2002; Oskooi and and Perdersen, 2005; Toh et al., 2006; among others). Here we describe extensions to this code, which we illustrate with tests on synthetic and real data.

We first briefly summarize WSINV3DMT; see Siripunvaraporn et al. (2005) for more technical details. The algorithm used is based on the classic Occam's inversion introduced by Constable et al. (1987) for the one-dimensional (1-D) MT and DC resistivity sounding problems. The Occam inversion seeks a minimum structure model (as defined by some model norm which penalizes roughness) subject to an appropriate fit to the data. The minimization is accomplished with a modified Gauss–Newton algorithm, in which the regularization parameter (which controls the tradeoff between model roughness and data fit) is also used for step length control (Parker, 1994). The main advantages of the Occam approach are its stability and robustness, and the fact that the scheme often converges to the desired misfit in a relatively small number of iterations (e.g., Siripunvaraporn and Egbert, 2000). Occam was extended to treat two-dimensional MT data by deGroot-Hedlin and Constable (1990), but for multi-dimensional inversion the originally proposed scheme can be computationally impractical, as the system of normal equations is explicitly formed and solved in the model space.

Siripunvaraporn and Egbert (2000) transformed the inverse problem into the data space (e.g., Parker, 1994). If the number of data (N) is small compared to the number of model parameters (M), as will typically be the case in 3-D, the data space variant requires a fraction of the CPU time and memory compared to a model space scheme. This data space Occam scheme forms the basis for the WSINV3DMT algorithm, which is summarized in Fig. 1.

The initial version of WSINV3DMT was only capable of inverting the impedance tensor **Z**, the 2 × 2 complex frequency dependent

[*] Corresponding author. Tel.: +662 201 5770; fax: +662 354 7159.
E-mail address: scwsp@mahidol.ac.th (W. Siripunvaraporn).

0031-9201/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.pepi.2009.01.013

Nomenclature

$\mathbf{d}$	observed data
$\mathbf{C}_d$	data error
$\mathbf{m}_0$	initial and prior model
$\mathbf{C}_m$	model covariance
$\mathbf{m}_k$	model at k iteration
$\mathbf{J}_k$	$N \times M$ sensitivity matrix forming from $\mathbf{m}_k$
$\mathbf{F}[\mathbf{m}_k]$	forward responses of $\mathbf{m}_k$
$\mathbf{\Gamma}_k$	data space cross product matrix
$\mathbf{R}_k$	representer for k iteration
λ	Lagrange multiplier
N_s	number of stations
N_m	number of modes
N_p	number of periods
N	number of data = $N_s \times N_m \times N_p$
M	number of model parameters

transfer function relating electric to magnetic fields

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{bmatrix} \begin{bmatrix} H_x \\ H_y \end{bmatrix}. \tag{1}$$

The impedance tensor is frequently used by itself for 3-D conductivity imaging (e.g., Tuncer et al., 2006; Heise et al., 2008; Patro and Egbert, 2008). However, modern MT field practice typically includes measurement of vertical magnetic fields (particularly at long periods, where a tri-axial magnetometer is used), and thence computation of vertical field transfer functions (VTFs)

$$H_z = \begin{bmatrix} T_{zx} & T_{zy} \end{bmatrix} \begin{bmatrix} H_x \\ H_y \end{bmatrix}. \tag{2}$$

The vertical magnetic field is only produced when there are lateral or horizontal variations of conductivity. Researchers have often used VTFs in the form of induction vectors (Parkinson, 1959) to indicate or point to the source of conductivity anomalies and to establish or verify geoelectic strike directions (e.g., Bedrosian et al., 2004; Uyeshima et al., 2005; Tuncer et al., 2006). A number of 2-D inversion codes (e.g., REBOCC of Siripunvaraporn and Egbert, 2000; and NLCG of Rodi and Mackie, 2001) allow inversion of VTFs (or “Tipper”), and these are often included along with TE and TM impedances in 2-D interpretations of MT profile data (e.g., Wannamaker et al., 1989; Wannamaer et al., 2008). Berdichevsky et al. (2003) studied VTFs using analytical and modeling studies, and concluded that inclusion of these additional induction transfer functions can substantially improve the reliability of geoelectrical models, because they are not affected as strongly by local distortion as the impedance tensor is.

Here, we describe the implementation of VTF inversion for the WSINV3DMT inversion code, and apply this to inversion of real and synthetic datasets. In addition, we describe implementation of a parallel version of WSINV3DMT, using MPI and parallelizing over frequencies to help reduce program execution times, which can be quite long for realistic modern datasets (e.g., Patro and Egbert, 2008).

The paper is organized as follows. First, we summarize the modifications to WSINV3D, for the most part omitting technical details. Next, we illustrate and test the new features on the same synthetic datasets previously used in Siripunvaraporn et al. (2005). Here we illustrate the speedup obtained with the parallelization, and explore the effectiveness of VTF data for recovering conductivity structures, alone, and in conjunction with impedance data. We then test the VTF inversion on the EXTECH dataset (Tuncer et al., 2006), comparing inverted models from only VTF data, from only impedance data, and from a joint inversion of both data types.

2. Implementation of WSINV3DMT to include the vertical magnetic transfer function

There are only two major modifications to the WSINV3DMT codes required to allow inversion of VTFs: adding the VTF computation to the forward modeling routine, and the corresponding modifications for the sensitivities of the real and imaginary parts of the VTFs.

In WSINV3DMT, the electric fields are calculated by solving the second order Maxwell’s equation using a staggered grid finite difference numerical scheme (Siripunvaraporn et al., 2002). Magnetic field components can then be computed (on grid cell faces) from Faraday’s law $\nabla \times \mathbf{E} = i\omega\mu\mathbf{H}$, and interpolated to the observation locations, which in the modified version of WSINV3D can be at any location on the surface. In order to compute the impedance tensor $\mathbf{Z}$ the forward equations are solved for two polarizations, and $\mathbf{Z}$ is calculated from the combination of horizontal electric and magnetic fields from both polarizations, as described in Siripunvaraporn et al. (2005).

The only modification required for the VTF is that the vertical magnetic field must also be computed at the observation location. As for the horizontal magnetic components, this is accomplished using Faraday’s law, taking the curl of the horizontal $\mathbf{E}$ components on the model air–Earth interface, and interpolating the result (defined at cell centers) to the observation locations. Then, similarly to the impedance tensor, the vertical and horizontal magnetic fields computed from the solutions for both polarizations are combined to form the vertical magnetic field transfer function $\mathbf{T}$,

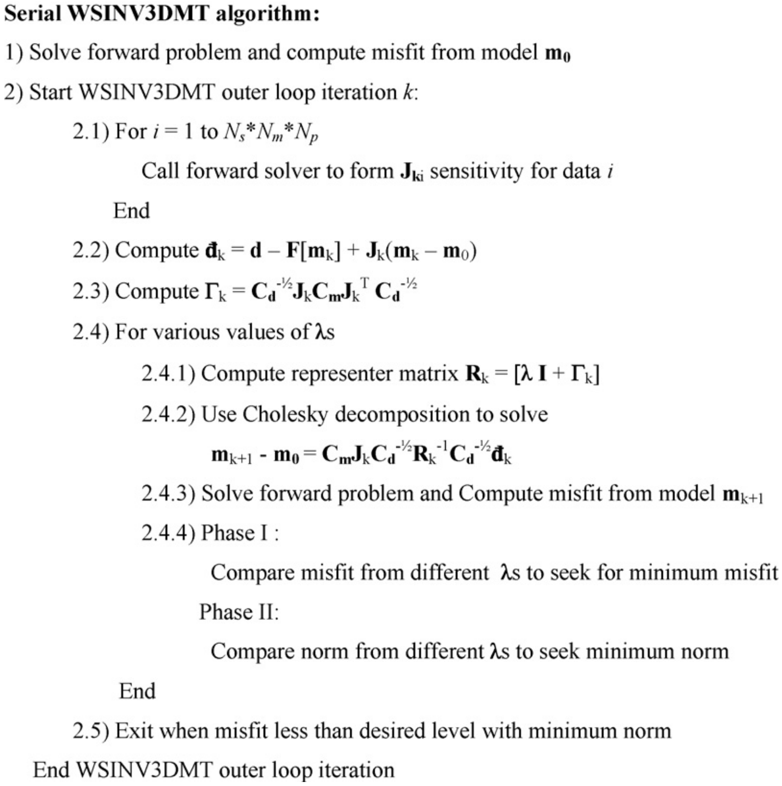
$$\begin{bmatrix} \boldsymbol{H}_z^1 & \boldsymbol{H}_z^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_{zx} & \boldsymbol{T}_{zy} \end{bmatrix} \begin{bmatrix} \boldsymbol{H}_x^1 & \boldsymbol{H}_x^2 \\ \boldsymbol{H}_y^1 & \boldsymbol{H}_y^2 \end{bmatrix} \tag{3}$$

Here $\boldsymbol{H}_z^1$ and $\boldsymbol{H}_z^2$ are the z-component of magnetic fields for the $\mathbf{E}_x$–$\mathbf{H}_y$ and $\mathbf{E}_y$–$\mathbf{H}_x$ polarizations, respectively, and similarly for other field components. For a joint inversion with impedance tensor, computing the vertical magnetic transfer function does not require any extra forward modeling calls, as all transfer functions are computed from the same solutions.

The sensitivity calculation for VTFs is essentially identical to that used for impedances, which is based on the reciprocity approach described in Rodi (1976), Newman and Alumbaugh (2000), and Siripunvaraporn et al. (2005). Briefly, the linearized data functional, which is represented by linear combinations of electric field solution components on cell edges surrounding the observation point, is used to force the adjoint equation, and the result is mapped to perturbations in the model parameter, as described in Siripunvaraporn et al. (2005). Only the first step requires modification, with the coefficients for the linearized functionals for $\mathbf{T}_{zx}$ and $\mathbf{T}_{zy}$ replacing those for $\mathbf{Z}_{xx}$ and $\mathbf{Z}_{xy}$. Details of this modification are straightforward, and are omitted here.

3. Parallel implementation with MPI

A major challenge in using WSINV3DMT, or for that matter, any 3-D MT inversion code, is that the program is very time consuming, especially when run with the sort of large dataset (and model domain) that justifies a 3-D interpretation. Run times exceeding a full month (on a single processor desktop computer, for the full inversion process, including multiple iterations of the outer loop of Fig. 1) have been reported when WSINV3D has been applied to even modest 3-D MT datasets (e.g., Patro and Egbert, 2008). These long run times primarily reflect the need for many forward modeling calls, each of which requires iterative



```
Serial WSINV3DMT algorithm:
1) Solve forward problem and compute misfit from model $\mathbf{m_0}$
2) Start WSINV3DMT outer loop iteration $k$:
    2.1) For $i = 1$ to $N_s*N_m*N_p$
            Call forward solver to form $\mathbf{J_{ki}}$ sensitivity for data $i$
         End
    2.2) Compute $\hat{\mathbf{d}}_k = \mathbf{d} - \mathbf{F}[\mathbf{m}_k] + \mathbf{J}_k(\mathbf{m}_k - \mathbf{m}_0)$
    2.3) Compute $\Gamma_k = \mathbf{C_d}^{-1/2}\mathbf{J}_k\mathbf{C_m}\mathbf{J}_k^T\,\mathbf{C_d}^{-1/2}$
    2.4) For various values of $\lambda$s
            2.4.1) Compute representer matrix $\mathbf{R}_k = [\lambda\,\mathbf{I} + \Gamma_k]$
            2.4.2) Use Cholesky decomposition to solve
                       $\mathbf{m}_{k+1} - \mathbf{m_0} = \mathbf{C_m}\mathbf{J}_k\mathbf{C_d}^{-1/2}\mathbf{R}_k^{-1}\mathbf{C_d}^{-1/2}\hat{\mathbf{d}}_k$
            2.4.3) Solve forward problem and Compute misfit from model $\mathbf{m}_{k+1}$
            2.4.4) Phase I :
                       Compare misfit from different $\lambda$s to seek for minimum misfit
                   Phase II:
                       Compare norm from different $\lambda$s to seek minimum norm
         End
    2.5) Exit when misfit less than desired level with minimum norm
 End WSINV3DMT outer loop iteration
```

Fig. 1. Pseudo-code for serial WSINV3DMT (after Siripunvaraporn and Egbert, 2007).

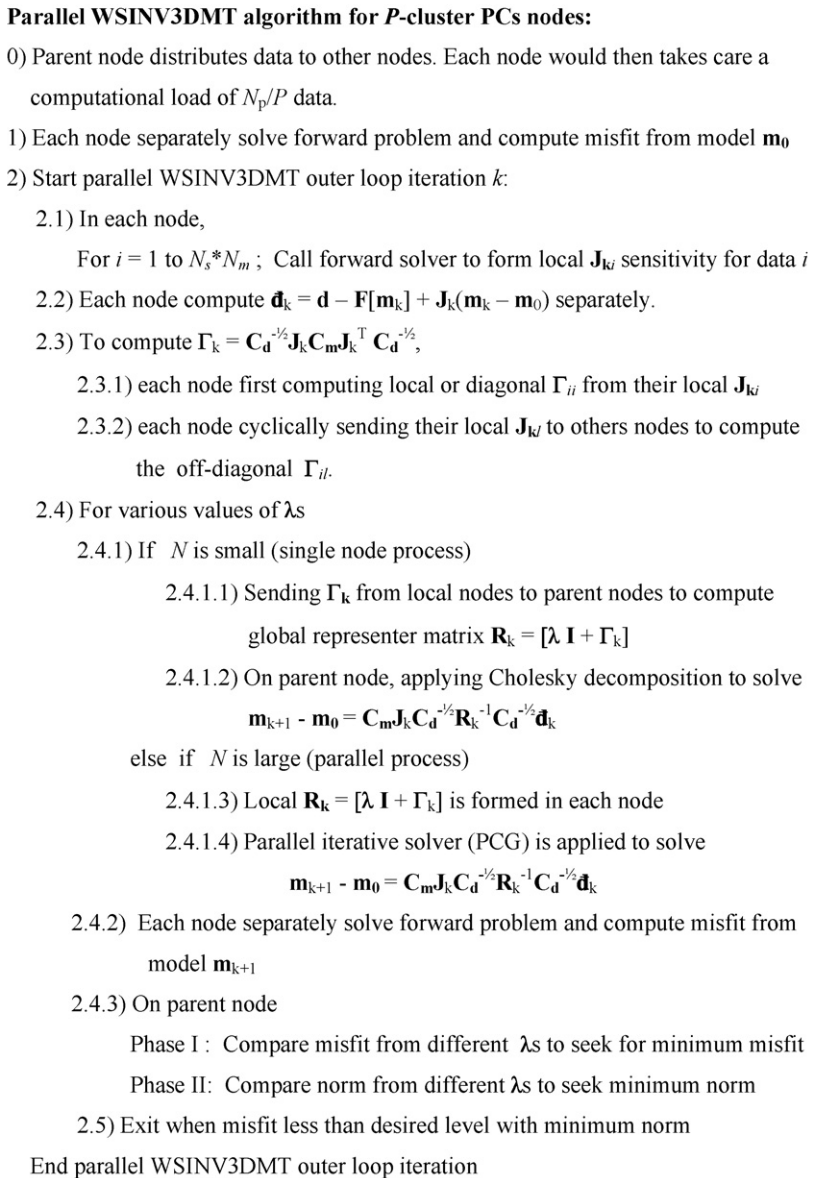
solution of the large sparse linear system arising from discretization of Maxwell's equations. WSINV3D was developed as a serial code, to run on a single processor. An obvious way to speed up execution is to parallelize the code, and make use of the multiple processors which are increasingly common even in desktop computers.

There are several ways to redesign the codes to run on parallel system, and the most appropriate approach will depend on system architecture. For supercomputers or large clusters to make effective use of hundreds of processors it would be necessary to rewrite parts of the forward solver—e.g., parallelizing the iterative solver and preconditioner (e.g., Newman and Alumbaugh, 2000), or domain decomposition. Here, we consider a parallelization approach appropriate to small systems with a few to several tens of processors. Such small clusters and multi-processor workstations are now readily affordable and more widely available than supercomputers. To adapt WSINV3DMT for this class of systems, we parallelize over frequencies, adding calls to MPI (Message Passing Interface) library routines to the existing codes. In this way, we do not have to alter the core forward modeling and sensitivity calculation routines in any way. The parallel algorithm is summarized in Fig. 2.

Forward modeling and sensitivity calculations for each period are sent to one processor (Steps 2.1 and 2.2 in Fig. 2). If there are fewer processors than periods, each processor performs calculations for more than one period. With this simple parallelization, which requires minimal inter-processor communication, the computational time should be theoretically reduced by a factor P, the number of processors available. This parallel implementation also distributes storage of the sensitivity matrix over the available nodes. The $N \times M$ sensitivity matrix $\mathbf{J}$ requires $8NM$ bytes (in double precision), and the need to store this in RAM limits the size of datasets and model grids that can be practically treated. With the parallelization, memory required on each node is reduced to about two times $8NM/P$ (including temporary storage for cross product computations), allowing WSINV3D to be run for larger models grids and datasets.

With the sensitivities distributed over processors, formation of the cross product matrix $\boldsymbol{\Gamma} = \mathbf{J}\mathbf{C}_m^{-1}\mathbf{J}^T$ also requires MPI calls. We have implemented this in a fairly simple way, breaking $\boldsymbol{\Gamma}$ into P^2 blocks to be computed on the P processors (Step 2.3 in Fig. 2). Diagonal blocks $\boldsymbol{\Gamma}_{ii}$ are computed on the local processor where the corresponding block $\mathbf{J}_i$ of the sensitivity matrix (corresponding to one or more frequencies) is computed and stored. The off-diagonal blocks ($\boldsymbol{\Gamma}_{ij}$) can only be formed by sharing blocks of $\mathbf{J}$ between nodes. Since $\boldsymbol{\Gamma}$ is symmetric, only upper off-diagonal blocks ($j > i$) need be formed. On step k block $\mathbf{J}_j$, where $j = \mathrm{mod}(i + k, P)$ is sent to node i to compute $\boldsymbol{\Gamma}_{ij}$ where this block is stored. With this simple scheme the load is balanced and the number of steps required is approximately $(N_p + 1)/2$. Although computing the cross products requires significant communication time to share sensitivities between nodes, it can still significantly reduce the total computing time required to form $\boldsymbol{\Gamma}$ compared to single node processing.

In the data space Occam scheme used by WSINV3D the system of normal equations (Eq. (6) in Siripunvaraporn et al., 2005) must be solved for a series of trial values of the regularization parameter (about 3–7 from our experience) to find the optimal (in terms of data misfit and model norm) model parameter update. In the serial version of WSINV3D these dense systems are solved by Cholesky decomposition (Step 2.4.2 in Fig. 1). Parallel Cholesky decomposition subroutines are available (e.g., Choi and Moon, 1997), but these are generally tailored to a specific parallel architecture and are not easily adapted. To develop code that will be portable, and reasonably efficient on a generic multi-processor system, we have thus pursued a different strategy, using the easily parallelized preconditioned conjugate gradient (PCG) algorithm to solve the normal equations (Step 2.4.1.2 in Fig. 2). The major computation in the



```
Parallel WSINV3DMT algorithm for $P$-cluster PCs nodes:
0) Parent node distributes data to other nodes. Each node would then takes care a
   computational load of $N_p/P$ data.
1) Each node separately solve forward problem and compute misfit from model $\mathbf{m_0}$
2) Start parallel WSINV3DMT outer loop iteration $k$:
  2.1) In each node,
       For $i$ = 1 to $N_s$*$N_m$ ; Call forward solver to form local $\mathbf{J_{k\mathit{i}}}$ sensitivity for data $i$
  2.2) Each node compute $\bar{\mathbf{d}}_k = \mathbf{d} - \mathbf{F}[\mathbf{m}_k] + \mathbf{J}_k(\mathbf{m}_k - \mathbf{m}_0)$ separately.
  2.3) To compute $\Gamma_k = \mathbf{C_d}^{-½}\mathbf{J}_k\mathbf{C_m}\mathbf{J}_k^{T}\,\mathbf{C_d}^{-½}$,
       2.3.1) each node first computing local or diagonal $\Gamma_{ii}$ from their local $\mathbf{J_{k\mathit{i}}}$
       2.3.2) each node cyclically sending their local $\mathbf{J_{k\mathit{l}}}$ to others nodes to compute
              the off-diagonal $\Gamma_{il}$.
  2.4) For various values of $\lambda$s
       2.4.1) If $N$ is small (single node process)
              2.4.1.1) Sending $\mathbf{\Gamma_k}$ from local nodes to parent nodes to compute
                       global representer matrix $\mathbf{R}_k = [\lambda\ \mathbf{I} + \Gamma_k]$
              2.4.1.2) On parent node, applying Cholesky decomposition to solve
                       $\mathbf{m}_{k+1}$ - $\mathbf{m_0} = \mathbf{C_m}\mathbf{J}_k\mathbf{C_d}^{-½}\mathbf{R}_k^{-1}\mathbf{C_d}^{-½}\bar{\mathbf{d}}_k$
         else if $N$ is large (parallel process)
              2.4.1.3) Local $\mathbf{R_k} = [\lambda\ \mathbf{I} + \Gamma_k]$ is formed in each node
              2.4.1.4) Parallel iterative solver (PCG) is applied to solve
                       $\mathbf{m}_{k+1}$ - $\mathbf{m_0} = \mathbf{C_m}\mathbf{J}_k\mathbf{C_d}^{-½}\mathbf{R}_k^{-1}\mathbf{C_d}^{-½}\bar{\mathbf{d}}_k$
       2.4.2) Each node separately solve forward problem and compute misfit from
              model $\mathbf{m}_{k+1}$
       2.4.3) On parent node
              Phase I : Compare misfit from different $\lambda$s to seek for minimum misfit
              Phase II: Compare norm from different $\lambda$s to seek minimum norm
  2.5) Exit when misfit less than desired level with minimum norm
End parallel WSINV3DMT outer loop iteration
```

Fig. 2. Pseudo-code for parallel WSINV3DMT for cluster PCs system.

PCG algorithm is matrix–vector multiplication. This is readily parallelized by dividing the vectors and matrix into blocks, spreading computations for individual blocks over processors, and then gathering the results back to the master node. To simplify the algorithm we have distributed the full matrix to all computational nodes.

The preconditioner, based on the diagonals of the coefficient matrix, is also trivially parallelized. Because the coefficient matrices are dense, the parallel PCG scheme may not be efficient when N is small, since communication and other overhead may exceed the serial computational time. For smaller N, we therefore retain the option of solving the normal equations with a serial Cholesky decomposition, after all blocks $\mathbf{\Gamma}_{ij}$ are sent back to the parent node. The optimal choice of solution scheme (parallel or serial) for a specific value of N will depend on the cluster architecture. We give examples below where each approach is more efficient.

Once the new model $\mathbf{m}_{k+1}$ is obtained, the parallelized forward solver is called to compute the responses of each period, with the results gathered to the parent node to compute misfits (Step 2.4.2 in Fig. 2). All steps are repeated until an acceptable misfit and norm are achieved

4. Synthetic data examples

To illustrate the efficiency of the parallelized WSINV3D, and the effectiveness of the VTF inversion, we first consider inversion of synthetic datasets, revisiting the two synthetic examples previously used by Siripunvaraporn et al. (2005), reproduced in Fig. 3. The results of these tests are consistent with those obtained for other synthetic examples. Our basic test configuration is the two-block model (Fig. 3a) consisting of two anomalies, 1 Ω m and 100 Ω m located next to each other within a 10 Ω m host. The spatially homogeneous layer, which extends from the surface to 10 km depth, is underlain by a 100 Ω m half space. To test the efficiency of our parallel codes, and the VTF inversion, we generated VTF and impedance data at 16 periods (from 0.1 to 1000 s) for a total of 40 sites in a regular grid, as illustrated in Fig. 3a. Gaussian noise (5% of the data magnitude) was added to the generated data. The inversions for this case are performed on a 21 × 28 × 21 (+7 air layers) mesh. The second model consists of a single conductive block (1 Ω m) buried in a 100 Ω m half-space (Fig. 3b), and responses were computed at 16 periods for 36 sites (Fig. 3b). The inversions

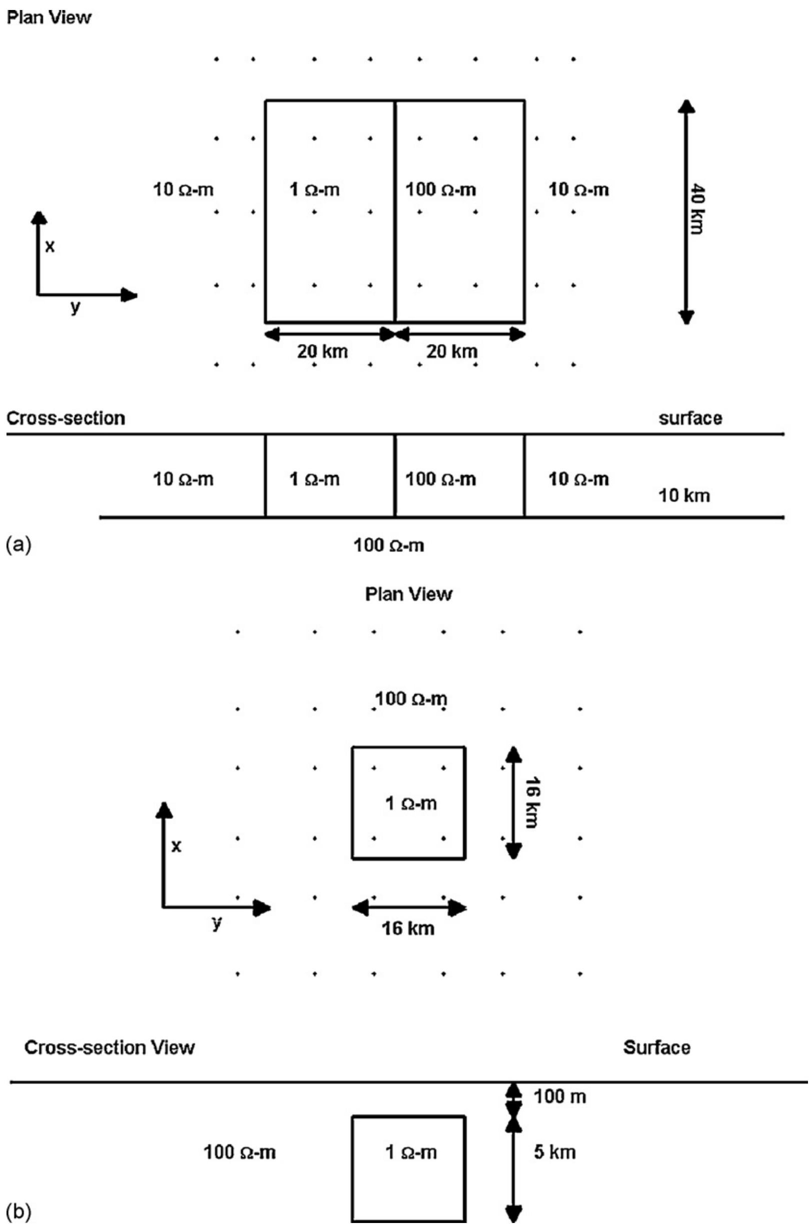


Fig. 3. Two synthetic models used to test our inversion. (a) Two-block synthetic model and (b) a single conductive block model. The solid dots indicate the observation sites. The cross-section view in the lower panel is a profile cutting across the middle of the model in the upper panel, and is not to scale (after Siripunvaraporn et al., 2005).

for the second case are performed on a 28 × 28 × 21 (+7 air layers) mesh.

We first demonstrate the efficiency of the parallel version of WSINV3D, using both VTF and joint VTF/impedance datasets for tests. We then consider the effectiveness of VTF data for recovering conductivity variations, both alone, and in conjunction with impedances.

4.1. Parallel efficiency

We tested WSINV3DMT by running on 1, 4, 8 and 16 nodes for the first synthetic test case (Fig. 3a), with the 16 periods divided evenly among nodes (e.g., with 4 nodes, each solves for 4 periods). Tests were conducted on a small PC-clusters and a supercomputer (SGI Altix 4700) at the Earthquake Research Institute, University of Tokyo. To quantify efficiency of the parallel code, we display the speedup, defined as $S = T_1/T_P$, where T_1 is the execution time of the sequential WSINV3DMT algorithm and T_P is the execution time of the parallel version, run on P processors. The idealized maximum speedup is P. However, due to computational overhead, the need for some computations to be performed only on the master node, and the time required to exchange information between nodes, S will always be less than P. Fig. 4 displays speedup versus the number of nodes. Inversions of all data (i.e., VTF + impedance, $N = 40 \times 12 \times 16 = 7680$) are plotted with solid lines. Inversions of the VTF only dataset ($N = 40 \times 4 \times 16 = 2560$, or one third the size of the joint inversion dataset) are plotted as dashed lines. We also compare speedups achieved with the two approaches for solving the normal equations: speedups obtained with the single processor Cholesky decomposition are plotted as solid symbols, while those obtained with the parallel PCG algorithm are plotted as open symbols.

For the inversion of the VTF dataset for this very small test problem, actual (wall clock) run times were about 186 min on a single node machine, 82 min on 4 nodes, 46 min on 8 nodes and 34 min on 16 nodes, resulting in speedups of about 2.2 for 4 nodes, 4 for 8

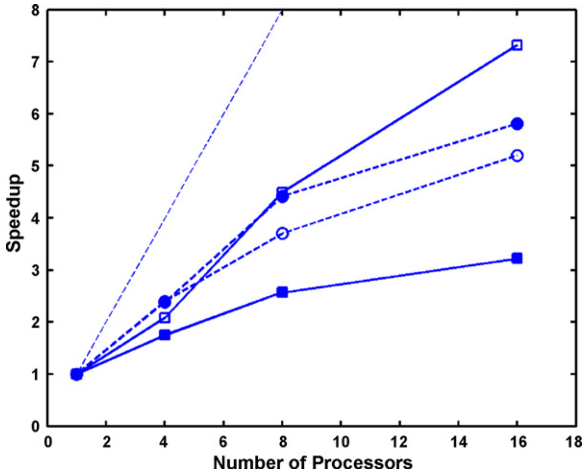


Fig. 4. Speedup versus the number of processors or nodes. Solid lines are the speedups from inversion using both VTF and impedance data ($N = 7680$). Dashed lines are the speedups from inversion using only VTF data ($N = 2560$). Results for the scheme which solves the normal equations by Cholesky decomposition on a single node (step 2.4.1.2 of Fig. 2) are plotted with solid symbols. The corresponding results with the parallel PCG solver (step 2.4.1.4 of Fig. 2) are plotted with open symbols. The thin-dashed line of slope one gives the ideal perfect speedup.

nodes and 5.4 for 16 nodes. Thus, as the number of nodes increases, the relative efficiency of additional nodes decreases. One reason for this is that the run time of the iterative forward modeling routine depends on the period of the data. Shorter periods typically require a larger number of iterations for convergence, and hence longer run times. Thus, some nodes are usually idle waiting for modeling computations to complete on other nodes, before moving on to the next step in the inversion. With fewer nodes some of the frequency-to-frequency variations average out, resulting in better balance.

Efficiencies are somewhat lower for the larger joint VTF/impedance dataset, when the serial Cholesky decomposition solver is used (solid line with solid square symbols of Fig. 4). Now the speedups are about 1.8, 2.6 and 3.2 for 4, 8 and 16 nodes, respectively, almost 50% below those achieved for the VTF only inversion. However, solving the normal equations with the parallel PCG solver (solid line with open square symbols in Fig. 4) significantly improves performance, increasing S to approximately 2, 4.5 and 7.3 for the three cases considered. In the VTF only case, where N is significantly smaller, both methods for solving the normal equations have similar performance (dashed lines in Fig. 4), and indeed the speedup is slightly greater when the single node Cholesky decomposition is used.

The difference between the two cases is readily understood. Operation counts for Cholesky decomposition scale as N^3 so computation times for the serial Cholesky decomposition in the all data case ($N = 7680$) are expected to be about 27 times greater than for the VTF only case ($N = 2560$). Other computational steps scale better with increasing N. For fixed model parameter size, total operation counts for the sensitivity calculations increase linearly in N, and formation of the cross product matrices increases as N^2. Thus, as the size of the dataset increases, run times required for the serial Cholesky decomposition step become increasingly significant, and at large enough N this step will control the overall runtime. Operation counts for a single iteration in the parallel PCG scheme scale as N^2, but overall runtimes will also depend on the number of iterations required. Although this should increase with N also, the dependence is weak, and so PCG becomes increasingly advantageous as N increases, particularly since computations for the PCG scheme can be distributed over the P processors.

The number of iterations for PCG also depends on the relative tolerance for the residual ($=||Ax - b||/||b||$) used to define convergence. We find that a tolerance of 10^{-4} results in models that are essentially identical to those obtained with the Cholesky decomposition technique. The number of iterations, and hence the run time of the parallel PCG scheme also depends on the condition number of the normal equations. For large values of the Lagrange multiplier (corresponding to a smoother model) the condition number is smaller, and the parallel solver thus converges in a small number of iterations. In contrast, when the Lagrange multiplier is very small (rough model) the parallel solver can require considerably more iterations, and solution times can exceed those for the serial Cholesky decomposition scheme. This occurred occasionally in our tests with the larger VTF/impedance dataset, but overall performance using the parallel PCG solver was much better when N is large enough.

We will not attempt to quantify more precisely how large N must be before the parallel approach to normal equation solution would be preferred. This will depend on the cluster architecture, especially on the sort of inter-processor communication used, since the parallel PCG solver requires substantial sharing of data.

In addition to reducing computational times, the parallel version also reduces the need for a large amount of memory on a single computer. Even for the small joint VTF/impedance inversion test example, about 1.5 GBytes are required for the representer and sensitivity matrices. In the parallel implementation, the required memory may be distributed over all of the nodes used. For example, with 16 nodes, each would require only 0.090 GBytes for storing the sensitivity matrix and forming cross products, almost 13 times less than required by the serial code. If the whole representer matrix is stored on a single processor (for the Cholesky decomposition, or to reduce the communication time between nodes for PCG) about 0.4 Gb are required on each node, still only a quarter required for a serial version.

The exact time speedup and per-node memory reduction factors will depend to some extent on the problem size, both in terms of model grid dimensions, and number of data. For larger problems, such as the real data EXTECH example considered below, similar performance gains were attained. For these larger problems, however, a speedup by a factor of roughly 7 means a run time that was perhaps 2–3 weeks on a single node is now reduced to 2–3 days, making inversion of realistic datasets considerably more practical. The practical impact of distributing memory is even greater. Total storage required by WSINV3D for the EXTECH example described below (joint inversion of the full impedance and VTFs) is at least 30 Gb, making this impractical on almost any shared memory machine.

4.2. *Vertical magnetic transfer function inversion*

We next consider the effectiveness of WSINV3DMT at correctly recovering resistivity when only VTF data are available. Because in practice one would not know *a priori* the correct background resistivity, we run the inversion using several prior (and starting) models. Inversion results for the synthetic VTF data from the test case of Fig. 3a are summarized in Figs. 5 and 6. Using a 50 Ω m half-space as a prior (this is intermediate between the true 10 Ω m upper layer background, and the 100 Ω m basement), inversion of VTF data reveals both the conductive body and the adjacent resistor, extending from near the surface to approximately 20 km depth. The calculated responses generated from the inverse solution of Fig. 5 fit the observed responses within 15% of the typical VTF amplitude (recall that 5% random noise was added to the synthetic data).

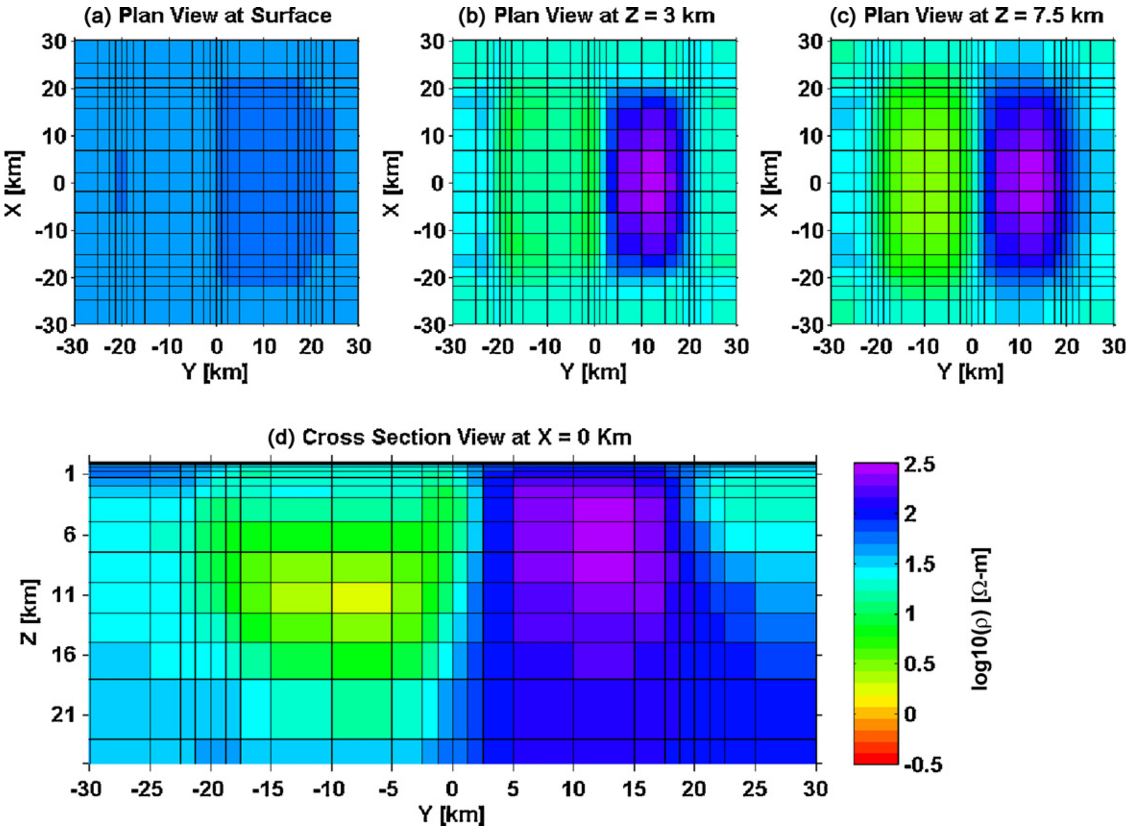


Fig. 5. An inverse solution from the VTF data alone after the 9th iterations with an RMS value of 1, fitting synthetic data generated from the model in Fig. 3a. The top panels (a)–(c) is a plan view at the surface, at 3 km and at 7.5 km depth, and the bottom panel (d) is a cross-section view cutting across the conductive block at $X = 0$ km. The solution is shown only in the central area around the anomalies, not for the full model domain.

Although both anomalies are detected in approximately the correct location, the true resistivities of Fig. 3a are not correctly estimated. However, calculating the average resistivity over the anomalous volumes we find for the inverse model of Fig. 5 an average resistivity of about 6.3 Ω m for the conductive anomaly, and of about 453 Ω m for the resistive body, while the background resistivity of the inverse model was changed only slightly from the 50 Ω m prior. Computing the volume average resistivity ratios from left to right in Fig. 5d, we obtain values of 7.9 (=50/6.3), 72 (=453/6.3) and 9 (=453/50), compared to the actual ratios (Fig. 3a) of 10 (=10/1), 100

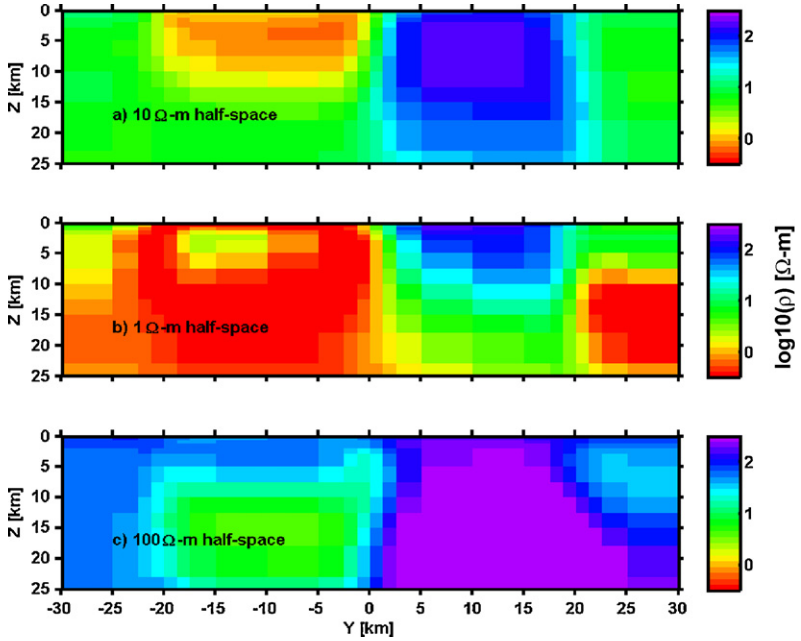


Fig. 6. Cross-sectional plots at $X = 0$ km (as in Fig. 5d) of the inverse solutions from VTF data alone, when the prior models are (a) 10 Ω m half-space, (b) 1 Ω m half-space and (c) 100 Ω m half-space.

(=100/1) and 10 (=100/10), respectively. The inversion thus results in roughly the correct structure, with approximately correct resistivity contrasts, but it does not recover the correct amplitude of either the background or the anomalies, or the actual depth extent of the anomalies.

To explore this issue further we ran the inversion on the same VTF dataset, using a range of values for the assumed half-space prior. Fig. 6 summarizes the results with cross-sectional plots of the inverse solutions at $X = 0$ km. When the prior model is the same as the correct background resistivity (i.e., a 10 Ω-m half-space in our example), the inversion quickly converges to the desired misfit within 4 iterations, even with error floors set to 5%. In this case, the inversion estimates the resistivity, and the depth extents, of the two anomalies quite well (Figs. 6a and 3a). However, the 100 Ω m basement resistivity (below 10 km depth in the synthetic test model of Fig. 3a) is not recovered—the prior resistivity of 10 Ω m remains unchanged at depth in the inverse solution. This again demonstrates that inversion of VTF data alone can only recover lateral resistivity contrasts, and is not effective at correcting resistivities, or their variations with depth.

Larger deviations of the prior model from the correct background result in even larger discrepancies in anomaly amplitudes and depths, but still generally allow the horizontal structure to be recovered. With a 1 Ω m half-space (Fig. 6b) data is fit to within 10%. Anomalies appear at very shallow depths (upper few km), with all features more conductive than their actual values. At greater depth, features with appropriate resistivity ratios are imaged, but the absolute levels are incorrectly estimated, and spurious structures appear. Using a 100 Ω m half-space as a prior, the VTF data can only be fit to within 20%. The basic structure is again recovered, but both anomalies are at greater depth (Fig. 6c) and have increased resistivity. The host resistivity is estimated to be slightly lower than the 100 Ω m starting value, but is still well above the correct value of 10 Ω m. As in the other cases, the basement resistivity remains the same as the prior model.

All of these experiments suggest that when only VTF data are available, to achieve the target misfit and recover correct amplitudes and depths, the inversion must be started with a prior model that is close to the correct host resistivity. However, even starting far from the correct background model, anomalies are recovered with the correct horizontal location and dimensions. This result is not unexpected since the vertical magnetic fields are generated where there are lateral discontinuities, but are not inherently sensitive to the profile of vertical conductivity structure.

In addition, resistivities of anomalous bodies scale with the assumed prior background (Fig. 6), and resistivity contrasts (i.e., ratios) can be close to actual values, especially if the assumed background resistivity is not too far off. However, the VTFs provide little intrinsic constraint on contrasts in the vertical direction, including the location of the top or the bottom of the anomalies. The inversion only gets these properties of the anomalies correct if something close to the correct background is used (Fig. 6a).

Performing similar experiments to those summarized in Fig. 6, but using impedance tensor data shows that these inversions are much less sensitive to the assumed prior model. This is consistent with the basic physics, as the ratio of electric to magnetic fields is intrinsically related to the resistivity profile. In spite of the well-known uncertainties in depth and absolute resistivity level that may result from local static distortions, there is by now ample evidence (e.g., Tuncer et al., 2006; Unsworth et al., 2000) that, with proper care, MT impedances can yield reliable information about conductivity-depth profiles. The same does not appear to be true in practice with VTF data, although theoretical analysis of idealized models suggests otherwise (Berdichevsky et al., 2003).

The above results suggest that VTF data will be most useful as an adjunct to impedance data, which can provide the necessary con-

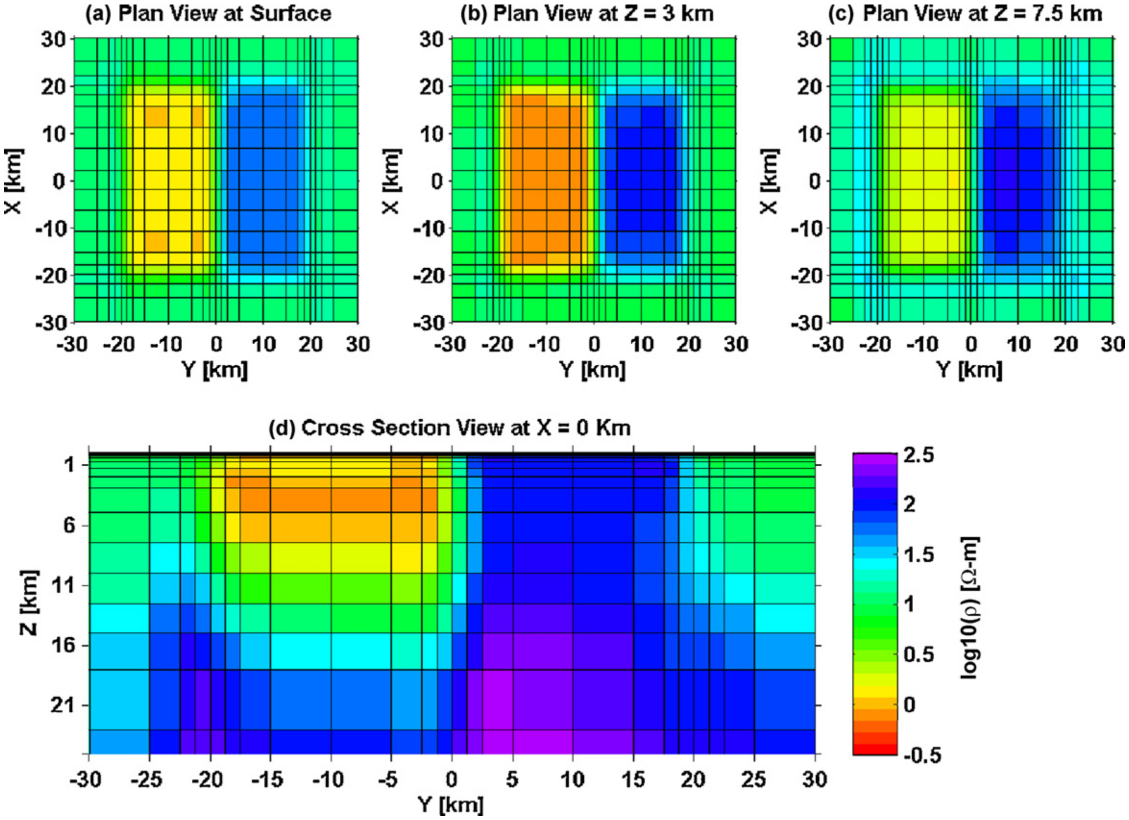


Fig. 7. Results from joint inversion of both VTF and impedance tensor data generated from the model in Fig. 3a. See caption of Fig. 4 for other details.

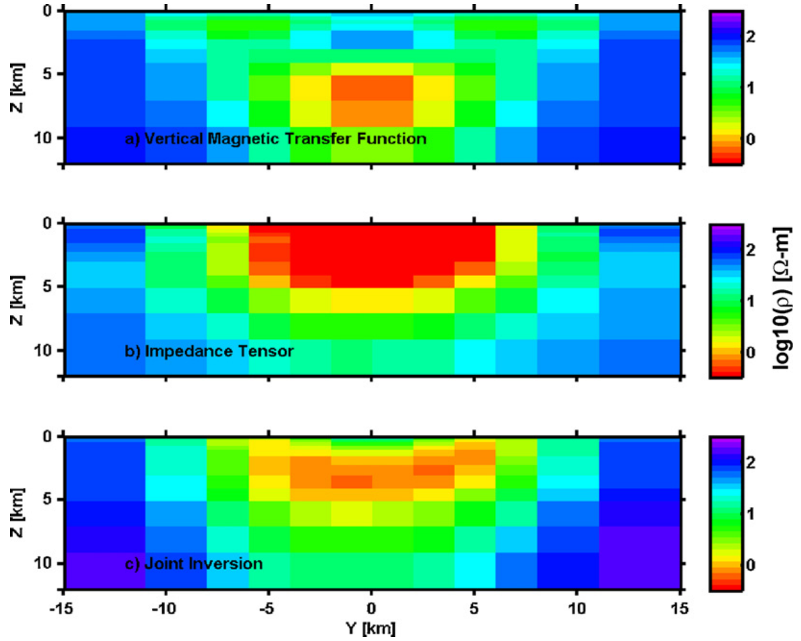


Fig. 8. Cross-sectional plots at $X = 0$ km of the inverse solution from (a) fitting the vertical magnetic transfer function alone, (b) fitting the impedance tensor alone, (c) fitting both data types. The data is generated from the synthetic model in Fig. 3b.

straint on background resistivities. As a first example, we consider joint inversion of VTF and the impedance tensor data derived for the synthetic model of Fig. 3a. As above we again tried a range of priori/initial models. Although in general the impedance tensor data can adjust the resistivity background, we still had difficulties getting the joint inversion to converge to the desired 5% misfit level, especially with priori models that differ greatly from the correct background resistivities. In this and other examples, we found that to achieve the target misfit for both data types, it was necessary to first fit the impedances to a half-space model, to determine a prior model for the joint inversion. Even with this additional step, we typically found it necessary to use increased error floors for the VTF data (but not the impedances) to achieve a normalized RMS of one.

Not surprisingly, a 50 Ω m half-space (as in Figs. 5 and 6 of Siripunvaraporn et al., 2005) yields a good fit to the synthetic impedance data for case 1. With error floors set to 15% for VTF data and 5% for impedance tensor data, the joint inversion converged to the target misfit in 5 iterations. In the final iteration (Fig. 7) the two anomalies are recovered with essentially correct background resistivities. In fact, in comparison with the inverse model obtained from inversion of just the impedance data (Fig. 6 of Siripunvaraporn et al., 2005), there is little difference. Clearly, the relatively simple structures in this synthetic example are well enough constrained already by the array of 40 MT sites that addition of the VTF data can add little. In any event, this example demonstrates the consistency of the two datasets, as both can be fit simultaneously with the same inverse solution.

Other synthetic examples demonstrate the potential benefit of joint inversion a bit more clearly. We performed three inversion tests on the second test case, with data generated for the synthetic model of Fig. 3b, as described above. Error floors were set at 10% and 5% for the VTF and the impedance data, respectively. Initial models for all runs are 50 Ω m half space. The first inversion was performed using just the VTF data, the second with just the impedance tensor, and the last with both data types. All inversion reaches the target misfit of 1 RMS. Fig. 8 displays cross-sectional plots at $X = 0$ km.

In all cases the conductor is recovered, although for the VTF case the burial depth is greater than what it should be (Fig. 8a). This again shows that the VTF data can primarily constrain the location of the conductor in the horizontal, but not the vertical. Inversion of the impedance tensor alone recovers the anomalous volume quite well (Fig. 8b), but the conductivity is noticeably above the correct value of 1 Ω m (Fig. 8b). The best results are obtained by the joint inversion, where the resistivity, shape, size and depth of the conductor are close to correct. It is not clear why this example demonstrates a benefit of including VTF data, and the other does not; possibly different results would be obtained if the experiment was repeated with different realizations of random noise added to the data, or if the locations of the MT sites were perturbed, or different initial or prior models were used. Clearly the need to satisfy additional data constraints reduces the effects of noise in the data, and is likely to improve the fidelity of the inverse solution. For more complex structure the value of additional constraints provided by the VTF inversion are even clearer, as we show next by consideration of an example with real data.

5. Numerical experiments on real data

We applied the VTF inversion to the EXTECH dataset (Tuncer et al., 2006), consisting of tensor audio-magnetotelluric (AMT) soundings for 131 stations around the McArthur River mine, Saskatchewan, Canada. The goal in this survey was to use electromagnetic data to detect and map low resistivity graphite which is indicative of unconformity-type uranium deposits. A full description of the survey, and an interpretation of this dataset based on 2-D and 3-D analysis (including inversion with WSINV3D), is given in Tuncer et al. (2006). Further efforts at 3-D interpretation are given in and recently Farquharson and Craven (2008).

Here, we invert VTF and impedance data from 16 periods (from 8000 Hz to 5 Hz) at 131 sites (Fig. 2 of Tuncer et al., 2006), comparing results obtained with the two sorts of responses, separately and in combination. We use a 1000 Ω m half-space as an initial and prior model for all runs, as previous inversion of the impedance tensor suggests that this is a reasonable average background, and should thus produce sensible results when inverting the VTF alone. For inversion of the VTF (T_{zx} and T_{zy}) only, minimum error bars were set at 15% of $(|T_{zx}|^2 + |T_{zy}|^2)^{1/2}$. The inversion required about 8 iterations

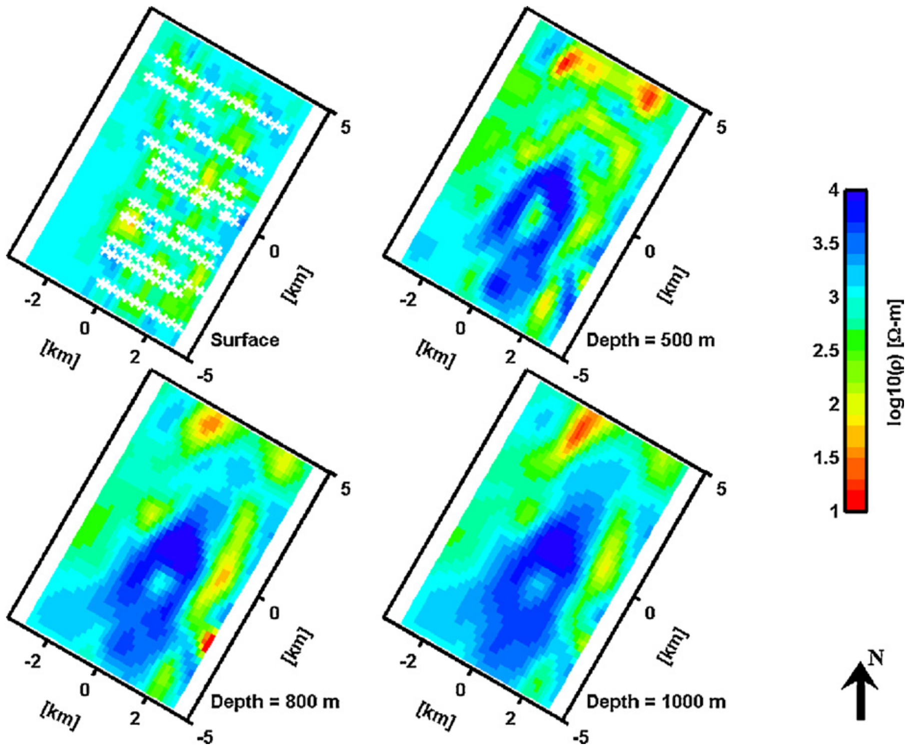


Fig. 9. The inverse solution at various depths from fitting the vertical magnetic transfer functions of the EXTECH dataset. The cross-symbols indicate the location of stations.

to converge to a minimum RMS of 1.2. Results for this inversion are given in Fig. 9.

For the second run we inverted the impedance tensor alone. In previous results using WSINV3D, reported in Tuncer et al. (2006) only the off-diagonal components (Z_{xy} and Z_{yx}) of the impedance were inverted. Here, we used all components including Z_{xx} and Z_{yy} also. The minimum error bar for this run was set at 5% of $|Z_{xy}^{1/2}Z_{yx}^{1/2}|$ for off-diagonal and 50% for diagonal terms. When the same error floors were tried for off-diagonal and diagonal terms, the misfit could not be reduced below 3 RMS. With the modified error floors, the inversion required 4 iterations to converge to the target level of 1 RMS. The resulting model is shown in Fig. 10. The last run was a joint

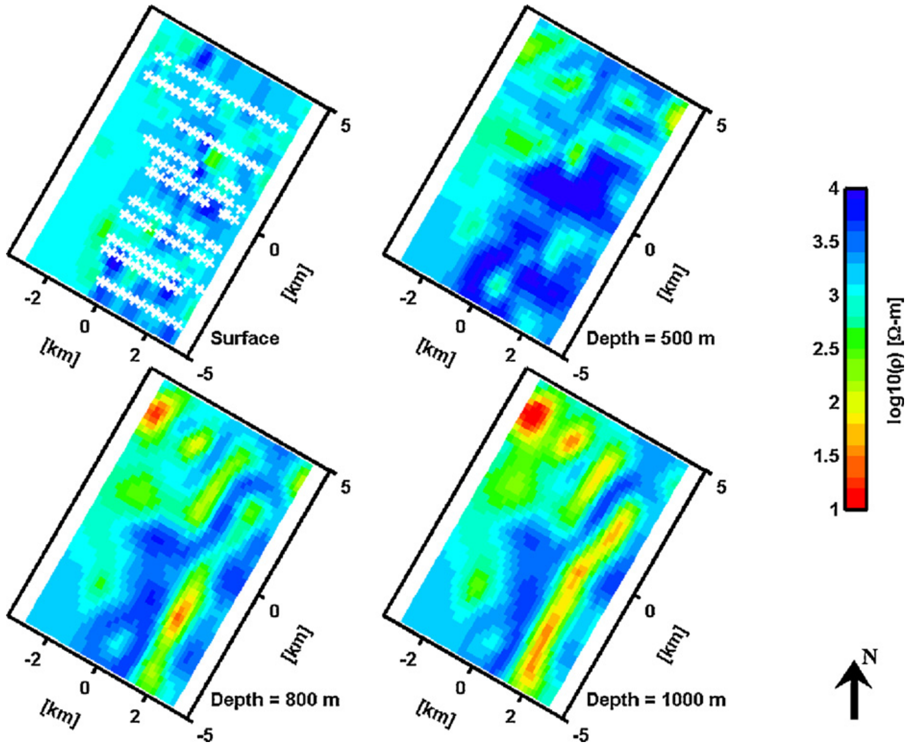


Fig. 10. The inverse solution at various depths from fitting all components of the impedance tensors of the EXTECH dataset.

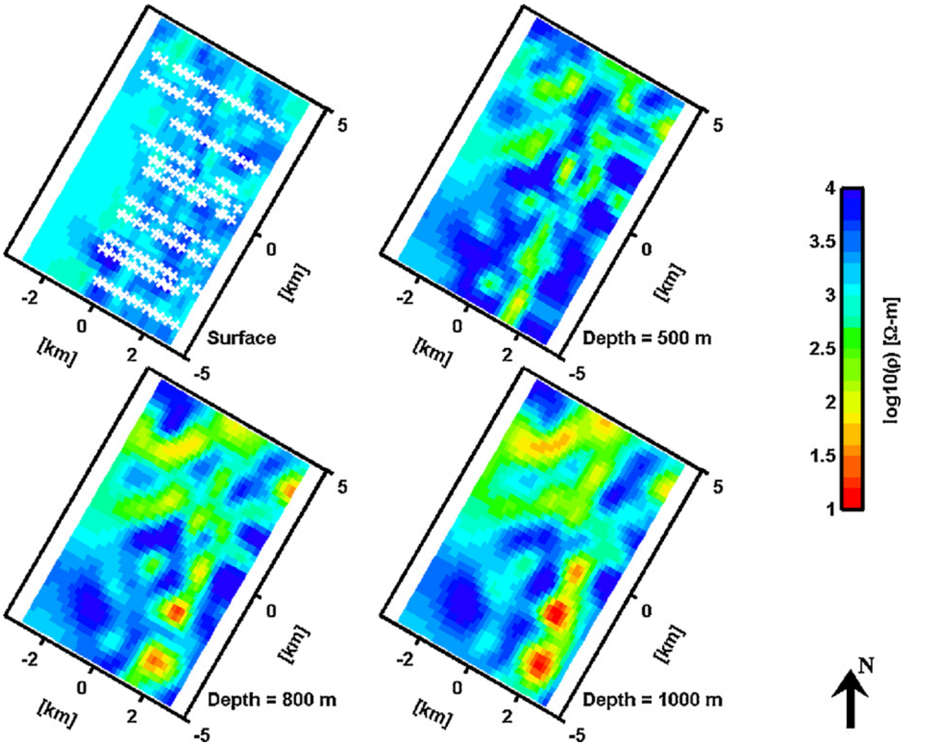


Fig. 11. The inverse solution at various depths from fitting both VTF and the impedance tensors of the EXTECH dataset.

inversion of the full impedance tensor and the vertical magnetic transfer function, with error floors set as in the first two runs. The inversion reduced the RMS misfit to 1.3 in 5 iterations. The model from the joint inversion is shown in Fig. 11.

Inverting just the impedance tensor (Fig. 10) reveals two main zones of high conductivity at 1000 m depth—an elongated feature of about 100 Ω m running perpendicular to the profiles on the east side of the model domain, and an area of variable (but

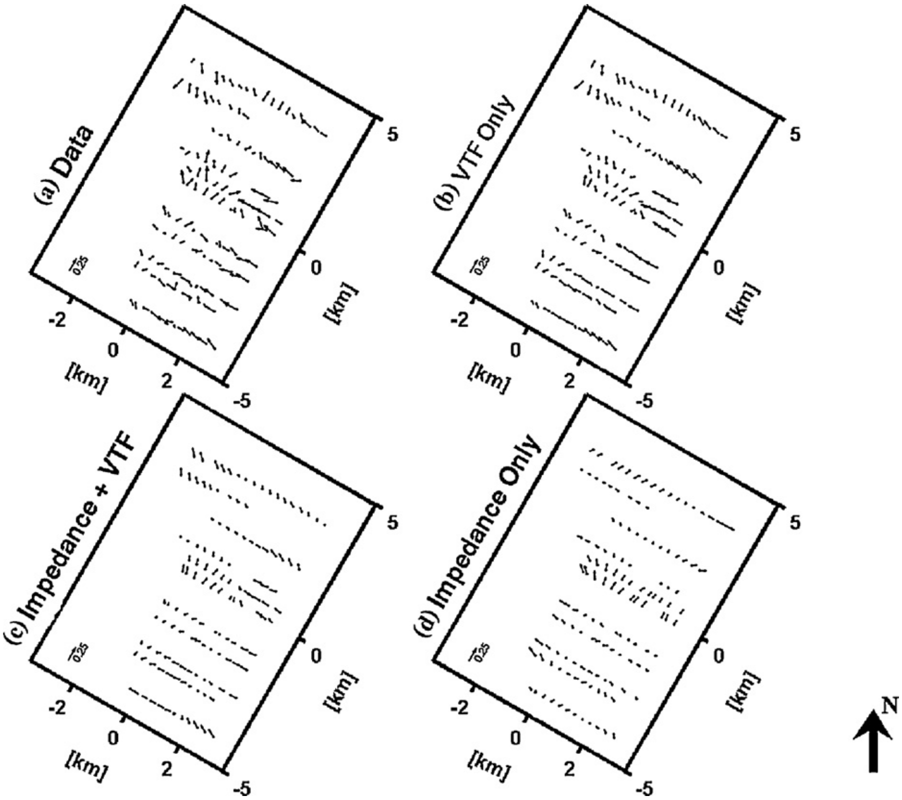


Fig. 12. The induction vectors at 100 Hz generated from (a) the observed VTF data, (b) the VTF inversion alone of Fig. 9, (c) the joint inversion of both impedance tensor and VTF data of Fig. 11, and (d) the impedance tensor inversion alone of Fig. 10. Notice that the calculated induction vectors in (d) fit the observed induction vectors more poorly.

generally higher) conductivity located in the northwest. The same features are evident, but somewhat weaker, in the 800 m layer. Similar features were obtained by inverting only the VTF data (Fig. 9). However, depth resolution appears poorer, as the inversion spreads the conductive features to shallower depths, particularly in the north, beyond the area covered by the MT profiles. The independent inversions of each data type confirm the lateral locations of the conductors. However, based on our experiments with synthetic data, the vertical position and extents of the conductive zones are almost certainly better constrained by the impedance tensor.

Results from joint inversion (Fig. 11) show increased conductivity in the same two general areas at 1000 m depth. However, the elongated conductor to the east now appears to be broken into segments, with patches of resistivity as low as 10 Ω m, separated by areas with resistivities of several hundred Ω m. In contrast, inverting impedances alone results in a more uniform (approximately 100 Ω m) continuous feature. Apparently, the VTFs cannot be fit by such a simple uniform conductor, but rather require significant along-strike variability (see Fig. 12). The feature to the north is also substantially modified by inclusion of both data types. Compared to the VTF only inversion, the depth of this feature is now clearly localized at around 1000 m, constrained by the impedance tensor. Inclusion of the VTF data also reduces peak conductivities in this area, and results in more linear conductive features which strike approximately east–west.

It is instructive to consider fits of the inverse solutions of Figs. 9–11 to the VTF data. Real induction vectors (with the Parkinson convention, so that arrows point toward conductors) are plotted in Fig. 12 for a frequency of 100 Hz, along with computed responses for the VTF only, impedance only, and joint inversions. The induction vectors are consistent with the presence of conductive features in the southeastern and northern parts of the array—e.g., note the clear reversal of vectors on most lines as they cross the elongated conductive feature at 1000 m depth (clearest in Fig. 10), and the reversal from South to North pointing vectors in the Northern corner of the study area. However, as noted by Tuncer et al. (2006) patterns in the observations are much more complex than can be reproduced by simple 3-D models. The VTF only inversion reproduces almost all of the complexity seen in the data (Figs. 12a and b). The joint inversion results in a smoother VTF response, and a slightly poorer fit to the data (Fig. 12; this is consistent with the larger error floor assumed in this case), but again, significant features in the data are reproduced in the fitted response. In contrast, the solution obtained from fitting the impedance tensor data alone (Fig. 12c) fits the VTF observations considerably less well, suggesting that the result from the joint inversion (Fig. 11) is more reliable than that from the impedance tensor alone (Fig. 10). A more detailed interpretation of this dataset is beyond the scope of this paper. See Tuncer et al. (2006) and Farquharson and Craven (2008) for further interpretation and discussion of the EXTECH data, and Craven et al. (2006) for comparison of inversion techniques using this data.

6. Conclusions

Experiments on both synthetic and real data show that inverting VTFs alone can recover anomalous structures, particularly if the prior model is close to the correct background or host value. In general, the qualities of the inverse solution obtained from VTF data alone are inferior to those obtained from inverting the impedance tensor alone. Vertical magnetic fields are generated whenever lateral conductivity gradients align with the normal inducing field. Thus, VTFs are sensitive to horizontal structures, and to some extent to resistivity contrasts, but not to depths or absolute values of resistivity. If some constraint on host resistivity can be provided, either *a priori*, or through inversion of impedances, the VTF data can result in accurate 3-D imaging of the anomalous structures. Joint inversion of VTFs and the impedance tensor can help constrain subsurface structures, as shown in both synthetic and real data examples. In cases with very simple structures which are already well resolved by the impedance data VTFs add little to the inverse solution. However, with more realistic levels of complexity, as exemplified by the EXTECH data, inclusion of VTF data results in significant modifications to the inverse solution. Because the joint inversion model fits both datasets, it is likely to be more reliable.

One issue that deserves further investigation is the inability of the inversion to fit synthetic VTF data to within the tolerance implied by the noise level, which of course is well known in synthetic tests. We speculate that the VTF data can only be fit perfectly when the background resistivity is correct—implying at least a weak sensitivity of this sort of data to the background, as the analysis of Berdichevsky et al. (2003) in fact showed. In the case of using the wrong background resistivity (for which the data have little sensitivity) no nearby model parameters can provide a better fit, perhaps after adjusting conductivities of the anomalous bodies to roughly fit the VTFs, the Occam inversion is stuck in a local minimum of the penalty functional, and cannot escape from. It would be useful to compare other search algorithms (e.g., NLCG) to see if they suffered from similar problems.

A significant drawback with WSINV3DMT has been the large amount of memory required to store the sensitivity matrix, and the extensive computational time required for forward and sensitivity solutions. These drawbacks can be ameliorated by adapting the code to run with MPI to on parallel systems. We have parallelized the computations over frequencies, requiring no significant changes to our forward modeling routine. This approach is probably most appropriate for small cluster type machines. To make efficient use of a cluster or supercomputer with more than a few tens of processors would require different approaches, such as decomposing the modeling domain for the forward solver. We have also parallelized computation of cross products, sharing rows of the sensitivity computed on separate nodes to compute blocks of the coefficient matrix needed for the Gauss–Newton normal equations. The resulting dense system of normal equations can be solved on the master node, or using a parallel solver based on iterative methods. The optimal choice here depends on the size of the data space, with the iterative parallel solver only efficient for large datasets. The speedup of the code on a test dataset with 16 periods is nearly linear (with a coefficient of roughly 0.5) for up to 8 processors, but rolls over for a further increase to 16 processors. Even so, the parallelization should make use of the code on realistic 3-D datasets significantly more practical.

Acknowledgments

This research has been supported by a grant from the Thailand Research Fund (RMU5080025) to W.S., and by grant DE-FG02-06ER15819 from the U.S. Dept. of Energy to G.E. Other partial supports for this work to W.S. are from Thai National Grid Center, Thailand and EMpulse, Canada. The authors would like to thank Jim Carven for allowing us to use EXTECH data sets, and Collin Farquharson, George Helffrich, and an anonymous reviewer for comments that improved the manuscript.

References

Bedrosian, P.A., Unsworth, M.J., Egbert, G.D., Thurber, C.H., 2004. Geophysical images of the creeping segment of the San Andreas fault: implications for the role of crustal fluids in the earthquake process. Tectonophysics 385, 137–158.

Berdichevsky, M.N., Dmitriev, V.I., Golubtsova, N.S., Mershchikova, N.A., Pushkarev, P.Yu., 2003. Magnetovariational sounding: new possibilities, Izvestiya. Physics of the Solid Earth 39, 701–727.

Choi, J., Moon, S., 1997. A parallel Cholesky factorization routine with a new version of PB-BLAS. In: International Conference on Parallel and Distributed Systems (ICPADS'97), pp. 52–58.

Constable, C.S., Parker, R.L., Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data. Geophysics 52, 289–300.

Craven, J.A., Farquharson, C., Mackie, R., Siripunvaraporn, W., Tuncer, V., Unsworth, M., 2006. A comparison of one-, two- and three-dimensional modeling of audiomagnetotellurics data collected at the world's richest uranium mine, Saskatchewan, Canada. In: 18th International Workshop on Electromagnetic Induction in the Earth, El Vendrell, Spain, 17–23 September.

deGroot-Hedlin, C., Constable, S., 1990. Occam's inversion to generate smooth, two-dimensional models from magnetotelluric data. Geophysics 55, 1613–1624.

Farquharson, C.G., Craven, J.A., 2008. Three-dimensional inversion of Magnetotelluric data for mineral exploration: an example from the McArthur River uranium deposit, Saskatchewan, Canada. J. Appl. Geophys., doi:10.1016/j.jappgeo.2008.02.002.

Heise, W., Caldwell, T.G., Bibby, H.M., Bannister, S.C., 2008. Three-dimensional modeling of magnetotelluric data from the Rotokawa geothermal field, Taupo volcanic zone, New Zealand. Geophys. J. Int. 173, 740–750.

Newman, G.A., Alumbaugh, D.L., 2000. Three-dimensional magnetotelluric inversion using non-linear conjugate gradients. Geophys. J. Int. 140, 410–424.

Oskooi, B., Perdersen, L.B., 2005. Comparison between VLF and RMT methods. A combined tool for mapping conductivity changes in the sedimentary cover. J. Appl. Geophys. 57, 227–241.

Parker, R.L., 1994. Geophysical Inverse Theory. Princeton University Press.

Parkinson, W.D., 1959. Directions of rapid geomagnetic variations. Geophys. J. Roy. Astronom. Soc. 2, 1–14.

Patro, Egbert, G., 2008. Regional conductivity structure of Cascadia: preliminary results from 3D inversion of USArray transportable array magnetotelluric data. Geophys. Res. Lett. 35, L20311, doi:10.1029/2008GL035326.

Pous, J., Heise, W., Schnegg, P., Munoz, G., Marti, J., Soriano, C., 2002. Magnetotelluric study of the Las Canadas caldera (Tenerife, Canary Islands): structural and hydrogeological implications. Earth Planet. Sci. Lett. 204, 249–263.

Rodi, W.L., 1976. A technique for improving the accuracy of finite element solutions for Magnetotelluric data. Geophys. J. Roy. Astr. Soc. 44, 483–506.

Rodi, W.L., Mackie, R.L., 2001. Nonlinear Conjugate Gradients Algorithm for 2-D magnetotelluric inversion. Geophysics 66, 174–187.

Siripunvaraporn, W., Egbert, G., 2000. An efficient data-subspace inversion method for 2-D magnetotelluric data. Geophysics 65, 791–803.

Siripunvaraporn, W., Egbert, G., Lenbury, Y., 2002. Numerical accuracy of magnetotelluric modeling: a comparison of finite difference approximations. Earth Planets Space 54, 721–725.

Siripunvaraporn, W., Egbert, G., Lenbury, Y., Uyeshima, M., 2005. Three-dimensional magnetotelluric inversion: data-space method. Phys. Earth Planet. Interiors 150, 3–14.

Siripunvaraporn, W., Egbert, G., 2007. Data space conjugate gradient inversion for 2-D magnetotelluric data. Geophys. J. Int. 170, 986–994.

Toh, H., Baba, K., Ichiki, M., Motobayashi, T., Ogawa, Y., Mishina, M., Takahashi, I., 2006. Two-dimensional electrical section beneath the eastern margin of Japan sea. Geophys. Res. Lett. 33, L22309.

Tuncer, V., Unsworth, M.J., Siripunvaraporn, W., Craven, J.A., 2006. Exploration for unconformity-type uranium deposits with audiomagnetotelluric data: a case study from the McArthur River mine, Saskatchewan, Canada. Geophysics 71, B201–B209.

Unsworth, M., Bedrosian, P., Eisel, M., Egbert, G., Siripunvaraporn, W., 2000. Along strike variations in the electrical structure of the San Andreas Fault at Parkfield, California. Geophys. Res. Lett. 27 (18), 3021–3024.

Uyeshima, M., Ogawa, Y., Honkura, Y., Koyama, S., Ujihara, N., Mogi, T., Yamaya, Y., Harada, M., Yamaguchi, S., Shiozaki, I., Noguchi, T., Kuwaba, Y., Tanaka, Y., Mochido, Y., Manabe, N., Nishihara, M., Saka, M., Serizawa, M., 2005. Resistivity imaging across the source region of the 2004 Mid-Niigata Prefecture earthquake (M6.8), central Japan. Earth Planets Space 57, 441–446.

Wannamaker, P.E., Booker, J.R., Jones, A.G., Chave, A.D., Filloux, J.H., Waff, H.S., Law, L.K., 1989. Resistivity cross section through the Juan de Fuca subduction system and its tectonic implications. J. Geophy. Res. 94 (B10), 14,114–14,127.

Wannamaer, P.E., Hasterok, D.P., Johnston, J.M., Stodt, J.A., Hall, D.B., Sodergren, T.L., Pellerin, L., Maris, V., Doerner, W.M., Groenewold, K.A., Unsworth, M.J., 2008. Lithospheric dismemberment and magmatic processes of the Great Basin-Colorado Plateau transition, Utah, implied from magnetotellurics. Geochem. Geophys. Geosyst. 9 (5), doi:10.1029/2007GC001886, Q05019.

Regional conductivity structure of Cascadia: Preliminary results from 3D inversion of USArray transportable array magnetotelluric data

Prasanta K. Patro[1,2] and Gary D. Egbert[1]

Received 11 July 2008; revised 27 September 2008; accepted 1 October 2008; published 31 October 2008.

[1] In conjunction with the USArray component of EarthScope, long period magnetotelluric (MT) data are being acquired in a series of arrays across the continental US. Initial deployments in 2006 and 2007 acquired data (10–10,000 s) at 110 sites covering the US Pacific Northwest, distributed with the same nominal spacing as the USArray seismic transportable array (~75 km). The most striking and robust features revealed by initial three-dimensional inversion of this dataset are extensive areas of high conductivity in the lower crust beneath all of southeastern Oregon, and beneath the Cascade Mountains, contrasting with very resistive crust in Siletzia and the Columbia Embayment. Significant variations in upper mantle conductivity are also revealed by the inversions, with the most conductive mantle beneath the Washington backarc, and the most resistive corresponding to subducting oceanic mantle. **Citation:** Patro, P. K., and G. D. Egbert (2008), Regional conductivity structure of Cascadia: Preliminary results from 3D inversion of USArray transportable array magnetotelluric data, *Geophys. Res. Lett.*, *35*, L20311, doi:10.1029/2008GL035326.

1. Introduction

[2] The USArray component of EarthScope is a continental-scale geophysical observational program that will provide new constraints on the structure and evolution of the North American continent. As an adjunct to the seismic transportable array (TA), which over the next decade will cover the continental US with temporary seismic observatories at an approximate spacing of 75 km, long period (10–20,000 s) magnetotelluric (MT) data will be acquired in selected areas, with comparable site densities. The first MT TA data were acquired at 30 sites in eastern Oregon in the summer of 2006, followed by 80 sites covering western Oregon, all of Washington State and western Idaho in summer 2007. In contrast to traditional MT surveys, where sites are concentrated along one or a few profiles, sites were widely spaced to provide quasi-uniform coverage of the entire area. This array configuration, and the geologic complexity of the study area (Figure 1), effectively demands a three-dimensional (3D) interpretation. Here we present the results of our preliminary efforts in this direction.

[3] The MT array traverses a wide range of geologic environments, from the subducting Juan de Fuca (JDF) plate in the west, across the Cascade volcanic arc, and into the Columbia Plateau, High Desert, western Snake River Plain and Northwest Basin and Range provinces to the east. The modern position of the subduction zone dates from approximately 48 Ma, when a large fragment of thickened oceanic lithosphere was accreted to the Pacific Northwest margin [*Madsen et al.*, 2006] near the end of Laramide orogeny. This accreted oceanic terrane, which fills the Columbia Embayment and forms the modern forearc basement in NW Oregon and SW Washington, is sometimes referred to loosely as Siletzia (e.g., E. Humphreys, Relation of flat subduction to magmatism and deformation in the western USA, submitted to *Backbone of the Americas*, 2008). From 17 to 12 million years ago, great flood basalts (over 200,000 km^3) erupted in Washington and Oregon, covering much of the Columbia Plateau [*Camp and Ross*, 2004]. These eruptions have been followed by age progressive silicic volcanism which continues to the present day and has resulted in the Snake River Plain (terminating in the east at Yellowstone [*Pierce and Morgan*, 1992]) and the High Lava Plains of eastern Oregon (terminating in the west at Newberry Volcano [*Jordan et al.*, 2004]).

[4] In a broader context, much of the crust in the Western US is rapidly deforming, with widespread extension in the Basin and Range (BR), which the southeast corner of the array intersects, and a broad zone of right-lateral shear to the south extending from the San Andreas Fault deep into the continental interior [*Humphreys and Coblentz*, 2007]. In contrast, Siletzia has retained sufficient strength to avoid deformation, accommodating right-lateral shear through clockwise block rotation which continues to this day [*Wells et al.*, 1998, *McCaffrey et al.*, 2007].

2. Magnetotelluric Data and Analysis

[5] MT data were acquired by a commercial contractor (GSY-USA) using conventional long period MT instruments based on fluxgate magnetometers. Time series data (typically of three weeks duration) were processed using a standard robust remote reference approach [*Egbert*, 1997], resulting in most cases in smooth response curves over the period range 10–10,000 s. Although there are significant site-to-site static shifts in apparent resistivities [e.g., *Bahr and Simpson*, 2005], spatial maps of phases are generally well behaved, and exhibit large scale coherent features (see auxiliary material[1], Figures S1 and S2).

[6] Induction vectors, which are computed from the ratio of vertical to horizontal magnetic field components, are indicative of lateral conductivity contrasts. For the Cascadia array these are strongly affected by the ocean in the western part of the array, but also reveal other substantial conductive

[1] College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon, USA.

[2] Now at National Geophysical Research Institute, Hyderabad, India.

0094-8276/08/2008GL035326

[1] Auxiliary materials are available in the HTML. doi:10.1029/2008GL035326.

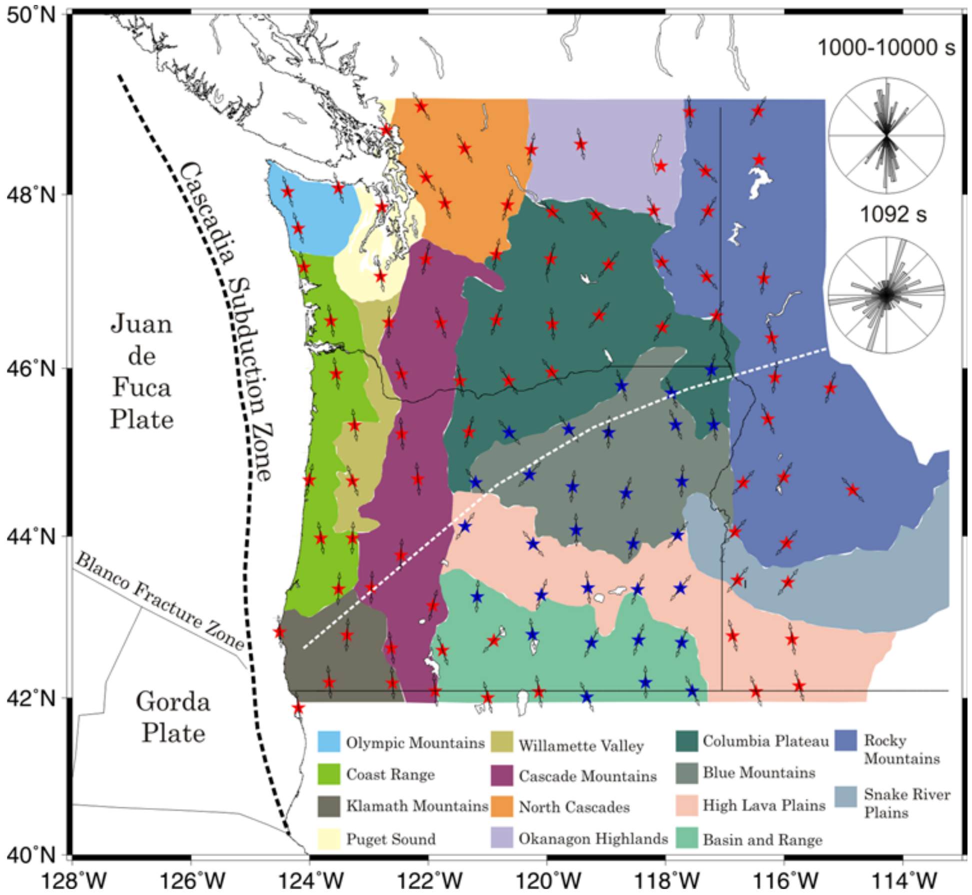


Figure 1. Location of MT sites collected in 2006 (blue stars) and 2007 (red stars) on a map of physiographic provinces [after *Rosenfeld*, 1985]. Black arrows give geo-electric strike directions determined by fitting the distortion model of *Smith* [1997] for periods of 1000–10000 s. The top inset shows the distribution of strike directions, which have a 90° ambiguity. The bottom inset is a rose diagram for real induction vectors, which point towards conductive features, for a period of 1092 s.

anomalies with varying orientations, as summarized in Figure 1 (bottom inset). Geo-electric strike analysis based on tensor decomposition [e.g., *Smith*, 1997] further confirms (Figure 1, top inset) that there is no consistent geo-electric strike that would allow two-dimensional interpretation of this dataset.

3. 3D inversion

[7] We used WSINV3D, a 3-D regularized inversion program [*Siripunvaraporn et al.*, 2005], to fit the 4 complex impedance tensor components for the 109 sites with data of acceptable quality. The model domain has total dimensions 1460 × 1590 × 550 km consisting of $N_x = 80$, $N_y = 78$ horizontal grid cells with nominal grid spacing in the central part of the domain approximately 12 km. The Pacific Ocean, with realistic bathymetry and conductivity (3.33 S/m), extends 480 km west of the coast. In the vertical the mesh has $N_z = 34$ layers, plus 7 additional layers for the air. Impedances for 8 periods (100–8000 s) were selected for inversion. Even for this relatively limited data subset and coarse model resolution the serial inversion code required 11 days per (outer loop) iteration on a single (2.8 GHz) processor PC, using essentially all of the 16 Gb of available RAM.

[8] Actual impedance estimation errors were used to normalize data misfits, and a half space of 100 ohm-m (except for the ocean) was used as a prior. As discussed in the auxiliary material, we experimented with several variants on the model covariance. The smoothest inverse solution, computed with larger horizontal decorrelation length scales, is shown in Figure 2. Alternate runs with less horizontal smoothing fit the data somewhat better (Figure S3), but all inverse solutions were qualitatively similar, particularly with regard to the large scale features emphasized below. The computed responses fit the observed signal well for periods up to a few thousand seconds (Figure S1). At longer periods fits are poorer, particularly for some sites near the coast (Figure S2). We also verified, by forward modeling, that the induction vectors were fit at least qualitatively by the inverse solutions.

4. Results and Interpretation

[9] The most prominent feature revealed by the inversion is an extensive lower crustal conductor (C1 in Figure 2), which occupies the triangular region southeast of the dotted white line extending from the coast near the California border to the eastern edge of the array, near the Oregon-Washington border (Figures 1 and 2). This conductive

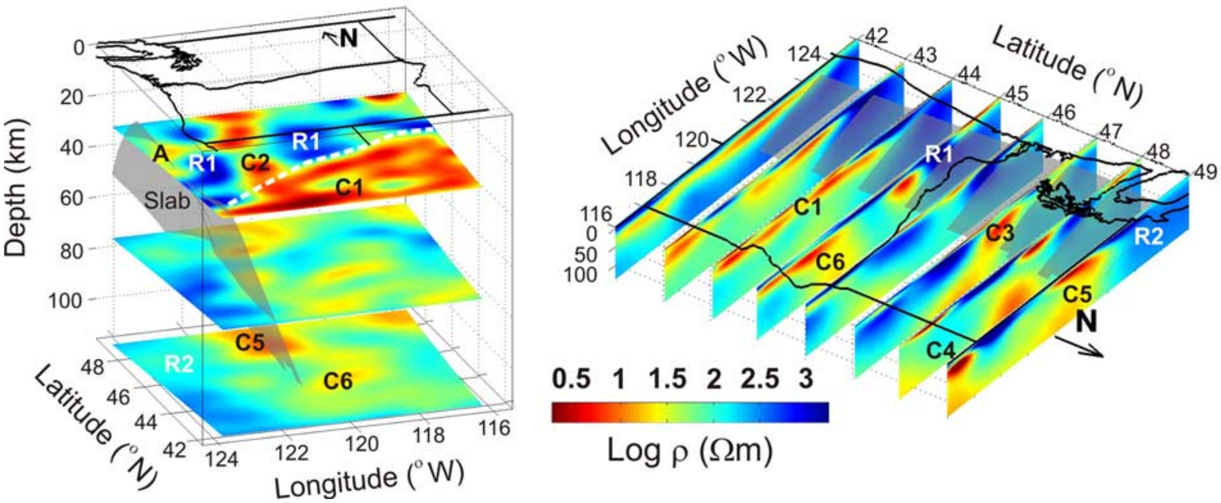


Figure 2. 3D resistivity image of the Pacific Northwest USA derived from the 3D inversion, plotted as slices of (left) constant depth and (right) constant latitude. Slab geometry from *McCrory et al.* [2003]. A: conductive zone in the forearc; C1–C6 and R1–R2 conductive and resistive features discussed in the text.

feature extends beneath the northwest BR, High Lava Plains, Western Snake River Plain, and Blue Mountains. The contact between C1 and more resistive crust to the northwest is interpreted as the southern boundary of the oceanic accreted terrane, i.e., Siletzia. Vertical conductivity profiles for selected physiographic provinces (Figure 1), computed by geographic averaging of the 3D inversion results, are given in Figure 3. The integrated conductance of the lower crustal layers is over 3000 S beneath the NWBR, and somewhat less to the north. In the NWBR in particular the zone of high conductivity appears to extend into the upper mantle, but further analysis will be required to verify this possibility.

[10] Based on higher frequency MT studies across the Cascade Range and surrounding geological provinces, *Stanley et al.* [1990] inferred a zone of high conductivity in southeastern Oregon below depths of about 15–20 km. Our results refine this picture significantly, revealing the broad spatial extent, approximate thickness, and high total conductance of this layer, which is quite similar in these respects to those seen in the lower crust elsewhere in the BR, and most probably reflects similar processes. *Wannamaker et al.* [2008] infer a lower crustal conductance in the eastern Great Basin and Transition Zone of ~3000 S and suggest that these high conductivities result from magmatic underplating associated with BR extension. As upward migrating magmas crystallize at the base of the crust several volume percent of H_2O-CO_2, highly saline brines are exsolved. With appropriate (interconnected) pore geometry these fluids (possibly with some contribution from partial melt near the moho) can easily account for the observed high conductivities [*Wannamaker et al.*, 2008].

[11] Note that the lower crustal conductor extends northward beyond what is normally considered to be the BR, albeit with reduced amplitude. This is consistent with the interpretation of Humphreys (submitted manuscript, 2008) that the interior shear zone in California broadens across NW Nevada and SE Oregon to accommodate rotation of the large strong crustal block that is Siletzia, resulting in faults of releasing orientation, and effective integration of this zone with Basin and Range extension (Humphreys, submitted manuscript, 2008). The zone of enhanced lower crustal conductivity thus correlates with weakened continental crust, and coincides with the zone of active crustal deformation.

[12] An elongate (N–S) conductive zone in the lower crust beneath the Cascade axis is also delineated (C2–C3 in Figure 2). This feature, which exhibits significant variability

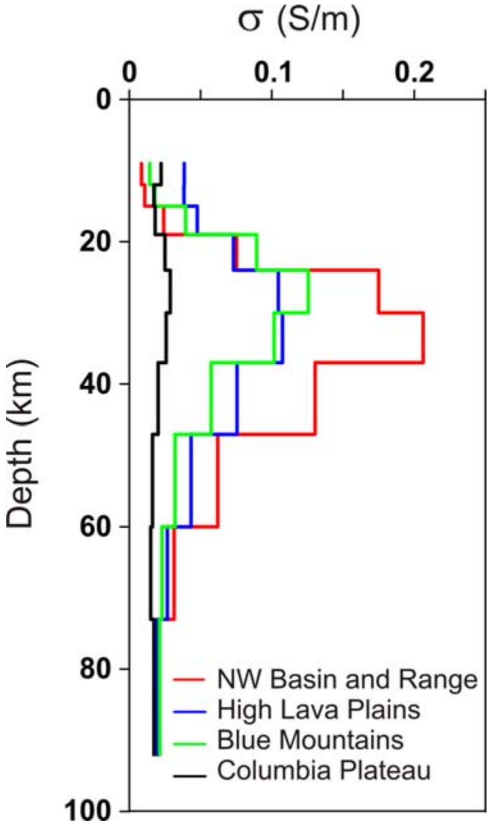


Figure 3. Vertical conductivity profiles for selected physiographic provinces.

along axis, also most likely reflects the presence of interconnected fluids, in this case from the subducting slab [*Wannamaker et al.*, 1989]. In central Oregon, near where the EMSLAB MT profile also imaged a zone of high conductivity in the lower crust beneath the high Cascades [*Wannamaker et al.*, 1989], C2 extends deeper, and merges with the lower crustal conductor to the east. In the north the Cascades conductive anomaly is more pronounced, and extends into the upper crust (C3 in Figure 2, right) where it coincides with the Southwest Washington Cascades Conductor (SWCC) [*Stanley et al.*, 1990; *Egbert and Booker*, 1993]. Similar to *Stanley et al.* [1990], the SWCC appears in the 3D model as an upper crustal feature just west of the Cascades, but then dips to the east, possibly even connecting to high conductivities in the upper mantle beneath the Columbia Plateau. The shallow part of the SWCC was interpreted by *Stanley et al.* [1990] to be a late Createaceous to early Eocene forearc basin and accretionary prism system sutured against pre-Eocene North America during accretion of Siletzia. It is likely that the deeper parts of the SWCC have a distinct cause (i.e., fluids associated with subduction and arc magmatism), given the near ubiquity of high conductivities beneath the arc.

[13] Most of the forearc is highly resistivity (R1) coinciding with the thick crust and high seismic velocities [*Parsons et al.*, 1999] of the Siletz terrane. There is some suggestion of higher conductivity above the slab further to the west, along much of the margin (A in Figure 2). This would be consistent with the zone of low resistivity imaged above the JDF plate by the EMSLAB MT profile which *Wannamaker et al.* [1989] inferred to be due to dewatering of subducted sediments (and possibly also mineral dehydration), but it may also result from accreted marine sedimentary rocks in the deformation front offshore. Given the wide station spacing and limited high frequency content of the data used for the 3D inversion, such (important) details are poorly resolved.

[14] There are several other zones of enhanced crustal conductivity evident in Figure 2. For example, a crustal conductor (C4) is evident near the northeast corner of the array. This feature appears to be shallower (upper crustal) near the Canadian border, but deepens as it extends to the southeast to at least 47.5N, where it perhaps then connects to more conductive features in the mantle. A similar crustal feature was identified by *Gough et al.* [1989] from EMSLAB magnetic variation array data as the southern termination of a prominent conductive feature (the Southern Alberta-British Columbia conductor) mapped in Canada with MV array data. High conductivity in the upper crust is also evident in the core rocks of the Olympic peninsula (47–48N, near the Pacific coast [*Aprea et al.*, 1998]).

[15] The oceanic mantle subducting beneath the North American continent is clearly more resistive (R2) than the adjacent continental mantle, to depths of at least 150 km. Perhaps the most striking feature in the mantle is a zone of high conductivity (C5) in the Washington backarc above the subducting JDF plate. This is consistent with elevated mantle conductivities reported for MT profiles just to the north by *Soyer and Unsworth* [2006], who suggested shallow convecting asthenosphere [*Currie et al.*, 2004] as the cause. To the extent that C5 continues at all into Oregon, this feature has reduced amplitude, and appears broken up and shifted to the east. There is also a circular conducting feature surrounded by a ring of more resistive mantle in central Oregon (C6). This pattern is qualitatively similar to variations in seismic velocities imaged by *Roth et al.* [2008] at similar mantle depths beneath Oregon. However, C6 is offset somewhat to the northeast relative to the lowest seismic velocities, which appear directly beneath Newberry Caldera, and were inferred to result from partial melts, concentrated in this area due to the combination of fluids released from the downgoing slab and elevated asthenospheric mantle temperatures beneath the High Lava Plains. These deeper mantle features in the resistivity images deserve more careful investigation, including further tests to verify that they are truly required of the data, and how well their position is resolved.

5. Conclusions

[16] In spite of the wide site spacing and limited control over near-surface distorting structures, a very sensible and coherent large scale picture of regional scale conductivity variations in the Pacific NW US results from 3D inversion of the USArray TA MT data. Major crustal features in the 3D inverse solution are generally consistent with previous higher resolution EM investigations in Cascadia [e.g., *Wannamaker et al.*, 1989; *Stanley et al.*, 1990], but the broad spatial coverage provides valuable new insights into the geoelectric structure of the region. *Gough et al.* [1989] inferred many of the same large scale features from the EMSLAB MV array, including the relatively conductive NWBR and Cascades. However, the interpretation by these authors was necessarily more qualitative, with very limited depth resolution—e.g., they inferred that the NWBR conductor was in the mantle, and they had no constraint on its conductance. The view of the mantle provided here, with a very clear delineation of more resistive oceanic mantle, and the variation of backarc conductivity from north to south, are in fact more novel, as previous MT investigations in this area have not had sufficient aperture to effectively explore to these depths. However, these deeper features have a more subtle expression in the MT data, and data fits are poorer at long periods. Further inversion studies, including exploration of issues of mantle anisotropy, are clearly warranted.

[17] While resolution of fine scale details, especially in the upper crust, will clearly be limited by the wide station spacing and the lack of high frequency data that will be collected by USArray, our initial inversion results are extremely encouraging. The regional scale MT array data that will be collected over the next few years, will, in conjunction with further development of 3D inversion capabilities, provide important new constraints on physical state and composition—in particular with regard to fluid content—of the North American crust and upper mantle.

[18] **Acknowledgments.** Thanks to Weerachai Siripunvaraporn for providing the 3D inversion code. Support for P. K. Patro while at Oregon State University was made possible through grants to G. D. Egbert from the National Science Foundation (NSF-EAR0345438) and the U.S. Department of Energy (DE-FG02-06ER15819).

References

Aprea, C., M. Unsworth, and J. Booker (1998), Resistivity structure of the Olympic Mountains and Puget Lowlands, *Geophys. Res. Lett.*, *25*, 109–112.

Bahr, K., and F. Simpson (2005), *Practical Magnetotellurics*, 270 pp., Cambridge Univ. Press, Cambridge, U. K.

Camp, V. E., and M. E. Ross (2004), Mantle dynamics and genesis of mafic magmatism in the intermontane Pacific Northwest, *J. Geophys. Res.*, *109*, B08204, doi:10.1029/2003JB002838.

Currie, C. A., K. Wang, R. D. Hyndman, and J. He (2004), The thermal effects of steady-state slab-driven mantle flow above a subducting plate: The Cascadia subduction zone and backarc, *Earth. Planet. Sci. Lett.*, *223*, 35–48.

Egbert, G. D. (1997), Robust multiple-station magntotelluric data processing, *Geophys J. Int.*, *130*, 475–496.

Egbert, G. D., and J. R. Booker (1993), Imaging crustal structure in southwestern Washington with small magnetometer arrays, *J. Geophys. Res.*, *98*(B9), 15,967–15,985.

Gough, D. I., D. M. McKirdy, D. V. Woods, and H. Geiger (1989), Conductive structures and tectonics beneath the EMSLAB land array, *J. Geophys. Res.*, *94*(B10), 14,099–14,110.

Humphreys, E. D., and D. D. Coblentz (2007), North American dynamics and western U.S. tectonics, *Rev. Geophys.*, *45*, RG3001, doi:10.1029/2005RG000181.

Jordan, B. T., A. L. Grunder, R. A. Duncan, and A. L. Deino (2004), Geochronology of age-progressive volcanism of the Oregon High Lava Plains: Implications for the plume interpretation of Yellowstone, *J. Geophys. Res.*, *109*, B10202, doi:10.1029/2003JB002776.

Madsen, J. K., D. J. Thorkelson, R. M. Friedman, and D. D. Marshall (2006), Cenozoic to Recent configuration in the Pacific Basin: Ridge subduction and slab window magmatism in western North America, *Geosphere*, *2*, 11–34, doi:10.1130/GES00020.1.

McCaffrey, R., A. I. Qamar, R. W. King, R. Wells, G. Khazaradze, C. A. Williams, C. W. Stevens, J. J. Vollick, and P. C. Zwick (2007), Fault locking, block rotation and crustal deformation in the Pacific Northwest, *Geophys. J. Int.*, *169*, 1315–1340.

McCrory, P. A., J. L. Blair, D. H. Oppenheimer, and S. R. Walter (2003), Depth to the Juan de Fuca slab beneath the Cascadia subduction margin: A 3-D model for sorting earthquakes [CD-ROM], *U.S. Geol. Surv. Digital Data Ser.*, *1*.

Parsons, T., R. E. Wells, M. A. Fisher, E. Flueh, and U. S. ten Brink (1999), Three-dimensional velocity structure of Siletzia and other accreted terranes in the Cascadia forearc of Washington, *J. Geophys. Res.*, *104*(B8), 18,015–18,039.

Pierce, K. L., and L. A. Morgan (1992), The track of the Yellowstone hot spot: Volcanism, faulting, and uplift, in *Regional Geology of Eastern Idaho and Western Wyoming*, edited by P. K. Link, M. A. Kuntz, and L. B. Platt, *Mem. Geol. Soc. Am.*, *179*, 1–53.

Rosenfeld, C. (1985), Landforms and geology, in *Atlas of the Pacific Northwest*, edited by A. J. Kimerling and P. L. Jackson, p. 40, Oreg. State Univ. Press, Corvallis, Oreg.

Roth, J. B., M. J. Fouch, D. E. James, and R. W. Carlson (2008), Three-dimensional seismic velocity structure of the northwestern United States, *Geophys. Res. Lett.*, *35*, L15304, doi:10.1029/2008GL034669.

Siripunvaraporn, W., G. Egbert, Y. Lenbury, and M. Uyeshima (2005), Three-dimensional magnetotelluric: Data space method, *Phys. Earth Planet. Inter.*, *150*, 3–14.

Smith, J. T. (1997), Estimating galvanic-distortion magnetic fields in magnetotellurics, *Geophys. J. Int.*, *130*, 65–72.

Soyer, W., and M. Unsworth (2006), Deep electrical structure of the northern Cascadia (British Columbia, Canada) subduction zone: Implications for the distribution of fluids, *Geology*, *34*, 53–56, doi:10.1130/G21951.1.

Stanley, W. D., W. D. Mooney, and G. S. Fuis (1990), Deep crustal structure of the Cascade range and surrounding regions from seismic refraction and magnetotelluric data, *J. Geophys. Res.*, *95*(B12), 19,419–19,438.

Wannamaker, P. E., J. R. Booker, A. G. Jones, A. D. Chave, J. H. Filloux, H. S. Waff, and L. K. Law (1989), Resistivity cross section through the Juan de Fuca subduction system and its tectonic implications, *J. Geophys. Res.*, *94*(B10), 14,127–14,144.

Wannamaker, P. E., et al. (2008), Lithospheric dismemberment and magmatic processes of the Great Basin–Colorado Plateau transition, Utah, implied from magnetotellurics, *Geochem. Geophys. Geosyst.*, *9*, Q05019, doi:10.1029/2007GC001886.

Wells, R. E., C. S. Weaver, and R. J. Blakely (1998), Fore arc migration in Cascadia and its neotectonic significance, *Geology*, *26*, 759–762.

G. D. Egbert, College of Oceanic and Atmospheric Sciences, Oregon State University, 104 CAS Administration Building, Corvallis, OR 97331, USA. (egbert@coas.oregonstate.edu)

P. K. Patro, National Geophysical Research Institute, Uppal Road, Hyderabad A.P., 500 007, India. (patrobpk@ngri.res.in)

Physics of the Earth and Planetary Interiors 187 (2011) 33–46

Contents lists available at ScienceDirect

Physics of the Earth and Planetary Interiors

journal homepage: www.elsevier.com/locate/pepi

Application of 3D inversion to magnetotelluric profile data from the Deccan Volcanic Province of Western India

Prasanta K. Patro[a,*], Gary D. Egbert[b]

[a] National Geophysical Research Institute, Council for Scientific and Industrial Research, Uppal Road, Hyderabad 500007, India
[b] College of Ocean and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331-5503, USA

ARTICLE INFO

Article history:
Received 26 October 2010
Received in revised form 1 April 2011
Accepted 14 April 2011
Available online 19 April 2011

Keywords:
Magnetotellurics
2D inversion
3D inversion
Deccan Volcanic Province

ABSTRACT

Using synthetic data Siripunvaraporn et al. (2005b) demonstrated possible advantages of interpreting single-profile MT data with a three-dimensional (3D) inversion program. Here we explore this idea further using real MT data from two profiles on the Indian subcontinent. The first profile (330 km long) cuts across the Deccan Volcanic Province of Peninsular India. The second (130 km long) is in the Narmada Son Lineament zone, approximately 100 km further north. Using the data-space Occam inversion code of Siripunvaraporn et al. (2005a) 3D inversion is carried out on each of these profiles independently, and results are compared with previously-published two-dimensional (2D) interpretations. In addition to inversion of the full impedance tensor, we consider 3D inversion of only the off-diagonal components. We also experiment with variants on the model covariance, in particular allowing for longer smoothing length scales along the geoelectric strike. Not surprisingly, the 3D inversion finds models that fit the data better than had been possible with the 2D programs. Many of the features inferred from these previous 2D interpretations are also present in the 3D inverse solutions, but the positions and amplitudes of individual conductive features are in some cases changed. The 3D models suggest substantial non-uniqueness in the single profile data. Even without explicit or special treatment, we find that the (relatively modest) near surface distortion effects in these datasets were well fit by the 3D inversion, by inserting small scale conductive and resistive features in surface layers, mostly off-profile.

1. Introduction

Due to both limitations in interpretation methods and the cost of data acquisition, magnetotelluric (MT) data have been traditionally obtained in profiles targeted to the geology, and then interpreted with two-dimensional inversion. In such an interpretation, one fits the off-diagonal impedances (Z_{xy} and Z_{yx}), generally after rotating the coordinate system so that the main diagonal components (Z_{xx} and Z_{yy}) are minimum, or at least small. It is seldom possible to find a single strike angle that is optimal for the full frequency range and for all sites, and possible impacts of off-profile structure must always be considered. Siripunvaraporn et al. (2005b) demonstrated the interpretation of MT profile data with a 3D inversion code using synthetic data examples. Results of that study suggest that inversion of even single profile MT data with the 3D algorithm results in more realistic images beneath the profile and, if the full tensor is fit, may even provide limited resolution of off-profile structures. With the availability of the 3D inversion code WSINV3DMT (Siripunvaraporn et al., 2005a), we were motivated to test this 3D interpretation approach on actual profile data from the Deccan Volcanic Province (DVP) of Western India. Here we present 3D inversion results for two MT profiles from this region and compare the results with previously published 2D interpretations.

2. Study region and the data

During 1998–1999 broad-band magnetotelluric studies were carried out on two profiles in the DVP, one of the great igneous provinces on Earth. Voluminous basalts were erupted around 65 Ma at the Upper Cretaceous–Tertiary boundary, widely believed to be due to the northward passage of the Indian plate over the Reunion hotspot (Duncan and Pyle, 1988). The study region (Fig. 1) is mainly covered by these flood basalts, and is an area associated with several continental scale and smaller rift zones (Biswas, 1982, 1987). The Narmada Son Lineament (NSL) is believed to have been a zone of weakness since the Precambrian times and regions north and south of the NSL have undergone vertical block movements (West, 1962). The Narmada valley represents a zone of

[*] Corresponding author.
E-mail addresses: patrobpk@ngri.res.in, patrobpk@rediffmail.com (P.K. Patro), egbert@coas.oregonstate.edu (G.D. Egbert).

0031-9201/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.pepi.2011.04.005

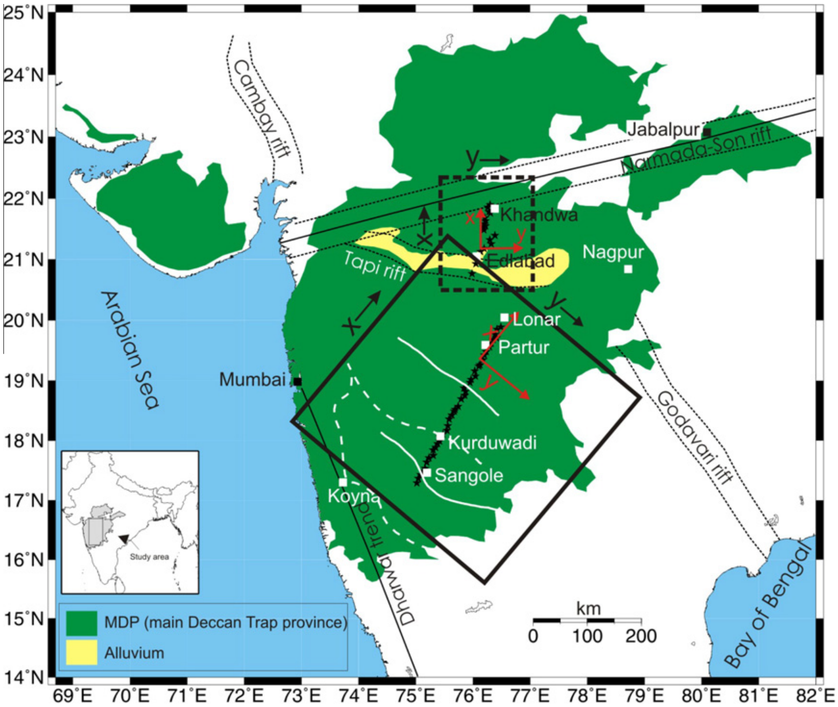


Fig. 1. Location of magnetotelluric stations plotted on the top of the geological map of the Deccan Volcanic Province (modified after Biswas, 1987; Peng et al., 1994). White lines are axes of gravity lows (dashed) and highs (solid) redrawn from Krishna Brahmam and Negi (1973). Major rift zones, such as the Cambay, Narmada Son and Godavari rifts are in contact with the DVP. The numerical 3D grid for the SP profile, oriented N45E, is shown as a solid black square. The grid for the EK profile is oriented N–S as shown by the dashed rectangle. Red arrows show orientations of *x* and *y* axes used in the text in discussions of 3D inversion results. The black solid and dotted lines shows the average geologic strike of the Narmada Son Lineament region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tectonic truncation of regional structural trends and is bounded by the Narmada North and Narmada South fault systems (Acharya et al., 1998).

The first profile considered here consists of 41 sites covering a period range 0.001–1000 s along the 330 km long NNE–SSW trending Sangole–Partur (SP) profile (station spacing was 5–10 km). The second, the NNE–SSW trending Edlabad–Khandwa (EK) profile, included a total of 18 MT sites, with data in the period range 0.001–1000 s, and station spacing ranging from 5 to 15 km. For both profiles Metronix GMS 05 broad-band MT data acquisition systems were used, and the data are generally of good quality. The time series were edited manually to remove sections contaminated by the most severe noise, and then analyzed using the Metronix robust processing code (PROCMT).

For the determination of regional strike direction the approaches of Smith (1997) and McNeice and Jones (2001) were used for the SP profile data. Both these approaches resulted in nearly the same strike angle for most of the sites, covering a range from 45° to 65° W of N, with the largest variations observed at sites near Kurduwadi (see Fig. 1). For the 2D interpretation reported in Patro et al. (2005a) and Patro and Sarma (2009) sites were divided into two groups, with impedances from the southern sites (1–18) rotated to 50° W of N, with the remaining sites rotated to 65° W of N. This coordinate system is consistent with the strike (NW–SE) evident in the regional geology (Arya et al., 1995; Peshwa and Kale, 1997). For the EK profile, the regional strike direction was computed using the tensor decomposition techniques of Smith (1997) and Becken and Burkhardt (2004). Both these analysis gave consistent results with a geoelectric strike of 75° E of N, in good agreement with the strike of the Narmada Son Lineament (Patro et al., 2005b). However, the recovered geoelectric strike direction is slightly oblique to the strike of Tapti rift, which the southern end of the EK profile crosses.

3. Review of 2D inversion results

We first summarize briefly the 2D inversion results presented in Patro et al. (2005a,b) and Patro and Sarma (2009), and relate these to the principal features evident in pseudo-sections for the two profiles. The 2D inversion of TE and TM data was carried out using the nonlinear conjugant gradient (NLCG) algorithm of Rodi and Mackie (2001). The period range used for the SP profile was 0.01–1000 s; for the EK profile data from 0.001 to 546 s were used. In both cases a uniform homogeneous half space of 100 Ω m was used as the prior (and starting) model. The final geoelectric models are shown in Fig. 2.

Observed pseudo-sections of apparent resistivities and phases are shown for the SP profile in Fig. 3a, and for the EK profile in Fig. 4a. In these figures we also present pseudo-sections of predicted data for the 2D inversions (Figs. 3b and 4b), and also for the 3D inversions (Figs. 3c, d and 4c, d) discussed below. Note that for the 3D inversions we have fit data from a subset of the available periods and sites, and for consistency only these responses are displayed in all of the pseudo-sections. A denser set of periods (42 for SP, 44 for EK) and sites (40 for SP, 18 for EK) were actually used for the 2D inversions. The most prominent feature that can be seen in the SP profile pseudo-sections (Fig. 3a) is the band of low phase above about 1 s in both modes. This feature results from approximately 1 km of moderately conductive overburden (presumably fractured basalts; barely visible in the inverse solution of Fig. 2a) over the highly resistive granitic basement. There are some spatial variations along profile in this upper band of low phases, reflecting variations of thickness and resistivity of the basalt layer. At longer periods there are several features in the pseudo-sections to note. The most prominent is the small decrease in ρ_{yx} (TE mode) at long periods at the northern end of the line beyond site 20, with a corresponding increase in phases centered at a period of approxi-

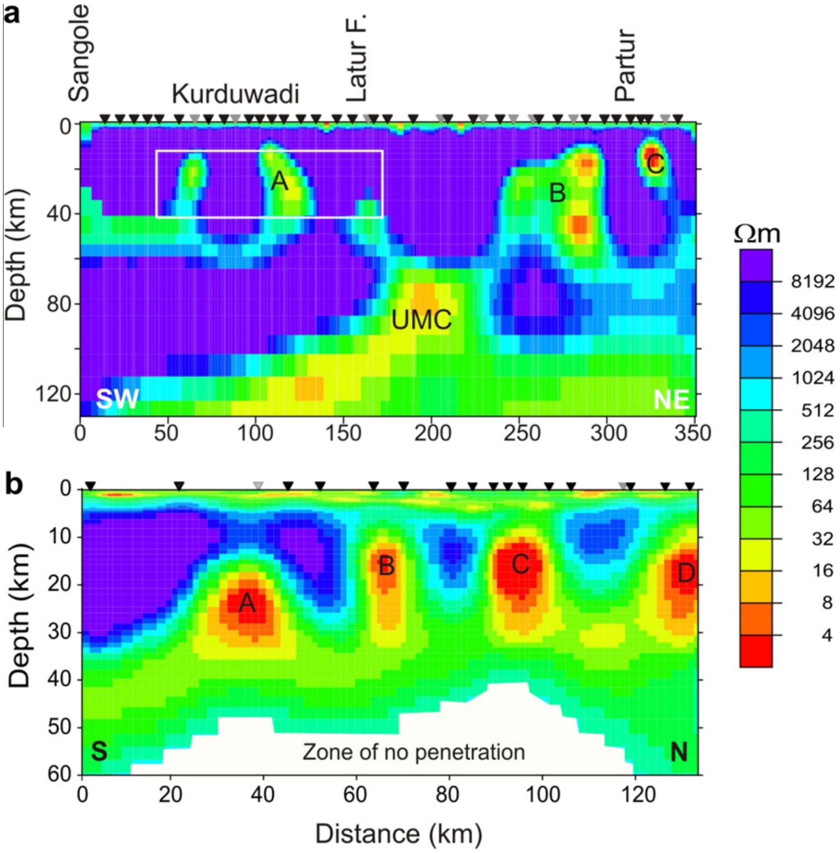


Fig. 2. (a) 2D geoelectric model derived from joint inversion of TE and TM data using the NLCG algorithm of Rodi and Mackie (2001) along Sangole–Partur profile (Patro and Sarma, 2009). A, B and C denotes deep crustal conductive features and UMC is the upper mantle conductor. (b) 2D geoelectric model derived in a similar manner for the Edlabad–Khandwa profile by Patro et al. (2005b) who interpreted conductive features A, B, C and D as the electrical signatures of Gavligarh fault, Tapti fault, Barwani-Sukta fault and Narmada south fault, respectively. The sites that were used for 3D inversion are marked with black triangles and omitted sites are marked with the gray symbols.

mately 30 s. This data feature corresponds to two patches of increased crustal conductivity (features B and C in Fig. 2a) at the northeast end of the SP profile.

Other features in the data for the SP profile are more subtle. Again in the *yx*-mode (TE) apparent resistivity is slightly reduced in patches at 10 s period centered at site 10, and also at the longest periods centered near site 20. The latter feature, which is also clearly visible in the slightly elevated phases at 300–400 s, corresponds to the upper mantle conductor (UMC) near and below 80 km depth in the 2D inverse solution (Fig. 2a). The *xy*-mode (TM) exhibits fewer spatial variations, though there is a phase increase on the southwest end of the profile at the longest periods. More importantly, there is a constant phase difference of around 10° between the *xy*- and *yx*-modes for periods of 50–100 s over most of the profile. This difference, with the TM mode exhibiting reduced phase, suggests alternating conductive and resistive features along the profile at crustal depths (e.g., in the box labeled A in Fig. 2a).

The data along the EK profile (Fig. 4a) are noisier, but also exhibit more substantial lateral variations, consistent with the complex geological terrain of the Narmada Son Lineament zone. The bottom of the conductive overburden (alluvium and fractured basaltic layer) is again marked most clearly by the band of low phase, now at longer periods due to the much greater thickness. Compared to the SP pseudo-sections there is more along-profile variation at all periods, which is consistent with more complex deeper structure, relative to the uniform highly resistive layer underlying the first profile. The 2D inversion of data from the EK profile revealed four conductive structural features extending from mid to deep crustal levels (A, B, C and D in Fig. 2b).

The conductive features imaged in the 2D inverse solution of Fig. 2b are seen most clearly in the data as the band of high phase around 100 s in the *yx*-mode (Fig. 4a). This band is not uniform along profile, suggesting a series of localized conductive features. Again, the *xy*- and *yx*-mode phases around 100 s differ, except perhaps at the N end of the profile. The relatively higher phases in the *yx*-mode (TE) at these periods are again indicative of a gross structural anisotropy, with higher conductivity in the deep crust aligned along strike. While a series of individual conductors are suggested by the along profile variations of this phase high, it is not clear that they are so distinct to require the same four conductors (A, B, C and D in Fig. 2b) that were used by the 2D inversion to match this behavior in the phase data.

4. 3D inversion of profile data

We inverted the MT data along the SP and EK profiles (separately) using WSINV3DMT, a 3D minimum structure inversion algorithm which is based on a data-space variant of the Occam scheme (Siripunvaraporn et al., 2005a). In addition to inversion of the full impedance tensor (four complex components, Z_{xx}, Z_{xy}, Z_{yx}, Z_{yy}), we also ran the inversion with only the off-diagonal components (Z_{xy}, Z_{yx}) as input data, and we experimented with several variations on the model covariance. In particular, WSINV3DMT allows smoothing over different length scales for the *x*, *y* and *z* directions, and we use this to test the effect of forcing structures to have longer length scales in the along-strike (approximately cross-profile) direction. The smoothing scales in WSINV3DMT are defined in terms of grid cells rather than physical length units. The default

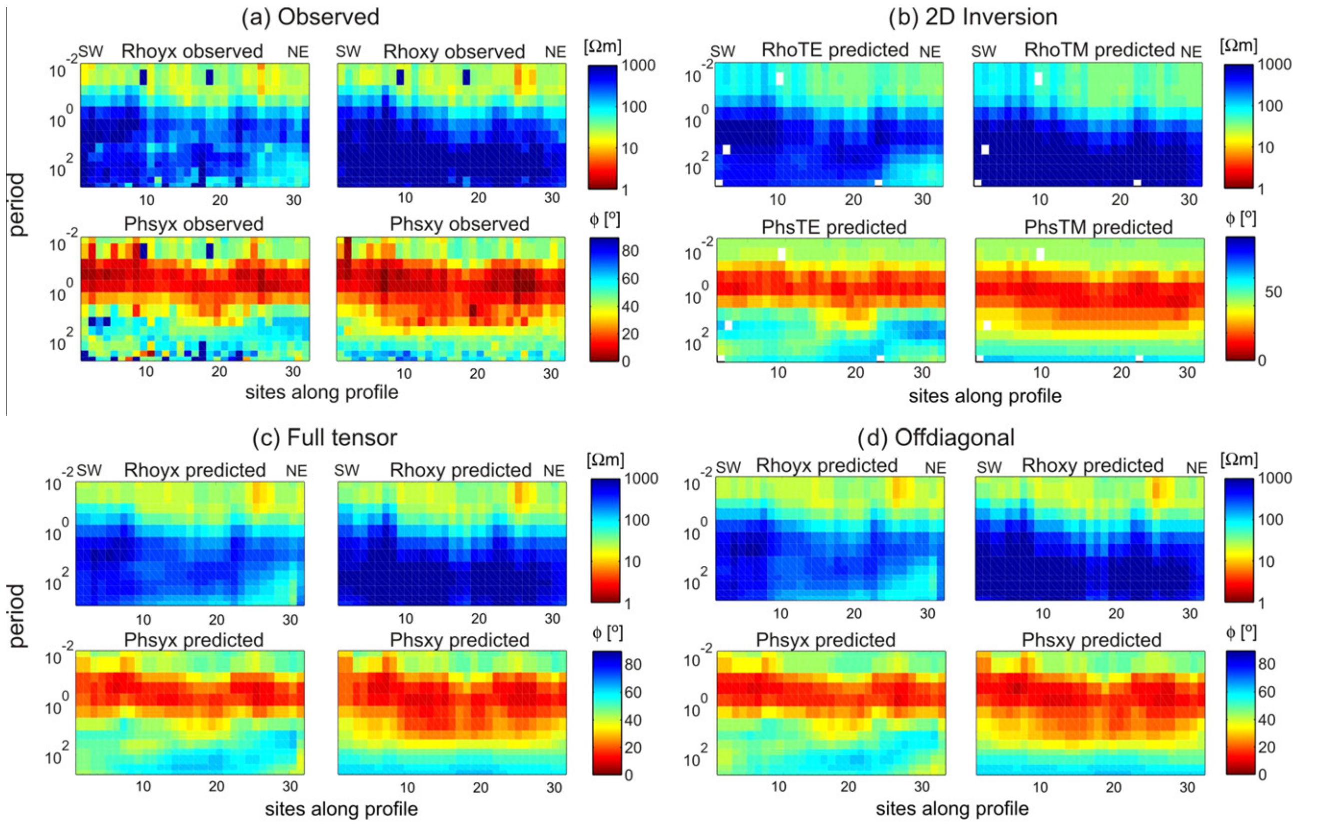


Fig. 3. (a) Observed pseudo-sections of apparent resistivity and phase along the Sangole–Partur profile, for the 13 periods and 32 sites used for the 3D inversion. (b) Predicted TE and TM mode pseudo-sections from the 2D inversion for the same periods and sites. Blank spaces show the data that were omitted from the 2D inversion. (c) Predicted pseudo-sections from the full-tensor 3D inversion. (d) As in (c) but for the 3D inversion fitting off-diagonal data only.

decorrelation scale, following the setup in the test case provided with the code, is $\sqrt{2}$ cells in all directions. We present results from two cases here: the default isotropic formulation with $\delta x = \delta y = \delta z = \sqrt{2}$, and an anisotropic covariance with $\delta x = \delta z = \sqrt{2}$, $\delta y = \sqrt{10}$, where y denotes the geoelectric strike direction inferred from the previously published 2D analysis.

Three-dimensional inversion is very expensive in terms of computation time and memory requirements. All the computations reported here were performed using a Sun Solaris system with 8 GB RAM. Because WSINV3DMT is a serial code, and forms the full Jacobian of the model parameter-data mapping, some compromises were required with regard to grid size and the number of periods fit. Given the relatively coarse grid spacing that we were forced to use (10 km in the central portion of the grid for the SP profile, and 4 km for EK) in some cases more than one site fell within a single cell of the 3D grid. Since the version of WSINV3DMT we used only allows data sites at the centers of grid cells, we generally selected the site nearest the center, considering also data quality and evidence for static shifts (see Fig. 7 of Patro et al., 2005a). This resulted in selection of 32 out of 41 sites for the SP profile. In the case of the EK profile 16 sites were chosen out of 18. The sites that were used for 3D inversion are marked with black triangles in Fig. 2; omitted sites are marked with the gray symbols. Considerations of computational practicality also forced us to select a subset of the available periods for the 3D inversion experiments. In both cases we used 13 periods; for the SP profile these were from 0.005 to 410 s, and for the EK profile from 0.009 to 911 s.

To allow different covariance length scales in along and across-strike directions, it is necessary to align the 3D numerical grid with the geoelectric strike. For the SP profile we thus rotated the grid, to coincide with the average geoelectric strike (55 degrees W of N) determined from the 2D analysis of Patro et al. (2005a). The impedances were of course also rotated to be consistent with this coordinate system. See Fig. 1 for orientations of the x and y axes, which are, respectively, oriented in cross-strike (roughly NE) and along-strike (SE) directions. Model grid dimensions for this profile were Nx = 36, Ny = 22 and Nz = 29 layers (plus 7 air layers). The mesh was created with a vertical factor of 1.33 with the top layer thickness being 60 m. Horizontal grid spacing was 10 km in the central part of the domain (including all of the MT sites) and total dimensions of the total model domain were 438 × 394 × 351 km. The prior model was a 100 Ω m half space. Note that the Arabian Sea, which is approximately 250 km west of the profile, is not included in the 3D model domain, as previous studies with 2D models showed that the ocean influence is minimal for the period range of MT data considered here (Patro et al., 2005a).

The EK profile (approximately N–S) is already nearly perpendicular to the geological strike, and for this dataset the grid was aligned so that the x-axis (cross-strike) is geographic N (Fig. 1). Now grid dimensions were Nx = 39, Ny = 21 and Nz = 28 layers, again with 7 air layers, and a vertical spacing factor of 1.4 was used. Dimensions of this model domain were 191 × 116 × 482 km in the x, y and z directions, respectively, in this case with a nominal resolution of 4 km in the grid core. The inversion was again started from a 100 Ω m half space.

With WSINV3DMT it is only possible to invert impedances, and it is thus impossible to assign different error floors to apparent resistivity and phase, as we did for the 2D inversions. Thus, for all 3D inversion tests we set errors as 5% of $\sqrt{|Z_{xy} \times Z_{yx}|}$. In general we used a two step procedure for the inversion results presented here. First, the model was run for 3–4 iterations, starting from the prior described above. Results from the iteration where the minimum RMS misfit was achieved were then taken as a new prior

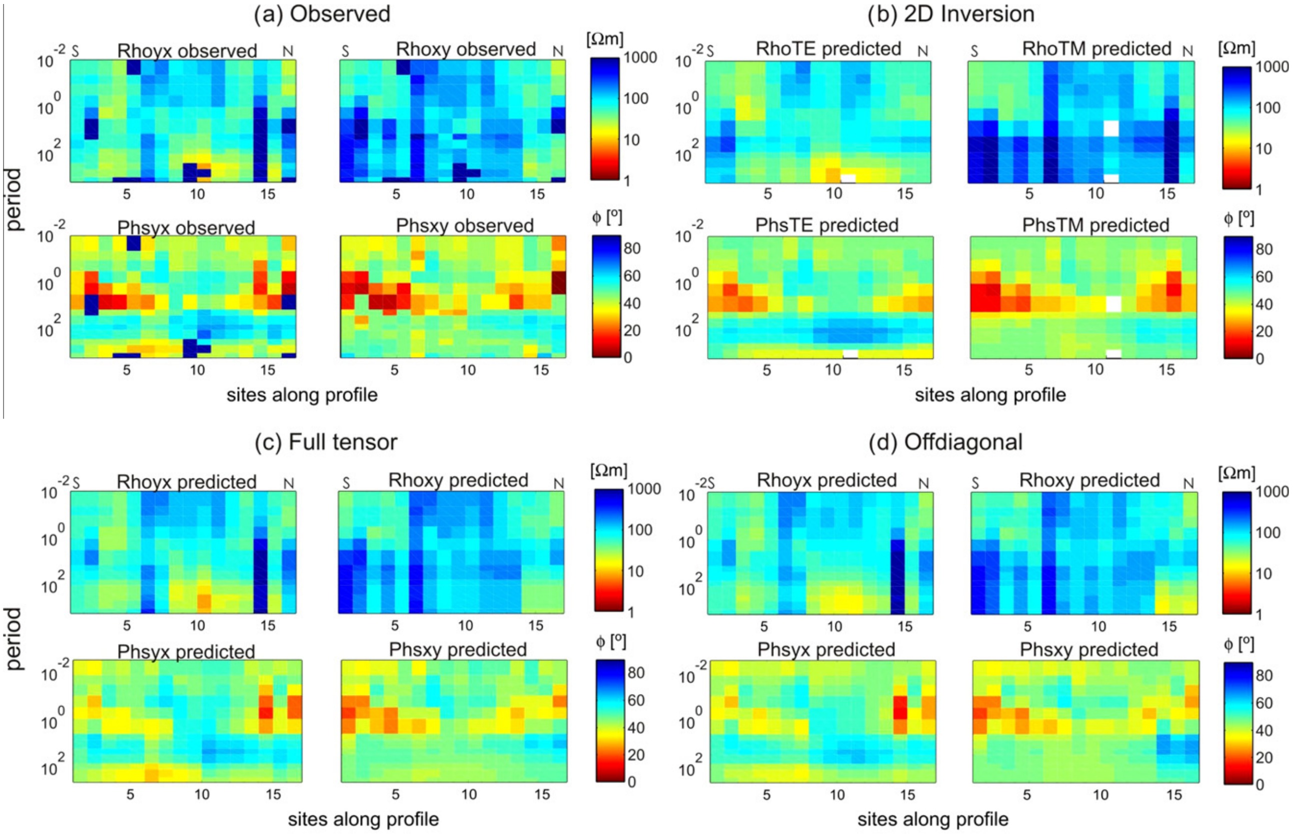


Fig. 4. As in Fig. 3, but for the Edlabad–Khandwa profile for which 13 periods and 16 sites were used for the 3D inversion.

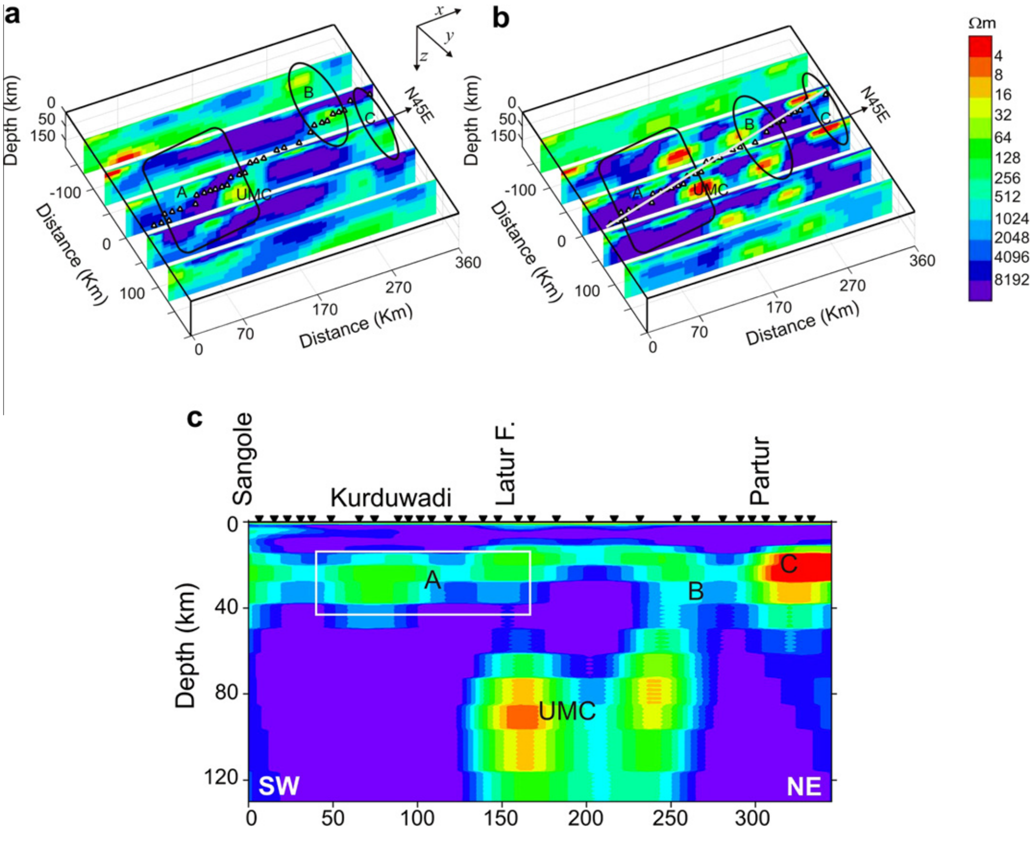


Fig. 5. Final model obtained from inversion of the full impedance tensors from the Sangole–Partur profile using (a) default model covariance and (b) anisotropic model covariance discussed in text. Features A, B and C correspond approximately to the crustal conductive features shown in Fig. 2a; UMC indicates the upper mantle conductor. (c) Profile cross section of the model (marked as gray color line in (b)) obtained from anisotropic covariance for Sangole–Partur profile.

model, and the inversion was restarted and run for an additional two iterations. This two step procedure was found to result generally in better fitting models than were obtained without restarting, but because only deviations from the prior model are penalized in the inversion, the final model does not represent a true minimum structure model, relative to the original prior.

5. Results

5.1. Isotropic and anisotropic model covariances

The final models obtained by inverting the full impedance tensor using both the default and modified anisotropic covariance described above, are presented in Figs. 5 and 6 for the SP and EK profiles, respectively. Profile cross section plots are presented in Fig. 5c and 6c for more direct comparison with Fig. 2. Although results obtained with the two covariances have many similarities there are also some significant differences, especially for the SP profile. For this profile the anisotropic covariance tends to result in structures that are more elongated along strike (i.e., in the direction of the *y*-axis). The upper mantle conductor (UMC), which appears in the 2D inversion (Fig. 2a) is found in both of the 3D inverse solutions, but with substantially different forms. For the isotropic case (Fig. 5a) the UMC is localized beneath the profile, centered near $x = 170$ km. When the anisotropic covariance is used the UMC shows much greater along strike continuity, and is significantly more conductive (<10 Ω m vs. ∼30 Ω m; Fig. 5b). In addition, a second zone of high conductivity appears in the upper mantle near $x = 250$ km, This is similar, but not identical, to the extension of the UMC to the north in the 2D inversion result. Note that in the 2D inversion data up to 1000 s were used, but for the 3D inversion only data up to 400 s are used. This may explain some of the differences between 2D and 3D results.

Subtle crustal conductive features are imaged on the southwest end of the profile by both the 2D and 3D inversions (box A in Figs. 2a and 5). In both cases these features peak at depths of 20–40 km, are of only moderately low resistivity (∼100 Ω m), and are broken into segments along profile. However, individual crustal conductive features from the 2D solution cannot be clearly matched to specific features in either of the 3D solutions. Again, results obtained with the anisotropic covariance exhibit greater along-strike continuity.

Much greater differences between the two 3D inversions are seen in crustal structure on the northern end of the profile. In particular, amplitudes are noticeably larger, and the model has a generally rougher appearance when the anisotropic covariance is used for regularization. Conductive features B and C from the 2D inversion (Fig. 2a) are not imaged as distinct anomalies in the 3D inversion with the isotropic covariance (Fig. 5a), but rather appear as an increase in crustal conductivity extending NE from near $x = 250$ km. These features appear as more distinct anomalies when the anisotropic covariance is used (Fig. 5b and c), but there are still significant differences from the 2D results, e.g., the peak in feature B is shifted to the southeast, where it connects to the NE branch of the UMC. Away from the profile there are often even greater differences in deep structure, e.g., with the areas of greatest resistivity shifted significantly between the two solutions, and from the 2D inversion results.

Differences between results obtained with the two covariances are less noticeable for the EK profile. All four of the conductive features seen in the 2D inverse solution (A–D in Fig. 2b) can be

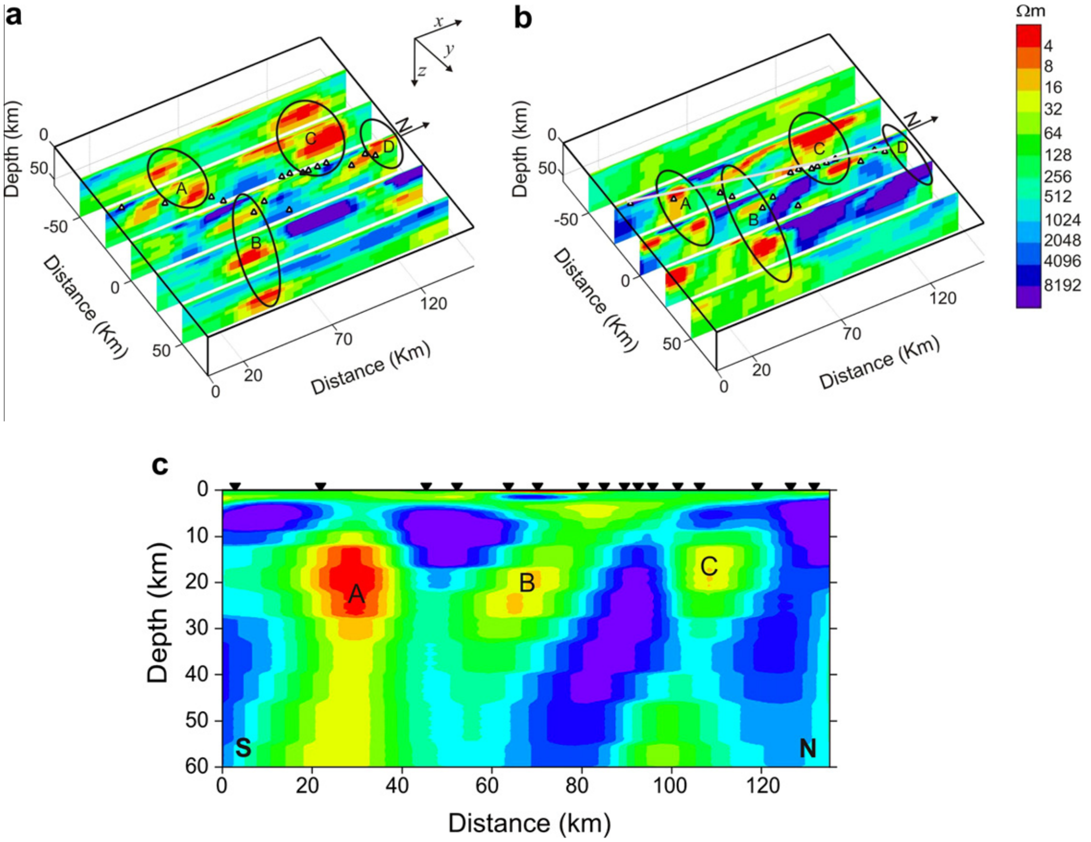


Fig. 6. Final model obtained from inversion of the full impedance tensors from the Edlabad–Khandwa data using (a) default model covariance and (b) anisotropic model covariance discussed in text. Features A, B, C and D correspond approximately to the crustal conductive features shown in Fig. 2b. (c) Profile cross section of the model (marked as gray color line in (b)) obtained from anisotropic covariance for Edlabad–Khandwa profile.

Table 1
Comparison of RMS misfits achieved for the 2D inversion, and for 3D inversion of full tensor, off-diagonal and main diagonal tensor elements of the MT data from SP and EK profiles for different model covariances.

	Off diagonal	Main diagonal	Total
SP profile—RMS			
Prior model misfit	10.89	2.49	7.90
2D inversion	2.93	2.49	2.72
Default covariance	2.20	1.75	1.99
Anisotropic covariance	2.30	1.80	2.06
Off-diagonal only with anisotropic covariance	2.27	3.27	2.81
EK profile—RMS			
Prior model misfit	6.54	4.28	5.53
2D inversion	2.82	4.28	3.62
Default covariance	2.04	1.53	1.80
Anisotropic covariance	2.22	1.67	1.96
Off-diagonal only with anisotropic covariance	2.24	5.35	4.10

matched to conductive features at similar positions along-profile in both of the 3D inverse solutions. However, now these features exhibit marked asymmetry with respect to the MT profile, with the northeast quadrant of the model domain particularly resistive. A conductive feature (labeled C in Figs. 2b and 6) appears in all of the inverse models, but is restricted to the west side of the profile in both 3D cases. The feature identified as D in Fig. 2b is not clearly a separate structure in either of the 3D inverse solutions. In the case of the isotropic covariance this feature appears more as a continuation of C, with a strike that is not perpendicular to the profile. Conductor B is mostly on the east side, though it is extended just across the profile when the anisotropic covariance is used. Overall, inversion results for the EK profile appear more starkly 3D, as might be expected from the geological complexity of this area. It is worth noting that the anisotropic covariance results in greater along-strike smoothing and extension of features in the case of the less clearly 3D SP profile. Thus, even when the off-diagonal components of the impedance might be reasonably consistent with

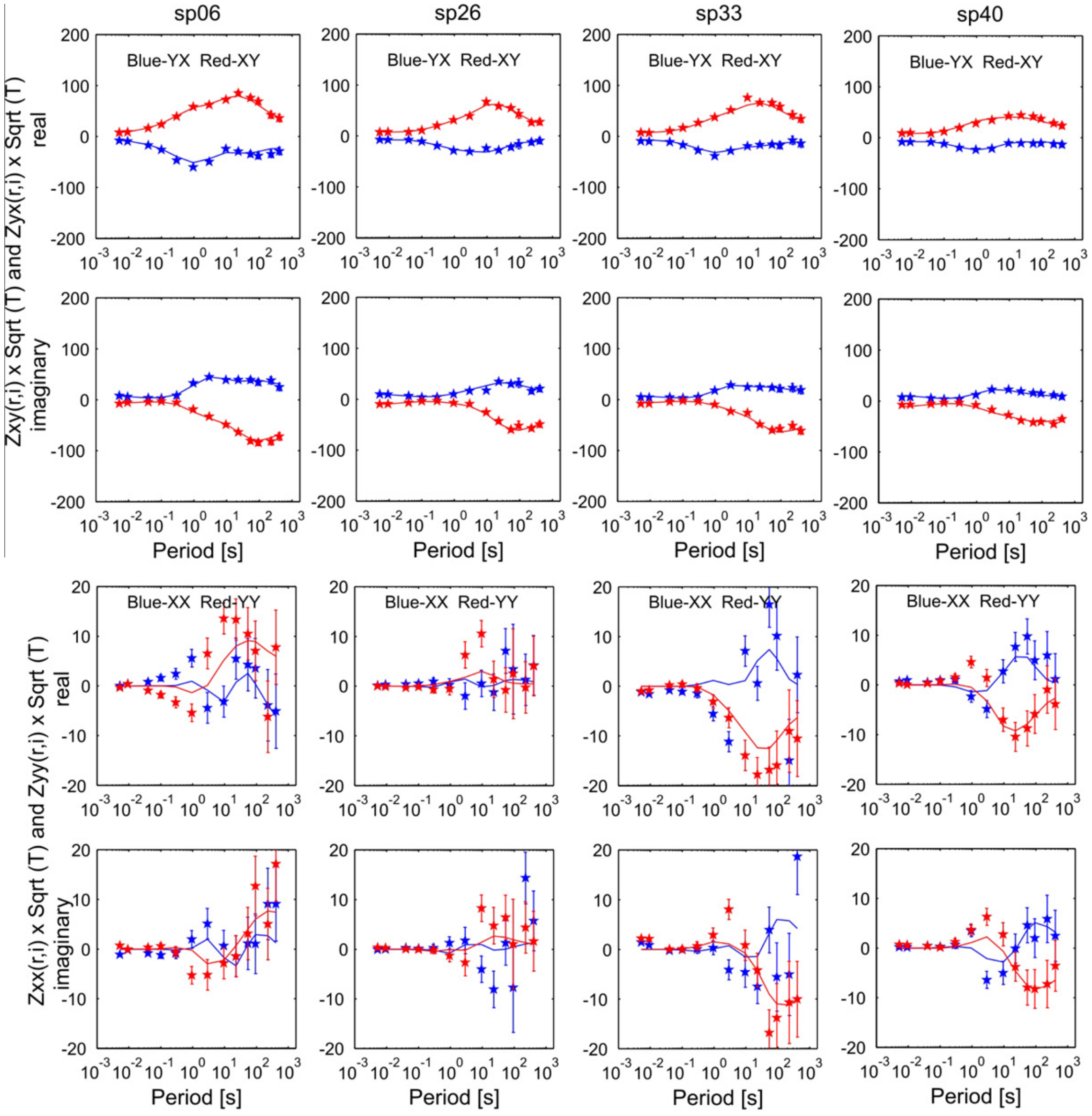


Fig. 7. Observed (star) and computed (line) responses of ρ_{xy}, ρ_{yx}, Φ_{xy}, Φ_{yx} (top) and $Z_{xx}(r,i)$, $Z_{yy}(r,i)$ (scaled by $\sqrt{T}$) (bottom) for Sangole–Partur profile at selected sites. Predicted data are shown for inversion of the full tensor, using the anisotropic model covariance.

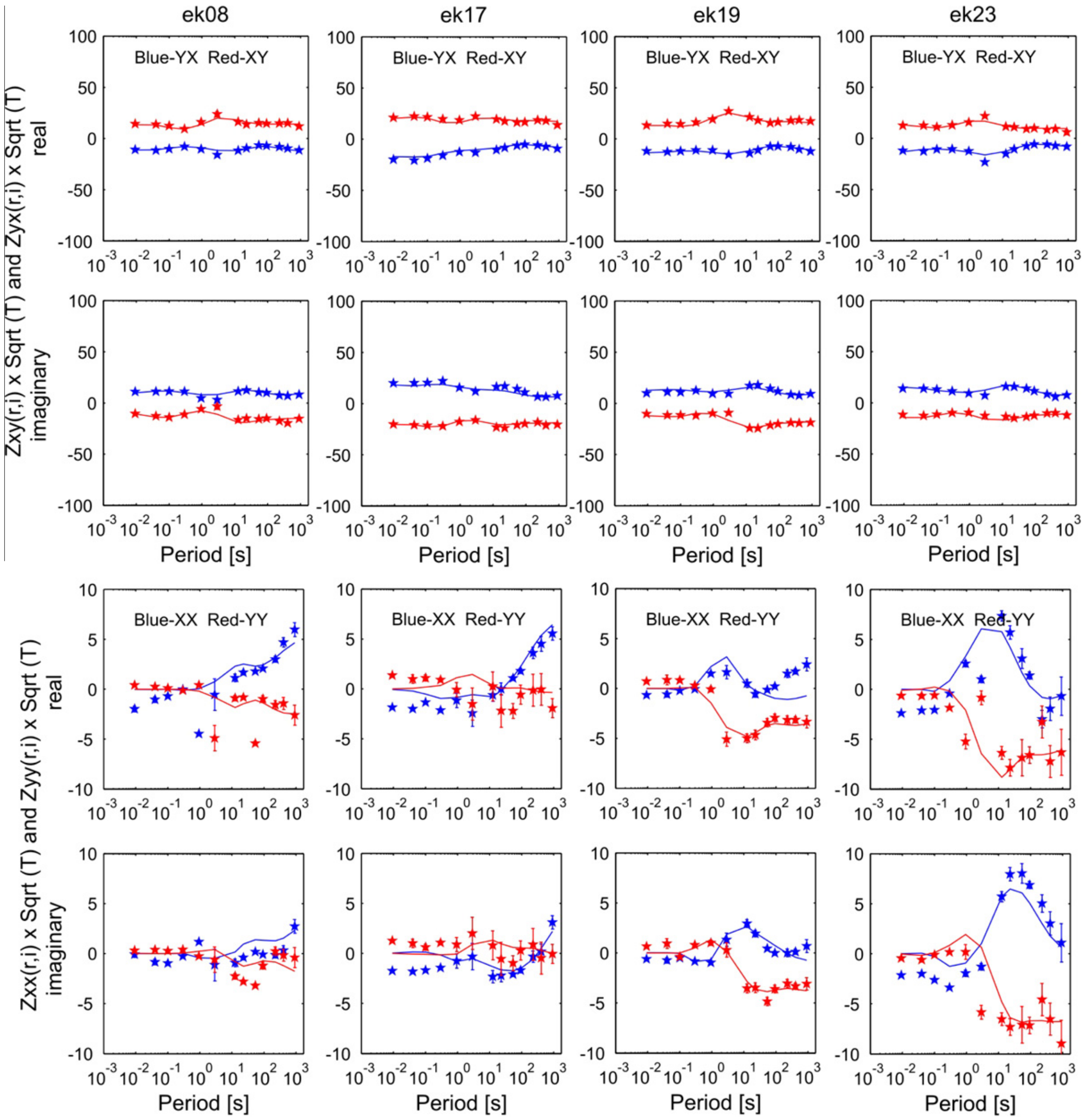


Fig. 8. Observed (star) and computed (line) responses of ρ_{xy}, ρ_{yx}, Φ_{xy}, Φ_{yx} (top) and $Z_{xx}(r,i)$, $Z_{yy}(r,i)$ (scaled by $\sqrt{T}$) (bottom) for Edlabad–Khandwa profile at selected sites. Predicted data are shown for inversion of the full tensor, using the anisotropic model covariance.

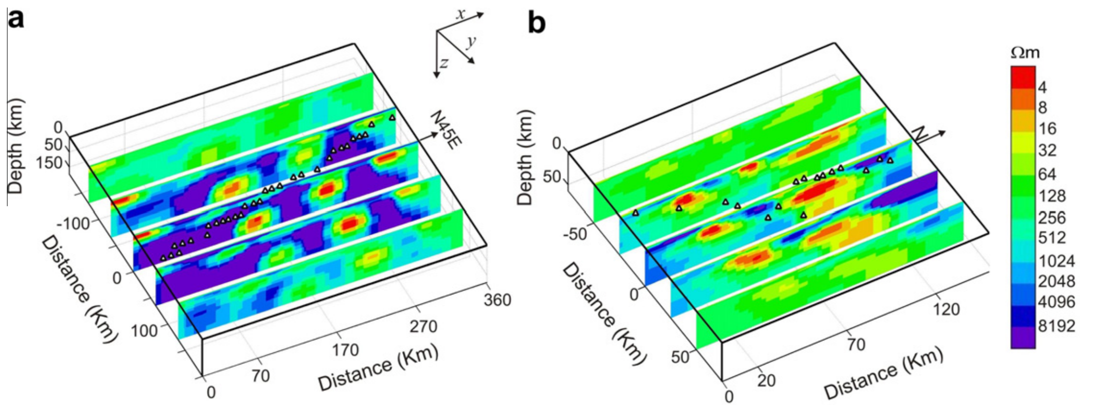


Fig. 9. Final models obtained from 3D inversion of off-diagonal tensor components, using the anisotropic model covariance, for (a) the Sangole–Partur profile and (b) the Edlabad–Khandwa data.

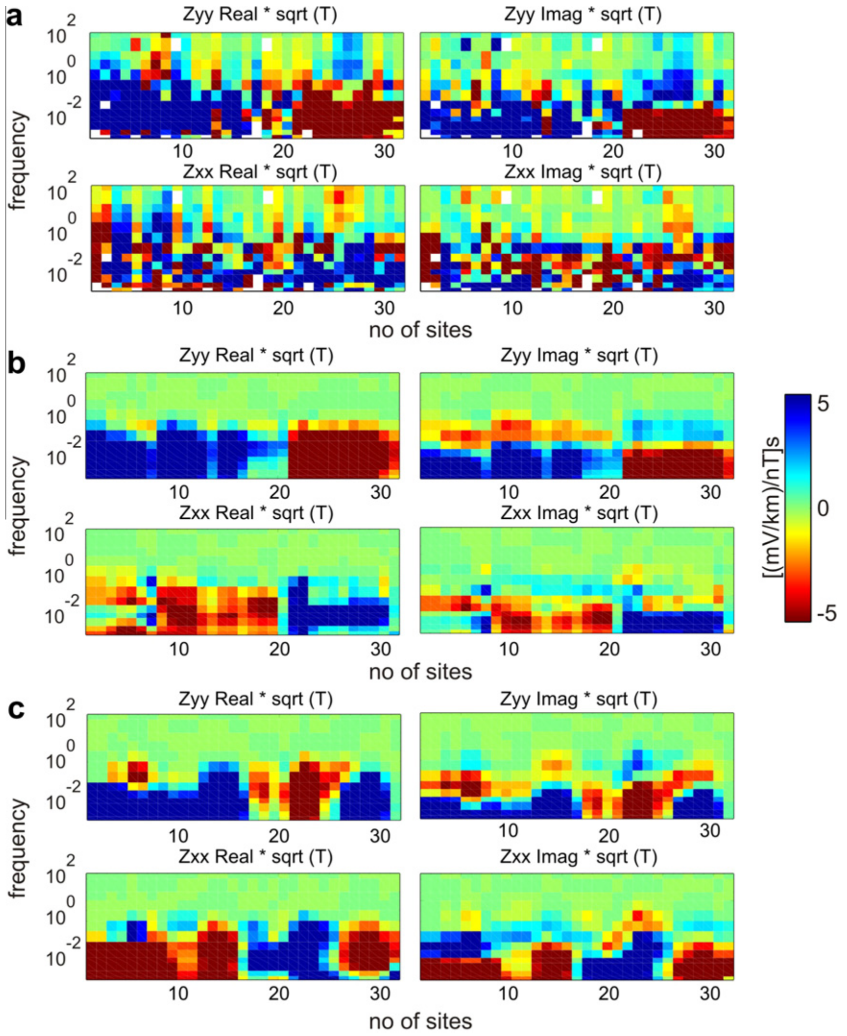


Fig. 10. Pseudo-sections of the real and imaginary parts (scaled by $\sqrt{T}$) of the main diagonal impedance components are plotted for the Sangole–Partur profile: (a) observed data; (b) predicted responses for inversion of full tensor and (c) predicted responses computed from the model derived by inverting only the off-diagonal components.

a 2D model (as for the EK profile; Patro et al., 2005b), fitting the full impedance may still require substantial local 3D structure.

A comparison of normalized RMS misfits achieved with the two covariances is presented in Table 1, along with the comparable misfits for the 2D inversions. For these comparisons we have used consistent datasets, and error bars, for computation of the normalized RMS for the 2D and 3D cases (i.e., with error floors set as 5% of $|Z_{xy}Z_{yx}|^{1/2}$). Different error normalizations were in fact used to specify the penalty functionals in the 2D inversion, so the numbers reported here are not directly comparable to results presented in Patro et al. (2005a,b). In the table the normalized RMS misfits are broken down into off-diagonal and main-diagonal components. For the SP profile the misfit of the prior model is strongly dominated by the off-diagonal components. For the EK profile the overall normalized RMS misfit of the prior model is smaller (primarily reflecting the lower signal-to-noise ratio in this dataset) but contributions from the main-diagonal component are significantly larger, consistent with the more strongly 3D character of the EK dataset.

For the SP profile the total RMS misfit (all impedance elements) obtained using the default model covariance was 1.99, only slightly below that achieved with the anisotropic covariance (RMS = 2.06). Both significantly improve the fit achieved by the 2D inversion (RMS = 2.72). The misfit is distributed roughly equally over both off-diagonal and main-diagonal elements for both 3D cases, as is the small increase in RMS with the modified covariance. Inversion with the anisotropic covariance can be viewed as an intermediate step between 3D inversion and enforcing a strictly 2D structure—along strike variations are penalized, but not prevented. With the less restrictive isotropic covariance the data can be fit more readily. Thus, even though the anisotropic solution has slightly larger misfit, it is rougher and has larger amplitude anomalies, suggestive of over-fitting within the constraints imposed by the model covariance.

The RMS misfit of the prior model is 7.9, so most of the signal is fit using either covariance, as can be seen in the pseudo-sections of Fig. 3c, and in Fig. 7, where observed and predicted curves for both the off-diagonal and diagonal impedance components are shown for selected sites. Note that in Fig. 7 the diagonal components, which are often very small and thus have a poorly defined phase, are plotted as real and imaginary parts of the impedance, multiplied by $\sqrt{T}$, where T is period in seconds. This compensates for the expected reduction in impedance amplitudes at long periods, e.g., for a half space this scaling would result in a flat (frequency independent) response for the off-diagonal impedance.

Similar analyses were carried out for the EK profile data. Again, misfits are increased slightly with the modified covariance (total RMS = 1.96 vs. 1.80). Both solutions reduce data misfit substantially compared to the 2D inversion results (RMS = 3.62). Much of the reduction results from improved fit of the main-diagonal components, where the 3D inversion reduces the normalized RMS misfit from 4.28 (prior model and 2D) to 1.67 and 1.53 (anisotropic and isotropic covariances). All of this is consistent with the results shown in Fig. 6 for the EK profile inverse solutions obtained with the two covariances are quite similar, in both cases exhibiting substantial 3D structure, as already suggested by the relatively larger main-diagonal components. Data fits for the EK profile, shown for apparent resistivity and phase pseudo-sections in Fig. 4c, and for selected sites in Fig. 8, are again reasonably good for both off-diagonal and diagonal impedance components.

5.2. 3D inversion of off-diagonal components

The inversion was also run using only the off-diagonal components (Z_{xy} and Z_{yx}) of the impedances as data. The frequencies and prior models are as given above for the two profiles, and the anisotropic covariance was used for regularization. Results are plotted in Fig. 9. For the SP profile (Fig. 9a) results are almost indistinguishable from those obtained with the full tensor inversion using the same covariance.

For the EK profile omitting the main diagonal components results in much greater differences in the inverse solution. The conductive feature between 20 and 40 km along profile ("A" in Figs. 2b, 6 and 9b) has similar position, size and amplitude in all inverse solutions. However, further north along the profile the pattern of conductive and resistive crustal features is noticeably different for the off-diagonal inversion, e.g., there is a conductor near where C is imaged in Figs. 2b and 6, but this feature is shifted to the south, and there is no separate structure near the profile corresponding to B. The absence of this conductor is compensated by the extension of C to the south, and the appearance of a small conductive feature to the west. Conductor D to the north of the profile is similar to that seen in the 2D and 3D anisotropic inversions (Figs. 2b and 6b). Overall, the 3D asymmetric character of the resistivity model is significantly reduced, the resistive area in the northeastern quadrant of the models of Fig. 6 is no longer clear, and conductive and resistive features are clearly oriented to the assumed strike, extending further across the profile. In some respects fitting only the off-diagonal components with the 3D inversion results in a solution that is more like the 2D result of Fig. 2b—e.g., conductors A and C are very similar. However, the smaller conductive features in the 2D solution (b and d) are not evident in 3D (Fig. 9b).

Another perspective on the diagonal components is provided by Figs. 10 and 11, where pseudo-sections of the real and imaginary parts of the diagonal impedance components (scaled by $\sqrt{T}$, as in Figs. 7 and 8) are plotted for the SP and EK profiles, respectively. Although the estimated diagonal components are relatively noisy, coherent features are evident at frequencies below about 1 Hz in

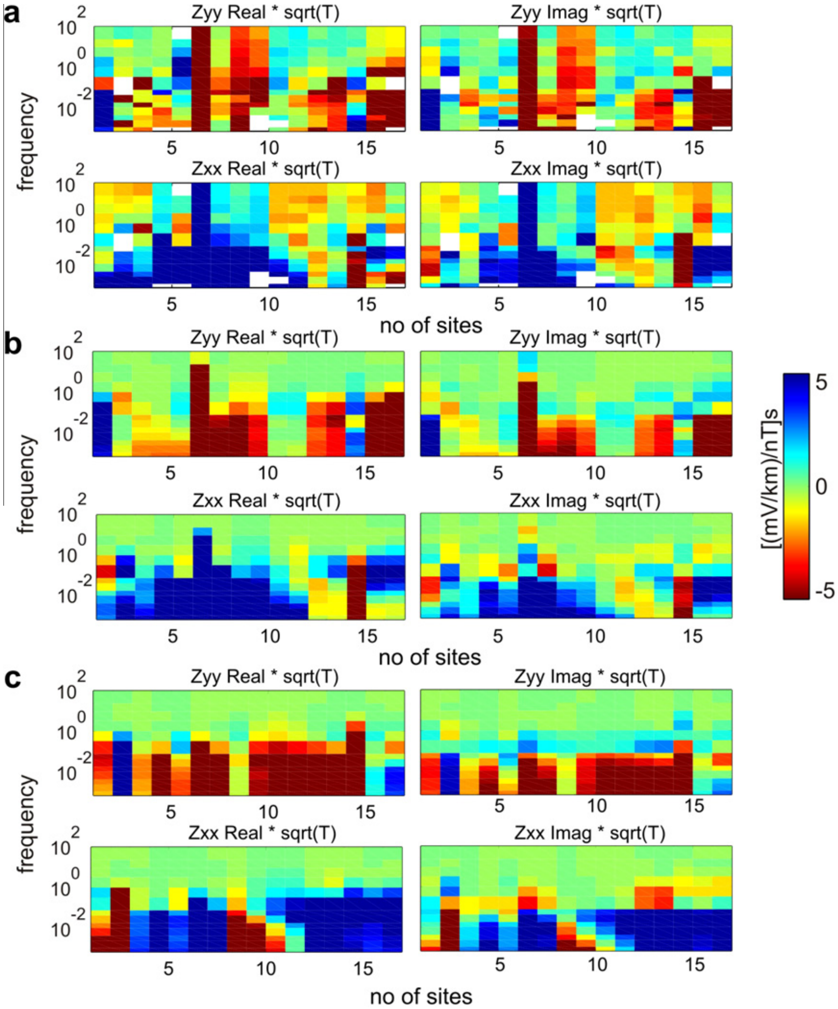


Fig. 11. Pseudo-sections of the real and imaginary parts (scaled by $\sqrt{T}$) of the main diagonal impedance components are plotted for the Edlabad–Khandwa profile: (a) observed data; (b) predicted responses for inversion of full tensor and (c) predicted responses computed from the model derived by inverting only the off-diagonal components.

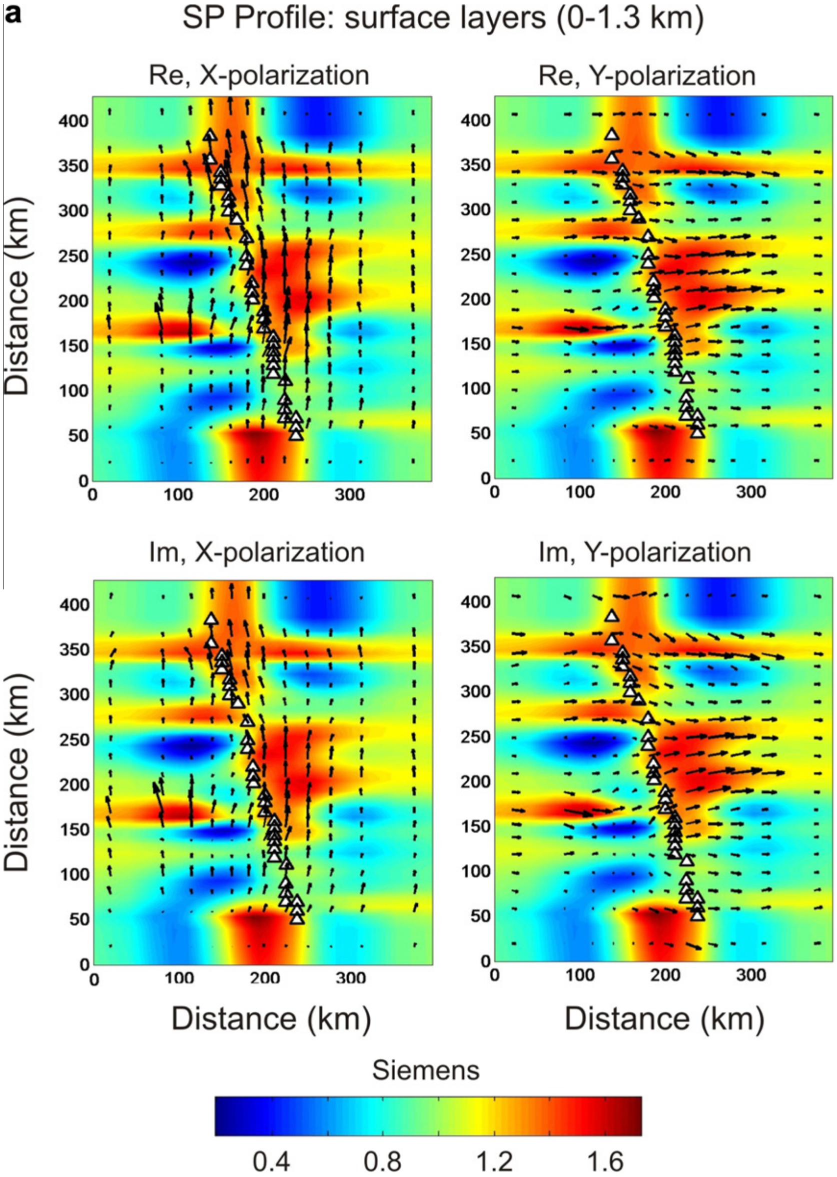


Fig. 12. Near surface conductance integrated over the top few model layers (0–1.3 km for SP profile; 0–1.4 km for EK profile) for the anisotropic covariance inverse solutions for SP and EK profile data. Electric current vectors (real and imaginary) modeled at a period of 50 s for N–S and E–W magnetic source polarizations are overlain, demonstrating how electric current flow is distorted by near-surface conductivity variations. The great similarity of the spatial pattern of real and imaginary parts is consistent with quasi-static galvanic distortion. Note that the SP profile numeric grid was oriented to coincide with the average geoelectric strike direction of 55° W of N.

the data sections for both profiles (Figs. 10a and 11a). These features are fit well by the full-impedance inverse solution (Figs. 10b and 11b), but not when the main diagonal elements are excluded (Figs. 10c and 11c). Indeed, for the SP profile the normalized RMS misfit (see Table 1) for the prior model (which, as a 1D model, has zero diagonal impedance response) is 2.49. This increases to 3.27 when the 3D inversion fits only the off-diagonal components, consistent with the impression given by Fig. 10, that the off-diagonal inversion actually degrades the fit to the main diagonal components. Similar results are obtained for the EK profile, where the normalized RMS misfit of the main diagonal components was 4.28 for the prior model, and increases to 5.35 when these components are not explicitly fit (see Table 1).

At higher frequencies the observed scaled diagonal impedance components for the SP profile are for the most part fairly small (see Fig. 10a). This is also seen in the example single-site curves of Fig. 7, and is consistent with the approximately 1D (and even largely undistorted) character of the TE and TM modes at high frequencies for most sites in this profile (Figs. 3a and 7). This can be explained by the relatively uniform weathered basaltic overburden in this area (Patro et al., 2005a).

Compared to the SP profile off-diagonal impedances for EK are significantly smaller (by a factor of more than two on average), while main diagonal amplitudes are as large or larger. The main diagonal components thus represent a significantly larger part of the total signal for the EK profile—e.g., note the much larger contribution of the main diagonal to prior misfits for this profile (Table 1). For the EK profile the observed scaled diagonal components are also relatively larger (compared to the SP profile) at high frequencies (see also Fig. 8). Although amplitudes are reduced compared to lower frequencies, there are now also spatially coherent features in the main-diagonal pseudo-sections even at short periods. This is again consistent with the more complicated 3D character of the MT data along this profile. For both profiles the

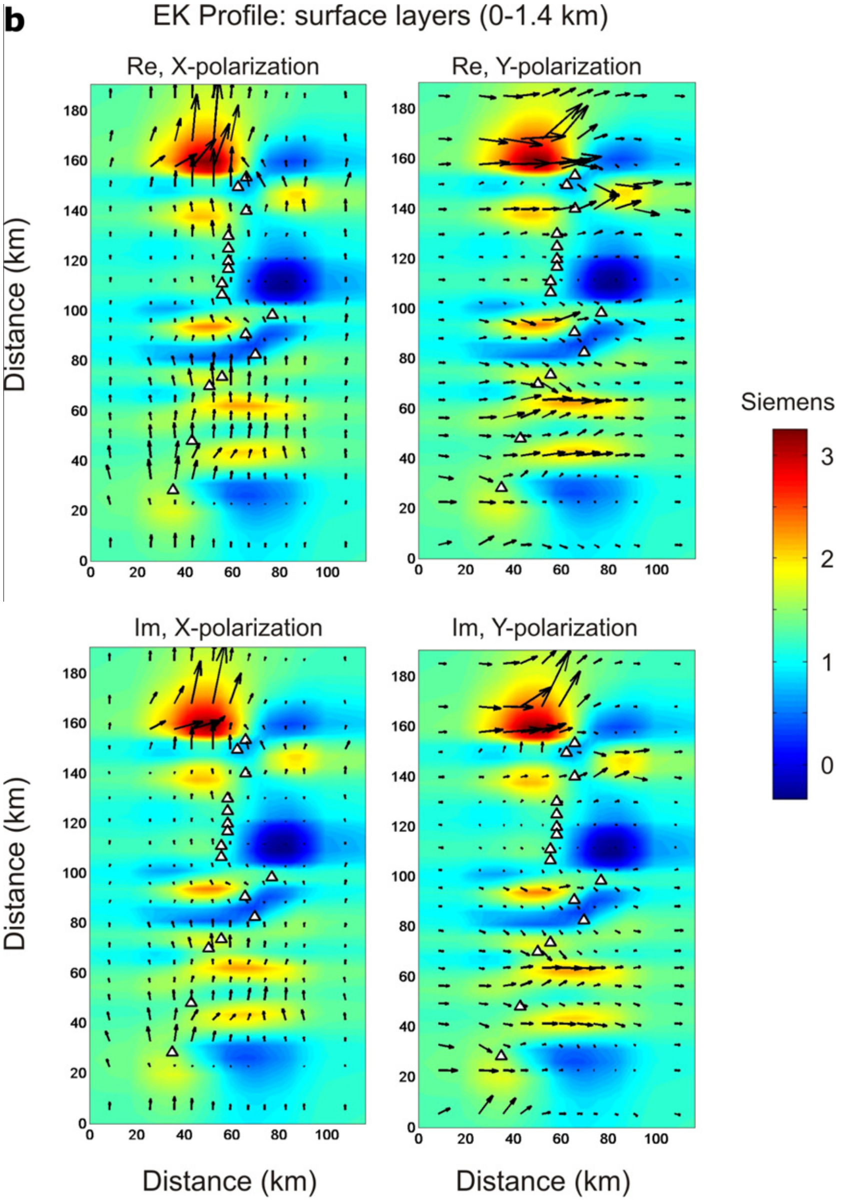


Fig. 12 (*continued*)

3D inverse solutions show little response at the highest frequencies. Features which appear to be clearly spatially coherent at frequencies above 1 Hz for the EK profile are not fit by the 3D inversion with either covariance.

5.3. Near surface structure and galvanic distortion

The main diagonal responses discussed above may result at least partly from galvanic distortion of the electric fields by near surface structure—the rapid site-to-site variations in diagonal components (Figs. 10 and 11) are particularly suggestive of such effects. With WSINV3DMT it is not possible to explicitly allow for galvanic distortion (e.g., by simultaneous fitting of a real frequency independent distortion matrix at each site or by increasing error floors for amplitudes but not phases). However one might hope that the inversion could account for these effects by inserting conductive and resistive features in the surface layers. That is, although our grid (and dataset) cannot resolve the actual near-surface distorting features, the distorted data may still be fit by inserting relatively shallow features near the profile. The effectiveness of this implicit treatment of distortion is likely reduced at higher frequencies, since the grid (horizontally 4/10 km) is certainly too coarse to allow modeling of purely galvanic distortion at the shortest periods considered (0.01 s). Thus, what we refer to as distortion here may not be purely galvanic over the full range of periods, but rather only for the longer periods relevant to the deeper structure emphasized by the vertical scale in Fig. 2.

To further explore how the inversion accommodates distortion at longer periods, we consider the impact of modeled shallow conductivity structure on near surface current flow. Surface conductance, computed by integration of model conductivity over the top seven layers (1.3 km for SP, 1.4 km for EK) is plotted for the anisotropic covariance inverse solutions for both profiles in Fig. 12. Electric current vectors computed from the inverse solutions by integrating over the same seven layers are overlain as arrows. For each profile four current maps are shown, i.e., real and imaginary parts for two source polarizations (corresponding to predominantly S–N and W–E current flow), all for a period of

50 s. The skin depth at this period greatly exceeds this layer thickness, so the effect of this layer on the modeled currents (and hence electric fields) is almost purely galvanic. The conductance variations clearly have a first-order effect on the magnitude and direction of the near-surface electric currents, which turn to avoid resistive patches and to flow into more conductive zones. The large scale patterns in the main diagonal components of Figs. 10 and 11 can be matched to the direction of current flow near the profile. Where currents (and hence electric fields) are strongly deflected from the dominant S–N or E–W direction of current flow for that mode, there will be a significant main diagonal impedance component. For example, for the SP profile with current flowing S–N (and hence the magnetic field in the y-direction) sites in the southern part of the profile (numbers below 20) show north-flowing currents deflected to the east; north of site 20 currents are deflected to the west, matching the spatial pattern seen in the Z_{yy} component predicted (and observed) data sections in Fig. 10.

Fig. 12 is instructive as to how the inversion accounts for surface distortion by inserting resistive or conducting zones near individual sites. Because we used an anisotropic covariance, these features are elongated along strike in Fig. 12, and they generally peak off the profile. Clearly these near surface layers are not resolved or interpretable—the actual distorting features are likely of much smaller scale.

At 50 s period, plots of the imaginary parts the electric field vectors (see Fig. 12) are nearly identical to those for the real parts, as would be expected if the spatial variations in electric field vectors were primarily controlled by galvanic distortion (and the regional impedance was relatively constant across the profile). At much shorter periods (not shown) inductive effects in the 1 km thick layer we have used to define the near surface are more significant, and real and imaginary parts become quite different. Even at 50 s period some differences between real and imaginary parts of the current vectors are evident for the EK profile, implying that deeper 3D structures responding inductively contribute more substantially to diagonal impedance components in this 3D case.

6. Discussion and conclusions

In this paper we have applied the data-space Occam inversion code of Siripunvaraporn et al. (2005a) to two MT profiles from peninsular India, and compared the resulting inverse solutions to previous 2D results. For the 3D inversion, in addition to the default isotropic model covariance, we tested an anisotropic covariance, with longer decorrelation length scales along previously determined geoelectric strikes. Models derived with this covariance are in some sense intermediate between those obtained with 2D inversion (with infinite length scales along strike) and those obtained with the default 3D covariance, which enforces no a priori strike preference. Given the substantial geological (Arya et al., 1995; Peshwa and Kale, 1997) and geophysical evidence for 2D structure in the areas surrounding these profiles—e.g., gravity (Krishna Brahmam and Negi, 1973; Tiwari et al., 2001) and seismic studies (Kaila et al., 1985)—such quasi-2D geoelectric models should perhaps be preferred, to the extent they provide an adequate fit to the data.

Our experiments with the 3D inversion suggest that data from the SP profile are much more nearly 2D than those from the EK profile to the north. Main diagonal components are smaller for the SP profile, and reduction in misfit achieved by the 3D inversion is modest. Use of the anisotropic covariance resulted in more significant changes to the inverse solution in the case of the more 2D SP profile, resulting in greater along strike extent and continuity of conductive features. But while the model was smoothed along strike, it became rougher across-strike, with amplitudes of most anomalies increased. Models obtained with the two covariances achieve nearly the same misfit but with structural complexity introduced in different ways. These results suggest that while the SP data might be consistent with a 2D interpretation there are probably finite strike length effects in the data. Moreover, our inversion results with the default isotropic covariance show that these data can be fit well by models with resistivity variations restricted to the near vicinity of the profile. As might be expected, along-strike extents of features imaged beneath the profile are poorly constrained by data from a single profile.

For the EK profile 3D structure is more clearly required near the profile, particularly to fit the diagonal impedance components. For this dataset the impact of changing the covariance was not so significant—although the anisotropic covariance favors features elongated along strike, the data do not allow such structures. These along-strike variations appear to be required by the main-diagonal impedance components. When these impedance components are not fit the 3D inverse solution for the EK profile changes substantially, becoming more nearly 2D with elongated conductive features extending along strike (at least when the anisotropic covariance is used). In contrast, for the more 2D SP profile omitting the diagonal components had almost no impact on the 3D inverse solution. Results for the two profiles thus suggest that the diagonal components of the impedance tensors for even a single profile at a minimum provide information about the need for nearby off-profile 3D structure. Perhaps some specific characteristics of the 3D structure might also be constrained (e.g., the need for more resistive crust in the NE quadrant of the EK profile) but this is an issue that requires further study.

In principal, near surface distortion can be accounted for directly in a fully 3D treatment—even if the site density is not sufficient to actually resolve the distorting structures, features can be inserted in the surface layers that allow distorted data to be fit. Our results suggest that the inversion did exactly this, inserting conductive and resistive features, generally with peak amplitudes off-profile, and restricted to the upper few model layers. These near surface features effectively fit site-to-site variations in impedance amplitudes, and also "twisting" of the electric field to produce diagonal impedance components. However, it should be born in mind that neither of the two datasets considered here included impedances with substantial distortion. To fit MT data that are more seriously distorted (e.g., out of quadrant phases, or mode splits of several orders of magnitude), and to allow for distortion at the shortest periods, a more explicit treatment of near surface distortion may in fact be required. This is another issue deserving of further investigation.

For both profiles similar conductive and resistive features appear at similar depths, and at similar locations along the profile in both 2D and 3D inverse solutions (see Figs. 2a, b, 5c and 6c). In particular, the 3D inversion results from the SP profile confirm the general areas of enhanced crustal conductivity (A, B and C), and the UMC delineated earlier from 2D interpretation. However, the amplitude of the UMC in the 3D inverse solutions is quite variable, depending on how well the data are fit, the model covariance, and the assumed prior model. Furthermore, the six low resistivity (∼5–50 Ω m) features at mid-crustal depths identified in Patro et al. (2005a) for the SP profile (see Fig. 2a) are not consistently resolved in the 3D images.

The 3D inversion results from the EK profile all reveal the presence of conductive Gavligarh fault, Tapti fault, Barwani-Sukta fault and Narmada south fault (features A, B, C and D). However, individual conductive and resistive features along the northern half of the profile (Fig. 2b) are again not reliably reproduced in the various 3D tests. The marked variation in position of the imaged crustal conductors and resistors beneath both profiles implies that many of these individual features are not well resolved. All inverse solutions do share alternating conducive and resistive zones, implying an anisotropic electric fabric (at some scale), with current flowing more readily along-strike (TE mode) than cross-strike (TM mode), as can in fact be seen clearly in the data pseudo-sections of Figs. 3 and 4. This gross electric anisotropy may result from enhanced conductivity in a few isolated zones within the crust (as in the 2D interpretation of Patro et al. (2005b) in terms of specific mapped faults), or it may reflect anisotropic fabric at finer scale. The MT data considered here, does not, by itself, resolve which of these possible interpretations should be preferred.

The observed differences between solutions reflect the inherent non-uniqueness of the MT data, particularly for details with spatial scales at or near the site spacing. However, it should be born in mind that due to computational constraints we have used a relatively coarse model discretization, and reduced the number of sites and frequencies, which we expect should further limit along-profile resolution in the 3D inverse models. Perhaps if all sites and all frequencies could be used resolution could be improved and non-uniqueness could be reduced.

In conclusion, 3D inversion appears to be a useful, if still computationally challenging, tool for enhancing interpretation of even single profile MT data. Many of the features obtained from earlier 2D inversions also appear in the 3D inversions, suggesting that these features are robust. At the same time, the 3D inversion allowed us to explore a larger range of solutions, further clarifying the range of acceptable models. Efforts at 3D inversion also clarified the extent and nature of 3D effects in the data. This in particular is an useful adjunct to the standard suit of tools for 2D interpretation.

Acknowledgments

We thank Weerachai Siripunvaraporn for providing his 3D inversion code and many useful discussions. The data were acquired under the financial support from Department of Science and Technology, Govt. of India. Support for P.K. Patro while out at Oregon State University was made possible through grants to G.D. Egbert from the National Science Foundation (NSF-EAR0345438) and the U.S. Department of Energy (DE-FG02-06ER15819). P.K. Patro thanks Dr. V.P. Dimri, Director NGRI for granting sabbatical leave to Oregon State University.

References

Acharya, S.K., Kayal, J.R., Roy, A., Chaturvedi, R.K., 1998. Jabalpur earthquake of May 22, 1997: constraint from aftershock study. J. Geol. Soc. India 51, 195–304.

Arya, A.S., Murthy, T.V.R., Garg, J.K., Naraian, A., Baldev, Sahai., 1995. Lineament pattern and its possible relationship with Killari earthquake: a case study using IRS data. Geol. Surv. India Spec. Publ. 27, 211–214.

Becken, M., Burkhardt, H., 2004. An ellipticity criterion in magnetotelluric tensor analysis. Geophys. J. Int. 159, 69–82.

Biswas, S.K., 1982. Rift basins in western margin of India and their hydrocarbon prospects with special reference to Kutch basin. Am. Assoc. Petr. Geol. Bull. 66, 1497–1513.

Biswas, S.K., 1987. Regional tectonic framework, structure and evolution of the western marginal basins of India. Tectonophysics 135, 307–327.

Duncan, R.A., Pyle, D.G., 1988. Raid eruption of the Deccan Traps at the Cretaceous/Tertiary boundary. Nature 333, 841–843.

Krishna Brahmam, N., Negi, J.G., 1973. Rift valleys beneath the Deccan Trap (India). Geophys. Res. Bull. 11, 207–237.

Kaila, K.L., Reddy, P.R., Dixit, M.M., Koteswar Rao, P., 1985. Crustal structure across the Narmada-Son lineament, Central India from deep seismic soundings. J. Geol. Soc. India 26, 465–480.

McNeice, G.W., Jones, A.G., 2001. Multisite, multifrequency tensor decomposition of magnetotelluric data. Geophysics 66, 158–173.

Patro, B.P.K., Brasse, H., Sarma, S.V.S., Harinarayana, T., 2005a. Electrical structure of the crust below the Deccan Flood Basalts (India), inferred from magnetotelluric soundings. Geophys. J. Int. 163, 931–943.

Patro, B.P.K., Harinarayana, T., Sastry, R.S., Rao, M., Manoj, C., Naganjaneyulu, K., Sarma, S.V.S., 2005b. Electrical imaging of Narmada-Son Lineament Zone, Central India from magnetotellurics. Phys. Earth Planet. Inter. 148, 215–232.

Patro, P.K., Sarma, S.V.S., 2009. Lithospheric electrical imaging of the Deccan trap covered region of western India. J. Geophys. Res. 114, B01102.

Peng, Z.X., Mahoney, J., Hooper, P., Harris, C., Beane, J., 1994. A role for lower continental crust in flood basalt genesis? Isotopic and incompatible element study of the lower six formations of the western Deccan Traps. Geochim. Cosmochim. Acta 58, 267–288.

Peshwa, V.V., Kale, V.S., 1997. Neotectonics of the Deccan trap province: focus on Kurduwadi lineament. J. Geophys. 18, 77–86.

Rodi, W., Mackie, R.L., 2001. Nonlinear conjugate gradients algorithm for 2-D magnetotelluric inversion. Geophysics 66, 174–187.

Siripunvaraporn, W., Egbert, G., Lenbury, Y., Uyeshima, M., 2005a. Three-dimensional magnetotelluric: data space method. Phys. Earth Planet. Inter. 150, 3–14.

Siripunvaraporn, W., Egbert, G., Uyeshima, M., 2005b. Interpretation of 2-D magnetotelluric profile data with 3-D inversion: synthetic examples. Geophys. J. Int. 160, 804–814.

Smith, J.T., 1997. Estimating galvanic-distortion magnetic fields in magnetotellurics. Geophys. J. Int. 130, 65–72.

Tiwari, V.M., Vyaghraswara Rao, M.B.S., Mishra, D.C., 2001. Density inhomogeneities beneath Deccan Volcanic Province, India, as derived from gravity data. J. Geodyn. 31, 1–17.

West, W.D., 1962. The line of Narmada-Son valley. Curr. Sci. 31, 143–144.

Geophysical Journal International

Geophys. J. Int. (2012) **189,** 251–267 doi: 10.1111/j.1365-246X.2011.05347.x

GJI Geomagnetism, rock magnetism and palaeomagnetism

Computational recipes for electromagnetic inverse problems

Gary D. Egbert and Anna Kelbert

College of Oceanic and Atmospheric Sciences, Oregon State University, 104 *COAS Admin Bldg., Corvallis, OR* 97331-5503, *USA.*
E-mail: egbert@coas.oregonstate.edu

Accepted 2011 December 18. Received 2011 December 13; in original form 2011 August 25

SUMMARY

The Jacobian of the non-linear mapping from model parameters to observations is a key component in all gradient-based inversion methods, including variants on Gauss–Newton and non-linear conjugate gradients. Here, we develop a general mathematical framework for Jacobian computations arising in electromagnetic (EM) geophysical inverse problems. Our analysis, which is based on the discrete formulation of the forward problem, divides computations into components (data functionals, forward and adjoint solvers, model parameter mappings), and clarifies dependencies among these elements within realistic numerical inversion codes. To be concrete, we focus much of the specific discussion on 2-D and 3-D magnetotelluric (MT) inverse problems, but our analysis is applicable to a wide range of active and passive source EM methods. The general theory developed here provides the basis for development of a modular system of computer codes for inversion of EM geophysical data, which we summarize at the end of the paper.

Key words: Numerical solutions; Inverse theory; Magnetotelluric; Geomagnetic induction.

1 INTRODUCTION

Over the past decade or so, regularized inversion codes have been developed for a range of three-dimensional (3-D) frequency-domain electromagnetic (EM) induction problems, including magnetotellurics (MT; e.g. Newman & Alumbaugh 2000; Siripunvaraporn *et al.* 2004) global geomagnetic depth sounding (Kelbert *et al.* 2008), and controlled source methods including cross-well imaging (e.g. Alumbaugh & Newman 1997) and marine controlled source EM (CSEM; e.g. Commer & Newman 2008). Generally, these efforts have been based upon minimization of a penalty functional, a sum of data misfit and model norm terms. Several distinct algorithms have been applied to solve the minimization problem, including Gauss–Newton (GN) schemes (Mackie & Madden 1993; Sasaki 2001; Siripunvaraporn *et al.* 2004) and direct gradient-based minimization schemes such as non-linear conjugate gradients (NLCG; e.g. Newman & Alumbaugh 2000) or quasi-Newton schemes (e.g. Newman & Boggs 2004; Avdeev & Avdeeva 2009). All of these various applications, and the different inversions algorithms that have been used, share many common elements. Here, we consider these commonalities explicitly, developing a general mathematical framework for frequency-domain EM inverse problems. Through this framework, we provide recipes for adapting previously developed inversion algorithms to new applications and for developing extensions to standard applications (e.g. new data types, model parametrizations and regularization approaches), and a basis for development of more efficient inversion algorithms.

Recently, Pankratov & Kuvshinov (2010) have given a general formulation for calculation of derivatives for 3-D frequency-domain EM problems. A general formalism for derivative calculation is also central to our development, so in principal there is considerable overlap between their development and what is presented here. However, in contrast to Pankratov & Kuvshinov (2010), we adopt a purely discrete approach, assuming from the outset that the forward problem has been discretized for numerical solution, so that all spaces (EM fields, model parameters and data) are finite dimensional. The penalty functional to be minimized is explicitly taken to be a discrete quadratic form, and derivatives, adjoints, etc. are all derived for this discrete problem. Similarly, we explicitly consider the need to use discrete interpolation operators to simulate the measurement process, and to represent dependence of the discrete model operator on the unknown parameters.

There has been considerable discussion in the ocean data assimilation literature concerning the virtues of discrete versus continuous formulations of inverse problems (e.g. Bennett 2002). Certainly, there are some issues in inverse problems (e.g. regularity and well-posedness) that can only be understood and discussed rigorously through consideration of the problem in continuous form (e.g. Egbert & Bennett 1996). However, for development of actual practical inversion algorithms there are good reasons to focus on the discrete formulation. In particular, only through a direct treatment of the discrete problem can symmetry (with respect to appropriate inner products) of the numerical implementation of adjoints be guaranteed. Furthermore, in discrete form many derivations are trivial, and the steps actually required for computations are often more clearly and explicitly laid out.

Of course, details about Jacobian calculations and discussions of the solution of EM inverse problems in discrete form have been given in many previous publications, both for specific EM methods (e.g. see references earlier), and with some degree of generality

(McGillivray *et al.* 1994; Newman & Hoversten 2000). One motivation for presenting this material again here, with a more abstract formulation and using homogeneous mathematical notation, is to provide a foundation for a system of modular computer codes for frequency-domain EM inverse problems that we have recently developed. We sketch key aspects of this system at the end of this paper, and provide a more detailed description elsewhere. Development of this modular system motivates us to clearly define all of the fundamental objects and methods required for a generic EM inverse problem, and to analyse the dependencies among these objects. This framework for the modular system, which we believe is unique both in its abstraction and completeness, is one of the key results presented in this paper. Another key result which emerges from our analysis concerns the structure of the Jacobian calculations in multifrequency and multitransmitter inverse problems. The rather obvious factorization of the Jacobian into receiver and transmitter components (previously used to improve efficiency in cross-well EM inversion by Newman & Alumbaugh 1997) is a simple consequence of our analysis. A less obvious consequence, which has not to our knowledge been previously noted or made use of, is that computations of sensitivities for problems with multicomponent transfer functions (TFs; e.g. 3-D MT) can also be factored, reducing required computations by a factor of 4 relative to a more naive approach. Our abstract treatment of Jacobian calculations thus provides a basis for developing more efficient computational strategies for specific problems.

This paper is organized as follows: In Section 2, we summarize some common linearized EM inversion methods based on gradient-based minimization of a penalty functional, demonstrating at a coarse level the basic objects used for EM inversion methods. A key component in all methods is the Jacobian of the mapping from model parameters to data; we derive general expressions for this linear operator in Section 3. In Section 4, we consider more explicitly the discretization of the governing differential equations, and the dependence of the discrete equation coefficients on the model parameter. Here, we introduce specific examples of EM inverse problems (2-D and 3-D MT) which we will follow throughout the remaining development. These EM inverse problems are sufficiently different to motivate and illustrate much of the abstraction required of a flexible modular system. With these examples as motivation, we then show in Section 5 how operations with the Jacobian can be factored into reusable components, and we consider how these components depend on each other, and on details of the EM method (e.g. sources and receivers), model parametrization and numerical discretization. In Section 6, we consider more explicitly the form of the Jacobian when there are multiple frequencies and multiple, possibly coupled, source geometries. Some new results on possible computational efficiencies are given here. In Section 7, we provide a brief overview and illustration of the modular system of Fortran 95 computer codes that we have developed based on the framework for general frequency-domain inversion developed in the preceding sections.

2 LINEARIZED EM INVERSION: OVERVIEW

We consider regularized inversion through gradient-based minimization of a penalty functional of the form

$$\Phi(\mathbf{m},\mathbf{d}) = (\mathbf{d}-\mathbf{f}(\mathbf{m}))^T\mathbf{C}_\mathbf{d}^{-1}(\mathbf{d}-\mathbf{f}(\mathbf{m})) + \nu(\mathbf{m}-\mathbf{m}_0)^T\mathbf{C}_\mathbf{m}^{-1}(\mathbf{m}-\mathbf{m}_0) \tag{1}$$

to recover, in a stable manner, $\mathbf{m}$, an M-dimensional Earth's conductivity model parameter vector, which provides an adequate fit to a data vector $\mathbf{d}$ of dimension N_d. In (1), $\mathbf{C_d}$ is the covariance of data errors, $\mathbf{f}(\mathbf{m})$ defines the forward mapping, $\mathbf{m}_0$ is a prior or first guess model parameter, ν is a trade-off parameter, and $\mathbf{C_m}$ (or more properly $\nu^{-1}\mathbf{C_m}$) defines the model covariance or regularization term. In practice, $\mathbf{C_d}$ is always taken to be diagonal, so by a simple rescaling of the data and forward mapping ($\mathbf{C}_\mathrm{d}{}^{-1/2}\mathbf{d}$, $\mathbf{C}_\mathrm{d}{}^{-1/2}\mathbf{f}$), we may eliminate $\mathbf{C}_\mathbf{d}^{-1}$ from the definition of Φ.

The prior model $\mathbf{m}_0$ and model covariance $\mathbf{C_m}$ can also be eliminated from (1) by the affine linear transformation of the model parameter $\tilde{\mathbf{m}} = \mathbf{C}_\mathbf{m}^{-1/2}(\mathbf{m}-\mathbf{m}_0)$, and forward mapping $\tilde{\mathbf{f}}(\tilde{\mathbf{m}}) = \mathbf{f}(\mathbf{C}_\mathbf{m}^{1/2}\tilde{\mathbf{m}}+\mathbf{m}_0)$, reducing (1) to

$$\begin{aligned}\Phi(\tilde{\mathbf{m}},\mathbf{d}) &= (\mathbf{d}-\tilde{\mathbf{f}}(\tilde{\mathbf{m}}))^T(\mathbf{d}-\tilde{\mathbf{f}}(\tilde{\mathbf{m}})) + \nu\tilde{\mathbf{m}}^T\tilde{\mathbf{m}}\\ &= ||\mathbf{d}-\tilde{\mathbf{f}}(\tilde{\mathbf{m}})||^2 + \nu||\tilde{\mathbf{m}}||^2.\end{aligned} \tag{2}$$

After minimizing (2) over $\tilde{\mathbf{m}}$, the model parameter in the untransformed space can be recovered as $\mathbf{m} = \mathbf{C}_\mathbf{m}^{1/2}\tilde{\mathbf{m}}+\mathbf{m}_0$. Note that this model space transformation is in fact quite practical if instead of following the usual practice of defining $\mathbf{C}_\mathbf{m}^{-1} = \mathbf{D}^T\mathbf{D}$, where $\mathbf{D}$ is a discrete representation of a gradient or higher order derivative operator, the regularization is formulated directly in terms of a smoothing operator (i.e. model covariance) $\mathbf{C_m}$. It is relatively easy to construct computationally efficient positive definite discrete symmetric smoothing operators for regularization (e.g. Egbert 1994; Siripunvaraporn & Egbert 2000; Chua 2001). Although the resulting covariance matrix $\mathbf{C_m}$ will not generally be sparse and may not be practical to invert, all of the computations required for gradient evaluations and for minimization of the transformed penalty functions require only multiplication by the smoothing operator $\mathbf{C}_\mathbf{m}^{1/2}$ (i.e. half of the smoothing of $\mathbf{C_m}$). It is also straightforward to define model covariances that can be inverted (i.e. so that multiplication by both $\mathbf{C_m}$ and $\mathbf{C}_\mathbf{m}^{-1}$ are practical.) In the following, we focus on the simplified 'canonical' penalty functional (2), with tildes omitted.

We begin by summarizing some standard approaches to gradient-based minimization of (2) using a consistent notation. Siripunvaraporn & Egbert (2000), Rodi & Mackie (2001), Newman & Boggs (2004), Avdeev (2005) provide further details and discussion on these and related methods. A key component in all of these linearized search schemes is the $N_\mathrm{d}\times M$ Jacobian, or sensitivity matrix, which we denote $\mathbf{J}$. This gives the derivative of $\mathbf{f}$ with respect to the model parameters, with $J_{ij} = \partial f_i/\partial m_j$. Newman & Alumbaugh (1997); Spitzer (1998); Rodi & Mackie (2001) provide detailed expressions for $\mathbf{J}$ for some specific EM inverse problems, and we will consider the general case extensively below (Section 3).

Search for a minimizer of (2) using $\mathbf{J}$ is iterative, as, for example, in the classical GN method. Let $\mathbf{m}_n$ be the model parameter at the nth iteration, $\mathbf{J}$ the sensitivity matrix evaluated at $\mathbf{m}_n$ and $\mathbf{r} = \mathbf{d} - \mathbf{f}(\mathbf{m}_n)$ the data residual. Then, linearizing the penalty functional in the vicinity of $\mathbf{m}_n$ for small perturbations $\delta\mathbf{m}$ leads to the $M\times M$ system of normal equations

$$(\mathbf{J}^T\mathbf{J}+\nu\mathbf{I})\delta\mathbf{m} = \mathbf{J}^T\mathbf{r}-\nu\mathbf{m}_n, \tag{3}$$

which can be solved for $\delta\mathbf{m}$ to yield a new trial solution $\mathbf{m}_{n+1} = \mathbf{m}_n + \delta\mathbf{m}$. As discussed in Parker (1994), this basic linearized scheme generally requires some form of step length damping for stability (e.g. a Levenberg–Marquardt approach; Marquardt 1963; Rodi & Mackie 2001).

There are many variants to this basic algorithm. For example, in the Occam approach (Constable *et al.* 1987; Parker 1994), (3) is rewritten as

$$(\mathbf{J}^T\mathbf{J} + \nu\mathbf{I})\mathbf{m} = \mathbf{J}^T\hat{\mathbf{d}}, \tag{4}$$

where $\hat{\mathbf{d}} = \mathbf{d} - \mathbf{f}(\mathbf{m}_n) + \mathbf{J}\mathbf{m}_n$. Although $\mathbf{m}_{n+1}$ is obtained directly as the solution to (4) the result is exactly equivalent to solving (3) for the change in the model at step $n + 1$, and adding the result to $\mathbf{m}_n$. A more substantive difference is that in the Occam scheme step length control is achieved by varying ν, computing a series of trial solutions, and choosing the regularization parameter so that data misfit is minimized. An advantage of this approach is that ν is determined as part of the search process, and at convergence one is assured that the solution attains at least a local minimum of the model norm $||\mathbf{m}|| = (\mathbf{m}^T\mathbf{m})^{1/2}$, subject to the data fit attained (Parker 1994). The Occam scheme can also be implemented in the data space (Siripunvaraporn & Egbert 2000; Siripunvaraporn *et al.* 2005). The solution to (4) can be written as

$$\mathbf{m}_{n+1} = \mathbf{J}^T\mathbf{b}_n; \quad (\mathbf{J}\mathbf{J}^T + \nu\mathbf{I})\mathbf{b}_n = \hat{\mathbf{d}}, \tag{5}$$

as can be verified by substituting (5) into (4) and simplifying. This approach requires solving an $N_\mathrm{d} \times N_\mathrm{d}$ system of equations in the data space instead of the $M \times M$ model space system of equations (4), and can thus be more efficient if the model is heavily overparametrized.

Computing the full Jacobian $\mathbf{J}$ required for any direct GN algorithm is a very demanding computational task for multidimensional EM problems, since (as we shall see in Section 3) the equivalent of one forward solution is required for each row (or column) of $\mathbf{J}$. An alternative is to solve the normal eqs (4) or (5) with a memory efficient iterative Krylov-space solver such as conjugate gradients (CG). This requires computation of matrix-vector products such as $[\mathbf{J}^T\mathbf{J} + \nu\mathbf{I}]\mathbf{m}$, which can be accomplished without forming or storing $\mathbf{J}$ at the cost of two forward solutions (e.g. Mackie & Madden 1993). Mackie & Madden (1993), Zhang *et al.* (1995), Newman & Alumbaugh (1997), Rodi & Mackie (2001) and others have used CG to solve (3), whereas Siripunvaraporn & Egbert (2007) have applied the same approach to the corresponding data space equations of (5).

The GN scheme requires solving normal equations derived from a quadratic approximation to (1). Alternatively, the penalty functional can be directly minimized using a gradient-based optimization algorithm such as NLCG (e.g. Rodi & Mackie 2001; Newman & Boggs 2004; Avdeev 2005; Kelbert *et al.* 2008). With this NLCG approach, one must evaluate the gradient of (1) with respect to variations in model parameters $\mathbf{m}$

$$\left.\frac{\partial\Phi}{\partial\mathbf{m}}\right|_{\mathbf{m}_n} = -2\,\mathbf{J}^T\mathbf{r} + 2\,\nu\mathbf{m}_n. \tag{6}$$

The gradient is then used to calculate a new 'conjugate' search direction in the model space. After minimizing the penalty functional along this direction using a line search which requires at most a few evaluations of the forward operator, the gradient is recomputed. NLCG again uses essentially the same basic computational steps as required for solving the linearized equations (3). In particular, the forward problem must be solved to evaluate $\mathbf{f}(\mathbf{m})$ and the residual $\mathbf{r}$ must be multiplied by $\mathbf{J}^T$. Quasi-Newton schemes (e.g. Nocedal & Wright 1999; Newman & Boggs 2004; Haber 2005; Avdeev & Avdeeva 2009) provide an alternative approach to NLCG for direct minimization of (1), with similar advantages with regard to storage and computation of the Jacobian, and similar computational requirements.

All of these schemes for minimizing (1) can be abstractly expressed in terms of a small number of basic 'objects' (data and model parameter vectors, $\mathbf{d}$ and $\mathbf{m}$), and operators (the forward mapping $\mathbf{f}(\mathbf{m})$, multiplication by the corresponding Jacobian $\mathbf{J}$ and its transpose $\mathbf{J}^T$ (and, implicitly, the data and model covariances $\mathbf{C}_\mathbf{m}$ and $\mathbf{C}_\mathbf{d}$). Given modular computer codes which implement these basic objects, any of the inversion algorithms outlined here, as well as many variants, are readily implemented. In the next sections, we analyse further the discrete forward problem, and provide a finer grained general formulation of the modules required to implement essentially any linearized inverse scheme for any EM problem. In particular, we provide 'recipes' for $\mathbf{J}$ in terms of more basic objects associated with the model parametrization, the forward solver and the numerical simulation of the observation operators.

3 DATA SENSITIVITIES

The EM forward operator $\mathbf{f}(\mathbf{m})$ generally involves two steps: (1) Maxwell's equations, with conductivity defined by the parameter $\mathbf{m}$ are solved numerically with appropriate boundary conditions and sources; (2) the resulting solution is used to compute predicted data—for example, an electric or magnetic field component, TF or apparent resistivity—at a set of site locations. For the first step, we write the numerical discretization of the frequency-domain EM partial differential equation (PDE) generically as

$$\mathbf{S}_\mathbf{m}\mathbf{e} = \mathbf{b}, \tag{7}$$

where the vector $\mathbf{b}$ gives appropriate boundary and forcing terms for the particular EM problem, $\mathbf{e}$ is the N_e-dimensional vector representing the discretized electric and/or magnetic fields (or perhaps potential functions), and $\mathbf{S}_\mathbf{m}$ is an $N_\mathbf{e} \times N_\mathbf{e}$ coefficient matrix which depends on the M-dimensional model parameter $\mathbf{m}$. We take $\mathbf{e}$ to represent both interior and boundary components of the discrete solution vector, so that any boundary conditions required for the problem are included in $\mathbf{b}$. The second step then takes the form

$$f_j(\mathbf{m}) = \psi_j(\mathbf{e}(\mathbf{m}), \mathbf{m}), \tag{8}$$

where ψ_j is some generally non-linear, but usually simple, function of the components of $\mathbf{e}$ (and possibly $\mathbf{m}$).

With this general setup we have, by the chain rule,

$$J_{jk} = \frac{\partial f_j}{\partial m_k} = \sum_l \frac{\partial\psi_j}{\partial e_l}\frac{\partial e_l}{\partial m_k} + \frac{\partial\psi_j}{\partial m_k}. \tag{9}$$

Letting $\mathbf{F}$, $\mathbf{L}$, $\mathbf{Q}$ be the partial derivative matrices

$$F_{lk} = \left.\frac{\partial e_l}{\partial m_k}\right|_{\mathbf{m}_0} \quad L_{jl} = \left.\frac{\partial\psi_j}{\partial e_l}\right|_{\mathbf{e}_0,\mathbf{m}_0} \quad Q_{jk} = \left.\frac{\partial\psi_j}{\partial m_k}\right|_{\mathbf{e}_0,\mathbf{m}_0}, \tag{10}$$

where $\mathbf{e}_0$ is the solution to (7) for model parameter $\mathbf{m}_0$, the Jacobian at $\mathbf{m}_0$ can be written in matrix notation as

$$\mathbf{J} = \mathbf{L}\mathbf{F} + \mathbf{Q}. \tag{11}$$

The jth row of $\mathbf{L}$ represents the linearized data functional, which is applied to the perturbation in the EM solution to compute the perturbation in d_j. These row vectors are generally very sparse, supported only on a few nodes surrounding the corresponding data site. When the observation functionals are independent of the model parameters (as they often are) $\mathbf{Q} \equiv \mathbf{0}$. When $\mathbf{Q}$ is required it is also typically sparse, but this depends on the specific nature of the model parametrization. Although, as we show below, derivation of expressions for $\mathbf{L}$ and $\mathbf{Q}$ can be quite involved for realistic EM data functionals, calculation of $\mathbf{F}$ presents the only real computational burden.

To derive a general expression for $\mathbf{F}$, differentiate (7) at $\mathbf{m}_0$ with respect to the model parameters $\mathbf{m}$. We assume that $\mathbf{b}$ is constant, independent of $\mathbf{m}$, although, as discussed in Appendix A (see also Newman & Boggs 2004) some subtle issues related to this point may arise with specific solution approaches. Then, letting $\mathbf{e}_0$ be the solution of (7) at $\mathbf{m}_0$, and noting that the EM solution $\mathbf{e}$ varies as $\mathbf{m}$ is varied, we obtain

$$\mathbf{S}_{\mathbf{m}_0}\left[\left.\frac{\partial \mathbf{e}}{\partial \mathbf{m}}\right|_{\mathbf{m}=\mathbf{m}_0}\right] = -\left.\frac{\partial}{\partial \mathbf{m}}(\mathbf{S}_{\mathbf{m}}\mathbf{e}_0)\right|_{\mathbf{m}=\mathbf{m}_0}, \tag{12}$$

or

$$\mathbf{S}_{\mathbf{m}_0}\mathbf{F} = \mathbf{P}. \tag{13}$$

The $N_e \times M$ matrix $\mathbf{P}$ depends on details of both the numerical model implementation and the conductivity parametrization (as discussed later), but is in general inexpensive to calculate, once the solution $\mathbf{e}_0$ has been computed. Putting together (11) and (13), we obtain an expression for the numerical Jacobian (or sensitivity matrix) $\mathbf{J}$ for general EM problems

$$\mathbf{J} = \mathbf{L}\mathbf{S}_{\mathbf{m}_0}^{-1}\mathbf{P} + \mathbf{Q}. \tag{14}$$

Computing all of $\mathbf{J}$ would appear to require solving the induction equation M times (i.e. applying the inverse operator $\mathbf{S}_{\mathbf{m}_0}^{-1}$ to each of the columns of $\mathbf{P}$.) However, simply taking the transpose of (14) we obtain

$$\mathbf{J}^T = \mathbf{P}^T[\mathbf{S}_{\mathbf{m}_0}^T]^{-1}\mathbf{L}^T + \mathbf{Q}^T, \tag{15}$$

so the sensitivity matrix can in fact be obtained by solving the transposed discrete EM system N_d times (once for each column of $\mathbf{L}^T$), the usual 'reciprocity' trick for efficient calculation of sensitivities (e.g. Rodi 1976; de Lugao *et al.* 1997). It should also be emphasized that for many of the inversion algorithms described in Section 2, $\mathbf{J}$ is not explicitly calculated. Instead, a series of multiplications of model space vectors by $\mathbf{J}$ and/or data space vectors by $\mathbf{J}^T$ are required. These, in turn, require implementation of the component operators $\mathbf{P}$, $\mathbf{L}$, $\mathbf{Q}$ and the solver $\mathbf{S}_{\mathbf{m}_0}^{-1}$, together with the adjoints (or transposes) of these operators.

The EM equations are self-adjoint (except for time reversal) with respect to the usual L^2 inner product (i.e. reciprocity holds). For now leaving aside complications regarding boundary conditions (these are discussed in Appendix B), this implies that on a uniform grid operator $\mathbf{S}_{\mathbf{m}}$ is symmetric. For more general grids, the fact that the EM operator is self-adjoint implies

$$\mathbf{S}_{\mathbf{m}}^T = \mathbf{V}\mathbf{S}_{\mathbf{m}}\mathbf{V}^{-1}, \tag{16}$$

where $\mathbf{V}$ is a diagonal matrix of integration volume elements for the natural discrete representation of the L_2 integral inner product on the model domain (see Appendix B). Eq. (16) implies $\mathbf{S}_{\mathbf{m}}^T\mathbf{V} = \mathbf{V}\mathbf{S}_{\mathbf{m}}$ is a symmetric (though not Hermitian) matrix. It is easier to compute solutions to this symmetrized problem, so solutions to the forward problem are generally computed as $\mathbf{e} = (\mathbf{V}\mathbf{S}_{\mathbf{m}})^{-1}\mathbf{V}\mathbf{b}$ (e.g. see Uyeshima & Schultz 2000). The solution for the adjoint problem can also be written in terms of the symmetrized inverse operator as $\mathbf{e} = (\mathbf{S}_{\mathbf{m}}^T)^{-1}\mathbf{b} = \mathbf{V}(\mathbf{V}\mathbf{S}_{\mathbf{m}})^{-1}\mathbf{b}$; the principal difference from the forward case is thus the order in which multiplication by the diagonal matrix $\mathbf{V}$ and the symmetrized solver are called. In general, the adjoint solver $(\mathbf{S}_{\mathbf{m}}^T)^{-1}$ for EM problems is trivial to implement, once a suitably general forward solver is available.

Before proceeding, two general points require discussion. First, we note that many of the computations in frequency-domain EM problems are most efficiently implemented (and described) using complex arithmetic, but the model conductivity parameter $\mathbf{m}$ is real. Data might be complex (e.g. in MT a complex impedance, formed as the ratio of electric and magnetic fields) or real (e.g. an apparent resistivity or phase, derived from the MT impedance). As already implicit in our formulation of the penalty functional (1), we formally assume that all data are real, that is, real and imaginary parts of a complex observation are separate elements of the real data vector $\mathbf{d}$, and that the basic operators $\mathbf{f}$ and $\mathbf{J}$ have been recast as real mappings from model parameter to data vector. However, we will frequently use complex notation and we will often be somewhat casual in moving between real and complex variables. Thus, for example, the frequency domain forward problem (7) will generally be formulated and solved in terms of complex variables. $\mathbf{P}$ will then be a mapping from the real parameter space to the complex space of forcings for the forward problem, whereas $\mathbf{L}$ will be a mapping from a complex, back to a real space. Both $\mathbf{P}$ and $\mathbf{L}$ can be most conveniently represented by complex matrices, with the convention that for $\mathbf{L}$ only the real part of the matrix-vector product is retained. We discuss this more explicitly in Appendix C.

Secondly, in most cases EM data are obtained for a large number of distinct sources, that is, different frequencies and/or different current source geometries. For example, in the case of MT, there are data for two source polarizations and a wide range of frequencies, whereas for controlled source problems there may be a multiplicity of transmitter geometries or locations. Each of these distinct sources, which we refer to in general as 'transmitters', requires solving a separate forward problem. In most, but not all, cases these forward problems are decoupled, so the data vector and forward modelling operator can be decomposed into $t = 1, \ldots, N_T$ blocks, one for each transmitter

$$\mathbf{d} = \begin{pmatrix} \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_{N_T} \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} \mathbf{f}_1 \\ \vdots \\ \mathbf{f}_{N_T} \end{pmatrix}. \tag{17}$$

Here $\mathbf{d}_t$ gives the data associated with a group of 'receivers', consisting of possibly multiple components, at multiple locations, all making observations of fields generated by transmitter t. Thus, with multiple (decoupled) transmitters the Jacobian can be partitioned into N_T blocks in the obvious way, and each block can be represented as in (14), so that the full sensitivity matrix can be expressed as

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_{N_T} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_1\mathbf{S}_{1,\mathbf{m}}^{-1}\mathbf{P}_1 + \mathbf{Q}_1 \\ \vdots \\ \mathbf{L}_{N_T}\mathbf{S}_{N_T,\mathbf{m}}^{-1}\mathbf{P}_{N_T} + \mathbf{Q}_{N_T} \end{pmatrix}. \tag{18}$$

The matrices $\mathbf{P}_t$ and $\mathbf{Q}_t$ generally depend on the solution for transmitter t. If the transmitter t only specifies the source geometry, the differential operator for the PDE $\mathbf{S}_{t,\mathbf{m}}$ will be independent of the transmitter; however, in general the transmitter will also define the forward problem to solve. An obvious example is the MT case, where the operator depends on frequency. Only $\mathbf{L}_t$ and $\mathbf{Q}_t$ depend on the configuration of receivers; these also in general depend on the forward solution, and thus on transmitter t.

There is an important complication to the simple form of (18), most clearly illustrated by the case of 3-D MT. Here, evaluation of the forward operator for an impedance tensor element requires solutions for the pair of transmitters associated with two uniform source polarizations. Thus, for 3-D MT, the rows of the Jacobian corresponding to a single frequency are formed from components for two transmitters, corresponding to N–S and E–W polarized uniform sources, coupled through the linearized measurement operators

L and **Q**,

$$\mathbf{J} = \sum_{t=1}^{2} \left[\mathbf{L}_t \mathbf{S}_{\mathbf{m}}^{-1} \mathbf{P}_t + \mathbf{Q}_t \right]. \tag{19}$$

The same complication would arise for any plane wave source TF (e.g. vertical, or intersite magnetic). Other examples of multiple polarization EM inverse problems can be imagined, for example, allowing for a combination of horizontal spatial gradients and uniform sources (e.g. Egbert 2002; Schmücker 2004; Semenov & Shuman 2009) would require allowing for five coupled sources. Pankratov & Kuvshinov (2010) discuss the general multiple polarization problem from a theoretical perspective, although to our knowledge, no actual applications of the theory to inversion of real data sets have yet been reported, beyond the standard two polarization uniform source case.

We return to the coupled multiple polarization case in Section 5.2, where we discuss measurement operators in more detail. Then, in Section 6, we consider the general multiple transmitter case further, and show more explicitly how the source and receiver configuration can result in special structure for the Jacobian, which can be exploited to improve computational efficiency. For the next few sections, we focus on the simpler case of a single transmitter, as we develop the basic building blocks for more complex and realistic problems.

4 DISCRETIZATION OF THE FORWARD PROBLEM

To derive more explicit expressions for the operators **L**, **P** and **Q**, and hence for the full Jacobian, more specific assumptions about the numerical implementation of the forward problem (7) are required. To motivate the general development, we consider as examples two specific cases in detail: inversion of 2-D and 3-D MT data. We discuss most explicitly a finite difference (FD) modelling approach, though most of the results obtained are more broadly applicable.

Numerical schemes for solving Maxwell's equations are often most elegantly formulated in terms of a pair of vector fields defined on conjugate grids. For example, the space of primary fields which we denote as $\mathcal{S}_\mathrm{P}$ may represent the electric fields, whereas the space of dual fields, denoted $\mathcal{S}_\mathrm{D}$, represents the magnetic fields. Even when the coupled first-order system (i.e. Maxwell's equations) is reduced to a second-order equation involving only the primary field, it is worthwhile to explicitly consider the dual field also. Most obviously, in many applications both electric and magnetic field components are required to evaluate the data functionals. Furthermore, depending on the model formulation, the dependence of the discrete PDE operator coefficients on the model parameter can generally be represented most explicitly through a mapping $\pi(\mathbf{m})$ from the model parameter space $\mathcal{M}$—sometimes to $\mathcal{S}_\mathrm{P}$, but in other cases to $\mathcal{S}_\mathrm{D}$, and a general treatment should allow for both cases. Boundary conditions are of course a critical part of the formulation of the forward problem. These are included implicitly in our generic formula of the forward problem (7). In the following, we omit technical details concerning implementation of boundary conditions, leaving discussion of these issues to Appendix B.

As a first illustration, we consider FD modelling of the 3-D quasi-static Maxwell's equations appropriate for MT. In the frequency domain (assuming a time dependence of $e^{i\omega t}$), the magnetic fields can be eliminated, resulting in a second-order elliptic system of PDEs in terms of the electric fields alone,

$$\nabla \times \nabla \times \mathbf{E} + i\omega\mu\sigma\mathbf{E} = 0, \tag{20}$$

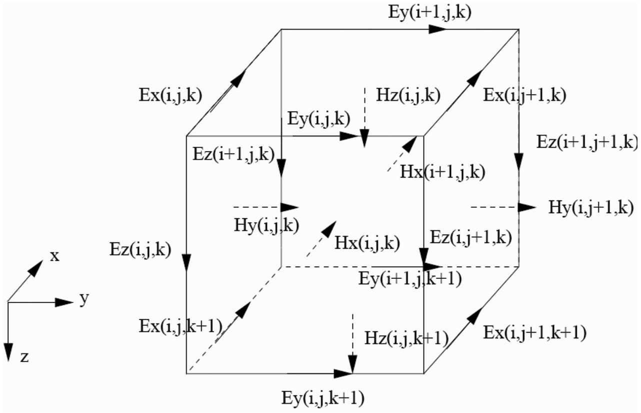


Figure 1. Staggered finite difference grid for the 3-D MT forward problem. Electric field components defined on cell edges are the primary EM field component, which the PDE is formulated in terms of. The magnetic field components can be defined naturally on the cell faces; these are the secondary EM field in this numerical formulation.

where ω is the angular frequency, μ is magnetic permeability and σ is electrical conductivity, with the tangential components of **E** specified on all boundaries. To solve (20) numerically in 3-D, we consider an FD approximation on a staggered grid of dimension $N_x \times N_y \times N_z$, as illustrated in Fig. 1 (e.g. Yee 1966; Smith 1996; Siripunvaraporn *et al.* 2002). In the staggered grid formulation, the discretized electric field vector components are defined on cell edges (Fig. 1). In our terminology, the primary field space $\mathcal{S}_\mathrm{P}$ is the space of such finite-dimensional cell edge vector fields. A typical element will be denoted by **e**. As illustrated in Fig. 1, the magnetic fields, which in continuous form satisfy $\mathbf{H} = (-i\omega\mu)^{-1}\nabla \times \mathbf{E}$, are naturally defined on the discrete grid of cell faces. The dual-field space $\mathcal{S}_\mathrm{D}$ is thus the space of discrete vector fields defined on faces. A typical element of this space will be denoted by **h**.

In the staggered grid FD discretization used for (20), the discrete magnetic and electric fields are related via

$$\mathbf{h} = (-i\omega\mu)^{-1}\mathbf{C}\,\mathbf{e}, \tag{21}$$

where $\mathbf{C} : \mathcal{S}_\mathrm{P} \mapsto \mathcal{S}_\mathrm{D}$ is the discrete approximation of the curl of cell edge vectors, and (20) can be expressed in its discrete form as

$$[\mathbf{C}^\dagger\mathbf{C} + \mathrm{diag}(i\omega\mu\sigma(\mathbf{m}))]\mathbf{e} = 0. \tag{22}$$

Here, diag(**v**) denotes a diagonal matrix with the components of the vector **v** on the diagonal, and $\mathbf{C}^\dagger : \mathcal{S}_\mathrm{D} \mapsto \mathcal{S}_\mathrm{P}$ is the discrete curl mapping interior cell face vectors to interior cell edges. As the notation indicates this operator is the adjoint of **C**, relative to appropriate (i.e. volume weighted) inner products on the spaces $\mathcal{S}_\mathrm{D}$ and $\mathcal{S}_\mathrm{P}$. Although **e** is the full solution vector (including boundary components), (22) provides equations only for the interior nodes. Additional equations are required to constrain **e** on the boundary, and to complete specification of the discrete forward operator $\mathbf{S}_\mathbf{m}$. These details, and further discussion of **C** and its adjoint, are provided in Appendix B. The key point here is that the dependence of the operator coefficients on the model parameter (which we take to be an element of some finite-dimensional space $\mathcal{M}$) is made explicit through the mapping $\sigma : \mathcal{M} \mapsto \mathcal{S}_\mathrm{P}$ in (22).

The 3-D EM induction forward problem can also be formulated in terms of magnetic fields

$$\nabla \times \rho\nabla \times \mathbf{H} + i\omega\mu\mathbf{H} = 0, \tag{23}$$

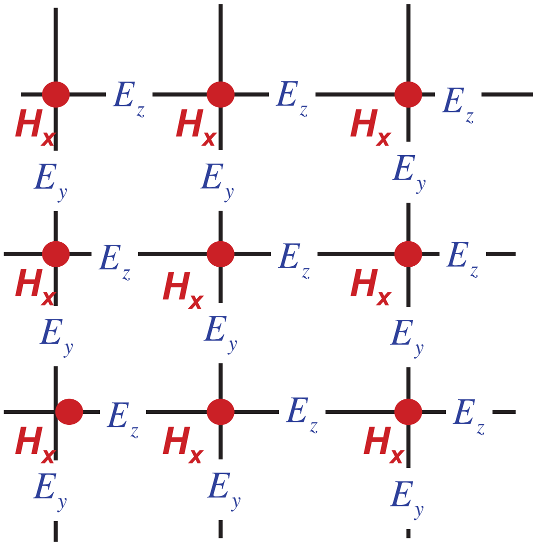


Figure 2. Finite difference grid for the 2-D MT TM mode forward problem. The scalar H_x magnetic field, defined on 2-D cell corners is the primary field. The secondary field components are E_y and E_z, defined on vertical and horizontal cell edges, respectively.

where ρ is electrical resistivity, now with the tangential component of the magnetic fields specified on boundaries. With this formulation (e.g. Mackie *et al.* 1994; Uyeshima & Schultz 2000), **H** would be the primary field, and the electric field $\mathbf{E} = \rho\nabla \times \mathbf{H}$ would be the dual field. Using an analogous staggered grid FD discretization, with magnetic field components defined on cell edges, and electric field components defined on cell faces, the discrete induction equation now takes the form

$$[\mathbf{C}^\dagger \operatorname{diag}(\rho(\mathbf{m}))\mathbf{C} + \operatorname{diag}(\mathrm{i}\omega\mu)]\mathbf{e} = 0. \tag{24}$$

In this case, the dependence of the coefficients on model parameter **m** is made explicit through the mapping to the dual-field space $\rho : \mathcal{M} \mapsto \mathcal{S}_\mathrm{D}$. Note that for both the electric and magnetic field formulations **e** represents the primary field, and **h** the dual field. Thus, in (24), **e** represents the discrete magnetic field **H**, and **h** would represent the discrete electric field.

It is also instructive to consider the 2-D MT inverse problem. Now there are effectively two distinct modelling problems: for transverse electric (TE) and transverse magnetic (TM) modes, with electric and magnetic fields, respectively, parallel to the geologic strike. The TE mode case is essentially identical to the 3-D electric field formulation of (20)–(22). The TM mode case, which is solved in terms of the magnetic field instead of the electric field, is more instructive with regard to generalization. In the TM mode, the magnetic field parallels the geological strike (x) and (23) can be reduced to a scalar PDE in the y-z plane

$$\partial_y \rho \partial_y H_x + \partial_z \rho \partial_z H_x + \mathrm{i}\omega\mu H_x = 0, \tag{25}$$

with H_x specified on boundaries.

As for the 3-D problems, for the discrete 2-D problem we can define finite-dimensional spaces of primary ($\mathcal{S}_\mathrm{P}$) and dual ($\mathcal{S}_\mathrm{D}$) EM fields. Now the primary field is H_x, defined on the nodes (corners) of the 2-D grid, and the dual fields are the electric field components E_y and E_z defined on the vertical and horizontal cell edges (Fig. 2). A natural centred FD approximation of (25) can be written in terms of a discrete 2-D gradient operator $\mathbf{G} : \mathcal{S}_\mathrm{P} \mapsto \mathcal{S}_\mathrm{D}$ and a 2-D divergence operator $\mathbf{D} : \mathcal{S}_\mathrm{D} \mapsto \mathcal{S}_\mathrm{P}$. Using $\mathbf{e} \in \mathcal{S}_\mathrm{P}$ to denote the primary discrete EM field solution (H_x), we have a more explicit form of (7) for this discrete TM mode implementation,

$$[\mathbf{D} \operatorname{diag}(\rho(\mathbf{m}))\, \mathbf{G} + \mathrm{i}\omega\mu\mathbf{I}]\, \mathbf{e} = 0, \tag{26}$$

with additional equations again required to specify boundary conditions.

In most other FD or finite volume modelling approaches, for example, with Maxwell's equations cast in terms of vector potentials, similar (although potentially more complicated) sets of conjugate spaces can be defined, the differential operator can be decomposed into discrete approximations to first-order linear differential operators which map between conjugate grids, and the dependence of discrete operator coefficients on an abstract model parameter space can be described explicitly through a mapping $\pi : \mathcal{M} \mapsto \mathcal{S}_\mathrm{P,D}$. Finite-element approaches to EM modelling will result in similar structures. For example, the space of linear edge elements (or more properly, the degrees of freedom associated with these elements; Nedelec 1980) can be taken as the primary space, representing the discrete electric field. The natural dual space is then the space of face elements, representing the discrete magnetic field (e.g. Rodrigue & White 2001). The natural model parameter mapping then defines conductivity associated with each edge degree of freedom.

5 COMPONENTS OF THE JACOBIAN

5.1 Matrix P

We can give an explicit expression for the operator **P**, assuming the forward operator $\mathbf{S_m}$ can be written in the general form

$$\mathbf{S_m e} \equiv \mathbf{S}_0\mathbf{e} + \mathbf{U}\,(\pi(\mathbf{m}) \circ \mathbf{Ve}), \tag{27}$$

where $\mathbf{S}_0$, **U** and **V** are some linear operators that do not depend on the model parameter vector **m**, $\pi(\mathbf{m})$ is a (possibly non-linear) operator that maps the model parameter space $\mathcal{M}$ to the primary or dual grid, and ($\circ$) denotes the component-wise multiplication of the two vectors in $\mathcal{S}_{P,D}$ (also known as the Hadamard product). Note that on an FD grid, $\mathbf{S_m}$ (and hence $\mathbf{S}_0$ and **V**) act on a full solution vector that includes both the interior and boundary edges (see Appendix B). All of the examples outlined earlier are special cases of (27), as we will discuss.

Assuming (27) and recalling the definition of **P** from (12) and (13), we find

$$\mathbf{P} = -\frac{\partial}{\partial \mathbf{m}}(\mathbf{S_m e}_0)\bigg|_{\mathbf{m}_0} = -\mathbf{U}\left(\frac{\partial \pi}{\partial \mathbf{m}}\bigg|_{\mathbf{m}_0} \circ \mathbf{Ve}_0\right), \tag{28}$$

$$= -\mathbf{U}\left(\mathbf{Ve}_0 \circ \frac{\partial \pi}{\partial \mathbf{m}}\bigg|_{\mathbf{m}_0}\right), \tag{29}$$

$$= -\mathbf{U}\operatorname{diag}(\mathbf{Ve}_0)\,\frac{\partial \pi}{\partial \mathbf{m}}\bigg|_{\mathbf{m}_0}. \tag{30}$$

Writing $\Pi_{\mathbf{m}_0}$ for the Jacobian of the (in general, non-linear) model parameter mapping $\pi(\mathbf{m})$ evaluated at the background model parameter $\mathbf{m}_0$, we have

$$\mathbf{P} = -\mathbf{U}\operatorname{diag}(\mathbf{Ve}_0)\Pi_{\mathbf{m}_0}, \tag{31}$$

$$\mathbf{P}^T = -\Pi^T_{\mathbf{m}_0}\operatorname{diag}(\mathbf{Ve}_0)\mathbf{U}^T. \tag{32}$$

Note that only the operator $\Pi_{\mathbf{m}}$ depends on the details of the model parametrization—the other terms depend only on the numerical discretization of the governing equations. Eqs (31) and (32) provide broadly applicable recipes for implementation of the operators $\mathbf{P}$ and $\mathbf{P}^T$, as illustrated in the following examples. If the dependence of the forward operator on the model parameter cannot be cast as a special case of (27), similar formal steps could almost certainly be used to derive appropriate expressions for these operators.

5.1.1 Example: 2-D MT

For the 2-D TM problem (26), the PDE coefficients depend on the model parameters through $\rho : \mathcal{M} \mapsto \mathcal{S}_{\mathrm{D}}$, that is, the resistivity $\rho(\mathbf{m})$ defined on the dual grid, cell edges. To be specific, we consider the simplest model parametrizations, with conductivity or log conductivity for each cell in the numerical grid an independent parameter. From physical considerations, it is most reasonable to compute the required edge resistivities from cell conductivities by first transforming to resistivity, and then computing the area weighted average of resistivities of the two cells on either side of the edge. Representing the averaging operator from 2-D cells to cell sides as $\mathbf{W}_{\mathrm{TM}}$, and letting $(\mathbf{m})^{-1}$ denote the component-wise inverse of the model parameter vector, we then have

$$\rho(\mathbf{m}) = \mathbf{W}_{\mathrm{TM}}(\mathbf{m})^{-1}, \tag{33}$$

$$\rho(\mathbf{m}) = \mathbf{W}_{\mathrm{TM}}\mathbf{exp}(-\mathbf{m}), \tag{34}$$

for linear and log conductivity, respectively. The model operator of (26) can be cast in the general form of (27) with the identifications $\mathbf{S}_0 \equiv -\mathrm{i}\omega\mu\mathbf{I}$, $\mathbf{U} \equiv \mathbf{D}$, $\mathbf{V} \equiv \mathbf{G}$ and $\pi(\mathbf{m}) \equiv \rho(\mathbf{m})$, where $\mathbf{D}$ and $\mathbf{G}$ are the discrete 2-D divergence and gradient operators defined in Section 4. Thus, we obtain the expressions for $\mathbf{P}$ and $\mathbf{P}^T$ in the 2-D TM mode case.

$$\mathbf{P} = -\mathbf{D}\mathrm{diag}(\mathbf{G}\mathbf{e}_0)\Pi_{\mathbf{m}_0}, \tag{35}$$

$$\mathbf{P}^T = -\Pi_{\mathbf{m}_0}^T\mathrm{diag}(\mathbf{G}\mathbf{e}_0)\mathbf{D}^T, \tag{36}$$

where $\Pi_{\mathbf{m}_0} = -\mathbf{W}_{\mathrm{TM}}[\mathrm{diag}(\mathbf{m}_0)\ ]^{-2}$ for the parametrization in terms of linear conductivity, or $\Pi_{\mathbf{m}_0} = -\mathbf{W}_{\mathrm{TM}}[\mathrm{diag}(\mathbf{exp}(-\mathbf{m}_0))\ ]$ for log conductivity.

5.1.2 Example: 3-D MT

We again assume the simplest model parametrization, with conductivity, or the natural logarithm of conductivity, specified independently for each of the $M = N_xN_yN_z$ cells in the numerical grid. The discrete operator of (22) requires conductivity defined on cell edges, where the electric field components are defined. For physical consistency (current should be conserved), the edge conductivities should represent the volume weighted average of the surrounding four cells. Let $\mathbf{W}$ be the $N_{\mathrm{e}} \times M$ matrix representing this weighted averaging operator, a mapping from $\mathcal{M}$ to $\mathcal{S}_{\mathrm{P}}$. Then, the conductivity parameter mapping is given by $\sigma(\mathbf{m}) = \mathbf{W}\mathbf{m}$ or $\sigma(\mathbf{m}) = \mathbf{W}\mathbf{exp}(\mathbf{m})$, for the cases of linear and log conductivity, respectively.

Eq. (22) can be seen to be a special case of (27) with the identifications $\mathbf{S}_0 \equiv \mathbf{C}^\dagger\mathbf{C}$, $\mathbf{U} \equiv \mathrm{i}\omega\mu\mathbf{I}$, $\mathbf{V} \equiv \mathbf{I}$ and $\pi(\mathbf{m}) \equiv \sigma(\mathbf{m})$, and we have

$$\mathbf{P} = \mathrm{diag}(-\mathrm{i}\omega\mu\mathbf{e}_0)\ \Pi_{\mathbf{m}_0}, \tag{37}$$

$$\mathbf{P}^T = \ \Pi_{\mathbf{m}_0}^T\mathrm{diag}(\mathrm{i}\omega\mu\mathbf{e}_0), \tag{38}$$

where $\Pi_{\mathbf{m}_0} = \mathbf{W}$ for linear conductivity, and $\Pi_{\mathbf{m}_0} = \mathbf{W}[\mathrm{diag}(\mathbf{exp}(\mathbf{m}_0))]$ for logarithmic conductivity. Note that the transposes of the averaging operators $\mathbf{W}$ and $\mathbf{W}_{\mathrm{TM}}$ represent mappings from cell edges to cells, a weighted sum of contributions from all edges that bound a cell.

5.2 Matrices L and Q

We turn now to the matrices $\mathbf{L}$ and $\mathbf{Q}$, which represent the linearized observation process, as it is applied to the discrete numerical forward solution.

*5.2.1 **L**: general case*

The very simplest sort of EM data is an observation of the primary field at a single location (e.g. $\epsilon = E_y(\mathbf{x})$), which can be represented as a local average of the modelled primary field

$$\epsilon = (\lambda^{\mathrm{P}})^T\mathbf{e}. \tag{39}$$

Here $\lambda^{\mathrm{P}} \in \mathcal{S}_{\mathrm{P}}$ is a sparse vector of interpolation coefficients, averaging from the discrete primary grid to the observation point $\mathbf{x}$. A point observation of the dual field (e.g. $\eta = H_x(\mathbf{x})$) is only slightly more complicated. Assuming, as will generally be the case, that the dual fields can be written as $\mathbf{h} = \mathbf{T}\mathbf{e}$, where $\mathbf{T} : \mathcal{S}_{\mathrm{P}} \mapsto \mathcal{S}_{\mathrm{D}}$ is a discrete differential operator (e.g. see 21), we have

$$\eta = (\lambda^{\mathrm{D}})^T\mathbf{T}\mathbf{e}, \tag{40}$$

where $\lambda^{\mathrm{D}} \in \mathcal{S}_{\mathrm{D}}$ is again a sparse vector of interpolation coefficients, now representing averaging on the dual grid. For some problems, $\mathbf{T} \equiv \mathbf{T}_{\pi(\mathbf{m})}$ will depend on the model parameter through $\pi(\mathbf{m})$ (see Section 5.2.4 for an example). It is also possible for the interpolation coefficients λ^{P} and/or λ^{D} to depend on the model parameter $\mathbf{m}$. We will return to these complications, which are accounted for in the operator $\mathbf{Q}$, below.

Note that for a finite-element formulation, where the solution is represented in terms of a discrete set of basis functions, field component evaluation functionals would have the same form (sparse vectors defined on the primary or dual space), but would have a slightly different interpretation—that is, the non-zero components of the evaluation functional for any location would be computed by evaluating (at this point) the basis functions for all degrees of freedom associated with the containing element.

Together, (39) and (40) give the basic evaluation functionals for the fundamental observables (point measurements of magnetic and electric fields) in any EM problem. For controlled source problems, where the data are typically just point measurements of the primary or dual field, these evaluation functionals are already the rows of $\mathbf{L}$. More generally, EM data are functions of both electric and magnetic components, at one or more locations. The most obvious example is the impedance, the local ratio of electric and magnetic fields. Other examples include interstation magnetic TFs, network MT accounting for the geometry of long dipoles (Siripunvaraporn *et al.* 2004), or horizontal spatial gradient methods based on array data (Schmücker 2003; Semenov & Shuman 2009). Inevitably, real data must be based on a discrete set of observations of the magnetic and electric fields, so the general EM data functional can be represented as

$$\psi_j(\mathbf{e}(\mathbf{m}), \mathbf{m}) \equiv \gamma_j(\epsilon_1(\mathbf{m}), \ldots, \epsilon_{K_{\mathrm{P}}}(\mathbf{m}),$$
$$\eta_1(\mathbf{e}(\mathbf{m})), \ldots, \eta_{K_{\mathrm{D}}}(\mathbf{e}(\mathbf{m}))), \tag{41}$$

where ϵ_k, $k = 1, \ldots, K_\mathrm{P}$ and η_k, $k = 1, \ldots, K_\mathrm{D}$ are sets of primary and dual components computed at one or several points in the model domain, as $\epsilon_k = (\lambda_k^\mathrm{P})^T \mathbf{e}$ and $\eta_k = (\lambda_k^\mathrm{D})^T \mathbf{Te}$, respectively.

From (10), the *j*th row of $\mathbf{L}$ is then given by

$$\mathbf{l}_j = \left.\frac{\partial \psi_j}{\partial \mathbf{e}}\right|_{\mathbf{e}_0,\mathbf{m}_0} = \sum_{k=1}^{K_\mathrm{P}} \frac{\partial \gamma_j}{\partial \epsilon_k} \left.\frac{\partial \epsilon_k}{\partial \mathbf{e}}\right|_{\mathbf{e}_0,\mathbf{m}_0} + \sum_{k=1}^{K_\mathrm{D}} \frac{\partial \gamma_j}{\partial \eta_k} \left.\frac{\partial \eta_k}{\partial \mathbf{e}}\right|_{\mathbf{e}_0,\mathbf{m}_0}, \tag{42}$$

$$= \sum_{k=1}^{K_\mathrm{P}} a_{jk}^\mathrm{P} (\lambda_k^\mathrm{P})^T + \sum_{k=1}^{K_\mathrm{D}} a_{jk}^\mathrm{D} (\lambda_k^\mathrm{D})^T \mathbf{T}, \tag{43}$$

where a_{jk}^P and a_{jk}^D are the partial derivatives of the *j*th data functional with respect to the *k*th local field components. These coefficients depend only on the details of the data functional formulation, and the background EM solution $\mathbf{e}_0$. Eq. (43) thus implies that we can decompose $\mathbf{L}$ into two sparse matrices as

$$\mathbf{L} = \mathbf{A}^T \Lambda^T, \tag{44}$$

with

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_\mathrm{P} \\ \mathbf{A}_\mathrm{D} \end{bmatrix} \quad \text{and} \quad \Lambda = [\, \Lambda_\mathrm{P} \ \mathbf{T}^T \Lambda_\mathrm{D} \,]. \tag{45}$$

Here, $\mathbf{L}$ is a sparse $N_\mathrm{d} \times N_\mathrm{e}$ matrix that maps the EM solution to the data space, as in Eq. (14). $\mathbf{A}$ is a $K \times N_\mathrm{d}$ sparse matrix ($K = K_\mathrm{P} + K_\mathrm{D}$), such that the non-zero elements in its *j*th column are the coefficients a_{jk}, the derivatives of the data functionals with respect to each of the relevant local magnetic or electric field components. Finally, Λ is an $N_\mathrm{e} \times K$ sparse matrix, with columns $\lambda_k \in \mathcal{S}_\mathrm{P}$ containing the field component evaluation functionals, that is, the interpolation coefficients required to compute the *k*th electric/magnetic field component at a point from the primary EM field.

Thus, Λ depends only on the observation locations for each of the K local field components (and possibly on the model parameter $\mathbf{m}_0$). Observation functionals (non-linear or linearized) for any sort of EM data will be constructed from the same field component functionals, which are closely tied to the specific numerical discretization scheme used. $\mathbf{A}$, however, depends on details of the observation functionals (e.g. impedance versus apparent resistivity), and will also depend, in general, on the background EM solution used for linearization, $\mathbf{e}_0$. However, $\mathbf{A}$ (which is essentially a linearization of γ) does not depend on the details of the numerical implementation of the forward problem.

5.2.2 *L: multivariate TFs*

Multivariate TFs are an important special case of non-linear data functionals which deserve a closer look. Plane wave source TFs provide the most important (and, in fact, only widely applied) example. In this case, two independent sources are assumed, corresponding to spatially uniform sources of a fixed frequency polarized in the *x*- and *y*-directions. As a consequence of the linearity of the induction equations, under this assumption any point observation of the EM fields can be linearly related to two reference components, through a frequency-dependent TF. Examples include the rows of the impedance tensor, such as

$$E_x = Z_{xx} H_x + Z_{xy} H_y, \tag{46}$$

vertical field TFs, and intersite magnetic TFs. TF components such as Z_{xx} and Z_{xy}, which are estimated from time-series of electric and magnetic fields observed at a single site, provide the basic input data for 3-D MT inversion.

For completeness, we consider the general case where a generic predicted component, which we denote as Y, is related to N_p predicting variables $X_1, \ldots, X_{N_\mathrm{p}}$ via the TF

$$Y = \theta_1 X_1 + \cdots + \theta_{N_\mathrm{p}} X_{N_\mathrm{p}}. \tag{47}$$

To evaluate the components of the complex TF vector $\boldsymbol{\Theta} = \left(\theta_1, \ldots, \theta_{N_\mathrm{p}}\right)^T$ it is necessary to solve forward problems for each of the assumed source configurations—that is, forward solutions $\mathbf{e}_1, \ldots, \mathbf{e}_{N_\mathrm{p}}$ for N_p transmitters are required. To compute the TF, we must evaluate Y and X_j, $j = 1, \ldots, N_\mathrm{p}$ for each of these forward solutions. Here, we represent this as

$$Y_i = \lambda_Y^T \mathbf{e}_i \quad X_{ij} = \lambda_{X_j}^T \mathbf{e}_i \quad i = 1, \ldots, N_\mathrm{p}. \tag{48}$$

Then, if $\mathbf{Y}$ denotes the vector of predicted components for the N_p transmitters and $\mathbf{X}$ denotes the corresponding $N_\mathrm{p} \times N_\mathrm{p}$ matrix of predicting variables, the TF can be computed as

$$\boldsymbol{\Theta} = \mathbf{X}^{-1}\mathbf{Y}. \tag{49}$$

Note that, in general, the evaluation functionals $\lambda_{X_j} \in \mathcal{S}_\mathrm{P}$ might be more complicated than the simple interpolation operators considered previously—for example, for the usual plane wave source case the predicting components are typically taken to be magnetic fields at the local site, which for the 3-D MT example we have considered would require computation of the secondary field (multiplication by the operator $\mathbf{T}$) followed by interpolation. And for more exotic cases such as the generalized horizontal spatial gradient (HSG) TF (Egbert 2002; Schmücker 2003, 2004; Semenov & Shuman 2009; Pankratov & Kuvshinov 2010) the predicting components would involve magnetic fields measured at multiple sites, used to form some sort of estimate of uniform and gradient field components. We thus assume only that these are sparse vectors representing linear functionals defined on $\mathcal{S}_\mathrm{P}$.

Taking partial derivatives of $\boldsymbol{\Theta}$ with respect to $\mathbf{e}_i$ we find, after some simplification

$$\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{e}_i} = \mathbf{X}_0^{-1} \left[\frac{\partial \mathbf{Y}}{\partial \mathbf{e}_i} - \frac{\partial (\mathbf{X}\boldsymbol{\Theta}_0)}{\partial \mathbf{e}_i} \right]. \tag{50}$$

In (50), the subscript zero denotes TFs and predicting components evaluated for the background forward solution. Note that the expression in brackets is a matrix of size $N_\mathrm{p} \times N_\mathrm{e}$ (N_e = dimension of $\mathbf{e}_i$), but only the *i*th row is non-zero (only the *i*th component of Y and row of $\mathbf{X}$ depend on solution $\mathbf{e}_i$). This row takes the form

$$\lambda_Y^T - \theta_1 \lambda_{X_1}^T - \cdots - \theta_{N_\mathrm{p}} \lambda_{X_{N_\mathrm{p}}}^T, \tag{51}$$

which is independent of the source polarization index *i*.

As we noted at the end of Section 3, rows of $\mathbf{L}$ for TF components couple the terms $\mathbf{S}_\mathbf{m}^{-1}\mathbf{P}_i$ for multiple transmitters. We can now give an explicit form for this coupling, considering only a single predicted component Y, so that there are N_p complex rows of the matrix $\mathbf{L}$, one for each component of the TF. $\mathbf{L}$ can also be divided into N_p blocks of columns, one for each transmitter as in (19). From (50) and (51), $\mathbf{L}$ can be written in terms of $\mathbf{X}_0^{-1}$ and block diagonal matrices as

$$\mathbf{L} = \mathbf{X}_0^{-1} \begin{bmatrix} \Psi & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Psi \end{bmatrix} \begin{bmatrix} \Lambda^\mathrm{T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Lambda}^T \end{bmatrix}, \tag{52}$$

where

$$\Psi = [\, 1 \ -\boldsymbol{\Theta}^T \,] \qquad \boldsymbol{\Lambda} = \left[\, \lambda_Y^T \ \lambda_{X_1}^T \ \cdots \ \lambda_{X_{N_\mathrm{p}}}^T \,\right]. \tag{53}$$

The product of the first two matrices corresponds to $\mathbf{A}^T$ (and the rightmost of course to Λ) in (44). The more explicit form here more clearly defines the coupling between transmitters, and has important implications for efficient calculation of the full Jacobian, as we discuss further in Section 6.

5.2.3 *Matrix* $\mathbf{Q}$

When either the evaluation functionals or the field transformation operator $\mathbf{T}$ have an explicit dependence on the model parameter (denoted in the latter case by $\mathbf{T}_{\pi(\mathbf{m})}$) there is an additional term in the sensitivity matrix, which we have denoted $\mathbf{Q}$. The jth row of this matrix is given by

$$\mathbf{q}_j = \left.\frac{\partial \psi_j}{\partial \mathbf{m}}\right|_{\mathbf{e}_0,\mathbf{m}_0} = \left[\sum_{k=1}^{K_\mathrm{P}} a_{jk}^{\mathrm{P}} \left.\frac{\partial \left(\lambda_k^{\mathrm{P}}\right)^T \mathbf{e}_0}{\partial \pi}\right|_{\pi(\mathbf{m}_0)} + \sum_{k=1}^{K_\mathrm{D}} a_{jk}^{\mathrm{D}} \right.$$
$$\left. \times \left(\left.\frac{\partial \left(\lambda_k^{\mathrm{D}}\right)^T \mathbf{T}_{\pi(\mathbf{m}_0)}\mathbf{e}_0}{\partial \pi}\right|_{\pi(\mathbf{m}_0)} + \left(\lambda_k^{\mathrm{D}}\right)^T \left.\frac{\partial \mathbf{T}_{\pi(\mathbf{m})}\mathbf{e}_0}{\partial \pi}\right|_{\pi(\mathbf{m}_0)} \right)\right] \Pi_{\mathbf{m}_0}, \tag{54}$$

where $\pi(\mathbf{m})$ is the (possibly non-linear) model parameter mapping to the dual or primary grid, and $\Pi_{\mathbf{m}_0}$ is the Jacobian of this mapping. Defining

$$\tilde{\mathbf{T}}_{\pi(\mathbf{m}_0),\mathbf{e}_0} = \left.\frac{\partial}{\partial \pi}[\mathbf{T}_{\pi(\mathbf{m})}\mathbf{e}_0]\right|_{\pi(\mathbf{m}_0)}, \tag{55}$$

$$\tilde{\Lambda}_\mathrm{P}^T = \left.\frac{\partial}{\partial \pi}[(\Lambda_\mathrm{P})^T\mathbf{e}_0]\right|_{\pi(\mathbf{m}_0)} \tag{56}$$

and

$$\tilde{\Lambda}_\mathrm{D}^T = \left.\frac{\partial}{\partial \pi}[(\Lambda_\mathrm{D})^T\mathbf{T}_{\pi(\mathbf{m}_0)}\mathbf{e}_0]\right|_{\pi(\mathbf{m}_0)}. \tag{57}$$

Eq. (54) can be given in matrix notation

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_\mathrm{P}^T & \mathbf{A}_\mathrm{D}^T \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{\Lambda}}_\mathrm{P}^T \\ \tilde{\mathbf{\Lambda}}_\mathrm{D}^T + \mathbf{\Lambda}_\mathrm{D}^T \tilde{\mathbf{T}} \end{bmatrix} \Pi_{\mathbf{m}_0}. \tag{58}$$

If the interpolation coefficients are independent of the model parameters (as will be most often the case) this reduces to

$$\mathbf{Q} = \mathbf{A}_\mathrm{D}^T \mathbf{\Lambda}_\mathrm{D}^T \tilde{\mathbf{T}}_{\pi(\mathbf{m}_0),\mathbf{e}_0} \Pi_{\mathbf{m}_0}. \tag{59}$$

5.2.4 *Example: 2-D MT*

For 2-D MT, the fundamental observation is an impedance, the ratio E/B of orthogonal components of the electric and magnetic fields. For the TE mode, E_x corresponds to the primary (modelled) field $\mathbf{e}$, whereas H_y is the secondary field, which is computed as $\mathbf{h} = \mathbf{T}_E\mathbf{e}$. The secondary field mapping can be given explicitly as $\mathbf{T}_E = (-\mathrm{i}\omega\mu)^{-1}\mathbf{OG}$, where $\mathbf{O}$ is a diagonal matrix with entries $+1$ and -1 for components corresponding to y- and z-edges, respectively. Columns of Λ_P and Λ_D now represent bilinear spline interpolation from the 2-D grid nodes and edges, respectively, to the data sites. These are independent of the model parameter, so $\mathbf{Q} \equiv 0$.

The impedance can be written explicitly as

$$Z \equiv \gamma_j(\mathbf{e}) = \frac{\lambda_E^T\mathbf{e}}{\lambda_H^T\mathbf{T}_E\mathbf{e}}, \tag{60}$$

where $\mathbf{e}$ is the (primary) electric field, and λ_E and λ_H are, respectively, columns of Λ_P and Λ_D, and represent bilinear spline interpolation functionals on node (primary) and edge (dual) spaces. From (42), the row of $\mathbf{L}$ corresponding to an impedance is found to be

$$\mathbf{l}_j \equiv \mathbf{l}_Z = \left(\lambda_H^T\mathbf{T}_E\mathbf{e}_0\right)^{-1}\lambda_E^T - \left[\lambda_E^T\mathbf{e}_0/\left(\lambda_H^T\mathbf{T}_E\mathbf{e}_0\right)^2\right]\lambda_H^T\mathbf{T}_E. \tag{61}$$

For the TM mode, Λ_P and Λ_D are the same as in the TE case, but the roles of primary and dual fields are reversed, so that

$$Z \equiv \gamma_j(\mathbf{e}) = \frac{\lambda_E^T\mathbf{T}_H\mathbf{e}}{\lambda_H^T\mathbf{e}}, \tag{62}$$

$\mathbf{e}$ now denoting the (primary) magnetic field. Also the field transformation operator is now $\mathbf{T}_H = \mathrm{diag}[\rho(\mathbf{m})]\mathbf{OG}$, and thus depends on the model parameter, so $\mathbf{Q}$ will be non-zero. Row j of $\mathbf{L}$ is now

$$\mathbf{l}_j \equiv \mathbf{l}_Z = -\left[\lambda_E^T\mathbf{T}_H\mathbf{e}_0/\left(\lambda_H^T\mathbf{e}_0\right)^2\right]\lambda_H^T + \left(\lambda_H^T\mathbf{e}_0\right)^{-1}\lambda_E^T\mathbf{T}_H, \tag{63}$$

whereas the corresponding row of $\mathbf{Q}$ is found to be

$$\mathbf{q}_j \equiv \mathbf{q}_Z = (\lambda_H^T\mathbf{e}_0)^{-1}\lambda_E^T\mathrm{diag}[\mathbf{OGe}_0]\Pi_{\mathbf{m}_0}. \tag{64}$$

Note that the expressions for the scalar impedance for 2-D MT can also be derived as a special (degenerate) case of the multivariate TFs considered earlier.

Linearized data functionals for apparent resistivity and phase are discussed in Appendix C.

5.2.5 *Example: 3-D MT*

For the 3-D MT problem formulated in terms of the electric fields (Section 5.1.2), the discrete operator $\mathbf{T} = (\mathrm{i}\omega\mu)^{-1}\mathbf{C}$ maps from edges to faces, computing magnetic fields through application of the discrete curl operator. Interpolation from edges and faces to an arbitrary location within the 3-D staggered grid model domain can be based on something simple such as trilinear splines. In this case, both Λ and $\mathbf{T}$ are independent of $\mathbf{m}$, and so $\mathbf{Q} \equiv 0$.

$\mathbf{L}$ can be readily derived as a special case of the multivariate TF with $N_\mathrm{p} = 2$. Each row of the 2×2 impedance tensor is a separate TF—that is, Y in the general development of Section 5.2.2 corresponds to E_x for the first row and E_y for the second. The predictor variables X_1, X_2 correspond to the local horizontal magnetic field. Thus, $\lambda_{X_i}^T = \lambda_{Hi}^T\mathbf{T}, i = 1, 2$ are functionals for computing the two magnetic field components and $\lambda_Y^T = \lambda_{Ek}$ for rows $k = 1, 2$ of the impedance. The 2×2 matrix $\mathbf{X}$ thus has elements $X_{ij} = \lambda_{Hi}^T\mathbf{Te}_j$ (same for both rows of the impedance). From (52) and (53), the row of the (complex) $\mathbf{L}$ corresponding to impedance element ki is

$$\begin{aligned} &\left[\mathbf{X}_{i1}^{-1}\left(\lambda_{Ek}^T - Z_{k1}\lambda_{H1}^T\mathbf{T} - Z_{k2}\lambda_{H2}^T\mathbf{T}\right)\right. \\ &\quad\left.\mathbf{X}_{i2}^{-1}\left(\lambda_{Ek}^T - Z_{k1}\lambda_{H1}^T\mathbf{T} - Z_{k2}\lambda_{H2}^T\mathbf{T}\right)\right] \end{aligned} \tag{65}$$

where the components of $\mathbf{X}$, and the impedance components Z_{kj} are calculated from the background solution. Note that this row of $\mathbf{L}$ has two blocks (each of length N_e), which multiply perturbations to the two polarizations $\delta\mathbf{e}_j, j = 1, 2$, and are summed to compute the total perturbation δZ_{ki} to the impedance element. Rows of $\mathbf{L}$ for vertical field TFs, which relate H_z to the two local horizontal components of the magnetic field, have the same form, with λ_{Hz} replacing λ_{Ek} and the two components of the vertical field TF replacing Z_{kj}, $j = 1, 2$.

6 MULTIPLE TRANSMITTERS

We now give a more explicit discussion of how all of the pieces of **J** fit together in the case of multiple transmitters, allowing for the sort of coupling that occurs with multivariate TFs. In general, there will be N_T transmitters, corresponding to different source geometries and/or different frequencies. There will also be a total of N_R measured components of the EM field at some location. Note that these would correspond to the actual field components observed. Some or all of the data actually used for the inversion would be constructed from these, for example, through TFs, with possible further transformation to apparent resistivity and phase. In general, subsets of receiver locations may be used for each transmitter. The full Jacobian for all data can then be written

$$\mathbf{J} = \mathbf{A}^T \begin{bmatrix} \Lambda^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Lambda^T \end{bmatrix} \begin{bmatrix} \mathbf{S}_1^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_{N_T}^{-1} \end{bmatrix} \times \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_{N_T} \end{bmatrix} + \begin{bmatrix} \mathbf{Q}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}_{N_T} \end{bmatrix}, \tag{66}$$

where **Λ** is the measurement operator, evaluating solutions for each transmitter for all of the N_R receivers. The matrices $\mathbf{P}_t$ and $\mathbf{Q}_t$ are generally different for each transmitter t, as they depend on the forward solution $\mathbf{e}_t$, computed for the reference model parameter used for the Jacobian calculation (see 12). Two transmitters, indexed by t_1, t_2, which only differ in the geometry of the source will typically have identical forward operators, that is, $\mathbf{S}_{t_1} = \mathbf{S}_{t_2}$ (though this is not true for the 2-D MT case, where the two source polarizations are decoupled, and different forward problems are solved for TE and TM modes). The solution operators will always be different for transmitters corresponding to different frequencies. Complications such as the possibility that not all receiver/transmitter pairs are observed, coupling between transmitters through TFs, and further non-linear transformations of data are embedded in the matrix **A**. This matrix will generally be very sparse, with diagonal blocks coupling at most a few transmitters.

Perhaps the simplest specific example of (66) is the controlled source cross-well imaging problem (e.g. Alumbaugh & Newman 1997). In this case, transmitters are point magnetic dipoles in one well, and observations are point measurements of the magnetic field in another well. Assuming all transmitter–receiver pairs are observed, the total number of data is $N_\mathrm{d} = N_T N_R$, and we may take $\mathbf{A} = \mathbf{I}$. Assuming further that all data are taken at a single frequency, the forward operators are all identical, $\mathbf{S}_t \equiv \mathbf{S}$. Then (assume $\mathbf{Q} \equiv 0$) the transpose of the full Jacobian can be computed as

$$\mathbf{J}^T = \left[\mathbf{P}_1(\mathbf{S}^T)^{-1}\mathbf{\Lambda} \quad \mathbf{P}_2(\mathbf{S}^T)^{-1}\mathbf{\Lambda} \quad \ldots \mathbf{P}_{N_T}(\mathbf{S}^T)^{-1}\mathbf{\Lambda} \right]. \tag{67}$$

Thus, any of the $N_T N_R$ rows of $\mathbf{J}^T$ can be constructed from N_T forward solutions (required to form $\mathbf{P}_t$, $t = 1, \ldots, N_T$), and N_R adjoint solutions (one for each column λ_r of Λ). At the same time, the gradient of the data misfit can be written in terms of the residual vector (as in 6)

$$\mathbf{J}^T\mathbf{r} = \sum_t \mathbf{P}_t(\mathbf{S}^T)^{-1}\Lambda\mathbf{r}_t, \tag{68}$$

where $\mathbf{r}_t$ are the components of the residual for transmitter t. Thus, calculation of the gradient (as required for each step in an NLCG or quasi-Newton search scheme) will require N_T adjoint solutions (and again N_T forward solutions, for $\mathbf{P}_t$, $t = 1, \ldots, N_T$). When $N_T \approx N_R$ (as, e.g., in the cross-well EM imaging example) the full Jacobian can thus be had for the same cost (at least in terms of calls to the forward/adjoint solver) as the gradient alone. Although storing the full Jacobian (of size $N_T N_R \times M$) might be prohibitively expensive in terms of memory, by computing and saving the N_T forward and N_R adjoint solutions, a GN scheme can be implemented, solving the normal equations with CG as in Alumbaugh & Newman (1997). This seems certain to be more practical and efficient than direct optimization schemes such as NLCG and quasi-Newton. Extensions of the simple case discussed here, to allow for multiple frequencies or more complex sampling patterns with only some transmitter–receiver pairs, would be straightforward.

In the simple cross-well example, the Jacobian 'factors' into components dependent on the transmitter and receiver with the sensitivity for data $d_{t,r}$ (where t and r are, respectively, the transmitter and receiver indices) is $\mathbf{P}_t(\mathbf{S}^T)^{-1}\lambda_r$. A similar factorization will apply to any problem where there are transmitters with a single frequency (more precisely, with identical forward solvers), but multiple source geometries. Many active source problems, in particular marine CSEM, would fall into this category.

This source–receiver factorization also applies to the case of multivariate TFs, and more complicated data derived from them. Consider the N_p rows of **J** associated with the components of a single TF **Θ**. These rows of **J** can be represented in the general form (66), with a single forward operator $\mathbf{S}_t \equiv \mathbf{S}$. From (52), we thus have

$$\mathbf{J}_\Theta = \mathbf{X}^{-1} \times \begin{bmatrix} \Psi\Lambda^\mathbf{T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Psi\Lambda^T \end{bmatrix} \begin{bmatrix} \mathbf{S}^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}^{-1} \end{bmatrix} \times \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_{N_\mathrm{p}} \end{bmatrix} \tag{69}$$

or, for the transpose

$$\mathbf{J}_\Theta^T = \left[\mathbf{P}_1^T(\mathbf{S}^T)^{-1}\mathbf{\Lambda}\Psi^T \quad \ldots \quad \mathbf{P}_{N_p}^T(\mathbf{S}^T)^{-1}\mathbf{\Lambda}\Psi^T \right] \left(\mathbf{X}^{-1}\right)^T. \tag{70}$$

Thus, all N_p rows of **J** require only a single adjoint solution, which must then be multiplied by each of the matrices $\mathbf{P}_t$, $t = 1, \ldots, N_p$. The resulting model space vectors are then coupled, through the $N_\mathrm{p} \times N_\mathrm{p}$ matrix $(\mathbf{X}^{-1})^T$ to form the N_p rows of $\mathbf{J}_\Theta$. For multivariate TF problems, there will generally be several predicted components at a single site, each associated with an N_p component TF. Each of these TFs will require a separate adjoint solution, $(\mathbf{S}^T)^{-1}\mathbf{\Lambda}_j\mathbf{\Psi}_j^T$ since Λ and Ψ will be different for each TF, but all share the same transmitter dependent matrices $\mathbf{P}_t$, and the same coupling matrix.

In the context of the 3-D MT problem, one has two TFs, corresponding to the two rows of the impedance tensor, and hence two adjoint solutions are required to compute sensitivities for the full impedance tensor. If vertical field TFs are also included, there would be a third TF, and a third adjoint solution would be required. A direct calculation of sensitivities through the transposed eq. (15), taking account of (19) but ignoring the factorization of **L** used to derive (69) and (70), would suggest that two adjoint solutions are required for each of the four components of the impedance tensor. This would imply a total of eight adjoint solutions to evaluate the full sensitivity for an impedance tensor at one location/frequency. Thus, the more careful analysis given here suggests substantial efficiencies, reducing the total number of adjoint solutions for a full calculation of the Jacobian by a factor of 4.

Sensitivities for any data derived from impedance tensor components can obviously be constructed from the adjoint solutions $(\mathbf{S}^T)^{-1}\mathbf{\Lambda}_j\mathbf{\Psi}_j^T$, $j = 1, 2$ essentially as in (70) but with a modified coupling matrix, analogous to $(\mathbf{X}^{-1})^T$. An example would be the four components of the phase tensor (e.g. Caldwell *et al.* 2004), which is a non-linear function of the full impedance.

7 MODULAR IMPLEMENTATION

The mathematical developments of previous sections provide a framework for implementation of a general modular system for inversion of frequency-domain EM data. Here, we provide an overview of the organization and principal features of such a system, which we have developed in Fortran 95. A more detailed description of this modular system (hereinafter referred to as ModEM) will be provided in a future publication. Although a purist might argue that it is not strictly possible to write object-oriented code in Fortran 95 we have based our development on this programming paradigm, following approaches appropriate for the Fortran language as described in Akin (2003). We also use the terminology of this approach in our discussion here. As with most object-oriented programming our goals in ModEM are code reuse for multiple related applications, and providing templates for rapid development of new applications.

As discussed in detail earlier, the basic data objects which are manipulated in any inversion scheme include model (**m**) and data (**d**) vectors, and EM solution and source fields (**e** and **b**). These are treated in ModEM as essentially ‘abstract data types’, encapsulated data structures with details of the internal representation effectively hidden from higher level routines which manipulate them. For each of these classes, a standard series of methods must be defined (creation, destruction, vector space methods, dot products, etc.) with standardized interfaces. The inversion algorithms then apply operators such as $\mathbf{f}$, $\mathbf{J}$, $\mathbf{S}^{-1}$, $\mathbf{L}$, $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{C}_{\mathrm{m}}$, which are implemented as methods that interact with the basic objects **m**, **d**, **e**, **b**. Standardizing type names and interfaces allows multiple instances of these operators and objects to be used interchangeably within the inversion system, and at the same time, simplifies development of any inversion algorithm that can be described in terms of these components.

Components in ModEM can be usefully organized into three layers, as illustrated in Fig. 3. On the left-hand side of the figure are components which define the basic discretization and numerical solution approach used for the forward problem, whereas the components on the right-hand side are more generic, constructed to be directly applicable to a wide range of EM inverse problems. These are separated by an interface layer, which serves to hide problem and implementation specific details from the more generic inversion modules. Each layer in the figure contains several boxes (representing modules or groups of modules in our actual implementation) which are worth distinguishing at the level of this overview.

Two boxes represent the core of the numerical implementation layer. The first includes the grid, data structures that define the primary and dual-field spaces $\mathcal{S}_{\mathrm{P}}$, $\mathcal{S}_{\mathrm{D}}$, field component interpolation functionals (Λ), and the primary to dual mapping **T**—everything needed to define the discrete formulation of the forward problem. The second provides the actual solver for these discretized equations. To be useful for the inversion system this solver, which will be used for both forward and sensitivity calculations, must allow for general sources and boundary conditions, and for solution of the transposed or adjoint system, as well as the usual forward problem. As noted earlier, the PDEs of EM are intrinsically symmetric, so supporting adjoint solutions is typically almost trivial, although there are some details (e.g. associated with non-uniform grids) that may require some care (e.g. Kelbert *et al.* 2008).

No specific data type or procedure names from the core numerical implementation modules are referenced by more generic components of ModEM, so there is a great deal of flexibility in actual implementation at this base level. We have so far used ModEM with three distinct numerical models: the 3-D (electric field) and 2-D (TE and TM mode) Cartesian coordinate FD models discussed earlier, and a 3-D spherical coordinate FD model for global induction studies formulated in terms of the magnetic fields. Source code from previously developed applications were used for the 2-D MT and spherical models, which are described by Siripunvaraporn & Egbert (2000) and Uyeshima & Schultz (2000), respectively. Relatively minor modifications to these codes were required to ensure the required generality of the solver, and to simplify interfacing with other components of ModEM.

The model space is also placed on the left-hand side of Fig. 3, as important components of this module—in particular the mappings π and Π—are strongly dependent on details of the numerical formulation of the forward problem. At the same time, the model space is heavily used by higher level components of ModEM, including the generic inversion modules, and possibly the data functionals (see Section 5.2). Thus, any implementation of the model space module must follow certain conventions to maintain consistency with the rest of the system, for example, providing methods with standardized names and interfaces for linear algebra, dot products and covariance operators. We view the model parametrization and regularization (also part of this module) as something that should be very easy to extend and modify to accommodate a diversity of interpretation problems. For example, the simple conductivity parametrizations discussed earlier could be modified to enforce bounds on conductivities, for example, by replacing the logarithm by a different conductivity transformation as in Avdeev & Avdeeva (2009), or additional parameters to allow explicitly for near-surface distortion (de Groot-Hedlin 1995) could be added. Completely different model parametrizations (e.g. in terms of interface positions between bodies of known conductivity; Smith *et al.* 1999; de Groot-Hedlin & Constable 2004) or regularization approaches may be appropriate in specific situations. To simplify modification and extension, we adopt a strict object-oriented approach for the model parameter space module, hiding all details of a specific instantiation from the rest of the modular system (i.e. in Fortran 95 all attributes of **m** are ‘private’). Note that only the model parameter mappings depend explicitly on the numerical discretization of the EM fields; the rest of the model space implementation is independent of these details and could in principal be used with multiple numerical modelling approaches.

The generic inversion layer is represented by the three boxes on the right-hand side in Fig. 3. The inversion box represents the actual search algorithms, which are written in a generic way using methods from data space, model space and sensitivity modules. Several of the algorithms discussed in Section 2 have been implemented, including the NLCG scheme (e.g. Rodi & Mackie 2001) and the Data space CG scheme of Siripunvaraporn & Egbert (2007). Other inversion

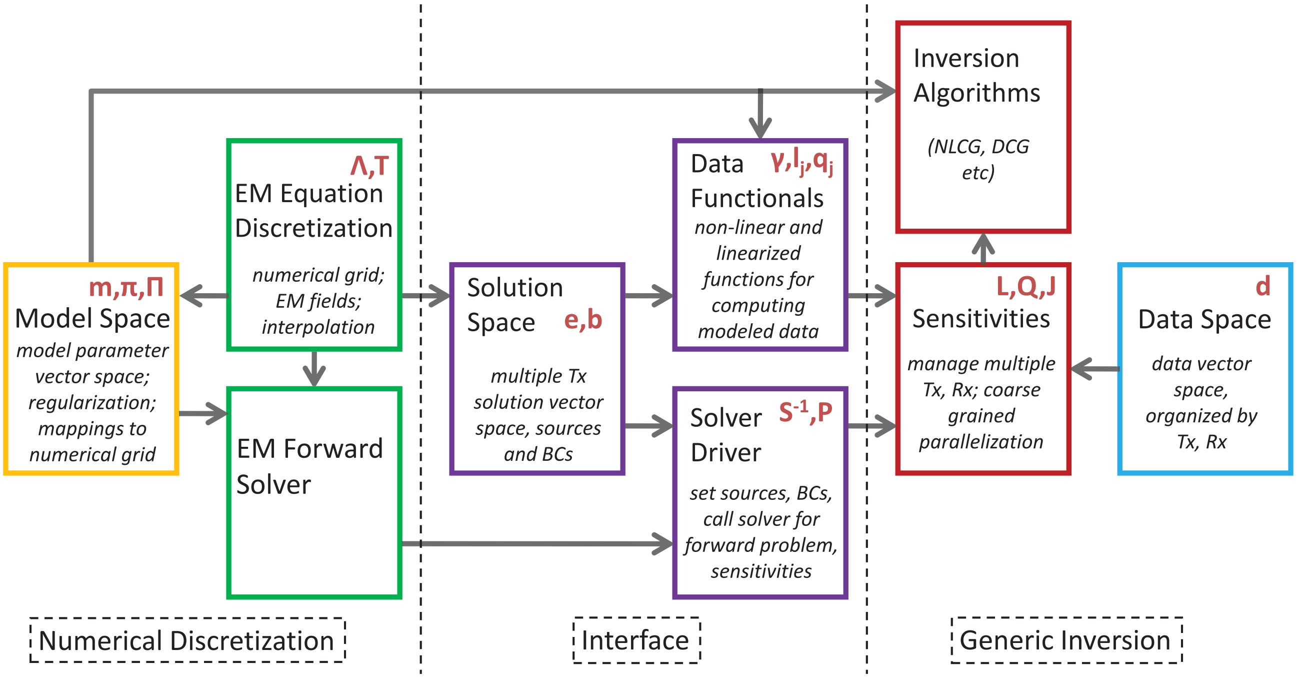

Inversion Algorithms
(NLCG, DCG etc)
Λ,T
EM Equation Discretization
numerical grid; EM fields; interpolation
Data γ, l_j, q_j
Functionals
non-linear and linearized functions for computing modeled data
m,π,Π
Model Space
model parameter vector space; regularization; mappings to numerical grid
Solution Space e,b
multiple Tx solution vector space, sources and BCs
L,Q,J
Sensitivities
manage multiple Tx, Rx; coarse grained parallelization
d
Data Space
data vector space, organized by Tx, Rx
EM Forward Solver
Solver S^{-1},P
Driver
set sources, BCs, call solver for forward problem, sensitivities
Numerical Discretization
Interface
Generic Inversion

Figure 3. Schematic overview of the Modular Electromagnetic Inversion (ModEM) system. Boxes represent modules (or groups of modules, in actual implementation), with dependencies defined by arrows. Data objects and operators, as defined in Sections 3–5, are listed in the appropriate module, along with a brief summary of function. Tx, Rx denote the transmitter and receiver indices, respectively.

approaches can easily be added. Of course, the same inversion routines can be used for multiple problems: the NLCG code has been applied to 2-D and 3-D MT, simple controlled source EM, and global induction problems in spherical coordinates.

The data space is also part of the generic layer. This is organized, following the discussion of Sections 5.2 and 6, to allow for multicomponent data, observed with multiple receivers, and with sources generated by multiple transmitters. Elements of the data vector **d** are thus described by three attributes: *transmitter*, *data type* and *receiver*. *Transmitter* uniquely defines the forward problem that must be solved, including both the specific PDE as well as the sources and boundary conditions. *Receiver* defines, in conjunction with *data type*, the measurement process that must be applied to the forward solution to allow comparison between model and data. The three attributes are treated abstractly at the level of the generic inversion modules, with data vector components carrying only pointers to the actual metadata associated with these attributes (e.g. site location, source polarization, transmitter location, etc.) which are stored as entries in lists, or *dictionaries*. This approach allows a generic format for data storage, hides extraneous details from the inversion modules, and still provides enough information about the transmitter/receiver structure so that forward modelling and sensitivity computations can be organized efficiently.

These tasks are managed by routines in the sensitivity module, which implement the full forward mapping **f** and operations with the Jacobian **J** or its transpose. For example, the transmitter, receiver and data type attributes can be used to ensure that each required forward problem is solved once (and only once), and then used to compute predicted data (or implement appropriate sensitivity calculations) for all necessary receivers and data types. For some cases (CSEM, and even to some extent 3-D MT; see Section 6), computations with the Jacobian can be 'factored' for efficiency into components that depend on the receiver and on the transmitter. In ModEM, such efficiencies can be implemented through specialized versions of the sensitivity module. A coarse grained parallelization (over transmitters, or unique forward problems, similar to the approach used in Siripunvaraporn & Egbert 2009) is also implemented through the sensitivity module. This allows the parallel version to be used with only minor modifications for a wide range of different applications, and to some extent different search algorithms, including those to be developed in future.

The middle layer in Fig. 3 provides an interface between the generic inversion modules, and the problem and numerical implementation specific base modules. In particular, the EM solution and source terms **e** and **b** are defined at this level in the solution space module. These objects must always meet the interface standards of the generic layer, but the implementation of a particular instance of these objects will be problem-specific, and built on base layer routines. Source and receiver details for each specific application are also defined in this interface layer. Thus, inversions for different EM methods may be developed using the same base of numerical discretization modules (and of course the same inversion modules) through modifications to the interface layer.

For example, we have used the 3-D Cartesian FD code base for both MT and CSEM. The fundamental EM solution and source objects have distinct implementations for the two methods. For a single transmitter (frequency) in the 3D MT problem, **b** represents boundary conditions for two orthogonal plane wave sources, and **e** represents the corresponding pair of solutions, each a 3-D vector field. For the CSEM problem, **b** represents a single dipole source, and **e** is just a single vector field. These differences are implemented in the solution space module. A secondary field formulation (e.g. Alumbaugh *et al.* 1996), which is essential for accurate forward modelling for the CSEM problem, but less critical for MT, is readily implemented through the solver interface module. For the CSEM case, the interface includes routines to compute the primary and scattered fields and hence the source term needed for the FD solver; for the MT case appropriate boundary conditions are simply generated (and the secondary field approach is not used). In both cases, the same base-level FD solver is then called (once for CSEM, twice for 3-D MT) to do the core computations. Data for the MT and CSEM problems also differ, requiring modifications to the data functional module: for 3-D MT data are TFs or impedances, whereas for CSEM they are just simple observations of individual electric or magnetic field components.

A joint MT-CSEM inversion could also be implemented with very minor changes to the interface layer: solution space, solver driver and data functional modules for MT and CSEM can be merged, with the appropriate case (one or two source polarizations, secondary field solver or MT boundary value problem, impedances or field components) selected based on the transmitter index. This idea can be extended to develop joint inversion for EM with other sorts of geophysical data (seismic, gravity, etc.). In a joint inversion setting, the base layer might include two or more numerical discretization and forward solver modules, with physical parameters that define the forward problems coupled (explicitly or structurally) through a joint model parameter module. These forward problem solvers would then be interfaced to the generic inversion layer through merged solution space, data functional and solver driver modules. The structure of the data space module provides a good basis for developing modified inversion search algorithms as might be appropriate for joint inversion, such as allowing for control over trade-offs between fitting data of different types.

As a brief illustration of some of the capabilities of ModEM, we consider synthetic data inversion tests for three of the EM inverse problems discussed in previous sections: 2-D MT, 3-D MT and global induction. In all of the tests discussed here, we generated synthetic data using some variant on a 'checkerboard' conductivity distribution, of the sort often used for resolution tests in seismic tomography, added Gaussian random noise and used the NLCG algorithm implemented in ModEM for inversion.

For the 2-D MT tests, we inverted TE and TM mode data for 12 periods evenly spaced on a logarithmic scale from 0.3–3000 s. Data were generated for 30 sites, with error standard deviation 3 per cent of impedance magnitude. The conductivity model consisted of a checkerboard pattern of 10 and 1000 ohm-m blocks embedded in a 100 ohm-m half-space (Fig. 4a). The same grid ($N_y = 106$ with nominal resolution 1.5 km; $N_z = 40$ increasing logarithmically, starting from 0.5 km) was used for generating the synthetic data, and for the inversion. The covariance used was similar to that of Siripunvaraporn & Egbert (2000), and the prior model was a 100 ohm-m half-space. The NLCG inversion converged from a normalized root-mean-square (rms) misfit of 15.9 to below 1.05 in 68 iterations. The resulting solution, which fits the data to within the expected errors, and captures the main features of the synthetic model, is shown in Fig. 4(b).

For the 3-D MT tests, we used a 3-D variant on the checkerboard, as illustrated in Fig. 5(a). For data we used the full impedance (all four complex components), plus the vertical magnetic field TFs, for 12 periods logarithmically spaced between 10–10 000 s. Error levels were set at 3 per cent of $|Z_{xy}Z_{yx}|^{1/2}$ for all impedance components, and at 0.03 for the non-dimensional vertical magnetic TF components. The grid (again used both for computing the synthetic data and for inversion) was 67 × 67 × 60, with a nominal resolution in the core of 20 km horizontally (see Fig. 5a) . A total of 225 sites, on a 15 × 15 regular 80 km grid were used for the inversion. The covariance was similar to that used for the 2-D tests, and the prior was again a 100 ohm-m half-space. The NLCG algorithm converged from a normalized rms misfit of 12.32 to below 1.05 in 51 iterations, resulting in the inverse solution shown in Fig. 5(b). Again, major model features are well recovered, with some degradation in imaging capability evident below shallower conductive features, as would be expected.

As a final example, we show a simple global induction example. As discussed earlier, the ModEM implementation for this case is based on the spherical coordinate forward solver of Uyeshima & Schultz (2000), which is formulated in terms of the magnetic fields. Data for this global problem are so-called C-responses, ratios of the vertical (H_z) and north (H_x) components of the magnetic field, computed under the assumption that external sources can be approximated well by a zonal (geomagnetic coordinates) dipole. A stand-alone inversion code for this sort of data, based on the same solver, is described by Kelbert *et al.* (2008). An application of the inversion to observatory data is given in Kelbert *et al.* (2009). Here, we only demonstrate our new ModEM version, using a simple synthetic example based on a four-layer 1-D Earth (0–100 km depth:

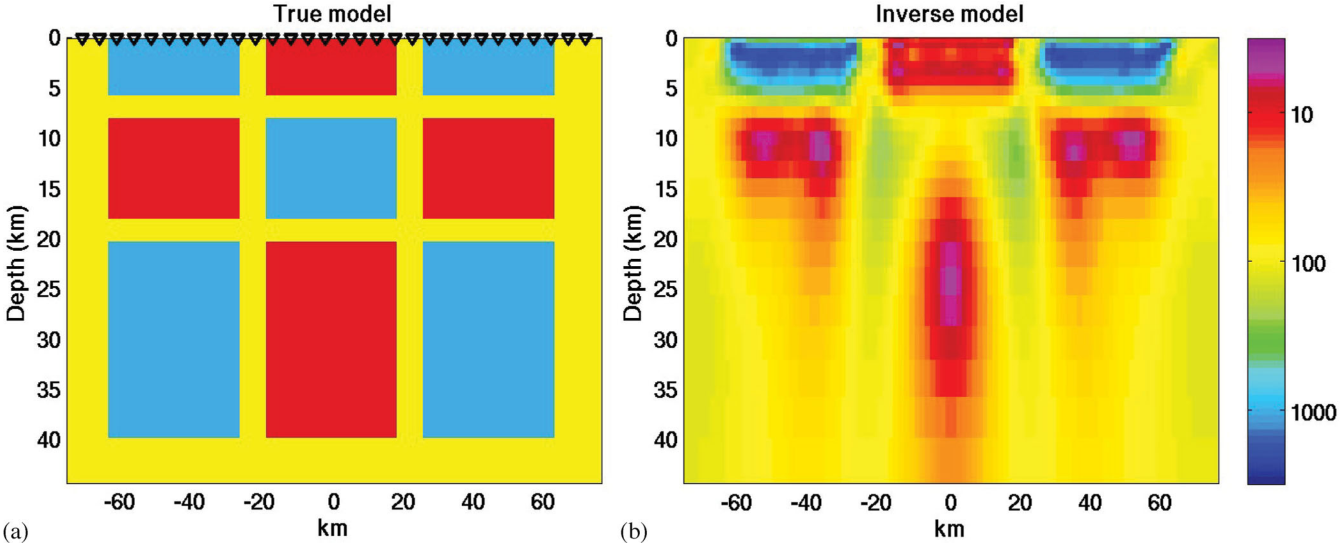


Figure 4. (a) Resistivity model used to generate synthetic data for 2-D MT test, with site locations shown at top. (b) Inverse solution obtained with ModEM, fitting TE and TM mode impedances with a normalized rms misfit of 1.05.

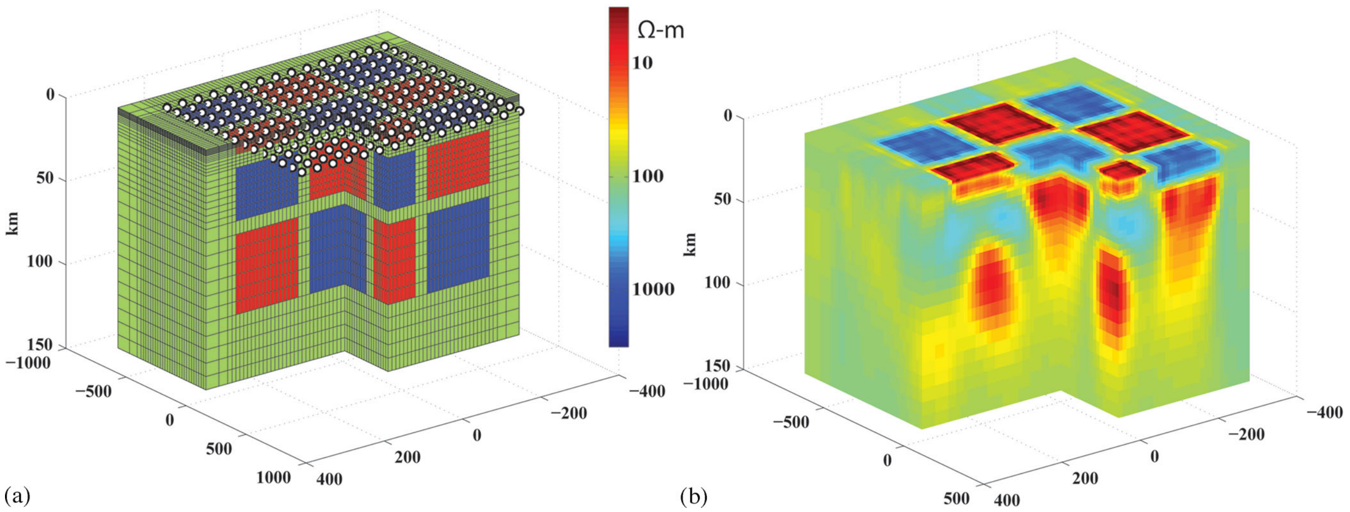


Figure 5. (a) Resistivity model used to generate synthetic data for 3D MT test. The centre of the model grid (used for generating data and for inversion) is shown, along with the regular grid of sites on the surface. (b) Inverse solution obtained with ModEM, fitting full impedance tensor plus vertical magnetic TFs to a normalized rms misfit of 1.05. Note that in the cut-away view the upper surface shown is at 2 km depth, but the structures shown extend to the surface.

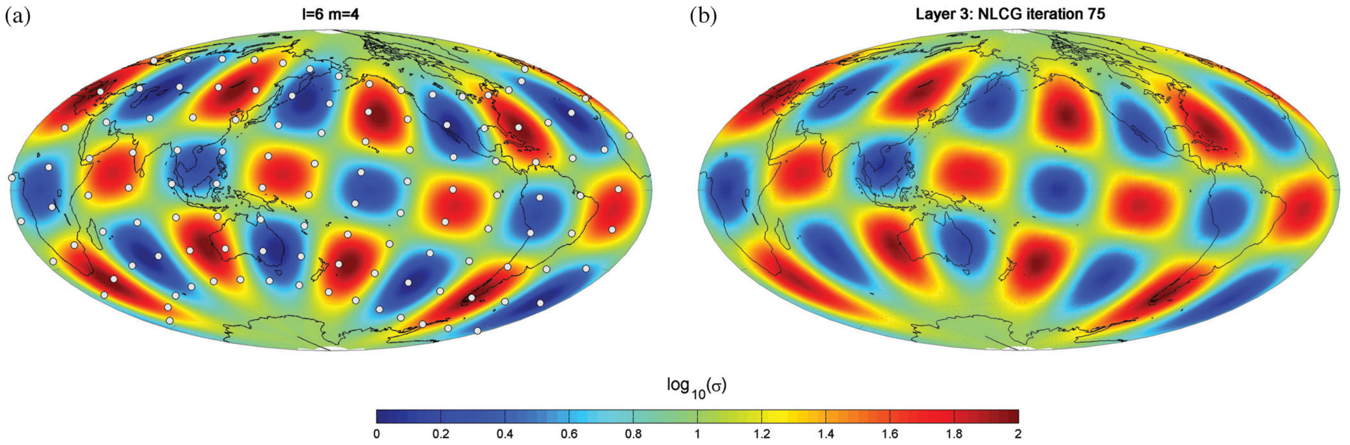


Figure 6. (a) Heterogeneous conductivity in layer 3 (400–650 km depth) of global model used to generate test data for the global induction inversion, with sites shown as filled circles. (b) Conductivity variations in the same layer recovered by the NLCG inversion, implemented with ModEM.

0.0001 S m^{-1}; 100–400 km: 0.01 S m^{-1}; 400–650 km: 0.1 S m^{-1}; 650–4000 km: 2.0 S m^{-1}) with an $l = 6, m = 4$ spherical harmonic perturbation (in geomagnetic dipole coordinates) imposed in layer 3 (400–650 km). The amplitude of the perturbation (Fig. 6a) is equivalent to one order of magnitude variation around the 0.1 S m^{-1} background.

Data were distributed on a regular spherical grid (eight latitudes, from 56S to 56N, 15 evenly spaced longitudes, 120 sites total), for four periods: 6 hr, 1, 4 and 16 d. The synthetic C-responses were computed on a 3° × 3° grid, and again 3 per cent Gaussian errors were added. The inversion assumed the same 1-D prior, and a relatively low-dimensional model parametrization: only the third layer was allowed to deviate from uniform, with variations parametrized by spherical harmonics up to degree and order 9. A diagonal (in the spherical harmonic domain) model error covariance was used for the inversion, which was run on a 5° resolution spherical grid. For this case, the inversion converged from a normalized rms misfit of 14.43 to 1.46 in 76 NLCG iterations. Although the fit is not quite to within the expected errors (presumably because of numerical errors associated with the coarser grid used for the inversion) conductivity variations in layer 3 (Fig. 6b) are recovered almost perfectly.

8 CONCLUSIONS

We have derived general recipes for the Jacobian calculations that are central to a wide range of EM inversion algorithms. Our analysis is based on the discrete formulation of the forward problem, including explicit treatment of parameter mappings and data functionals in the numerical implementation. Through this analysis, we show how the Jacobian can be decomposed into simpler operators, and we analyse the dependence of these operators on the specific EM problem (e.g. through the transmitter and receiver configuration), or on implementation specific details, such as the model parametrization or the nature of the numerical discretization. Based on the general formulation, we provide explicit expressions for Jacobian calculations for several example problems, including 2-D and 3-D MT, and 3-D controlled source problems with multiple transmitter locations. A key result of our general analysis is the ‘factorization’ of the Jacobian into components dependent only on transmitters, and on receivers. This has important implications for efficient implementation of inversion algorithms, which will be explored more thoroughly elsewhere. To the extent that we have discussed numerical and discretization details of the forward problem, we have focused on FD methods. However, much of our theory is more generally applicable—for example, the division of the Jacobian into components, and the dependencies of these components on details of the EM problem and model parametrization—and will provide a useful guide to development of inversion algorithms for any numerical implementations of the EM forward problem.

Building on the general theoretical framework, we have sketched our development of ModEM, a modular system of computer codes for EM inversion. ModEM allows inversion codes developed for one purpose to be rapidly adapted to other problems, and simplifies development of new capabilities. For example, the 3-D MT inversion discussed earlier can be extended to include intersite magnetic TFs through very minor modifications to the data functional module (essentially adding rows to the matrix **A** in 44). Only slightly greater modifications were required for initial development of an inversion for CSEM data, for which both sources and receivers are different. Flexibility and ease of modification of the model parametrization, and interchangeable inversion search algorithms are other noteworthy features of ModEM.

REFERENCES

Akin, J.E., 2003. *Object-Oriented Programming via Fortran 90/95,* Vol. 1317, Issue 1, Cambridge University Press, Cambridge, 348pp.

Alumbaugh, D.L. & Newman, G.A., 1997. Three-dimensional massively parallel electromagnetic inversion: II. Analysis of a crosswell electromagnetic experiment, *Geophys. J. Int.,* **128,** 355–363, doi:10.1111/j.1365-246X.1997.tb01560.x.

Alumbaugh, D.L., Newman, G.A., Prevost, L. & Shadid, J.N., 1996. Three-dimensional wide band electromagnetic modeling on massively parallel computers, *Radio Sci.,* **33,** 1–23.

Avdeev, D.B., 2005. Three-dimensional electromagnetic modelling and inversion from theory to application, *Surv. Geophys.,* **26**(6), 767–799.

Avdeev, D.B. & Avdeeva, A., 2009. 3D magnetotelluric inversion using a limited-memory quasi-Newton optimization, *Geophysics,* **74**(3), F45–F57, doi:10.1190/1.3114023.

Bennett, A.F., 2002. *Inverse Modeling of the Ocean and Atmosphere,* Cambridge University Press, Cambridge.

Caldwell, T.G., Bibby, H.M. & Brown, C., 2004. The magnetotelluric phase tensor, *Geophys. J. Int.,* **158**(2), 457–469.

Chua, B., 2001. An inverse ocean modeling system, *Ocean Modelling,* **3,** 137–165, doi:10.1016/S1463-5003(01)00006-3.

Commer, M. & Newman, G.A., 2008. New advances in three-dimensional controlled-source electromagnetic inversion, *Geophys. J. Int.,* **172,** 513–535, doi:10.1111/j.1365-246X.2007.03663.x.

Constable, S.C., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics,* **52**(3), 289–300.

Egbert, G.D., 1994. A new stochastic process on the sphere: application to characterization of long-period global scale external sources, in *14th Workshop on Electromagnetic Induction in the Earth and Moon,* Brest, France.

Egbert, G.D., 2002. Processing and interpretation of electromagnetic induction array data, *Surv. Geophys.,* **23**(2–3), 207–249.

Egbert, G.D. & Bennett, A.F., 1996. Data assim, *Modern Approaches to Data Assimilation in Ocean Modeling,* p. 147, Elsevier Science, Amsterdam.

de Groot-Hedlin, C., 1995. Inversion for regional 2-D resistivity structure in the presence of galvanic scatterers, *Geophys. J. Int.,* **122**(3), 877–888, doi:10.1111/j.1365-246X.1995.tb06843.x.

de Groot-Hedlin, C. & Constable, S.C., 2004. Inversion of magnetotelluric data for 2D structure with sharp resistivity contrasts, *Geophysics,* **69**(1), 78, doi:10.1190/1.1649377.

Haber, E., 2005. Quasi-Newton methods for large-scale electromagnetic inverse problems, *Inverse Probl.,* **21**(1), 305–323.

Kelbert, A., 2006. Geophysical inverse theory applied to reconstruction of large-scale heterogeneities in electrical conductivity of Earth's mantle, *PhD thesis*, Cardiff University.

Kelbert, A., Egbert, G.D. & Schultz, A., 2008. Non-linear conjugate gradient inversion for global EM induction: resolution studies, *Geophys. J. Int.,* **173**(2), 365–381, doi:10.1111/j.1365-246X.2008.03717.x.

Kelbert, A., Schultz, A. & Egbert, G.D., 2009. Global electromagnetic induction constraints on transition-zone water content variations, *Nature,* **460**(7258), 1003–1006, doi:10.1038/nature08257.

de Lugao, P., Portniaguine, O. & Zhdanov, M.S., 1997. Fast and stable two-dimensional inversion of magnetotelluric data, *J. Geomag. Geoelectr.,* **49**(11–12), 1437–1454.

McGillivray, P.R., Oldenburg, D.W., Ellis, R.G. & Habashy, T.M., 1994. Calculation of sensitivities for the frequency-domain electromagnetic problem, *Geophys. J. Int.,* **116**(1), 1–4, doi:10.1111/j.1365-246X.1994.tb02121.x.

Mackie, R.L. & Madden, T.R., 1993. 3-dimensional magnetotelluric inversion using conjugate gradients, *Geophys. J. Int.,* **115**(1), 215–229.

Mackie, R.L., Smith, J.T. & Madden, T.R., 1994. 3-dimensional electromagnetic modeling using finite-difference equations—the magnetotelluric example, *Radio Sci.,* **29**(4), 923–935.

Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters, *J. Soc. Ind. Appl. Math.,* **11,** 431–441.

Nedelec, J.C., 1980. Mixed finite elements in R3, *Numer. Math.,* **35**(3), 315–341, doi:10.1007/BF01396415.

Newman, G.A. & Alumbaugh, D.L., 1997. Three-dimensional massively parallel electromagnetic inversion I. Theory, *Geophys. J. Int.,* **128,** 345–354, doi:10.1111/j.1365-246X.1997.tb01559.x.

Newman, G.A. & Alumbaugh, D.L., 2000. Three-dimensional magnetotelluric inversion using non-linear conjugate gradients, *Geophys. J. Int.,* **140,** 410–424.

Newman, G.A. & Boggs, P.T., 2004. Solution accelerators for large-scale three-dimensional electromagnetic inverse problems, *Inverse Probl.,* **20,** 151–170, doi:10.1088/0266-5611/20/6/S10.

Newman, G.A. & Hoversten, G.M., 2000. Solution strategies for two- and three-dimensional electromagnetic inverse problems, *Inverse Probl.,* **16,** 1357–1375, doi:10.1088/0266-5611/16/5/314.

Nocedal, J. & Wright, S.J., 1999. *Numerical Optimization,* Springer-Verlag, New York, NY.

Pankratov, O. & Kuvshinov, A., 2010. General formalism for the efficient calculation of derivatives of EM frequency-domain responses and derivatives of the misfit, *Geophys. J. Int.,* **181**(1), 229–249.

Parker, R.L., 1994. *Geophysical Inverse Theory,* Princeton University Press, Princeton, NJ.

Rodi, W.L., 1976. A technique for improving the accuracy of finite element solutions for magnetotelluric data, *Geophys. J. Int.,* **44**(2), 483–506, doi:10.1111/j.1365-246X.1976.tb03669.x.

Rodi, W.L. & Mackie, R.L., 2001. Nonlinear conjugate gradients algorithm for 2-D magnetotelluric inversion, *Geophysics,* **66**(1), 174–187.

Rodrigue, G. & White, D., 2001. A vector finite element time-domain method for solving Maxwell's equations on unstructured hexahedral grids, *SIAM J. Sci. Comput.,* **23**(3), 683, doi:10.1137/S1064827598343826.

Sasaki, Y., 2001. Full 3-D inversion of electromagnetic data on PC, *J. appl. Geophys.,* **46,** 45–54, doi:10.1016/S0926-9851(00)00038-0.

Schmücker, U., 2003. Horizontal spatial gradient sounding and geomagnetic depth sounding in the period range of daily variation, in *Protokoll Aber das Kolloquium elektromagnetische Tiefenforschung,* Kolloquium: Konigstein, pp. 228–237.

Schmücker, U., 2004. Multivariate magneto-variational soundings (MVS), in *Proceedings of the 17th EM Induction Workshop,* Hyderabad.

Semenov, V.Y. & Shuman, V.N., 2009. Impedances for induction soundings of the Earth's mantle, *Acta Geophys.,* **58**(4), 527–542.

Siripunvaraporn, W. & Egbert, G.D., 2000. An efficient data-subspace inversion method for 2-D magnetotelluric data, *Geophysics,* **65**(3), 791–803.

Siripunvaraporn, W. & Egbert, G.D., 2007. Data space conjugate gradient inversion for 2-D magnetotelluric data, *Geophys. J. Int.,* **170,** 986–994, doi:10.1111/j.1365-246X.2007.03478.x.

Siripunvaraporn, W. & Egbert, G., 2009. WSINV3DMT: vertical magnetic field transfer function inversion and parallel implementation, *Phys. Earth planet. Inter.,* **173**(3–4), 317–329, doi:10.1016/j.pepi.2009.01.013.

Siripunvaraporn, W., Egbert, G.D. & Lenbury, Y., 2002. Numerical accuracy of magnetotelluric modeling: a comparison of finite difference approximations, *Earth Planets Space,* **54**(6), 721–725.

Siripunvaraporn, W., Uyeshima, M. & Egbert, G.D., 2004. Three-dimensional inversion for Network-Magnetotelluric data, *Earth Planets Space,* **56**(9), 893–902.

Siripunvaraporn, W., Egbert, G.D., Lenbury, Y. & Uyeshima, M., 2005. Three-dimensional magnetotelluric inversion: data-space method, *Phys. Earth planet. Inter.,* **150**(1–3), 3–14.

Smith, J.T., 1996. Conservative modeling of 3-D electromagnetic fields: 1. Properties and error analysis, *Geophysics,* **61**(5), 1308–1318.

Smith, T., Hoversten, M., Gasperikova, E. & Morrison, F., 1999. Sharp boundary inversion of 2D magnetotelluric data, *Geophys. Prospect.,* **47**(4), 469–486, doi:10.1046/j.1365-2478.1999.00145.x.

Spitzer, K., 1998. The three-dimensional DC sensitivity for surface and subsurface sources, *Geophys. J. Int.,* **134,** 736–746, doi:10.1046/j.1365-246x.1998.00592.x.

Uyeshima, M. & Schultz, A., 2000. Geomagnetic induction in a heterogeneous sphere: a new three-dimensional forward solver using a conservative staggered-grid finite difference method, *Geophys. J. Int.,* **140**(3), 636–650.

Yee, K., 1966. Numerical solution of inital boundary value problems involving Maxwell's equations in isotropic media, *IEEE Trans. Antennas Propag.,* **14,** 302–307, doi:10.1109/TAP.1966.1138693.

Zhang, J., Mackie, R.L. & Madden, T.R., 1995. 3-D resistivity forward modeling and inversion using conjugate gradients, *Geophysics,* **60**(5), 1313–1325.

APPENDIX A: DEPENDENCE OF SOURCE TERMS ON MODEL PARAMETERS

In some cases (in particular for active source problems), it is appropriate to use a so-called 'secondary field' approach to solve the forward problem (e.g. Alumbaugh *et al.* 1996). In this case, a background (typically 1-D) conductivity is assumed, allowing quasi-analytic computation of a background solution, with the 'secondary' field due to deviation from the background conductivity then computed numerically. More precisely, the total field solution is represented as $\mathbf{e} = \hat{\mathbf{e}} + \delta\mathbf{e}$, where the background field $\hat{\mathbf{e}}$ satisfies the 1-D equation defined by conductivity parameter $\hat{\mathbf{m}}$. It is readily verified that the secondary field $\delta\mathbf{e}$ satisfies the induction equation with a modified source. Assuming the 3-D operator can be expressed as in (27) this takes the form

$$\mathbf{S}_{\mathbf{m}}\delta\mathbf{e} = -\mathbf{U}[(\pi(\mathbf{m}) - \pi(\hat{\mathbf{m}})) \circ \mathbf{V}\hat{\mathbf{e}}]. \tag{A1}$$

The RHS in (A1) depends on the model parameter $\mathbf{m}$, suggesting that an additional term should be included in eq. (12).

However, if we differentiate both sides of (A1) with respect to $\mathbf{m}$ and use (27) again we find

$$\frac{\partial}{\partial\mathbf{m}}\left[\mathbf{S}_0\delta\mathbf{e} + \mathbf{U}(\pi(\mathbf{m}) \circ \mathbf{V}\delta\mathbf{e})\right] = -\frac{\partial}{\partial\mathbf{m}}\left[\mathbf{U}(\pi(\mathbf{m}) \circ \mathbf{V}\hat{\mathbf{e}}\right], \tag{A2}$$

implying

$$0 = \frac{\partial}{\partial\mathbf{m}}\left[\mathbf{S}_0\delta\mathbf{e} + \mathbf{U}(\pi(\mathbf{m}) \circ \mathbf{V}\mathbf{e})\right] = \frac{\partial}{\partial\mathbf{m}}\left[\mathbf{S}_{\mathbf{m}}\mathbf{e}\right], \tag{A3}$$

the last equality following from the fact that $\mathbf{S}_0\hat{\mathbf{e}}$ does not depend on $\mathbf{m}$. Thus, as long as the RHS of the original problem is independent of the model parameter, $\partial\mathbf{e}/\partial\mathbf{m} = \partial[\delta\mathbf{e}]/\partial\mathbf{m}$ satisfies (12) without any additional terms, even if the equation for the secondary field does depend on $\mathbf{m}$. Note also that even if the forward problem is solved with a secondary field approach, the Jacobian calculation (either through 14 or 15) involves only the standard discrete solver $\mathbf{S}_{\mathbf{m}}^{-1}$. Use of a secondary field approach only affects the derivative indirectly through its dependence on the forward solution.

APPENDIX B: 3-D STAGGERED GRID DETAILS

Here, we give a more precise definition of the discrete finite difference (FD) operator corresponding to $\nabla \times \nabla \times + \mathrm{i}\omega\mu\sigma$ and its adjoint, and clarify implementation of boundary conditions for the 3-D magnetotelluric (MT) problem. Similar considerations apply to other cases considered in the text. To do this, we need to distinguish more precisely between interior and boundary nodes in the grid. In the main text, $\mathcal{S}_{\mathrm{P}}$ ($\mathcal{S}_{\mathrm{D}}$) have been used to denote the space of discrete complex vector fields defined on all edges (faces) of the staggered grid. Here, we use the same symbols with tildes ($\tilde{\mathcal{S}}_{\mathrm{P}}$, $\tilde{\mathcal{S}}_{\mathrm{D}}$) to indicate the restriction to interior edges or faces. The discrete curl operator is naturally defined as a mapping from all edges to all faces, but we need only consider the partial mapping which computes the curl for interior faces (see e.g. Kelbert (2006) for details). Denote this as

$$\mathbf{C} : \mathcal{S}_{\mathrm{P}} \mapsto \tilde{\mathcal{S}}_{\mathrm{D}} \tag{B1}$$

and partition $\mathbf{e} \in \mathcal{S}_{\mathrm{P}}$ and $\mathbf{C}$ into interior and boundary edge components

$$\mathbf{e} = \begin{bmatrix} \tilde{\mathbf{e}} \\ \mathbf{e}_b \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} \tilde{\mathbf{C}} & \mathbf{C}_b \end{bmatrix}, \tag{B2}$$

so that $\tilde{\mathbf{C}} : \tilde{\mathcal{S}}_{\mathrm{P}} \mapsto \tilde{\mathcal{S}}_{\mathrm{D}}$ and $\mathbf{C}\mathbf{e} = \tilde{\mathbf{C}}\tilde{\mathbf{e}} + \mathbf{C}_b\mathbf{e}_b$.

To define adjoints precisely, we need to specify inner products. The natural inner products for the primary and dual spaces (interior nodes only) are

$$\langle \tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2 \rangle_{\mathrm{P}} = \tilde{\mathbf{e}}_1^* \mathbf{V}_{\mathrm{E}} \tilde{\mathbf{e}}_2 \qquad \left\langle \tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2 \right\rangle_{\mathrm{D}} = \tilde{\mathbf{h}}_1^* \mathbf{V}_{\mathrm{F}} \tilde{\mathbf{h}}_2. \tag{B3}$$

In (B3), $\mathbf{V}_{\mathrm{E}}$ and $\mathbf{V}_{\mathrm{F}}$ are real diagonal matrices of edge and face volume elements. Edge volumes, for example, are defined as one-fourth of the total volume of the four cells sharing the edge, so that the first discrete inner product in (B3) approximates the integral L_2 inner product for vector fields $\int\int\int \mathbf{E}_1^*(\mathbf{x}) \cdot \mathbf{E}_2(\mathbf{x})\mathrm{d}V$. The adjoint of the interior curl operator $\tilde{\mathbf{C}}^\dagger : \tilde{\mathcal{S}}_{\mathrm{D}} \mapsto \tilde{\mathcal{S}}_{\mathrm{P}}$ satisfies, by definition,

$$\left\langle \tilde{\mathbf{h}}, \tilde{\mathbf{C}}\tilde{\mathbf{e}} \right\rangle_{\mathrm{D}} = \left\langle \tilde{\mathbf{C}}^\dagger \tilde{\mathbf{h}}, \tilde{\mathbf{e}} \right\rangle_{\mathrm{P}} \qquad \forall \tilde{\mathbf{e}} \in \tilde{\mathcal{S}}_{\mathrm{P}}, \tilde{\mathbf{h}} \in \tilde{\mathcal{S}}_{\mathrm{D}}. \tag{B4}$$

Noting that that $\tilde{\mathbf{C}}$ is real, one then readily derives

$$\tilde{\mathbf{C}}^\dagger = \mathbf{V}_{\mathrm{E}}^{-1} \tilde{\mathbf{C}}^T \mathbf{V}_{\mathrm{F}}. \tag{B5}$$

From the definitions of $\mathbf{V}_{\mathrm{E}}$ and $\mathbf{V}_{\mathrm{F}}$ one can verify that $\tilde{\mathbf{C}}^\dagger$ indeed corresponds to the appropriate geometric definition of the curl operator defined on cell faces. Thus, the electric field eq. (22) with source $\mathbf{j}_s$

$$\nabla \times \nabla \times \mathbf{E} + \mathrm{i}\omega\mu\sigma\mathbf{E} = \mathbf{j}_s \tag{B6}$$

can be approximated on the discrete grid as

$$\tilde{\mathbf{C}}^\dagger \mathbf{C}\mathbf{e} + \mathrm{i}\omega\mu\sigma\tilde{\mathbf{e}} = [\tilde{\mathbf{C}}^\dagger\tilde{\mathbf{C}} + \mathrm{i}\omega\mu\sigma]\tilde{\mathbf{e}} + \tilde{\mathbf{C}}^\dagger\mathbf{C}_b\mathbf{e}_b = \tilde{\mathbf{b}}, \tag{B7}$$

where $\tilde{\mathbf{b}} \in \tilde{\mathcal{S}}_{\mathrm{P}}$ gives the discrete approximation for the source current $\mathbf{j}_s$ inside the domain; these currents (and hence $\tilde{\mathbf{b}}$) vanish for the 3-D MT example we have focused on. The discrete system (B7) has one equation for each of the $\tilde{N}_\mathrm{e}$ interior edges, but N_e (= total number of edges) unknowns. Boundary conditions are thus required, most simply specification of tangential electric field components on the boundary edges. Then, the full system of equations ($\mathbf{Se} = \mathbf{b}$) can be decomposed into interior and boundary components as

$$\begin{bmatrix} \tilde{\mathbf{C}}^\dagger\tilde{\mathbf{C}} + \mathrm{i}\omega\mu\sigma & \tilde{\mathbf{C}}^\dagger\mathbf{C}_b \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}} \\ \mathbf{e}_b \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{ii} & \mathbf{S}_{ib} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}} \\ \mathbf{e}_b \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{b}} \\ \mathbf{b}_b \end{bmatrix}, \tag{B8}$$

where $\mathbf{b}_b$ represents the specified boundary data. Eliminating the boundary edges results in a well-posed $\tilde{N}_\mathrm{e} \times \tilde{N}_\mathrm{e}$ problem for electric fields restricted to interior edges

$$[\tilde{\mathbf{C}}^\dagger\tilde{\mathbf{C}} + \mathrm{i}\omega\mu\sigma\mathbf{I}]\tilde{\mathbf{e}} = \mathbf{S}_{ii}\tilde{\mathbf{e}} = \tilde{\mathbf{b}} - \tilde{\mathbf{C}}^\dagger\mathbf{C}_b\mathbf{b}_b, \tag{B9}$$

with the RHS determined from the boundary data, and any source terms in the domain. Using (B5), we see that the discrete operator in (B9) can be written as $\mathbf{S}_{ii} = \mathbf{V}_\mathrm{E}^{-1}\tilde{\mathbf{C}}^T\mathbf{V}_\mathrm{F}\tilde{\mathbf{C}} + \mathrm{i}\omega\mu\sigma\mathbf{I}$. Thus, as sketched in Section 3, the system $\mathbf{Se} = \mathbf{b}$ can be reduced to symmetric form by eliminating the boundary nodes, and then multiplying both sides of the resulting eq. (B9) by $\mathbf{V}_\mathrm{E}$.

We emphasize that in our treatment of the discrete forward problem we take $\mathbf{e}$, $\mathbf{S}$ and $\mathbf{b}$ to include both interior and boundary nodes. Thus to be precise in our application of (27) to the 3-D FD equations considered here, we should take

$$\mathbf{S}_0 = \begin{bmatrix} \tilde{\mathbf{C}}^\dagger\tilde{\mathbf{C}} & \tilde{\mathbf{C}}^\dagger\mathbf{C}_b \\ 0 & \mathbf{I} \end{bmatrix}, \tag{B10}$$

and we should define $\pi(\mathbf{m}) \equiv \sigma(\mathbf{m}) \equiv 0$ on boundary edges. This is a general property of $\pi(\mathbf{m})$, since the boundary conditions do not depend on the model parameter. This implies that the columns of $\mathbf{P}$ corresponding to boundary nodes will all vanish. Also, accounting for the boundary conditions in the transpose of $\mathbf{S}$ we have, in the notation of (B8),

$$\mathbf{S}^T\mathbf{e} = \begin{bmatrix} \mathbf{S}_{ii} & 0 \\ \mathbf{S}_{ib}^T & \mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}} \\ \mathbf{e}_b \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{b}} \\ \mathbf{b}_b \end{bmatrix}. \tag{B11}$$

The transposed solution operator $(\mathbf{S}^T)^{-1}\mathbf{b}$, which appears extensively throughout the main text, can thus be interpreted as solution of the homogeneous problem (for interior nodes)

$$\mathbf{S}_{ii}^T\tilde{\mathbf{e}} = \tilde{\mathbf{b}} \tag{B12}$$

followed by computation of the boundary terms

$$\mathbf{e}_b = \mathbf{b}_b - \mathbf{S}_{ib}^T\tilde{\mathbf{e}}. \tag{B13}$$

In fact, solutions to the adjoint problem $(\mathbf{S}^T)^{-1}\mathbf{b}$ are always multiplied by $\mathbf{P}^T$, and because the rows of $\mathbf{P}^T$ corresponding to boundary components are zero, the boundary terms in (B13) are never actually required for our purposes.

APPENDIX C: TRANSFORMATION OF JACOBIAN TO REAL FORM

To allow for the fact that the model parameter $\mathbf{m}$ is typically real, and in some cases data are also real, we have assumed that $\mathbf{d}$ and $\mathbf{J}$ are real, with any complex observations (e.g. an impedance) represented as two real elements of the data vector. However, throughout the text, we have used complex notation for $\mathbf{L}$, $\mathbf{S}_{\mathbf{m}_0}^{-1}$, $\mathbf{P}$ and $\mathbf{Q}$, so $\mathbf{J}$ computed from (14) would also be complex. In fact, for complex observations it is readily verified that the real and imaginary parts of a row of the complex expression for the Jacobian give the sensitivity (a vector in the real model parameter space) for the corresponding real and imaginary parts of one observation. Thus, to keep the Jacobian and the data vector strictly real, we can set

$$\bar{\mathbf{d}} = \begin{bmatrix} \Re(\mathbf{d}) \\ \Im(\mathbf{d}) \end{bmatrix} \quad \bar{\mathbf{J}} = \begin{bmatrix} \Re(\mathbf{J}) \\ \Im(\mathbf{J}) \end{bmatrix} = \Re\left[\begin{bmatrix} \mathbf{L} \\ -\mathrm{i}\mathbf{L} \end{bmatrix}\mathbf{S}^{-1}\mathbf{P} + \begin{bmatrix} \mathbf{Q} \\ -\mathrm{i}\mathbf{Q} \end{bmatrix}\right] \tag{C1}$$

with the convention that for any observations that are intrinsically real the rows corresponding to the imaginary component are omitted. From (C1), $\bar{\mathbf{J}}^T\bar{\mathbf{d}} = \Re(\mathbf{J}^T)\Re(\mathbf{d}) + \Im(\mathbf{J}^T)\Im(\mathbf{d})$. It is easily seen that

$$\bar{\mathbf{J}}^T\bar{\mathbf{d}} = \Re[\mathbf{J}^T\mathbf{d}^*] = \Re[\mathbf{P}^T{\mathbf{S}^T}^{-1}\mathbf{L}^T\mathbf{d}^* + \mathbf{Q}\mathbf{d}^*], \tag{C2}$$

where the superscript asterisk denotes the complex conjugate. Thus, the complex component matrices can be used to construct the real Jacobian $\bar{\mathbf{J}}$, and to implement multiplication by this matrix and its transpose. Note also that while we assume the data vector is real, real and imaginary parts of sensitivities for a complex observation are computed (e.g. via 15) with a single adjoint solution.

Apparent resistivity and phase provide examples of observations that are intrinsically real. In terms of the impedance, the apparent resistivity is defined as

$$\rho_a = (\omega\mu)^{-1}|Z|^2 = (\omega\mu)^{-1}\left[Z_\mathrm{r}^2 + Z_\mathrm{i}^2\right], \tag{C3}$$

where Z_r and Z_i are real and imaginary parts of the impedance Z, and ω is angular frequency. Applying the chain rule,

$$\frac{\partial\rho_a}{\partial\mathbf{m}} = \frac{\partial\rho_a}{\partial Z_\mathrm{r}}\frac{\partial Z_\mathrm{r}}{\partial\mathbf{m}} + \frac{\partial\rho_a}{\partial Z_\mathrm{i}}\frac{\partial Z_\mathrm{i}}{\partial\mathbf{m}} = \frac{2}{\omega\mu}\left[Z_\mathrm{r}\frac{\partial Z_\mathrm{r}}{\partial\mathbf{m}} + Z_\mathrm{i}\frac{\partial Z_\mathrm{i}}{\partial\mathbf{m}}\right] \tag{C4}$$

$$= \frac{2}{\omega\mu}\left[Z_\mathrm{r}\Re\frac{\partial Z}{\partial\mathbf{m}} + Z_\mathrm{i}\Im\frac{\partial Z}{\partial\mathbf{m}}\right] = \Re\left[\frac{2Z^*}{\omega\mu}\frac{\partial Z}{\partial\mathbf{m}}\right]$$
$$= \Re\left[\frac{2Z^*\mathbf{l}_Z^T}{\omega\mu}\frac{\partial\mathbf{e}}{\partial\mathbf{m}}\right]. \tag{C5}$$

Thus, $\mathbf{l}_\rho = 2Z^*\mathbf{l}_Z^T/\omega\mu$ gives the (complex) row of $\mathbf{L}$ for an apparent resistivity, again with the convention that the real part of the product in (14) is taken for the corresponding row of the real Jacobian. Similarly for the phase $\phi = \tan^{-1}(Z_\mathrm{r}/Z_\mathrm{i})$, we find that the row of $\mathbf{L}$ takes the form $\mathbf{l}_\phi = \mathrm{i}Z^*\mathbf{l}_Z^T/|Z|^2$.

Geophysical Journal International

Geophys. J. Int. (2012) **190,** 255–266 doi: 10.1111/j.1365-246X.2012.05523.x

Hybrid conjugate gradient-Occam algorithms for inversion of multifrequency and multitransmitter EM data

Gary D. Egbert

College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA. E-mail: egbert@coas.oregonstate.edu

Accepted 2012 April 24. Received 2012 March 14; in original form 2011 September 02

SUMMARY

We describe novel hybrid algorithms for inversion of electromagnetic geophysical data, combining the computational and storage efficiency of a conjugate gradient approach with an Occam scheme for regularization and step-length control. The basic algorithm is based on the observation that iterative solution of the symmetric (Gauss-Newton) normal equations with conjugate gradients effectively generates a sequence of sensitivities for different linear combinations of the data, allowing construction of the Jacobian for a projection of the original full data space. The Occam scheme can then be applied to this projected problem, with the tradeoff parameter chosen by assessing fit to the full data set. For EM geophysical problems with multiple transmitters (either multiple frequencies or source geometries) an extension of the basic hybrid algorithm is possible. In this case multiple forward and adjoint solutions (one each for each transmitter) are required for each step in the iterative normal equation solver, and each corresponds to the sensitivity for a separate linear combination of data. From the perspective of the hybrid approach, with conjugate gradients generating an approximation to the full Jacobian, it is advantageous to save all of the component sensitivities, and use these to solve the projected problem in a larger subspace. We illustrate the algorithms on a simple problem, 2-D magnetotelluric inversion, using synthetic data. Both the basic and modified hybrid schemes produce essentially the same result as an Occam inversion based on a full calculation of the Jacobian, and the modified scheme requires significantly fewer steps (relative to the basic hybrid scheme) to converge to an adequate solution to the normal equations. The algorithms are expected to be useful primarily for 3-D inverse problems for which the computational burden is heavily dominated by solution to the forward and adjoint problems.

Key words: Inverse theory; Magnetotelluric; Geomagnetic induction.

1 INTRODUCTION

Among the most widely applied, and practical, approaches to inversion of electromagnetic (EM) geophysical data (e.g., magnetotellurics; MT) in two and three dimensions are regularized schemes based on minimizing a penalty functional of the form

$$\Phi(\mathbf{m},\mathbf{d}) = (\mathbf{d}-\mathbf{f}(\mathbf{m}))^{\mathrm{T}}\mathbf{C}_{\mathbf{d}}^{-1}(\mathbf{d}-\mathbf{f}(\mathbf{m})) + \lambda(\mathbf{m}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{C}_{\mathbf{m}}^{-1}(\mathbf{m}-\mathbf{m}_0), \tag{1}$$

(e.g. see Avdeev (2005) and Siripunvaraporn (2012) for reviews). In (1) $\mathbf{C_d}$ and $\mathbf{C_m}$ are data and model covariances; as these are not central to our focus we assume the simplest form for both ($\mathbf{C_d} = \mathbf{I}, \mathbf{C_m} = \mathbf{I}$), and we take the *a priori* model parameter $\mathbf{m}_0 = 0$. Treatment of the more general case complicates notation, but presents no essential difficulty for the ideas discussed here (see the Appendix for details). We consider in particular methods for minimization of (1) based on linearization of the non-linear model-data mapping $\mathbf{f}(\mathbf{m})$, that is, that make use of the derivative of $\mathbf{f}$, the $N \times M$ Jacobian $\mathbf{J}$ (so $J_{ij} = \partial f_i/\partial m_j$; N = #data; M = # model parameters). Two general approaches, each with many variants, can be distinguished. In a Gauss-Newton approach (e.g. Parker 1994) the full Jacobian is used to approximate the second-order (Taylor series) expansion of the penalty functional around a current estimate of the model solution. The resulting quadratic form is then minimized, leading to a standard linear least-squares problem, defined (at least formally) by the system of normal equations

$$(\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\mathbf{I})\delta\mathbf{m} = \mathbf{J}^{\mathrm{T}}(\mathbf{d} - \mathbf{f}(\mathbf{m}_n)) - \lambda\mathbf{m}_n, \tag{2a}$$

which can be solved for the model update

$$\mathbf{m}_{n+1} = \mathbf{m}_n + \delta\mathbf{m}. \tag{2b}$$

The whole procedure must be iterated, with the Jacobian recomputed for the updated model parameter, to achieve the minimum of (1). As described in Parker (1994) some form of step-length control is required (e.g. setting $\mathbf{m}_{n+1} = \mathbf{m}_n + \mu\delta\mathbf{m}$ with $0 < \mu \leq 1$ determined by line search). The second approach is epitomized by non-linear conjugate gradients (NLCG; e.g. Rodi & Mackie 2001):

the minimum of (1) is found by direct optimization, computing the gradient of the penalty functional

$$\frac{1}{2}\left.\frac{\partial \Phi}{\partial \mathbf{m}}\right|_{\mathbf{m}_n} = -\mathbf{J}^{\mathrm{T}}(\mathbf{d} - \mathbf{f}(\mathbf{m}_n)) + \lambda \mathbf{m}_n, \tag{3}$$

and using this to define a search direction in the model space. Φ is then minimized along this search direction, the model parameter is updated to $\mathbf{m}_{n+1}$ and the whole process is repeated. Both approaches are reviewed and compared in the context of EM geophysics problems of the sort considered here in Rodi & Mackie (2001). Limited memory quasi-Newton (Liu & Nocedal 1989) represents an alternative direct optimization approach, which has also been used for EM inverse problems (e.g. Avdeev & Avdeeva 2009).

The forward problem $\mathbf{f}(\mathbf{m})$ for a frequency-domain EM induction problem, such as 2-D or 3-D MT, involves solving elliptic partial differential equations (PDEs), derived from Maxwell's equations. For example, in quasi-static 3-D EM problems the governing equations formulated in terms of the electric field $\mathbf{E}$ are:

$$\nabla \times \nabla \times \mathbf{E} - \mathrm{i}\omega\mu\sigma\mathbf{E} = \mathbf{s}. \tag{4}$$

The forward mapping $\mathbf{f}(\mathbf{m})$ requires solving (4) subject to appropriate boundary conditions, and using the solution, evaluated at observation locations, to compute predicted data. In (4) σ is the spatially varying electrical conductivity of the medium, which we assume is defined through the unknown discrete model parameter $\mathbf{m}$, ω is angular frequency and $\mathbf{s}$ represents the sources (which may vanish, as for MT where the system is forced through the boundary conditions). In most realistic problems data are available for N_f frequencies, and N_s source geometries, so a total of $N_f N_s$ PDEs must be solved to evaluate $\mathbf{f}(\mathbf{m})$ for a single model parameter. As shown in general by Newman & Hoversten (2000), Pankratov & Kuvshinov (2010), and Egbert & Kelbert (2012) and previously for numerous specific examples referenced therein, computing one row (or one column) of $\mathbf{J}$ requires solving the governing PDE (or more precisely, its adjoint; though (4) is essentially self-adjoint) once. Evaluating a matrix-vector product such as $\mathbf{J}^{\mathrm{T}}\mathbf{r}$ (e.g. in the gradient of data misfit used in (3)) requires essentially the same computations as one forward problem.

A GN approach would appear at first blush to be much less efficient than NLCG: to implement (2) directly, one must apparently first compute all of $\mathbf{J}$ (requiring $N = N_f N_s N_r$ (where N_r is the number of receivers) solutions of the appropriate PDE, one for each row of the Jacobian), and then form and solve the $M \times M$ system of equations. In contrast, a single iteration with (3) requires a single gradient computation, followed by a line search to minimize over the search direction (generally requiring 2–4 additional solutions of the forward problem). However, as Rodi & Mackie (2001) show, NLCG requires many more iterations (typically 50–100 or more) compared to a GN scheme (typically 5–10 or less; see examples below). Furthermore, for 'multitransmitter' problems (i.e. with multiple frequencies and/or source geometries) each forward solution or gradient evaluation actually requires solving the governing PDE $N_f N_s$ times. Accounting for the significantly greater number of iterations needed for convergence (each requiring a line search) direct minimization with NLCG may require as many or more PDE solutions as a GN scheme based on full calculation of $\mathbf{J}$ (Siripunvaraporn & Egbert 2007; Siripunvaraporn & Sarakorn 2011). However, NLCG still avoids forming and solving the large system of normal equations of (2), so this and related approaches are now used in almost all implementations of 3-D inversion (e.g. Commer & Newman 2008; Avdeev & Avdeeva 2009); the efforts of Sasaki (2001), Siripunvaraporn *et al.* (2005) and Siripunvaraporn & Egbert (2009) are exceptions.

It is of course possible to use a GN approach without explicitly forming the normal equations of (2a), but instead solve this symmetric linear system of equations iteratively using conjugate gradients (CG). This approach, which has been used fairly extensively for EM inversion (e.g. Mackie & Madden 1993; Alumbaugh & Newman 1997; Rodi & Mackie 2001) is a variant on the truncated Newton approach to optimization (e.g. Dembo *et al.* 1982; Nash 2000), with the Hessian replaced by the GN approximation (e.g. Newman & Hoversten 2000).

To be concrete, and to set the stage for coming developments, we consider a variant on the GN equations of (2):

$$(\mathbf{J}\mathbf{J}^{\mathrm{T}} + \lambda\mathbf{I})\mathbf{b} = \hat{\mathbf{d}} = \mathbf{d} - \mathbf{f}(\mathbf{m}_n) + \mathbf{J}\mathbf{m}_n \tag{5a}$$

$$\mathbf{m}_{n+1} = \mathbf{J}^{\mathrm{T}}\mathbf{b}. \tag{5b}$$

This data space scheme (e.g. Siripunvaraporn & Egbert 2000; Siripunvaraporn & Sarakorn 2011), which requires solving the $N \times N$ system of normal equations in the data space (instead of the $M \times M$ system in the model parameter space), can be shown to be equivalent to (2). Instead of actually making the full dense matrix, one can again use CG, which requires multiplying an arbitrary vector by the coefficient matrix $(\mathbf{J}\mathbf{J}^{\mathrm{T}} + \lambda\mathbf{I})$. This in turn requires multiplication of data space vectors by $\mathbf{J}^{\mathrm{T}}$ and model space vectors by $\mathbf{J}$, essentially the same sort of computations as required by NLCG. This approach also avoids calculation of the full Jacobian and eliminates the need to form the normal equations. As shown in Siripunvaraporn & Egbert (2007) the total number of PDE solutions is, however, still typically comparable to that required for a full calculation of $\mathbf{J}$. And the CG scheme has an apparent disadvantage: once the full Jacobian is calculated, solving (5a) for different values of the tradeoff parameter λ is fairly fast—in particular no further PDE solutions are required.

The Occam approach (Constable *et al.* 1987; see also Parker 1994) exploits this efficiency, varying λ both for step length control, and as a damping parameter, to search for minimum norm inverse solutions, which fit the data to a prescribed tolerance. Once $\mathbf{J}$ is computed (5) is used to compute a series of trial solutions corresponding to a range of λ, and the forward problem is then solved for each to evaluate the actual data misfit achieved as a function of λ. Initially, λ is chosen to minimize data misfit; as the scheme converges λ is chosen to minimize the model norm while keeping the misfit constant at the target value (Constable *et al.* 1987; Parker 1994). With this approach λ is determined as part of the search process, and at convergence one is assured that the solution attains at least a local minimum of the model norm, subject to the data fit attained (Parker 1994). With a straightforward application of CG all of the PDE solution steps must be repeated for each new trial value of λ (Siripunvaraporn & Egbert 2007). The same situation holds for NLCG: the penalty functional is minimized with λ fixed, and the entire (or at least much of) the iterative process must be repeated with each new trial value. Thus, if one has to vary the regularization parameter—and often this is critical, even if one does not have the precise information about data error levels required to rigorously provide an *a priori* target misfit—GN schemes based on full calculation of $\mathbf{J}$ would appear to have some advantages.

We make two points in this paper. The first is in fact rather obvious: at the cost of a modest increase in memory requirements, CG schemes can be easily modified to allow the Occam approach to be implemented without computing the full Jacobian. The idea is closely related to the so-called hybrid algorithms, which have previously been discussed in the context of damped least-squares problems (O'Leary & Simmons 1981; Kilmer & O'Leary 2001; Hanke 2001). It can also be viewed as a special case of the subspace inversion methods of Oldenburg *et al.* (1993), in which a small and effective model subspace is generated by the iterative CG solver.

Our second point is more novel, and is specific to multitransmitter inverse problems where computational costs are dominated by the need for multiple expensive forward solutions. Such problems are the norm in EM geophysics, and arise also in other geophysical problems, such as full waveform seismic inversion (e.g. Tape *et al.* 2010). We show that for such problems iterative solution of the data space normal eq. (5a) can be modified to achieve substantially more rapid convergence (in particular, with fewer required forward solutions). Both ideas follow from a more careful examination of the iterative CG algorithms used to solve (5a). We thus review the basis for the CG solution approach—that is, the Lanczos process—in Section 2, and demonstrate how the standard solution scheme can be easily modified to implement a hybrid CG-Occam scheme. In Section 3, we develop a modification to the Lanczos process that uses the multiplicity of forward and adjoint solutions required in multitransmitter EM geophysical inverse problems to accelerate convergence of the solution to the normal equations, leading to a modified hybrid CG-Occam algorithm. In Section 4, we demonstrate the efficacy of the new schemes using the 2-D MT inverse problem as a simple illustrative example. Although this simple problem is sufficient to demonstrate the effectiveness of the new algorithms, we stress that these schemes are likely to be most useful for 3-D problems where computational costs are dominated by expensive forward and adjoint solutions required for gradient calculations. Results and possible extensions are discussed in Section 5.

2 A HYBRID CG-OCCAM SCHEME

To motivate and describe the hybrid schemes we begin with a review of the Lanczos bi-diagonalization algorithm of Paige & Saunders (1982a), which forms the basis for standard CG solution methods. Here the algorithm 'BIDIAG1' is applied to the Jacobian $\mathbf{J}$, with the ultimate objective of solving the system of normal eq. (5a), initially taking $\lambda = 0$. In the first step of the Lanczos process unit vectors in the data and model space are computed

$$\beta_1 \mathbf{u}_1 = \hat{\mathbf{d}} \qquad \|\mathbf{u}_1\| = 1 \tag{6a}$$

$$\alpha_1 \mathbf{v}_1 = \mathbf{J}^{\mathrm{T}} \mathbf{u}_1 \qquad \|\mathbf{v}_1\| = 1. \tag{6b}$$

A key point to note here is that the model space vector $\alpha_1 \mathbf{v}_1 = \mathbf{J}^{\mathrm{T}}\mathbf{u}_1$ is just the sensitivity of a particular linear combination of data components, namely $\mathbf{u}_1^{\mathrm{T}}\mathbf{d}$ (ignoring noise $\partial \mathbf{u}_1^{\mathrm{T}}\mathbf{d}/\partial\mathbf{m} = \partial \mathbf{u}_1^{\mathrm{T}}\mathbf{f}/\partial\mathbf{m} = \mathbf{u}_1^{\mathrm{T}}\partial\mathbf{f}/\partial\mathbf{m} = \mathbf{u}_1^{\mathrm{T}}\mathbf{J} = \alpha_1\mathbf{v}^{\mathrm{T}}$). Next compute

$$\mathbf{J}\mathbf{v}_1 = \alpha_1 \mathbf{u}_1 + \beta_2 \mathbf{u}_2, \tag{7}$$

where $\mathbf{u}_2$ is orthogonal to $\mathbf{u}_1$. Now if $\beta_2 = 0$, $\mathbf{J}\mathbf{J}^{\mathrm{T}}[\beta_1/\alpha_1^2]\mathbf{u}_1 = \beta_1\mathbf{u}_1 = \hat{\mathbf{d}}$ so $\mathbf{b}_1 = (\beta_1/\alpha_1^2)\mathbf{u}_1$ would be an exact solution to (5a). In general, this will not be the case, and this initial estimate of $\mathbf{b}$ must be refined. We thus continue for $k = 2, \ldots, K$(where K is determined by the stopping criterion discussed below)

$$\beta_k \mathbf{u}_k = \mathbf{J}\mathbf{v}_{k-1} - \alpha_{k-1}\mathbf{u}_{k-1} \qquad \|\mathbf{u}_k\| = 1 \tag{8a}$$

$$\mathbf{J}^{\mathrm{T}}\mathbf{u}_k - \beta_k \mathbf{v}_{k-1} = \alpha_k \mathbf{v}_k \qquad \|\mathbf{v}_k\| = 1, \tag{8b}$$

generating sequences of data and model space vectors (which can be saved as orthogonal matrices $\mathbf{U}_K = [\mathbf{u}_1 \cdots \mathbf{u}_K]$ and $\mathbf{V}_K = [\mathbf{v}_1 \cdots \mathbf{v}_K]$, respectively) and scalars α_k, β_k which can be organized as the bi-diagonal matrix

$$\mathbf{B}_K = \begin{bmatrix} \alpha_1 & \beta_2 & \cdots & 0 \\ 0 & \alpha_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \beta_K \\ 0 & \cdots & 0 & \alpha_K \end{bmatrix}. \tag{9}$$

Then (6)–(8) can be expressed in matrix notation as

$$\mathbf{J}^{\mathrm{T}}\mathbf{U}_K = \mathbf{V}_K \mathbf{B}_K \tag{10}$$

$$\mathbf{J}\mathbf{V}_K = \mathbf{U}_K \mathbf{B}_K^{\mathrm{T}} + \beta_{K+1}\mathbf{u}_{K+1}\hat{\mathbf{e}}_K^{\mathrm{T}}, \tag{11}$$

where $\hat{\mathbf{e}}_K$ is the unit vector for coordinate K in $\mathbb{R}^K$ and $\mathbf{J}\mathbf{v}_K = \alpha_K\mathbf{u}_K + \beta_{K+1}\mathbf{u}_{K+1}$. The original system $\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{b} = \hat{\mathbf{d}}$ can be solved approximately by first projecting into the K-dimensional data subspace spanned by the columns of $\mathbf{U}_K$; i.e.

$$\mathbf{U}_K^{\mathrm{T}}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{U}_K\tilde{\mathbf{b}}_K = \mathbf{U}_K^{\mathrm{T}}\hat{\mathbf{d}} = \beta_1\hat{\mathbf{e}}_1. \tag{12}$$

The last equality follows from (6a) and orthonormality of the columns of $\mathbf{U}_K$. On the other hand, from the orthonormality of $\mathbf{V}_K$ and (10), the system to be solved can be seen to be symmetric, positive definite and tri-diagonal

$$\mathbf{B}_K^{\mathrm{T}}\mathbf{B}_K\tilde{\mathbf{b}} = \beta_1\hat{\mathbf{e}}_1, \tag{13}$$

and hence easily solved. The vector $\mathbf{b}_K = \mathbf{U}_K\tilde{\mathbf{b}}$ then provides an approximate solution to the original system. Indeed we have from (10–13) and (6a)

$$\begin{aligned} \mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{b}_K &= \mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{U}_K\tilde{\mathbf{b}} = \mathbf{J}\mathbf{V}_K\mathbf{B}_K\tilde{\mathbf{b}} = \left[\mathbf{U}_K\mathbf{B}_K^{\mathrm{T}} + \beta_{K+1}\mathbf{u}_{K+1}\hat{\mathbf{e}}_K^{\mathrm{T}}\right]\mathbf{B}_K\tilde{\mathbf{b}} \\ &= \mathbf{U}_K\mathbf{B}_K^{\mathrm{T}}\mathbf{B}_K\tilde{\mathbf{b}} + \beta_{K+1}\mathbf{u}_{K+1}\left[\hat{\mathbf{e}}_K^{\mathrm{T}}\mathbf{B}_K\tilde{\mathbf{b}}\right] \\ &= \beta_1\mathbf{U}_K\hat{\mathbf{e}}_1 + \beta_{K+1}\left[\alpha_K\hat{\mathbf{e}}_K^{\mathrm{T}}\tilde{\mathbf{b}}\right]\mathbf{u}_{K+1} \end{aligned} \tag{14}$$

$$= \beta_1\mathbf{u}_1 + \alpha_K\beta_{K+1}\tilde{b}_K\mathbf{u}_{K+1} = \hat{\mathbf{d}} + \alpha_K\beta_{K+1}\tilde{b}_K\mathbf{u}_{K+1}, \tag{15}$$

where $\tilde{b}_K$ is the Kth component of the vector $\tilde{\mathbf{b}}$.

In standard implementations of CG the system (12) is not actually formed and solved. Rather, the approximate solution $\mathbf{b}_K$ is updated 'on the fly', starting from $\mathbf{b}_1 = (\beta_1/\alpha_1^2)\mathbf{u}_1$. Iterations can be terminated when the residual in the solution to (eq. 5a; i.e. $\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{b}_K - \hat{\mathbf{d}} = \alpha_K\beta_{K+1}\tilde{b}_K\mathbf{u}_{K+1}$) is sufficiently reduced, for example, when $\|\alpha_K\beta_{K+1}\tilde{b}_K\mathbf{u}_{K+1}\|/\|\hat{\mathbf{d}}\| < \varepsilon$. More generally, memory efficient and numerically stable schemes for damped least squares problems (e.g., LSQR) have been developed based on Lanczos bi-diagonalization (Paige & Saunders 1982b). With these approaches memory requirements are minimal—only the most recent $\mathbf{u}_k, \mathbf{v}_k$ need be retained, and solutions (and residuals) are updated at each step k. However, by actually saving all of $\mathbf{U}_K$, $\mathbf{V}_K$ and $\mathbf{B}_K$ (or in fact $\mathbf{U}_K$ and $\mathbf{J}^{\mathrm{T}}\mathbf{U}_K$) it is possible to form and solve the small $(K \times K)$ system $[\mathbf{U}_{\mathrm{k}}^{\mathrm{T}}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{U}_K + \lambda\mathbf{I}]\mathbf{b}_\lambda = \mathbf{U}_K^{\mathrm{T}}\hat{\mathbf{d}}$ (analogous to (12)) for any value of the regularization parameter λ. It is readily verified that the same error estimate (15) applies to this modified system. This approach, which allows an efficient implementation of the Occam scheme, is an example of a hybrid algorithm, of the sort previously discussed extensively in the numerical linear algebra literature (O'Leary & Simmons 1981; Kilmer & O'Leary 2001; Hanke 2001).

A hybrid Occam-CG scheme is thus obvious: (1) Apply Lanczos bi-diagonalization to $\mathbf{J}$, saving the orthonormal matrix $\mathbf{U}_K$, and the K model space vectors $\mathbf{J}^{\mathrm{T}}\mathbf{U}_K$. (2) Use these to form the $K \times K$

```
$\mathbf{m}_{prior}$ = prior model        $\mathbf{m}_0$ = starting model
Outer loop: For $n$ = 0, 1, 2, ...
    $\hat{\mathbf{d}}_n = \mathbf{d} - \mathbf{f}(\mathbf{m}_n) + \mathbf{J}[\mathbf{m}_n - \mathbf{m}_{prior}]$

    BIDIAG1:
    $\beta_1 \mathbf{u}_1 = \hat{\mathbf{d}}$        $\|\mathbf{u}_1\| = 1$
    $\alpha_1 \mathbf{v}_1 = \mathbf{J}^T \mathbf{u}_1$        $\|\mathbf{v}_1\| = 1$
    for $k = 2, 3, ...$
        $\beta_k \mathbf{u}_k = \mathbf{J}\mathbf{v}_{k-1} - \alpha_{k-1}\mathbf{u}_{k-1}$        $\|\mathbf{u}_k\| = 1$
        if $|\beta_k \alpha_{k-1}| / \|\hat{\mathbf{d}}\| < \varepsilon$ exit loop
        $\mathbf{J}^T \mathbf{u}_k - \beta_k \mathbf{v}_{k-1} = \alpha_k \mathbf{v}_k$        $\|\mathbf{v}_k\| = 1$
    end BIDIAG1 (save $\mathbf{U}_K = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_K]$, $\mathbf{J}^T \mathbf{U}_K$ )

    Optimize $\lambda$: for trial values of $\lambda$
        Solve $K \times K$ system $[\mathbf{U}_K^T \mathbf{J}\mathbf{J}^T \mathbf{U}_K + \lambda \mathbf{I}]\mathbf{b}_\lambda = \mathbf{U}_K^T \hat{\mathbf{d}}$
        $\mathbf{m}_\lambda = \mathbf{J}^T \mathbf{U}_K \mathbf{b}_\lambda$
        Phase I:
            choose $\lambda$ to minimize $\|\mathbf{d} - \mathbf{f}(\mathbf{m}_\lambda)\|^2$
        Phase II:
            choose $\lambda$ so that $\|\mathbf{d} - \mathbf{f}(\mathbf{m}_\lambda)\|^2 = Tol$
end outer loop
```

Figure 1. Pseudo-code for hybrid Occam-DCG.

cross-product matrix $\mathbf{R} = \mathbf{U}_K^{\mathrm{T}}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{U}_K$. This matrix is in principal tridiagonal, but round-off error will cause increasing large deviations as K increases, so it is best to retain and work with the matrix $\mathbf{J}^{\mathrm{T}}\mathbf{U}_K$. (3) Optimize the regularization parameter by solving the projected system $[\mathbf{R} + \lambda\mathbf{I}]\mathbf{b}_\lambda = \mathbf{U}_K^{\mathrm{T}}\hat{\mathbf{d}}$ for a series of values of λ, and minimize the misfit $\|\mathbf{d} - \mathbf{f}(\mathbf{m}_\lambda)\|^2$, where $\mathbf{m}_\lambda = \mathbf{J}^{\mathrm{T}}\mathbf{b}_\lambda$; after the target misfit is achieved, choose λ to minimize the penalty functional subject to achieving the target misfit. Pseudo-code for the scheme is given in Fig. 1. This hybrid scheme effectively uses the Lanczos process to generate a subset of sensitivities (i.e. the columns of $\mathbf{U}_K^{\mathrm{T}}\mathbf{J}$), corresponding to the data subspace spanned by $\mathbf{U}_K$. The Occam scheme is then applied to this projected problem, with the tradeoff parameter chosen by assessing fit to the full data set.

The basic hybrid algorithm solves the linear subproblem in the model subspace spanned by the columns of $\mathbf{V}_K$, and can thus be viewed as a special case of the subspace inversion methods discussed in Oldenburg *et al.* (1993). Although we have focused on a data space Occam approach, the same ideas are readily adapted to alternative G-N formulations in the model space, for example, to solve (2) for $\delta\mathbf{m}$. From this perspective the Lanczos bi-diagonalization can be viewed as a scheme for generating a particular model subspace, which approximates the row span of $\mathbf{J}$, and thus should be particularly efficient for finding approximate solutions to the full system of normal equations (with any value of the regularization parameter). Note that the Lanczos process already generates the sensitivity matrix-model parameter products $\mathbf{J}\mathbf{V}_K$ needed to generate the reduced normal equations for the subspace inversion approach (see Oldenburg *et al.* 1993), so a subspace inversion based on saving the full set of Lanczos vectors would be quite efficient.

3 A MODIFIED HYBRID SCHEME

In most EM inverse problems data are available for multiple frequencies, or more generally, with multiple transmitters (different frequencies and/or different source geometries). In this case the data vector and Jacobian can be decomposed into J (= number of transmitters) blocks as

$$\mathbf{d} = \begin{pmatrix} \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_J \end{pmatrix}, \tag{16}$$

$$\mathbf{J}^{\mathrm{T}} = \begin{bmatrix} \mathbf{J}_1^{\mathrm{T}} & \cdots & \mathbf{J}_J^{\mathrm{T}} \end{bmatrix}. \tag{17}$$

A product such as $\mathbf{J}^{\mathrm{T}}\mathbf{r} = \sum_j \mathbf{J}_j^{\mathrm{T}}\mathbf{r}_j$ actually entails separate computations for each transmitter (each requiring solution of the governing PDE appropriate for that frequency), followed by summing the results (a sequence of J model space vectors). Details of the Jacobian calculation are somewhat different for the case of multiple transmitters (with a common frequency), but the decompositions of (16) and (17) remain valid and the required number of forward and adjoint solutions remains the same. Each one of the model space vectors $\mathbf{J}_j^{\mathrm{T}}\mathbf{r}_j$ gives the sensitivity of a linear combination of data for a single transmitter. The basic idea behind the modified hybrid algorithm is to save all of these separate sensitivities, and use these to solve a projected data-space system analogous to (12). The key modification is actually to the Lanczos scheme, which we will show in Section 4 results in convergence of the normal equations in fewer iterations (i.e. with fewer matrix-vector products $\mathbf{J}^{\mathrm{T}}\mathbf{u}$ and $\mathbf{Jv}$).

The scheme is motivated by the observation that if we saved the individual transmitter data-space vectors and corresponding sensitivities $\mathbf{u}_{kj}, \mathbf{J}_j^{\mathrm{T}}\mathbf{u}_{kj}$ generated by the Lanczos bi-diagonalization discussed earlier, we could project (5) into a much larger (JK-dimensional) data subspace (which would contain the K-dimensional space spanned by the vectors $\mathbf{u}_k$), perhaps leading to a more accurate solution—or rather, allowing an equally accurate solution for a smaller value of K. We have found a modification to this simple idea to be significantly more effective.

As for the standard Lanczos bi-diagonalization, the modified scheme generates a sequence of data space vectors which we denote as $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_K$, subdivided into individual transmitter components as $\tilde{\mathbf{u}}_k^{\mathrm{T}} = [\mathbf{w}_{1k}^{\mathrm{T}} \cdots \mathbf{w}_{Jk}^{\mathrm{T}}]$, with each now normalized separately so $\|\mathbf{w}_{jk}^{\mathrm{T}}\| = 1$. As for standard Lanczos schemes the process is started from the right hand side $\hat{\mathbf{d}}$ of the system (5), but now with each block of the data vector normalized separately $\mathbf{w}_{j1} = \hat{\mathbf{d}}_j/\|\hat{\mathbf{d}}_j\|$. Leaving aside for the moment how the vectors $\mathbf{w}_{jk}$ are generated for $k > 1$, let $\Omega_{jk} = [\mathbf{w}_{j1} \cdots \mathbf{w}_{jk}]$ be the matrix constructed from the first $k$subvectors for transmitter$j$, and define the block diagonal matrix

$$\mathbf{W}_k = \mathbf{diag}\,(\Omega_{1k}, \ldots, \Omega_{Jk})\,. \tag{18}$$

Then the columns of $\mathbf{J}_j^{\mathrm{T}}\Omega_{jk}$ are model parameter vectors corresponding to the sensitivity for the k linear combinations of data defined by $\Omega_{jk}^{\mathrm{T}}\mathbf{d}_j$, and

$$\mathbf{J}^{\mathrm{T}}\mathbf{W}_K = \left[\mathbf{J}_1^{\mathrm{T}}\Omega_{1K} \cdots \mathbf{J}_J^{\mathrm{T}}\Omega_{JK}\right], \tag{19}$$

is the $M \times (KJ)$ matrix containing all of the sensitivities generated by the first K steps. We show by induction that, with the scheme for generating $\tilde{\mathbf{u}}_k$ described next, $\Omega_{jk}^{\mathrm{T}}\Omega_{jk} = \mathbf{I}$ for all k (i.e. for fixed j the vectors $\mathbf{w}_{jk}, k = 1, \ldots, K$ are orthonormal) so that $\mathbf{W}_K^{\mathrm{T}}\mathbf{W}_K = \mathbf{I}$.

Orthonormality of $\mathbf{W}_K$ certainly holds for $K = 1$. Supposing it holds also for K, we can use the computed sensitivities to solve the projected problem

$$\left(\mathbf{W}_K^{\mathrm{T}}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{W}_K + \lambda_0\mathbf{I}\right)\tilde{\mathbf{b}} = \mathbf{W}_K^{\mathrm{T}}\hat{\mathbf{d}} \tag{20}$$

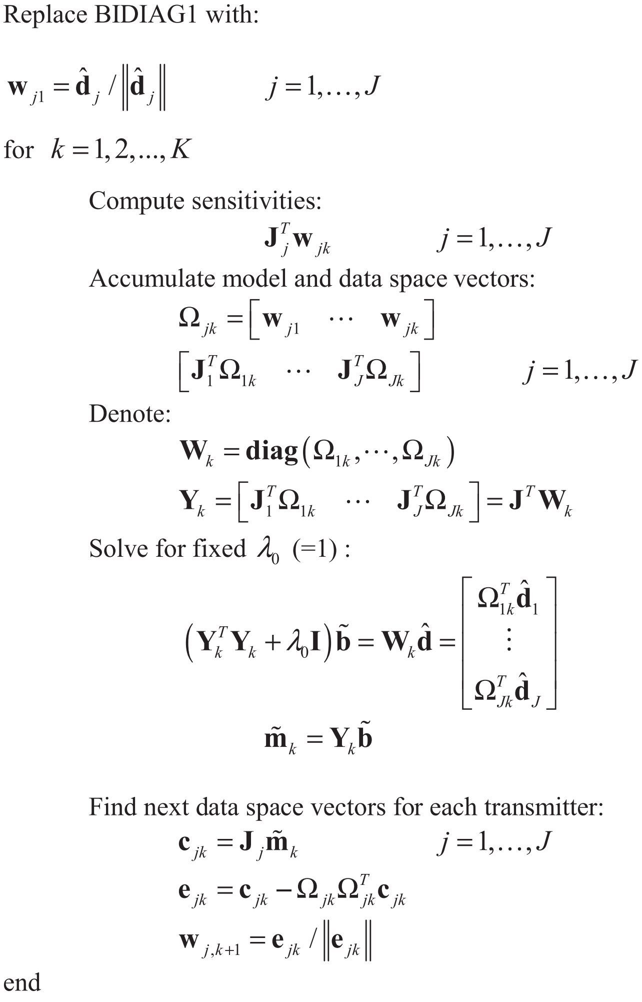
for any fixed λ_0. This is analogous to (12), but the matrix $\mathbf{W}_K$ has KJ instead of K columns, so the projected problem is solved in a larger subspace. Given the solution to (20) we next compute $\tilde{\mathbf{m}}_K = \mathbf{J}^{\mathrm{T}}\mathbf{b}_K = \mathbf{J}^{\mathrm{T}}\mathbf{W}_k\tilde{\mathbf{b}}$. If iterative solution of the linear subproblem (5a) were truncated at this point, $\tilde{\mathbf{m}}_K$ would be the model update for the next iteration given in eq. (5b). To continue iterations we compute

$$\mathbf{J}\tilde{\mathbf{m}}_K = \mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{W}_K\tilde{\mathbf{b}}_K = \mathbf{W}_K\mathbf{W}_K^{\mathrm{T}}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{W}_K\tilde{\mathbf{b}}_K + \mathbf{e}_{K+1}, \tag{21}$$

where $\mathbf{e}_{K+1}$ is orthogonal to all of the columns of $\mathbf{W}_K$, that is, $\mathbf{W}_K^{\mathrm{T}}\mathbf{e}_{K+1} = 0$. But then we have $\Omega_{jK}^{\mathrm{T}}\mathbf{e}_{j,K+1} = 0$, so setting $\mathbf{w}_{j,K+1} = \mathbf{e}_{j,K+1}/\|\mathbf{e}_{j,K+1}\|$, this vector is orthogonal to $\mathbf{w}_{jk}, k = 1, \ldots, K$. Thus, $\Omega_{j,K+1}, j = 1, \ldots, J$, and hence $\mathbf{W}_{K+1}$, are all orthonormal matrices, as claimed. Note that $\mathbf{e}_{K+1}$ is analogous to $\mathbf{u}_{K+1}$ in (8a)—that is, it represents the next data-space search direction, but blocks for each transmitter will be used separately.

Note that $\mathbf{W}_K\mathbf{W}_K^{\mathrm{T}}\hat{\mathbf{d}} = \hat{\mathbf{d}}$, and thus (20–21) imply that

$$\left(\mathbf{J}\mathbf{J}^{\mathrm{T}} + \lambda_0\mathbf{I}\right)\mathbf{b}_K = \hat{\mathbf{d}} + \mathbf{e}_{K+1}, \tag{22}$$



```
Replace BIDIAG1 with:
$\mathbf{w}_{j1} = \hat{\mathbf{d}}_j / \|\hat{\mathbf{d}}_j\|$        $j = 1,\ldots,J$
for $k = 1,2,...,K$
    Compute sensitivities:
        $\mathbf{J}_j^T\mathbf{w}_{jk}$        $j = 1,\ldots,J$
    Accumulate model and data space vectors:
        $\Omega_{jk} = \left[\mathbf{w}_{j1} \;\cdots\; \mathbf{w}_{jk}\right]$
        $\left[\mathbf{J}_1^T\Omega_{1k} \;\cdots\; \mathbf{J}_J^T\Omega_{Jk}\right]$        $j = 1,\ldots,J$
    Denote:
        $\mathbf{W}_k = \mathbf{diag}\left(\Omega_{1k},\cdots,\Omega_{Jk}\right)$
        $\mathbf{Y}_k = \left[\mathbf{J}_1^T\Omega_{1k} \;\cdots\; \mathbf{J}_J^T\Omega_{Jk}\right] = \mathbf{J}^T\mathbf{W}_k$
    Solve for fixed $\lambda_0$ (=1) :
        $\left(\mathbf{Y}_k^T\mathbf{Y}_k + \lambda_0\mathbf{I}\right)\tilde{\mathbf{b}} = \mathbf{W}_k\hat{\mathbf{d}} = \begin{bmatrix}\Omega_{1k}^T\hat{\mathbf{d}}_1 \\ \vdots \\ \Omega_{Jk}^T\hat{\mathbf{d}}_J\end{bmatrix}$
        $\tilde{\mathbf{m}}_k = \mathbf{Y}_k\tilde{\mathbf{b}}$
    Find next data space vectors for each transmitter:
        $\mathbf{c}_{jk} = \mathbf{J}_j\tilde{\mathbf{m}}_k$        $j = 1,\ldots,J$
        $\mathbf{e}_{jk} = \mathbf{c}_{jk} - \Omega_{jk}\Omega_{jk}^T\mathbf{c}_{jk}$
        $\mathbf{w}_{j,k+1} = \mathbf{e}_{jk} / \|\mathbf{e}_{jk}\|$
end
```

Figure 2. Pseudo-code for modified hybrid scheme.

so that $\mathbf{b}_K$ provides a good approximate solution to (5a) provided $\|\mathbf{e}_{K+1}\|$ is small enough. This can thus serve as a stopping criterion. If the residual is not sufficiently reduced, $\mathbf{e}_{K+1}$ can be used to generate the data space vectors $\mathbf{w}_{j,K+1}, j = 1, \ldots, J$, along with the corresponding model space vectors $\mathbf{J}_j^{\mathrm{T}}\mathbf{w}_{j,K+1}$ for the next iteration.

Pseudo-code for the modified hybrid scheme is given in Fig. 2. A key point to note is that a full solution to the projected linear subproblem is required at each step—that is, (20) is solved, and $\tilde{\mathbf{m}}_K$ formed for the computation of (21). In fact this is what is required to verify that the solution in the projected subspace (i.e. $\mathbf{b}_K$) solves the unprojected system (22) with sufficiently small residual. However, forming and solving the projected normal equations will not represent a serious computational challenge as long as KJ remains a small fraction of the total number of data. In particular, for 3-D problems where computational effort is dominated by solving the 3-D forward and adjoint problems, these extra steps will typically be negligible.

Note that the modified scheme depends on the initial tradeoff parameter selected λ_0. This is because the intermediate solution $\tilde{\mathbf{m}}_k$, which is used through (21) to compute the next data space search vectors $\mathbf{e}_{k+1}, \mathbf{w}_{j,k+1}$, depends on λ_0. The trade-off parameter should scale with the eigenvalues of the matrix $\mathbf{J}\mathbf{J}^{\mathrm{T}}$, and we can very roughly estimate this scale from

$$\mathbf{Tr}\left[\mathbf{W}_1^{\mathrm{T}}\mathbf{J}\mathbf{J}^{\mathrm{T}}\mathbf{W}_1\right] = \sum\nolimits_j \|\mathbf{J}_j^{\mathrm{T}}\mathbf{w}_{j1}\|^2/J = \eta_0, \tag{23}$$

which is computed in the first step of the modified Lanczos process, before λ_0 is required.

In our tests we have taken $\lambda_0 = 0.01\eta_0$ for the first loop of the Occam scheme, and then used the optimal λ from the previous iteration for subsequent iterations.

As with the standard scheme, the modified Lanczos scheme generates a model subspace, now spanned by the KJ columns of $\mathbf{J}^{\mathrm{T}}\mathbf{W}_K$. However, the connection to the model subspace inversion methods of Oldenburg *et al.* (1993) is now somewhat weaker than for the standard hybrid scheme. To solve the model space equations of (2) projected into this subspace it would be necessary to compute $\mathbf{Jv}$ for all KJ model space vectors, but in the modified Lanczos scheme the Jacobian is only applied to the K intermediate model solution vectors $\tilde{\mathbf{m}}_k$. The modified hybrid scheme developed here is thus more clearly rooted in the data space perspective, with the inverse problem solved for a projection of the full data vector. It is also worth noting that the sequence of data subspaces that we solve the problem in are not the usual Krylov subspaces generated by repeated application of $\mathbf{JJ}^{\mathrm{T}}$. Indeed the actual sequence of projected spaces depends to some extent on λ_0.

4 EXAMPLE: 2-D MT

As an illustration of the above ideas we consider the 2-D MT inverse problem. While the more complicated modified hybrid scheme for multitransmitter problems can hardly be justified for such a problem, where computational costs associated with forward and adjoint calculations are relatively low, this simple problem is sufficient to demonstrate the two main points we wish to make: (1) even for modest values of K the hybrid schemes essentially reproduce results obtained with a full Occam scheme based on a full Jacobian calculation, and (2) the modified hybrid scheme accomplishes this with fewer iterations of the Lanczos schemes, and hence fewer forward and adjoint solutions.

For 2-D MT electrical conductivity is assumed to be a function of depth z and 'cross-strike' distance y, with no variation along the x-direction (e.g. Fig. 3a), and source magnetic fields are assumed to be uniform (constant in x and y at $z = -\infty$). The magnetic source can be polarized either perpendicular or parallel to strike, corresponding to TE and TM modes, with induced currents flowing along and across strike, respectively (i.e. in the x direction and in the y–z plane). Data for this problem are complex impedances ($Z_{xy} = E_x/B_y$ for TE mode; $Z_{yx} = E_y/B_x$ for TM mode) observed at a series of N_s y-locations at the surface $z = 0$, for a set of N_f frequencies. With this setup the total number of 'transmitters', each requiring solution of a separate forward problem, is $J = 2N_f$, and the total number of (complex) data is $N = 2N_f N_s$. We have implemented the inversion schemes outlined above (a standard data space Occam approach, plus the hybrid scheme of Fig. 1 and the modified hybrid scheme of Fig. 2) and tested these on a range of synthetic data sets; we show results from two cases here. Forward and adjoint problems were solved numerically using a finite difference approach, essentially identical to that used in Siripunvaraporn & Egbert (2000), with the actual inversion procedures implemented in Matlab. In our implementation the regularization term was essentially as in (1), with deviations from a prior model ($\mathbf{m} - \mathbf{m}_0$) penalized using a model space covariance similar to that described by Siripunvaraporn & Egbert (2000). See the Appendix for further details.

The test case I (Fig. 3a) is fairly simple, consisting of a series of blocks (three relatively conductive, one resistive) buried in a 100 ohm-m half-space. TE and TM mode data were generated for $N_s = 40$ sites evenly spaced between $-30 \leq y \leq 30$ km, at $N_f = 16$ frequencies logarithmically spaced between 0.00033 and 3.3 Hz. We thus have a total of 1280 complex (2560 real) synthetic observations, to which we add 5 per cent random noise. Results of applying the data space Occam inversion scheme, using a 100 ohm-m half-space as a prior (and starting) model, are shown in Fig. 3(b). The algorithm converges in four outer loop iterations, and structures in the synthetic model are recovered accurately. Trade-off curves, showing misfit as a function of the regularization parameter λ used in (5), are shown for each of the four outer-loop Occam iterations in Fig. 4(a). Note that the minimum in the trade-off curve occurs because of nonlinearity of the inverse problem; for a linear problem the misfit would converge to zero as λ is reduced.

Test case II (Fig. 5a) presents greater challenges, with a more complex pattern of near-surface heterogeneity, and more spatially extensive deep conductivity variations. Synthetic data for this model were generated in the same way as for case one ($N_s = 40$, $N_f = 16$, 5 per cent noise added). Starting from the 100 ohm-m prior the initial misfit is much greater (650 vs. 23.5 normalized rms), and the Occam scheme does not quite achieve the target misfit, stalling with a normalized rms of 2.4 after eight iterations (Fig. 6a). The model achieving this misfit is shown in Fig. 5(b). Many features are recovered (e.g. the alternating pattern of conductive and resistive near-surface blocks, the deep vertical conductor in the middle of

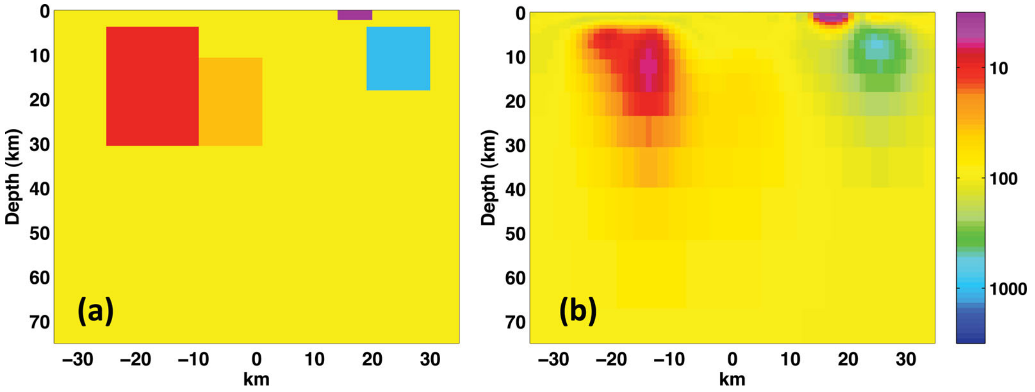


Figure 3. Synthetic 2-D MT test case I. (a) Resistivity model used to generate synthetic data. (b) Resistivity model recovered by Occam algorithm, based on full Jacobian calculation. Results obtained with the hybrid and modified hybrid schemes are indistinguishable, and are not plotted.

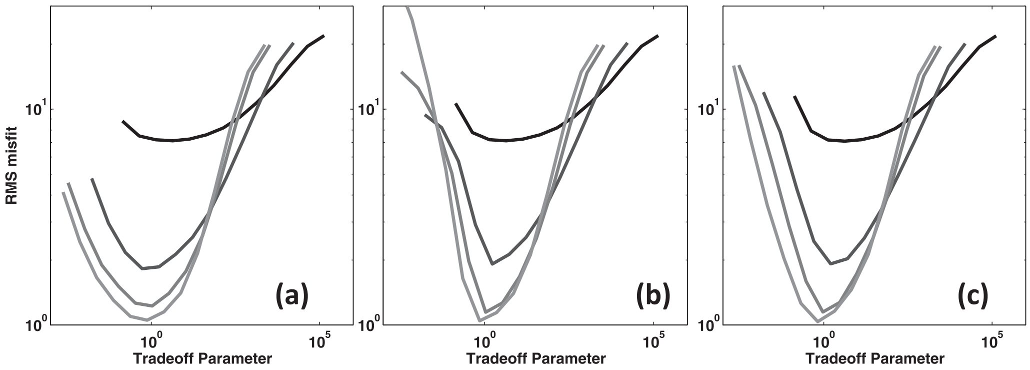


Figure 4. Trade-off curves for test case I, for (a) full Occam scheme; (b) hybrid scheme; (c) modified hybrid scheme.

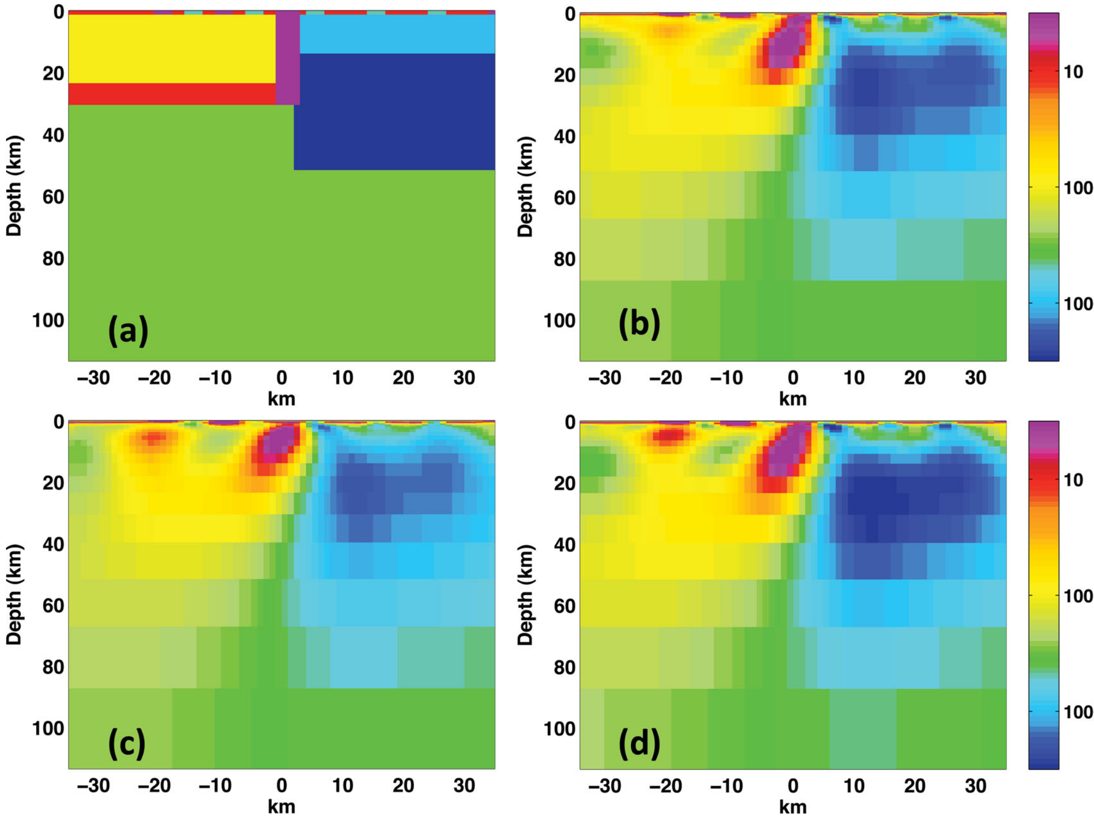


Figure 5. Synthetic 2-D MT test case II. Resistivity models (a) used to generate synthetic data; (b) recovered by Occam algorithm; (c) with the Hybrid scheme, and (d) with modified hybrid scheme. Essentially the same solution is recovered in all cases.

the domain, lateral variations in deep resistivity), although in detail the result deviates somewhat from the model used to generate the data,

For the hybrid Occam scheme we terminated the inner-loop (BIDIAG1) algorithm when the relative error in the solution to (5) (i.e. $\|(\mathbf{J}\mathbf{J}^{\mathrm{T}} + \lambda\mathbf{I})\mathbf{b} - \hat{\mathbf{d}}\|/\|\hat{\mathbf{d}}\|$ dropped below $\varepsilon = 10^{-2}$, or the number of iterations exceeded $K_{\max} = 30$. Using these convergence criteria the hybrid scheme reproduced results obtained with the standard Occam scheme based on the full Jacobian for both test cases. The final hybrid-scheme solution for case II (also fitting to a normalized rms of 2.4) is shown in Fig. 5(c). For case I results from the hybrid scheme are indistinguishable from the full Occam solution, and are not shown. Trade-off curves for the hybrid scheme are shown in Figs 4(b) and 6(b) for the two synthetic test cases. The behaviour as a function of iteration is very similar to that obtained with the full Jacobian, though the minima of the trade-off curves become somewhat narrower with the hybrid scheme.

For each iteration the hybrid solution is constructed as a linear combination of the model space vectors $\mathbf{J}^{\mathrm{T}}\mathbf{u}_k$, $k = 1, \ldots, K$. The first three of these, computed for the first iteration of test case I (i.e. with the Jacobian calculated for a 100 ohm-m half-space) are plotted in Figs 7(a)–(c). Note this Jacobian depends only on the uniform distribution of sites (and the frequencies), and the spatial patterns that dominate the basis functions are determined by the data (which determine the data-space vectors $\mathbf{u}_k$). In particular, the large positive feature in Fig. 7(a) coincides with the near-surface conductor between kilometres 10 and 20 (Fig. 1a), which has severely distorted the synthetic data from nearby sites.

The basis for the modified hybrid scheme is illustrated through the two lower rows of Fig. 7, where some of the individual transmitter

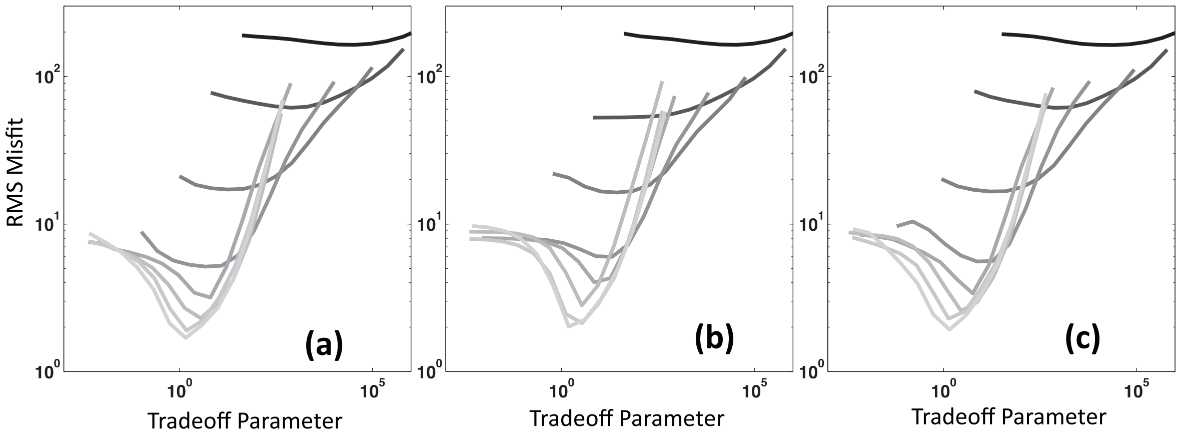


Figure 6. Trade-off curves for test case II, for (a) full Occam scheme; (b) Hybrid scheme; (c) Modified hybrid scheme.

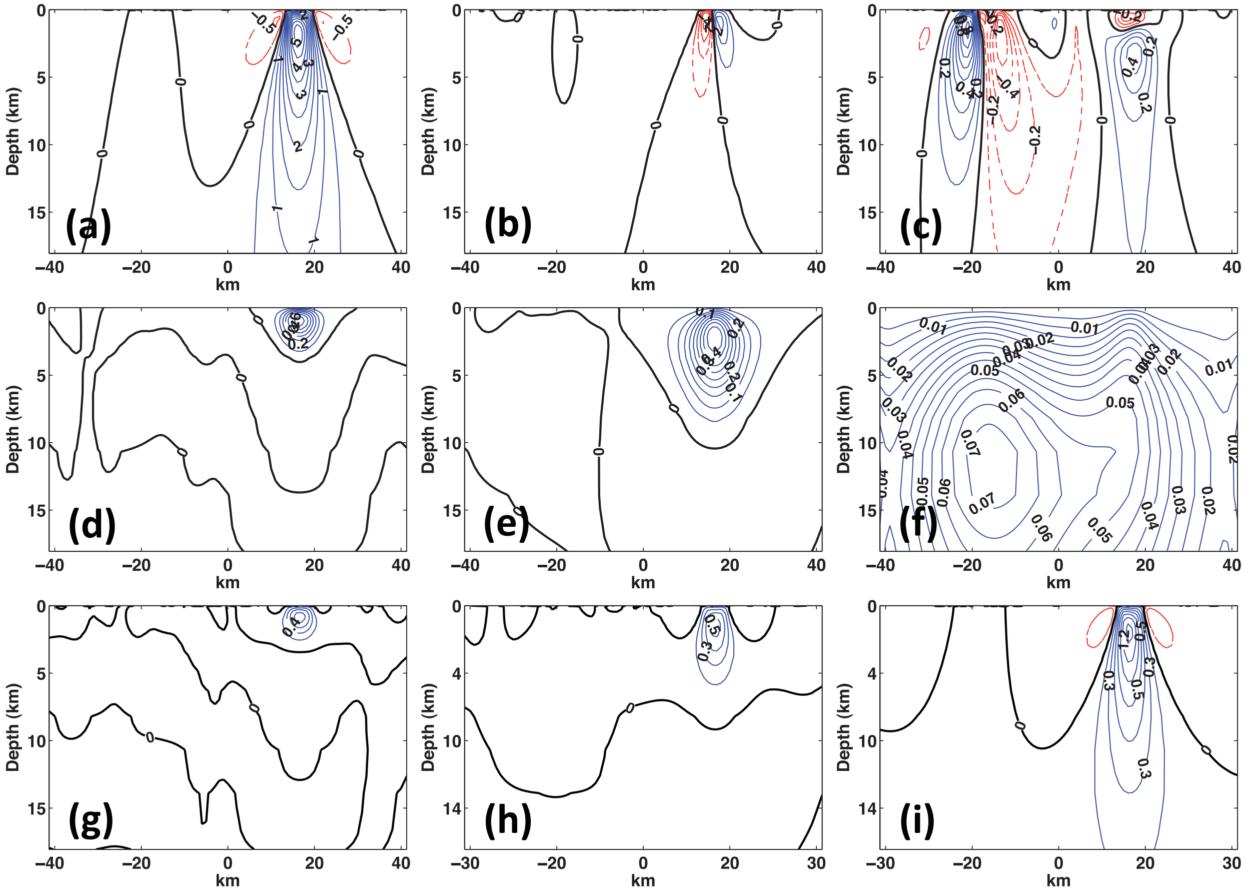


Figure 7. Sensitivity components $\mathbf{J}^{\mathrm{T}}\mathbf{u}$ generated by the hybrid and modified hybrid schemes. (a–c) First three model space vectors generated by BIDIAG1 on the first (outer loop) iteration for test case I (i.e. with the Jacobian calculated for a 100 ohm-m half-space). In the lower two rows selected individual transmitter component sensitivities (again for the first iteration test case) are plotted for three frequencies (3.3, 0.5, 0.02 Hz) for (d–f) the TE mode and (g–i) the TM mode.

component sensitivities are plotted. More specifically, the model space vector derived at the first step (plotted for test case I in Fig. 7a) can be expanded

$$\mathbf{J}^{\mathrm{T}}\mathbf{u}_1 = \sum_{j=1}^{J} \mathbf{J}_j^{\mathrm{T}}\mathbf{u}_{1j}, \tag{24}$$

where the index $j = 1, \ldots, J$ indicates transmitter number. For our examples, with TE and TM mode data for 16 periods, the total number of transmitters is $J = 32$. Component sensitivities for three frequencies (3.3, 0.5, 0.02 Hz) are plotted for the TE mode in Figs 7(d)–(f). Sensitivities for the same three frequencies for the TM mode are shown in Figs 7(g)–(i). Note that the sensitivities generally vary fairly smoothly with frequency. The lowest frequency TM mode sensitivities (e.g. Fig. 7i) are very similar by themselves to the sum of (24). Evidently, fitting the large static shifts associated with the near-surface conductor is the first priority in the iterative CG solution. Other model features are evident in the larger set of basis functions available to the multitransmitter scheme. In particular the large conductive block on the left side of the model at 5–30 km depth (Fig. 1a), corresponds to a clear peak in the same area in the longest period TE mode sensitivity (Fig. 7f). We anticipate that the additional model space basis functions will allow better approximation of the solution after fewer inner-loop steps.

This expectation is confirmed in Fig. 8, where we plot convergence to the solution of (5) for the inner loop of the standard and modified CG Occam schemes. For both test cases I and II the modified scheme converges more rapidly, with comparable reduction in the normal equation residual in roughly half the number of iterations required of the standard CG scheme. Convergence of the outer loop of the modified hybrid Occam scheme remains comparable to the standard data space Occam implementation based on the full Jacobian (Figs 4c and 6c). Final model results are also virtually identical for both case I (not shown) and case II (Fig. 5d).

It is well know that numerical round-off causes orthogonality of the sequence of vectors $\mathbf{u}_k, \mathbf{v}_k$ generated by the Lanczos process to break down as k increases, degrading convergence of the CG solver (e.g. Gollub & Van Loan 1989). The modified hybrid scheme explicitly enforces orthogonality of the sequence $\tilde{\mathbf{u}}_k$, and perhaps this is at least in part responsible for the more rapid convergence seen in Fig. 8. To test this we repeat the Lanczos bi-diagonalization, modified so that the sequence $\mathbf{u}_k, k = 1, \ldots, K$ remains exactly orthonormal, as in, for example, the generalized conjugate residual scheme of Eisenstat *et al.* (1983). For both cases I and II convergence of the CG scheme with explicit orthogonalization at each step shows significant improvement, but is still significantly slower compared to the modified scheme of Section 3 (Fig. 9).

In Fig. 10 we further compare the convergence behaviour of the hybrid schemes for a set of eight test cases (including cases I and II). Here we plot the number of inner-loop iterations required for each outer-loop step in the Occam scheme for the Lanczos bi-diagonalization with explicit orthogonalization, and for the modified multitransmitter scheme. Black filled symbols are used for the first scheme, and grey open symbols for the modified scheme,

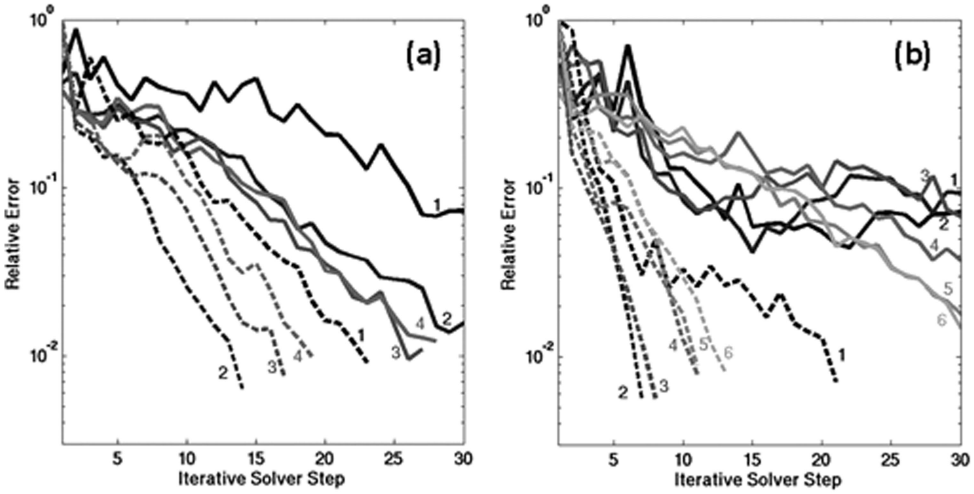


Figure 8. Convergence of the solution to the linear subproblem (5), for the hybrid scheme of Section 2 (solid lines) and for the modified hybrid scheme of Section 3 (dashed lines). Different line shadings correspond to different outer-loop Occam iterations, which are numbered near the end of each curve. Panels (a) and (b) give results for test cases I and II, respectively. In both cases the iterative solution is terminated when the relative error in the solution to eq. (5a; defined as $\|(\mathbf{JJ}^{\mathrm{T}} + \lambda_0\mathbf{I})\mathbf{b}_k - \hat{\mathbf{d}}\|/\|\hat{\mathbf{d}}\|$) drops below 10^{-2}, or k exceeds 30. In all cases the multifrequency scheme converges significantly faster than the standard CG iterative solution.

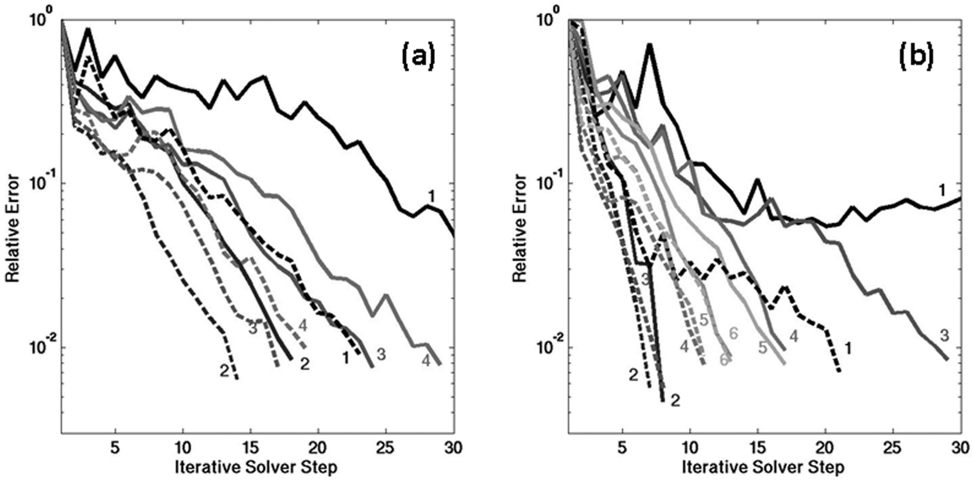


Figure 9. As in Fig. 8, but using a modified Lanczos bi-diagonalization scheme with explicit orthogonalization of all saved data-space vectors $\mathbf{u}_k$ for the standard hybrid scheme (solid lines). Dashed lines are as in Fig. 8. Explicit orthogonalization improves the convergence, although the multifrequency scheme still performs better.

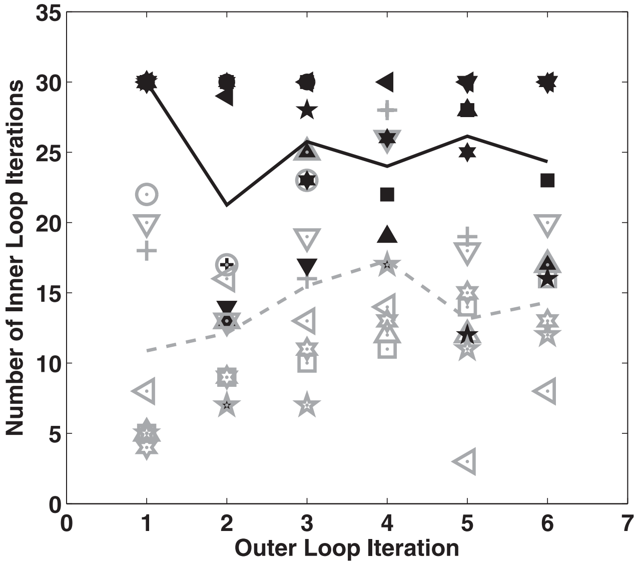


Figure 10. Total number of inner-loop iterations required for each outer-loop step in the Occam scheme for eight different synthetic model test cases. Different symbol styles are used for each case, with black filled and grey open symbols used for the standard (with explicit orthogonalization) and modified hybrid schemes, respectively. The lines give the averages (over test cases) for each outer-loop iteration: solid black line denotes standard hybrid scheme, dashed grey denotes modified scheme.

with different symbol styles used for each of the different synthetic model tests. The lines give the averages (over test cases) for each of the outer-loop iterations. The greatest increase in efficiency for the modified approach occurs on the first iteration, where the average number of steps decreases from 30 to 11. More modest, but still significant, improvement is seen for later iterations. Overall, the modified multitransmitter scheme reduces the total number of inner-loop iterations by a bit less than half, compared to Lanczos bi-diagonalization with explicit re-orthogonalization.

5 DISCUSSION AND CONCLUSIONS

We have discussed two hybrid schemes, which approximate the Occam scheme almost exactly without full calculation of the forward data mapping Jacobian. Both are based on the observation that iterative solution of the symmetric normal equations in the Gauss-Newton scheme effectively generates a sequence of sensitivities for different linear combinations of data, allowing construction of the Jacobian for a projection of the full data space. The Occam scheme can then be applied to this projected problem, with trade-off parameters chosen by assessing fit to the full data set. For EM geophysical problems with multiple transmitters (either multiple frequencies or source geometries) multiple forward solutions are required for a search step in the Lanczos process. Each of these solutions generates the sensitivity for a linear combination of data from the corresponding transmitter. From the perspective of the hybrid approach, with the Lanczos process generating an approximation to the full Jacobian, it is advantageous to save all of the component sensitivities, and use these to solve the projected problem in a larger subspace. This forms the basis for our second scheme, the modified hybrid algorithm.

Compared to standard CG schemes the proposed hybrid methods require substantially more storage, as the full sequence of data and model space vectors generated by the Lanczos process must be saved ($K(M + N)$ real numbers). For the modified approach storage requirements are even greater, as separate model space vectors are saved for each transmitter and each step in the solution process ($KJM + KN$ real numbers). However, as long as $KJ \ll N$ the additional memory required even for the modified scheme will be small compared to the MN real numbers required for storage of the full Jacobian. For our examples we have $KJ \approx 500$ while $N = 2560$.

Note also that a key component of the modified hybrid scheme is to explicitly solve the normal equations for the projected problem at each step in a modified Lanczos process, construct a 'trial' solution $\tilde{\mathbf{m}}$, and then apply the Jacobian $\mathbf{J}$ to this solution (i.e. compute $\mathbf{J}_j\tilde{\mathbf{m}}$, $j = 1, \ldots, J$) to generate the next set of data-space search vectors. Thus, additional computation is also required with the modified scheme [to compute $\tilde{\mathbf{m}}$; the equivalent multiplication by $\mathbf{J}$ is already required for the Lanczos process, e.g. in (8a)]. However, the projected system of normal eqs (12) or (20) will generally be small enough to be solved very rapidly—the largest system in our test cases was about 500 × 500, and the size of this system would not change significantly for a large 3-D inverse problem. Even so, this extra computation probably only makes sense when a single vector matrix multiply such as $\mathbf{J}^{\mathrm{T}}\mathbf{r}$ is sufficiently expensive, as it would be for something like the 3-D-MT inverse problem, where this single multiplication represents solving J independent 3-D PDEs. Indeed, for the 2-D-MT example we have used for illustration, solution of forward problems is sufficiently fast that justification for the modified scheme is at best marginal. Note also that one could apply the modified algorithm of Fig. 2 to any matrix $\mathbf{J}$, artificially divided into row blocks. However the extra computations required for each step of this scheme would in general overwhelm any saving due to reduction in the number of Lanczos steps that could be achieved.

The hybrid schemes described here are likely to be especially useful for joint inversion, for example, of MT and controlled source EM (e.g. Commer & Newman 2009), or EM and seismic traveltime data (e.g. Gallardo & Meju 2007). In the first place, multiple data types require running multiple forward models, and this can also be exploited within the framework developed here (as it was in the 2-D MT example, where TE and TM model solutions are computed). Furthermore, experience inverting multiple data types (e.g. Commer & Newman 2009) demonstrates that multiple trade-off parameters may be required to allow for differential weighting of disparate data types. And, one approach to joint inversion is to enforce structural similarity between two or more distinct physical parameters (e.g. conductivity and seismic velocity) by minimizing the norm of parameter gradient cross products (Gallardo & Meju 2004). Structural similarity defined in this way can be enforced by introducing another term into the penalty functional (1), with yet another adjustable weight. Efficient schemes for choosing these weights, as may be offered by hybrid schemes, are thus likely to prove valuable for joint inversion.

There are a number of potential extensions and refinements of the ideas presented here. First, we have focused on basic ideas, ignoring details that might make the schemes numerically more stable or efficient. For example, the cross product matrices in the projected normal eqs (12) and (20) need not be formed explicitly. Instead the singular value decomposition of the projected sensitivity matrix (e.g. $\mathbf{W}_K^{\mathrm{T}}\mathbf{J}$) could be used, both for efficient and stable solution of the normal equations, and to reduce storage requirements. And with the projected Jacobian saved, forming an approximation to the linearized resolution matrix would be straightforward (e.g. Minkoff 1996). In this application the additional sensitivity vectors provided by the modified scheme would improve the approximation of the resolution matrix (e.g. see discussion on approximations to the resolution matrix in Deal & Nolet 1996). Another possible extension worth exploring would be to use the approximated Jacobian computed in a hybrid scheme as a preconditioner for the next outer loop iteration of the Occam inversion scheme.

Finally, we have focused on making the data-space Occam scheme efficient for even very large problems. The basic idea behind this scheme could be adapted to a more general truncated Gauss–Newton scheme. More generally, it would be worth considering how (or if) the individual transmitter gradient components generated in each evaluation of the penalty functional gradient might be used in other search algorithms such as NLCG or quasi-Newton.

ACKNOWLEDGMENTS

This work was supported by DOE grant DE-FG0302ER15318. I thank Weerachai Siripunvaraporn and an anonymous reviewer for comments that helped improved clarity of the manuscript.

REFERENCES

Alumbaugh, D.L. & Newman, G.A., 1997. Three-dimensional massively parallel electromagnetic inversion: II. Analysis of a crosswell electromagnetic experiment, *Geophys. J. Int.,* **128,** 355–363, doi:10.1111/j.1365246X.1997.tb01560.x.

Avdeev, D., 2005. Three-dimensional electromagnetic modeling and inversion from theory to application, *Surv. Geophys.,* **26,** 767–799.

Avdeev, D. & Avdeeva, A., 2009. 3D Magnetotelluric inversion using a limited-memory quasi-Newton optimization, *Geophysics,* **74,** F45–F57.

Commer, M. & Newman, G.A., 2008. New advances in three-dimensional controlled-source electromagnetic inversion, *Geophys. J. Int.,* **172,** 513–535.

Commer, M. & Newman, G.A., 2009. Three-dimensional controlled-source electromagnetic and magnetotelluric joint inversion, *Geophys. J. Int.,* **178,** 1305–1316, doi:10.1111/j.1365-246X.2009.04216.x.

Constable, C.S., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics,* **52,** 289–300.

Deal, M.M. & Nolet, G., 1996. Comment on 'Estimation of resolution and covariance for large matrix inversions' by J. Zhang and G. A. McMechan, *Geophys. J. Int.,* **127,** 245–250, doi:10.1111/j.1365-246X.1996.tb01548.x.

Dembo, R.S., Eisenstat, S.C. & Steihaug, T., 1982. Inexact Newton methods, *SIAM J. Numer. Anal.,* **19,** 400–408.

Egbert, G.D. & Kelbert, A., 2012. Computational recipes for electromagnetic inverse problems, *Geophys. J. Int.,* **189,** 251–267, doi:10.1111/j.1365-246X.2011.05347.x.

Egbert, G.D., Bennett, A.F. & Foreman, M.G.G., 1994. TOPEX/POSEIDON Tides estimated using a global inverse model, *J. geophys. Res.,* **99,** 24 821–24 852.

Eisenstat, S.C., Elman, H.C. & Schultz, M.H., 1983. Variational iterative methods for nonsymmetric systems of linear equations, *SIAM J. Numer. Anal.,* **20,** 345–357.

Gallardo, L.A. & Meju, M.A., 2004. Joint two-dimensional inversion with cross-gradient constraints, *J. geophys. Res.,* **109,** B03311, doi:10.1029/2003JB002716.

Gallardo, L.A. & Meju, M.A., 2007. Joint two-dimensional cross-gradient imaging of magnetotelluric and seismic traveltime data for structural and lithologic classification, *Geophys. J. Int.,* **169,** 1261–1272.

Gollub, G.H. & Van Loan, C., 1989. *Matrix Computations,* 2nd edn, Johns Hopkins University Press, Baltimore, MD.

Hanke, M., 2001. On Lanczos based methods for the regularization of discrete ill-posed problems, *BIT,* **41,** 1008–1018.

Kilmer, M.E. & O'Leary, D.P., 2001. Choosing regularization parameters in iterative methods for ill-posed problems, *SIAM J. Matrix Anal. Appl.,* **22,** 1204–1221.

Liu, D.C. & Nocedal, J., 1989. On the limited memory method for large scale optimization, *Math. Prog. B,* **45,** 503–528.

Mackie, R.L. & Madden, T.R., 1993. Conjugate direction relaxation solutions for 3-D magnetotelluric modeling, *Geophysics,* **58,** 1052–1057.

Minkoff, S.E., 1996. A computationally feasible approximate resolution matrix for seismic inverse problems, *Geophys. J. Int.,* **126,** 345–359.

Nash, S.G., 2000. A survey of truncated-Newton methods, *J. Comput. appl. Math.,* 45–59.

Newman, G.A. & Hoversten, G.M., 2000. Solution strategies for two- and three-dimensional electromagnetic inverse problems, *Inverse Probl.,* **16,** 1357, doi:10.1088/0266-5611/16/5/314.

O'Leary, D.P. & Simmons, J.A., 1981. A bidagonalization-regularization procedure for large scale discretization of ill-posed problems, *SIAM J. Sci. Statist. Comput.,* **2,** 474–489.

Oldenburg, D.W., McGillivray, P.R. & Ellis, R.G., 1993. Generalized subspace methods for large-scale inverse problems, *Geophys. J. Int.,* **114,** 12–20, doi:10.1111/j.1365-246X.1993.tb01462.x.

Paige, C.C. & Saunders, M.A., 1982a. LSQR: An algorithm for sparse linear equations and sparse least squares, *ACM Trans. Math. Softw.,* **8,** 43–71.

Paige, C.C. & Saunders, M.A., 1982b. Algorithm 583 LSQR: sparse linear equations and least squares problems, *ACM Trans. Math. Softw.,* **8,** 195–209.

Pankratov, O. & Kuvshinov, A., 2010. General formalism for the efficient calculation of derivatives of EM frequency-domain responses and derivatives of the misfit, *Geophys. J. Int.,* **181,** 229–249.

Parker, R.L., 1994. *Geophysical Inverse Theory,* Princeton University Press, Princeton, NJ.

Purser, R.J., Wu, W.-S., Parrish, D.F. & Roberts, N.M., 2003a. Numerical aspects of the application of recursive filters to variational statistical analysis. part I: spatially homogeneous and isotropic Gaussian covariances, *Monthly Wea. Rev.,* **131,** 1524–1535.

Purser, R.J., Wu, W.-S., Parrish, D.F. & Roberts, N.M., 2003b. Numerical aspects of the application of recursive filters to variational statistical analysis. part II: spatially inhomogeneous and anisotropic general covariances, *Monthly Wea. Rev.,* **131,** 1536–1548.

Rodi, W.L. & Mackie, R.L., 2001. Nonlinear conjugate gradients algorithm for 2-D Magnetotelluric inversion, *Geophysics,* **66,** 174–187.

Sasaki, Y., 2001. Full 3-D inversion of electromagnetic data on PC, *J. appl. Geophys.,* **46,** 45–54.

Siripunvaraporn, W., 2012. Three-dimensional magnetotelluric inversion: an introductory guide for developers and users, *Surv. Geophys.,* **33,** 5–27, doi:10.1007/s10712-011-9122-6.

Siripunvaraporn, W. & Egbert, G., 2000. An efficient data-subspace inversion method for 2-D magnetotelluric data, *Geophysics,* **65,** 791–803.

Siripunvaraporn, W. & Egbert, G.D., 2007. Data space conjugate gradient inversion for 2-D magnetotelluric data, *Geophys. J. Int.,* **170,** 986–994.

Siripunvaraporn, W. & Egbert, G., 2009. WSINV3DMT: Vertical magnetic field transfer function inversion and parallel implementation, *Phys. Earth planet. Inter.,* **173,** 317–329.

Siripunvaraporn, W. & Sarakorn, W., 2011. An efficient data space conjugate gradient Occam's method for three-dimensional magnetotelluric inversion, *Geophys. J. Int.,* **186,** 567–579, doi:10.1111/j.1365-246X.2011.05079.x.

Siripunvaraporn, W., Egbert, G., Lenburi, Y. & Uyeshima, M., 2005. Three-dimensional magnetotelluric inversion: data space method, *Phys. Earth planet. Inter.,* **140,** 3–14.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2010. Seismic tomography of the southern California crust based on spectral-element and adjoint methods, *Geophys. J. Int.,* **180,** 433–462, doi:10.1111/j.1365-246X.2009.04429.x.

APPENDIX: INCLUDING MODEL AND DATA COVARIANCES

Here we briefly sketch treatment of the general form of the penalty functional (1) where model and data covariances are not the identity. We follow the approach used by Siripunvaraporn & Egbert (2000) where the model covariance $\mathbf{C_m} = \mathbf{C_m^{1/2}} \mathbf{C_m^{1/2}}$ is implemented as a positive definite symmetric smoothing operator; applying half the smoothing steps essentially provides the square root of the operator. Examples of such covariance operators are given in Egbert *et al.* (1994), Siripunvaraporn & Egbert (2000), and Purser *et al.* (2003a,b). Defining a transformed model parameter $\tilde{\mathbf{m}}$ implicitly through

$$\mathbf{m} = \mathbf{C}_{\mathbf{m}}^{1/2}\tilde{\mathbf{m}} + \mathbf{m}_0, \tag{A1}$$

and transforming the data vector in the usual way as $\tilde{\mathbf{d}} = \mathbf{C}_{\mathbf{d}}^{-1/2}\mathbf{d} = \mathbf{C}_{\mathbf{d}}^{-1/2}\mathbf{f} + \mathbf{C}_{\mathbf{d}}^{-1/2}\varepsilon$ (so that the data error covariance is the identity), the Jacobian for the transformed problem can be written as

$$\tilde{\mathbf{J}} = \frac{\partial\tilde{\mathbf{f}}}{\partial\tilde{\mathbf{m}}} = \mathbf{C}_{\mathbf{d}}^{-1/2}\frac{\partial\mathbf{f}}{\partial\mathbf{m}}\mathbf{C}_{\mathbf{m}}^{1/2} = \mathbf{C}_{d}^{-1/2}\mathbf{J}\mathbf{C}_{\mathbf{m}}^{1/2}, \tag{A2}$$

where $\mathbf{J}$ is the original sensitivity for the untransformed problem. Dropping the tildes the penalty functional (1) is reduced to the simpler form used throughout the paper, and the methods described can be applied to invert the transformed data for the transformed model parameter $\tilde{\mathbf{m}}$. This can be converted back to the physical model parameter using (A1). Note that only multiplication by the model covariance operator $\mathbf{C}_{\mathbf{m}}^{1/2}$ and the inverse data error covariance square root $\mathbf{C}_{\mathbf{d}}^{-1/2}$ are required; the inverse model covariance is never directly used. Although we have focused on data space solution methods, the same approach can be used for model space solution approaches, such as NLCG—that is, the penalty functional can be minimized with respect to $\tilde{\mathbf{m}}$, with the gradient derived from the transformed Jacobian $\tilde{\mathbf{J}}$.

The principal limitation of the approach described here is that multiplication by $\mathbf{C}_{\mathbf{m}}^{1/2}$ must be implemented, and for some classical regularization operators this may not be so straightforward. For example, if the regularization term is taken to be $\|\nabla^2\mathbf{m}\|^2$, applying the smoothing operator $\mathbf{C}_{\mathbf{m}}^{1/2}$ amounts to solving Poisson's equation; boundary conditions are a complicating, although not insurmountable, issue. Such details are beyond the scope of this paper.

Geology

Crust and upper mantle electrical conductivity beneath the Yellowstone Hotspot Track

A. Kelbert, G. D. Egbert and C. deGroot-Hedlin

Geology 2012;40;447-450
doi: 10.1130/G32655.1

Notes

Crust and upper mantle electrical conductivity beneath the Yellowstone Hotspot Track

A. Kelbert[1], G. D. Egbert[1], and C. deGroot-Hedlin[2]
[1]College of Earth, Ocean and Atmospheric Sciences, Oregon State University, 104 CEOAS Admin Building, Corvallis, Oregon 97331, USA
[2]Scripps Institution of Oceanography, University of California–San Diego, La Jolla, California 92037, USA

ABSTRACT

Combining long-period magnetotelluric data from the spatially uniform EarthScope USArray and higher-resolution profiles, we obtain a regional three-dimensional electrical resistivity model in the Snake River Plain and Yellowstone areas (Idaho and Wyoming, United States), and provide new constraints on the large-scale distribution of melt and fluids beneath the Yellowstone hotspot track. Contrary to what would be expected from standard mantle plume models, the electromagnetic data suggest that there is little or no melt in the lower crust and upper mantle directly beneath Yellowstone caldera. Instead, low mantle resistivities (10 Ωm and below), which we infer to result from 1%–3% partial melt, are found 40–80 km beneath the eastern Snake River Plain, extending at least 200 km southwest of the caldera, beneath the area of modern basaltic magmatism. The reduced resistivities extend upward into the mid-crust primarily around the edges of the Snake River Plain, suggesting upward migration of melt and/or fluid is concentrated in these areas. The anomaly also shallows toward Yellowstone, where higher temperatures enhance permeability and allow melts to ascend into the crust. The top of the conductive layer is at its shallowest, in the upper crust, directly beneath the modern Yellowstone supervolcano.

INTRODUCTION

The Snake River Plain–Yellowstone (Idaho and Wyoming, United States) volcanic province has long been associated with a stationary deep mantle plume source (e.g., Hadley et al., 1976; Geist and Richards, 1993). However, this simple model is difficult to reconcile with at least some important observations of the system, including the temporal persistence of basaltic volcanism and geochemistry of erupted magmas, and the spatial and temporal relationship of magmatism in the Snake River Plain (SRP), Yellowstone, and High Lava Plains, leading some to emphasize the role of shallower lithospheric convection (e.g., Humphreys et al., 2000; Christiansen et al., 2002; Leeman et al., 2009).

While body-wave tomography images for this area are in broad agreement, they have been interpreted to support both plume and "no plume" hypotheses. Early studies based on local arrays suggested a continuous low-velocity plume beneath Yellowstone, dipping to the northwest and extending to at least the transition zone (e.g., Yuan and Dueker, 2005). Resolution has been greatly improved with the deployment of the USArray (a component of the EarthScope project), revealing that this low-velocity feature extends into the lower mantle, but is discontinuous (e.g., Tian et al., 2011; James et al., 2011). USArray seismic data also provide a regional context for the Yellowstone low-velocity anomaly. Several studies have identified a broad volume devoid of fast anomalies extending through the transition zone into the uppermost lower mantle: a "slab gap" between segments of the subducting Juan de Fuca plate (Tian et al., 2011). Relatively low velocities within the gap have been interpreted as evidence for interaction with a deeper plume source (e.g., Obrebski et al., 2010; Tian et al., 2011), or mantle upwelling through the gap in response to a sinking slab segment (James et al., 2011).

Surface wave inversions of USArray data show fast anomalies beneath the eastern SRP and Yellowstone at mid- to lower crustal depths (except directly beneath and near the Yellowstone caldera), but a very pronounced low-velocity anomaly in the mantle between the Moho and 200 km depth, extending southwest from Yellowstone caldera in the direction of North American plate motion (Obrebski et al., 2010; Wagner et al., 2010; Gao et al., 2011; Yang et al., 2011).

MAGNETOTELLURIC DATA

Long-period magnetotelluric (MT) data have also been collected as part of USArray project, using the same 70 km site spacing as the seismic component. These data are highly sensitive to the presence of volatiles and partial melt, and thus offer potentially valuable additional constraints on the physical state of the crust and mantle in this tectonically and magmatically active area. Here we use recently developed three-dimensional (3-D) inversion methods to interpret long-period MT data from 91 USArray MT sites, covering much of Idaho and Wyoming, southern Montana, eastern Oregon, and northern Nevada, together with 32 sites from an earlier MT survey, collected in two denser profiles along (~40 km site spacing) and across (~10 km site spacing) the eastern SRP (see Fig. 1).

For 3-D inversion, we employed the Modular system for Electromagnetic Inversion (ModEM; Egbert and Kelbert, 2012), a flexible system for regularized inversion of electromagnetic data. In our application to the MT data, we regularized with a model covariance that penalizes deviations from a prior model, fitting all six MT data components from 123 sites, at 14 periods from 7.3 s to 5.2 h. Poor-quality data (~1%) were removed from the data set, and an error floor

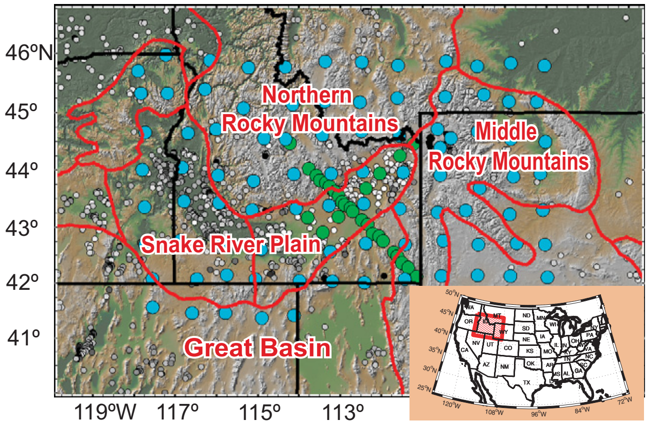


Figure 1. Topography of study area (see inset map for location within the United States), with physiographic provinces outlined in red. USArray magnetotelluric (MT) site locations used for this study are marked with blue dots; 32 sites from the earlier Snake River Plain profiles are denoted by green dots. Smaller gray dots indicate heat flow from Pollack et al. (1991), ranging from 0 (white) to >300 mW/m² (black) .

[1]GSA Data Repository item 2012118, inversion procedure and resolution tests, is available online at www.geosociety.org/pubs/ft2012.htm, or on request from editing@geosociety.org or Documents Secretary, GSA, P.O. Box 9140, Boulder, CO 80301, USA.

Geology, May 2012; v. 40; no. 5; p. 447–450; doi:10.1130/G32655.1; 3 figures; Data Repository item 2012118.

of 5% was imposed. See the GSA Data Repository[1] for details of our inversion procedure.

Multiple inverse solutions were obtained at 10 km nominal resolution, using a range of prior one-dimensional models and varying degrees of smoothing. The preferred solution (model 1; Figs. 2 and 3) used a 200 Ωm half space as the prior, and fit the data to a normalized root mean square misfit of 1.89. We also discuss results from two alternative models (models 2 and 3, shown in the Data Repository) in the following section.

RESULTS

The most striking feature in all of the inverse solutions is a large, interconnected conductive body extending from the Yellowstone caldera at least 200 km to the southwest, roughly parallel to the direction of North America absolute motion (Figs. 2C and 3A). The depth to the top of this conductor varies from 30 to 60 km along the SRP, except in localized areas, including directly beneath Yellowstone caldera, where it reaches into the upper crust (Fig. 3A), and around the edges of the eastern SRP where it shallows to 18 km or so (Figs. 2A and 3B). The thickness of the most pronounced conductive area is 30–40 km, mostly in the uppermost mantle, and all within 80–100 km of the surface. To the east, the mantle is significantly more resistive, over 600 Ωm. At greater depths beneath the study area the upper mantle has moderately low resistivity (100 Ωm or less).

Crustal thickness beneath the eastern SRP and Yellowstone is inferred to be 40–50 km (Yuan et al., 2010), with the thickest crust directly beneath Yellowstone. Thinner crust surrounds the SRP, except to the east, beneath the Rocky Mountains. The vertically integrated conductivity (conductance; SI unit Siemens; S) of the lower crust (16–42 km) is highly variable in the inverse solution, ranging from ~80 to over 10,000 S beneath and around the SRP, and averaging ~1000 S. The average conductance at 42–80 km is over 3000 S beneath the SRP, reducing to 30–300 S in the Wyoming craton and directly beneath Yellowstone.

Depth resolution of the MT data is limited, both by the diffusive propagation of the electromagnetic fields in the conducting Earth, and by the distorting effects of near-surface heterogeneity. Indeed, assuming a more conductive mantle a priori results in conductive features that, while very similar in plan view, are shifted upward by up to 10 km (model 2, shown in Figs. DR1 and DR2 in the Data Repository). We thus must entertain the possibility that the low resistivities imaged at the top of the mantle in model 1 might actually be above the Moho. To test this, we ran the inversion using a prior model in which the crust was less resistive (60 Ωm) than the upper mantle (200 Ωm), thus pushing the low resistivities into the crust as much as possible while still fitting the data adequately. The resulting inverse solution (model 3; Figs. DR3 and DR4) is noticeably rougher and has excessively conductive crust, averaging ~3000 S below the SRP, with peak values exceeding 10,000 S. Such high conductivities are difficult to explain other than with free saline fluids distributed throughout the mid- to lower crust, an inference which is difficult to reconcile with seismic surface wave studies that reveal normal to fast lower crustal velocities in this area (Gao et al., 2011; Yang et

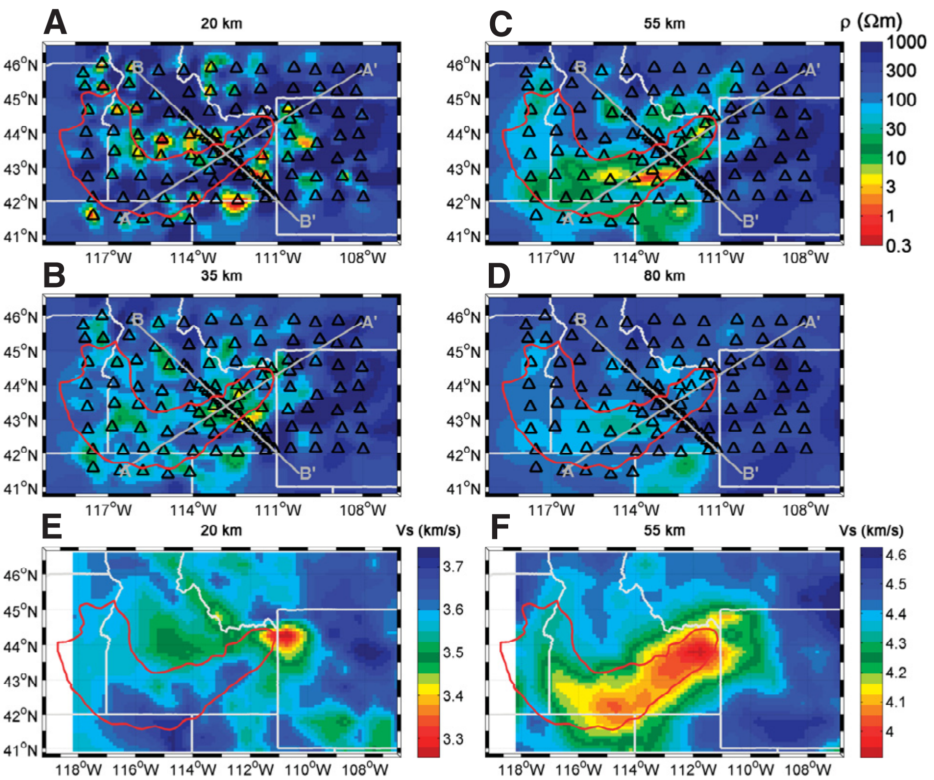


Figure 2. A-D: Preferred inverse model at representative depths. Lines indicate locations of profiles A–A′ and B–B′ shown in Figure 3. E-F: Seismic surface-wave velocity model of Yang et al. (2011) plotted at mid-crustal (20 km) and uppermost mantle (55 km) depths and interpolated to our grid for easier comparison. The contours of the Snake River Plain are plotted in red for reference.

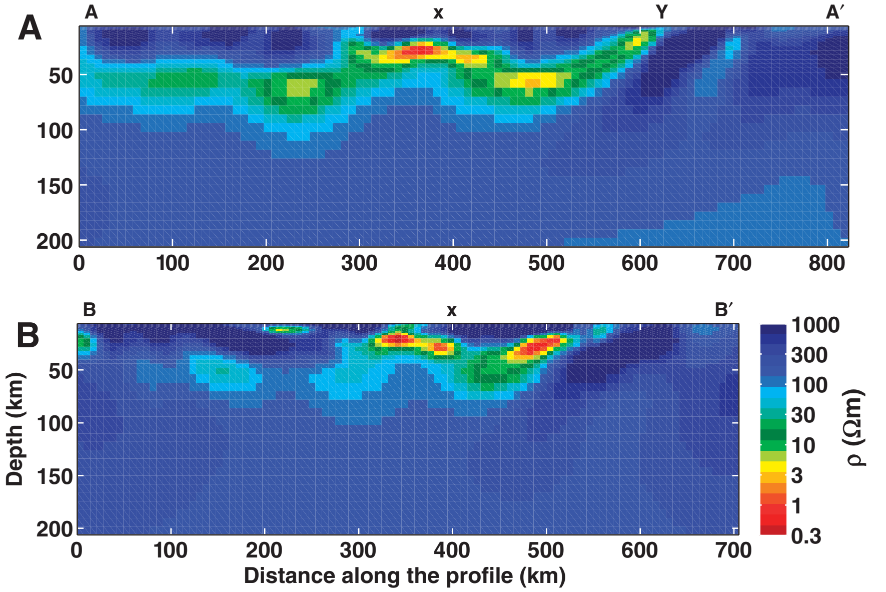


Figure 3. Cross sections from the preferred model along (A–A′) and across (B–B′) the eastern Snake River Plain. x—point of profile intersection; Y—Yellowstone caldera.

al., 2011; see also Fig. 2E). Furthermore, even with much of the anomaly pushed into the crust, model 3 is still anomalously conductive at the top of the mantle, with an average conductance of roughly 1000 S between 42 and 80 km depths beneath the SRP. We thus conclude that the MT data require elevated conductivities at the top of the upper mantle. Resistivities average no more than 40 Ωm beneath the SRP, and we consider the much lower values found in model 1 (Figs. 2 and 3) to be more likely.

In all inverse solutions, the upper mantle below ~100 km is more resistive beneath the Wyoming craton than beneath the SRP, but a range of resistivities, from 30 to over 100 Ωm, is recovered beneath the SRP depending on the prior model and regularization (Figs. DR3 and DR4). We conclude that deeper structures are shielded and confounded by the highly conductive and inhomogeneous crust and lithosphere, and thus concentrate our discussion on the lower crustal and uppermost mantle inhomogeneities that are resolved robustly.

DISCUSSION

Depth of Conductive Layers

High conductivities near the Moho have frequently been observed in the western United States, but these have most often been interpreted to be in the lower crust. For example, in the eastern Great Basin (Wannamaker et al., 2008) and the Pacific Northwest (Patro and Egbert, 2008), integrated lower crustal conductances of 3000 S or more, imaged by MT data, have been interpreted as saline fluids and partial melt associated with magmatic underplating. Indeed, elevated lower crustal conductivities for the eastern SRP were previously inferred by Stanley et al. (1977) from wide-band (0.001–500 s) MT profile data. Based on one-dimensional inversion of data from 12 sites, Stanley et al. (1977) inferred a low resistivity (~1 Ωm) layer with the top at ~7–9 km directly beneath the Yellowstone caldera system, deepening to ~25 km beneath the Island Park caldera, and then slightly shallowing further to the southeast. This is very similar to the shape of the anomaly we image, although our preferred solution (model 1; Figs. 2 and 3) locates the top of the conductive zone somewhat deeper.

The data of Stanley et al. (1977) were restricted to periods below 500 s, and could not image below the first highly conductive layer encountered in the crust. At the longer periods we have used (up to 20,000 s), the electromagnetic fields penetrate this layer, allowing resolution of deeper structure. However, as the variations between the three models discussed here demonstrate, the MT data by themselves do not always precisely constrain depths to specific features. This depth ambiguity results largely from the effects of near-surface conductive heterogeneity (Jones, 1988), which can distort electric fields, essentially multiplying impedances for each site by a different frequency-independent real factor. This translates into uncertainty in MT data amplitudes, which carry the information about depth and magnitude of a conductor. An extreme example of this ambiguity is perhaps provided by the recent study by Zhdanov et al. (2011), who inverted a subset of the USArray MT data considered here. Their interpretation emphasized a deep (~300 km), extremely conductive (1 Ωm or less) sub-horizontal mantle structure dipping to the southwest, with a footprint quite similar to the anomaly we image. However, Zhdanov et al. (2011) only fit phase data in their inversion, and these data provide little constraint on actual depths.

Correlation with the results from seismic imaging can reduce the depth uncertainties and allow us to choose among models that fit the MT data. The footprint of the conductive anomaly in all of models 1–3 coincides with that of the very prominent low in shear wave velocities inferred at the top of the mantle from surface wave tomography (Obrebski et al., 2010; Wagner et al., 2010; Gao et al., 2011; Yang et al., 2011; see also Fig. 2F). Although the low velocities extend to greater depth than the conductive anomaly seen in even model 1 (200 km versus 100 km), peaks in the seismic and conductivity anomalies are both between 40 and 80 km depth, suggesting a common physical explanation, partial melt.

Implications for Melt Porosity

Accounting for the uncertainties associated with model smoothing and with the near-surface heterogeneity distortions, resistivities of 10 Ωm or below over large areas in the uppermost mantle are robustly resolved by the MT data (see the Results section). Observed shear-wave velocity anomalies of 6%–8% (Wagner et al., 2010) in the uppermost mantle beneath the SRP suggest a melt porosity of 1%–2% (Hammond and Humphreys, 2000). Laboratory measurements of basaltic melt in an olivine matrix (Yoshino et al., 2010) suggest that ~1% melt results in bulk resistivities of 2–10 Ωm, when extrapolated to temperatures of 1350–1450 °C, appropriate for the SRP lithosphere (Leeman et al., 2009). These results are consistent with complete wetting of grain boundaries and a melt resistivity of roughly 0.1 Ωm (at 1350 °C). Other studies (Ni et al., 2011) have found somewhat higher values for resistivity of dry basaltic melts, 0.2–0.3 Ωm at 1350–1450 °C. Based on the Hashin-Shtrikman upper bound, this would require a melt fraction of ~3% for a bulk resistivity of 10 Ωm. Significantly lower resistivities are found just below the Moho (a few Ωm). Considering the possibility of imperfect melt connection, higher melt fractions, or some other explanation might be required. Although erupted SRP basalts are relatively dry (no more than ~1 wt% water; Leeman et al., 2009; Till et al., 2010), as little as 0.5 wt% water could reduce melt resistivity by a factor of two (Ni et al., 2011). Variations in composition could also increase the conductivity of the melt phase (Roberts and Tyburczy, 1999).

CONCLUSIONS

The MT and seismic results are consistent with the presence of a few percent partial melt between 40 and 80 km, depths that would normally be considered mantle lithosphere. The spatial coincidence of this region with the Yellowstone hotspot track suggests that passage of the North American plate over the plume has resulted in significant modification or thinning, perhaps leaving little or no lithospheric root beneath the eastern SRP. The presence of melt at the top of the upper mantle is consistent with the ongoing basaltic magmatism in the SRP, which has continued along the length of the hot-spot track since initial passage over the plume. Our images are also consistent with inferences from thermobarometry (Leeman et al., 2009) on shallow melt equilibration depths of 80–100 km or less, and with suggestions (Till et al., 2010) that similar basaltic magmas from the nearby High Lava Plains have equilibrated just below the Moho.

None of our inverse solutions show the SRP conductivity anomaly extending beneath Yellowstone at mantle depths. This stands in contrast to the seismic images (e.g., Wagner et al., 2010; Yang et al., 2011; see also Fig. 2F), which show substantial slow anomalies in the mantle immediately beneath Yellowstone. Both melt and high temperatures could contribute to these low seismic velocities, but the high resistivities rule out significant interconnection of any melt phase in the lithosphere directly beneath Yellowstone. Possibly the cratonic lithosphere in this area is largely intact, and still too thick to allow decompression melting (Leeman et al., 2009). In this scenario, the seismic anomaly would have a purely thermal explanation, with the lithosphere heated by a deep plume source. Alternatively, melt may be present in the lithosphere beneath Yellowstone, but at too low a concentration to be interconnected. Indeed, elevated temperatures beneath the active volcanic center would result in greater permeability, allowing magma to ascend to shallower depths and pool in the crust, instead of collecting in the mantle lithosphere, as beneath the SRP. We thus speculate that little melt is entering the system from below at present, perhaps due to intermittency of supply (as suggested by the apparent discontinuity with depth of the seismically imaged plume; e.g., James et al., 2011), while melt from earlier plume activity has mostly already ascended into the shallow crust, leaving behind only isolated pockets of residual melt in the mantle lithosphere and lower crust. These would be effective at reducing seismic shear wave velocities (in conjunction with elevated temperatures), but would not significantly reduce resistivity.

High conductivities occur at mid-crustal levels (e.g., Fig. 2A) almost exclusively around the edges of the SRP. In cross section, these shallower features connect to the deeper conductive structure below the Moho (Fig. 3B), much as the shallow Yellowstone caldera conductor dips to the southwest and connects into the deeper anomaly (Fig. 3A). This suggests that melt, and perhaps also fluids exsolved by magmatic underplating, ascend into the crust preferentially around the edges of the generally impermeable SRP. We note the coincidence of these mid-crustal low resistivities with the "tectonic parabola" (e.g., Humphreys et al., 2000) of late Cenozoic normal faults, which may help to provide a preferred pathway for fluid or melt migration. Note also that the highest heat flows in this region occur around the edges of the SRP, again coincident with the mid-crustal conductive anomalies (Pollack et al., 1991; see also Fig. 1), and that shear wave velocities are reduced in this area at 20 km depth (see Fig. 2F).

Finally, we emphasize that the MT data provide at best weak constraints on deeper structure. The vertical seismic anomaly inferred (Yuan and Dueker, 2005; Obrebski et al., 2010) to be the mantle plume would likely represent only a thermal anomaly of roughly 125–150 °C (Leeman et al., 2009), at depths below 100 km or so. This excess temperature would increase electrical conductivity of dry olivine, but only modestly compared to the effects of fluids and melt. Such a relatively subtle signal would be challenging to resolve given the highly variable features we image at shallower depths.

In summary, our conductivity images suggest a more complex pattern of melt beneath the SRP and Yellowstone than would be expected from a continuously supplied, classical mantle plume with head sheared to the southwest by North American plate motion. Collection of partial melts at the base of the SRP province, inferred from the MT data, can perhaps explain some of the distinct features of SRP and Yellowstone magmatism.

ACKNOWLEDGMENTS

Egbert acknowledges support from the U.S. Department of Energy grant DE-FG0302ER15318, and National Science Foundation grant EAR1053628. deGroot-Hedlin acknowledges support of National Science Foundation grant EAR0229814. Thanks to M. Moschetti, Y. Yang, and M. Ritzwoller for providing seismic velocity models, and Naser Meqbel, Anita Grunder, and William Leeman for helpful discussions.

REFERENCES CITED

Christiansen, R.L., Foulger, G.R., and Evans, J.R., 2002, Upper-mantle origin of the Yellowstone hotspot: Geological Society of America Bulletin, v. 114, p. 1245–1256, doi:10.1130/0016-7606(2002)114<1245:UMOOTY>2.0.CO;2.

Egbert, G. D., and Kelbert, A., 2012, Computational Recipes for Electromagnetic Inverse Problems: Geophysical Journal International, doi:10.1111/j.1365-246x.2011.05347.x.

Gao, H., Humphreys, E.D., Yao, H., and van der Hilst, R.D., 2011, Crust and lithosphere structure of the northwestern U.S. with ambient noise tomography: Terrane accretion and Cascade arc development: Earth and Planetary Science Letters, v. 304, p. 202–211, doi:10.1016/j.epsl.2011.01.033.

Geist, D., and Richards, M., 1993, Origin of the Columbia Plateau and Snake River plain: Deflection of the Yellowstone plume: Geology, 21, 789-+. doi:10.1130/0091-7613(1993)021<0789:OOTCPA>2.3.CO;2

Hadley, D.M., Stewart, G.S., and Ebel, J.E., 1976, Yellowstone: Seismic Evidence for a Chemical Mantle Plume: Science, v. 193, p. 1237–1239, doi:10.1126/science.193.4259.1237.

Hammond, W.C., and Humphreys, E.D., 2000, Upper mantle seismic wave velocity: Effects of realistic partial melt geometries: Journal of Geophysical Research, v. 105, p. 10,975–10,986, doi:10.1029/2000JB900041.

Humphreys, E.D., Dueker, K.G., Schutt, D.L., and Smith, R.B., 2000, Beneath Yellowstone: evaluating plume and nonplume models using teleseismic images of the upper mantle: GSA Today, v. 10, p. 1–7.

James, D.E., Fouch, M.J., Carlson, R.W., and Roth, J.B., 2011, Slab fragmentation, edge flow and the origin of the Yellowstone hotspot track: Earth and Planetary Science Letters, doi:10.1016/j.epsl.2011.09.007.

Jones, A.G., 1988, Static shift of magnetotelluric data and its removal in a sedimentary basin environment: Geophysics, v. 53, p. 967–978, doi:10.1190/1.1442533.

Leeman, W.P., Schutt, D.L., and Hughes, S.S., 2009, Thermal structure beneath the Snake River Plain: Implications for the Yellowstone hotspot: Journal of Volcanology and Geothermal Research, v. 188, p. 57–67, doi:10.1016/j.jvolgeores.2009.01.034.

Ni, H., Keppler, H., and Behrens, H., 2011, Electrical conductivity of hydrous basaltic melts: Implications for partial melting in the upper mantle: Contributions to Mineralogy and Petrology, v. 162, doi:10.1007/s00410-011-0617-4.

Obrebski, M., Allen, R.M., Xue, M., and Hung, S.-H., 2010, Slab-plume interaction beneath the Pacific Northwest: Geophysical Research Letters, v. 37, p. L14305, doi:10.1029/2010GL043489.

Patro, P.K., and Egbert, G.D., 2008, Regional conductivity structure of Cascadia: Preliminary results from 3D inversion of USArray transportable array magnetotelluric data: Geophysical Research Letters, v. 35, p. L20311, doi:10.1029/2008GL035326.

Pollack, H.N., Hurter, S.J., and Johnson, J.R., 1991, The New Global Heat Flow Compilation: Ann Arbor, Michigan, Department of Geological Sciences, University of Michigan, Technical Report.

Roberts, J.J., and Tyburczy, J.A., 1999, Partial-melt electrical conductivity: Influence of melt composition: Journal of Geophysical Research, v. 104, B4, doi:10.1029/1998JB900111.

Stanley, W.D., Boehl, J.E., Bostick, F.X., and Smith, H.W., 1977, Geothermal significance of magnetotelluric sounding in the eastern Snake River Plain-Yellowstone region: Journal of Geophysical Research, v. 82, p. 2501–2514, doi:10.1029/JB082i017p02501.

Tian, Y., Zhou, Y., Sigloch, K., Nolet, G., and Laske, G., 2011, Structure of North American mantle constrained by simultaneous inversion of multiple-frequency SH, SS, and Love waves: Journal of Geophysical Research, v. 116, p. B02307, doi:10.1029/2010JB007704.

Till, C., Grove, T., and Carlson, R., 2010, Message from the Moho: Petrologic clues to the origin of Quaternary basaltic lavas from oregon's High Lava Plains: Geological Society of America Abstracts with Programs, v. 42, no. 5, p. 343.

Wagner, L., Forsyth, D.W., Fouch, M.J., and James, D.E., 2010, Detailed three-dimensional shear wave velocity structure of the northwestern United States from Rayleigh wave tomography: Earth and Planetary Science Letters, v. 299, p. 273–284, doi:10.1016/j.epsl.2010.09.005.

Wannamaker, P.E., Hasterok, D.P., Johnston, J.M., Stodt, J., Hall, D.B., Sodergren, T.L., Pellerin, L., Maris, V., Doerner, W.M., Groenewold, K., and Unsworth, M.J., 2008, Lithospheric dismemberment and magmatic processes of the Great Basin-Colorado Plateau transition, Utah, implied from magnetotellurics: Geochemistry Geophysics Geosystems, v. 9, p. 1–38, doi:10.1029/2007GC001886.

Yang, Y., Shen, W., and Ritzwoller, M.H., 2011, Surface wave tomography on a large-scale seismic array combining ambient noise and teleseismic earthquake data: Earth Science, v. 24, p. 55–64, doi:10.1007/s11589-011-0769-3.

Yoshino, T., Laumonier, M., McIsaac, E., and Katsura, T., 2010, Electrical conductivity of basaltic and carbonatite melt-bearing peridotites at high pressures: Implications for melt distribution and melt fraction in the upper mantle: Earth and Planetary Science Letters, v. 295, p. 593–602, doi:10.1016/j.epsl.2010.04.050.

Yuan, H., and Dueker, K., 2005, Teleseismic P-wave tomogram of the Yellowstone plume: Geophysical Research Letters, v. 32, p. L07304, doi:10.1029/2004GL022056.

Yuan, H., Dueker, K., and Stachnik, J., 2010, Crustal structure and thickness along the Yellowstone hot spot track: Evidence for lower crustal outflow from beneath the eastern Snake River Plain: Geochemistry Geophysics Geosystems, v. 11, p. Q03009, doi:10.1029/2009GC002787.

Zhdanov, M.S., Smith, R.B., Gribenko, A., Cuma, M., and Green, M., 2011, Three-dimensional inversion of large-scale EarthScope magnetotelluric data based on the integral equation method: Geoelectrical imaging of the Yellowstone conductive mantle plume: Geophysical Research Letters, v. 38, p. L08307, doi:10.1029/2011GL046953.

Manuscript received 18 July 2011
Revised manuscript received 22 December 2011
Manuscript accepted 30 December 2011

Printed in USA