The Geosciences DeVL Experiment: new information generated from old magnetotelluric data of The University of Adelaide on the NCI High Performance Computing Platform

Nigel Rees*1
nigel.rees@anu.edu.au

Ben Evans1
ben.evans@anu.edu.au

Graham Heinson2
graham.heinson@adelaide.edu.au

Dennis Conway2
dennis.conway@adelaide.edu.au

Rui Yang1
rui.yang@anu.edu.au

Stephan Thiel3
stephan.thiel@sa.gov.au

Kate Robertson3
kate.robertson2@sa.gov.au

Kelsey Druken1
kelsey.druken@anu.edu.au

Bruce Goleby4
bruce.goleby@opmconsulting.com.au

Jingbo Wang1
jingbo.wang@anu.edu.au

Lesley Wyborn1
lesley.wyborn@anu.edu.au

Hoël Selillé6
hoel.seille@csiro.au

1National Computational Infrastructure, Australian National University, Canberra, ACT 2601
2The University of Adelaide, School of Physical Sciences, Adelaide, SA 5005
3Geological Survey of South Australia, Department of the Premier and Cabinet, Adelaide, SA 5000
4OPM Consulting, Canberra, ACT, Australia
5CSIRO, Deep Earth Imaging Future Science Platform, Kensington, WA 6152

SUMMARY

In recent years, magnetotelluric (MT) processing has become computationally intensive as the scale and size of MT surveys being run increases. Consequently, High Performance Computing (HPC) is now becoming a valuable tool for timely processing and modelling of these large MT datasets. As part of the MT component of the 2017-2019 Australian Research Data Commons (ARDC) funded Geoscience Data Enhanced Virtual Laboratory (DeVL) continuity project, The National Computational Infrastructure (NCI) at the Australian National University will enable MT datasets from The University of Adelaide to be added to the NCI HPC platform with the goal of creating a more Findable, Accessible, Interoperable, Reusable (FAIR) and open public resource. A focus will be on making the time series datasets more suitable for use on HPC and more interoperable with other Earth science disciplines, where High Performance Data (HPD) formats will allow for better scalability and performance. Metadata attributes, as defined by the Australian MT research community, will be added directly to the time series data files. Additionally, time series processing and 3D inversion codes are being optimised for HPD/HPC, with the end goal of rapid time series processing and 3D inversion. Making FAIR MT time series available on HPC can lead to a transformative change in the way MT data analysis is routinely conducted and such a change has the capacity to create new ways of doing collaborative and transparent MT analysis.

Key words: magnetotellurics, time series, high performance computing (HPC), high performance data (HPD), FAIR

INTRODUCTION

The magnetotelluric (MT) method is a long-established technique that has been around for almost 70 years, but it is only in the last 10 to 15 years that MT has been applied to image entire mineral systems, rather than just single deposits. From the 1960s to the 1990s, most electromagnetic surveys in Australia were undertaken using only magnetometers to give geomagnetic depth sounding information. Such surveys were used to map out large scale crustal conductivity anomalies that occurred in roughly the top 10 km of the subsurface of the Earth. However, geomagnetic depth soundings provided little information about depth variation of resistivity and the scaling and spacing of the instruments meant that the resolution was always quite poor. From the early 2000s, the priority switched to broadband or long-period MT surveys along transects, particularly following deep seismic reflection transects. Though modelling was typically done using 2D inversion methods, the results from Heinson et al. (2006) gave the first indications that ore deposit such as Olympic Dam had associated major conductive signals in the lower crust and that MT could be used to image entire mineral systems.

During this time, even more ambitious MT activities were commencing. Long-period MT arrays were beginning to be deployed (e.g. Aivazpourporgou et al., 2015; Kirkby et al., 2015) and culminated in the launching in 2013 of the Australian Lithospheric Architecture Magnetotelluric Project (AusLAMP) (Geoscience Australia, 2014; Duan et al., 2016; Robertson et al., 2018a). AusLAMP aims to acquire long-period (10-10,000s) MT data in a half degree interval grid (~55 km) over the whole of the Australian continent to map the electrical resistivity structure of the Australian lithosphere. The collection of the AusLAMP MT survey coincided with an increased use of 3D inversion methods, but because of limits in both software that could use this data and computational power, 3D inversion was done sparsely and with very coarse grids (e.g. Thiel and Heinson, 2013), and was both expensive and time consuming. The scaling of ModEM 3D inversion software (Egbert and Kelber, 2012; Kelbert et al., 2014) to High Performance Computing (HPC) systems has recently shown the improvements that can be made to these large province scale MT datasets (NCI, 2016; Robertson et al., 2016; Robertson et al., 2018b; Goleby et al., 2018).

These province-scale 3D inversions take considerable computational resources and require that MT modelling enter the HPC community, with even more scalability improvements needed as we move to larger continental datasets. For example, a typical 3D inversion of 100 sites and 21 periods per site (from 10 to 10,000s), using an improved parallelised 3D ModEM
inversion code available at the National Computational Infrastructure (NCI), with 400 GB of memory and 220 CPUs currently takes in the order of six days to run. To perform an inversion at continental scale, the MT community would need more memory, more CPUs and most importantly, further investigation into the optimisation of 3D inversion codes. A large part of the inversions is dependent on I/O performance and so effort in preparing for more 3D inversions will be hindered unless the MT data can be organised and tuned for better access in these HPC environments.

Furthermore, in Australia, since 1960, although considerable amounts of MT data have been collected with public funding in both the government agencies and university sectors (Figure 1), only a small amount by volume is currently openly accessible and available to the wider community online, and the datasets barely comply with the modern Findable, Accessible, Interoperable and Reusable (FAIR) principles of Wilkinson et al. (2016). What is there is only findable in multiple site-specific catalogues and accessible as file downloads for local processing; the time series datasets are rarely available at all. Further, the data formats for storing processed data were developed in the 1980’s (Wight, 1987) and there was limited interoperability both between various MT surveys and with other types of geophysical data. This relative lack in accessibility and consistency of formats limits reusability of MT data in modern HPC environments.

![Figure 1. Electromagnetic surveys conducted in Australia with public funding (government and university) from the 1960s to present that can be obtained. Although EDI files are available for most of them, very few of the time series datasets are findable through online catalogues or generally accessible.](image)

**THE CURRENT STATE**

MT surveys typically involve deploying long-period or broadband loggers at Earth’s surface which record naturally occurring time variations of electromagnetic fields as raw packaged data onto hard drives attached to the instrument. The acquired time series data generally include three components of the vector magnetic fields and two components of the horizontal electric field. The time series data are then time ordered, calibrated, cleaned through noise reduction techniques and resampled based on the desired MT application. Sophisticated time series processing methods are then deployed to estimate transfer functions (TFs) relating magnetic and electric field components: distortion and strike analysis reduce the TFs to a simpler set consistent with simplified one or two-dimensional models of Earth’s conductivity (Egbert, 2011). Finally, the TFs are used as input for inversion algorithms and the resulting conductivity models are used for interpretations of Earth’s subsurface structure. The different processing levels for MT datasets are outlined in Table 1.

Most data providers only release processed MT EDI files and model outputs as file downloads, and enabled their distribution of time series datasets on physical media. A lack of agreed community standards has meant that many processed EDI datasets from past surveys had inadequate metadata and it was difficult to determine exactly what processing steps had been undertaken. MT practitioners became reliant on the processing conducted by another MT scientist, which may or may not have met their target depth or processing requirements.

In general, the larger volume Level 1A calibrated time series (Table 1) datasets were not made routinely available and it was necessary to contact the author or the institution that collected the data to obtain it. The MT time series are unique and can be very expensive to acquire or re-acquire, especially at continental scale. Increasingly there is a demand for them to be more accessible than current practices allow, as they are required for replication, reanalysis and testing of new processing techniques. It is becoming essential to make calibrated MT time series datasets align with the FAIR principles to increase reusability of the data and allow for more targeted processing to specific user needs. Many different time series analysis and processing techniques exist (e.g. Chave and Thomson (1989, 2004), Larsen et al. (1996), Egbert and Eisel (1998), Eisel and Egbert (2001), Manoj and Nagarajan (2003)) and new techniques and codes will continue to be developed into the future as more computationally and data intensive capabilities become available.

**THE GEOSCIENCES DeVL PROJECT**

In 2017 the Australian Research Data Commons (ARDC) Initiative provided some seed funding for the Geosciences Data-enhanced Virtual Laboratory (DeVL) Project that sought to make significant academic geophysics data collections accessible online using the FAIR principles. The project was a collaboration between AuScope, NCI, CSIRO, The University of Adelaide, ARDC, the Research School of Earth Sciences at ANU and Curtin University and comprised four work packages: MT, Passive Seismic, the International Geo Sample Number (IGSN), and the AuScope Virtual Research Environment (AVRE) platform and portals.

The MT work package was a collaboration between The University of Adelaide and NCI, and sought to follow the exemplar of the EarthScope USArray MT project and take MT data ‘out of the drawer’ (Kelbert et al., 2018) and make both time series and processed data (EDI files, models) openly accessible in a sustainable and searchable manner. The Geosciences DeVL project involved transferring both time series (where available) and processed EDI files in their current state from the University of Adelaide and installing them along with a suite of MT processing codes onto the NCI HPC platform at the Australian National University as a proof of concept. Following the practice of the USArray MT project (Kelbert et al., 2018), each survey was assigned a globally unique digital object identifier (DOI) to enable attribution, to ensure it could become a citable scientific contribution, and ultimately to assist in usage tracking including facilitating discovering publications based on that dataset. The University of Adelaide data are now discoverable via the NCI catalogue [link], which includes a link to allow users to access the datasets either via the THREDDS data server.
NEW OPPORTUNITIES FOR MT IN HPC

A number of new opportunities are emerging in the way MT processing and modelling is currently done. Many of the time series processing codes were written decades ago in an era before a need for parallelisation and HPC. With the size of the MT data and processing, these codes must now be modernised and adapted for HPC so that whole surveys could be processed or reprocessed in a matter of minutes. Inversion codes could be advanced to run more efficiently on supercomputers which could lead to drastically reduced run times.

Time series datasets must also be revamped into HPD formats, allowing for much better scalability and performance and lowering the barrier for interoperability between different scientific disciplines. Processing pipelines and workflows could be developed and referenced in a communal environment, allowing for much more transparency in the way MT processing and modelling is conducted and ultimately leading to reproducible science. By reducing the computational barrier, HPC can lead to a step change in modelling sophistication, including better representation of uncertainty (Bryan, 2013).

HPC can allow us to do more interesting modelling with our MT data. For example, there is growing interest in probabilistic modelling of MT data which can provide interesting ways of modelling time-lapse data (Rosas-Carbajal et al., 2015), exploration of the solution space (Rosas-Carbajal et al., 2013; Conway et al., 2018; Mandolesi et al., 2018), and the ability to more accurately model the errors in the MT response (Sielie and Visser, 2018). However, such techniques are computationally burdensome and could greatly benefit from HPC. In addition, there is also some interest in applying machine learning techniques, particularly deep learning, to MT modelling (Conway et al., 2019), making the modelling process more efficient. This will require a large amount of high quality samples to create good training data that can then be used to provide accurate results. Using modern HPC deep learning libraries, the processing of these training samples is simple to parallelise and generate the inferencing model from these datasets.

All these opportunities could lead to a transformative change in the way MT data analysis is routinely conducted and such a change has the capacity to create new ways of doing collaborative and transparent MT analysis. However, to make MT datasets more interoperable and reusable, particularly on HPC, requires greater international agreement on data formats and consensus on what metadata attributes are to be collected both at the time of acquisition and during processing. To be really useful, data must be accompanied with information about how they are captured, processed, analysed and validated and other information that enables interpretation and use (Hills et al., 2015). The Australian proposal for metadata attributes (Kirkby et al., 2019) needs to be socialised with the international community. The international MT community already acknowledges that there is a need for change to modern formats (e.g. Kelbert et al., 2018) and although they have yet to agree on a preferred format for either a time series archival storage format or on a replacement for the EDI format, discussions are currently underway:

https://groups.google.com/forum/#!forum/em-data-formats
https://gitext.gfz-potsdam.de/orrit/MET-Data-Exchange

THE FUTURE OF MT

The major developments over the last 20 years which have made continental scale surveys possible are:

1. The improvement in instrumentation;
2. The ability to store and process large volumes of data;
3. The routine processing of large raw time series data to a common standard of dataset; and
4. The ability to run higher resolution 3D inversions on HPC.

There is now a better opportunity to prepare for future developments, such as more distributed arrays and deploying instruments by drones. Ideally, we would want to record more data in more places to provide better constrained models. In the next decade we will likely see significant improvements in the way data

(http://dapds00.nci.org.au/thredds/catalog/my80/catalog.html) or directly from computers attached to NCI’s filesystem. Additionally, because the catalogue metadata records from NCI, Geoscience Australia, CSIRO are harvested into the Research Data Australia (RDA) portal it is now possible to locate MT datasets from these agencies from this single site. By allowing FAIR access to MT time series datasets and processing codes in NCI’s HPC environment, an MT scientist now has the ability to reprocess transfer functions to their desired standard and specific use case without having to rely on what another data provider had produced.

In 2019 the Geosciences DeVL received additional funding from the ARDC to enable more MT time series datasets from the University of Adelaide to be added to the NCI data repository and further work to be undertaken to enhance their use in HPC processing environments. Many of the new metadata attributes defined by the Australian MT research community (Kirkby et al., 2019) will be added directly to the time series data files.

To make the University of Adelaide MT time series datasets even more suitable for use in HPC environments and more interoperable with data from other disciplines (e.g. gravity, magnetics, etc.), High Performance Data (HPD) formats such as HDF5 or netCDF will be utilised. These HPD formats are advantageous as they are self-describing, and their performance scales better on HPC when compared with traditional ASCII files. These formats have significant user communities in many Earth science disciplines and already have data services created around them (e.g. Domenico et al., 2002; Larraundo et al., 2017). The University of Adelaide and NCI are also working on optimising time series processing codes for HPC, with the aim of reprocessing older time series using HPD modified algorithms. In particular, the Renmark (2009) time series survey (http://dx.doi.org/10.25914/5bea5867bb322) is being reprocessed using an optimised version of the Bounded Influence, Remote Reference Processing (BIRRP) code (Chave and Thomson (1989, 2004)) with tuned compiling flags, parallel processing and I/O. This new optimised version allows the input of the entire time series (110 million data points per electromagnetic input), which means none of the data is discarded during processing, leading to additional data points and potentially better data fits. Prior to optimising, the processing time of the BIRRP code was initially taking five hours to run per site using one CPU core. These long run times were mainly due to the serial nature of BIRRP and its writing of intermediate files. In contrast, the initial results show the new optimised version has drastically reduced run times, with 3 minutes per site, using 4GB/core memory and 32 CPU cores.

By having the time series available, we have been able to reprocess the Renmark survey to reveal much more information at longer periods (Figure 2). The processing steps that were undertaken can be captured and documented in Jupyter notebooks, thus allowing other MT (and non-MT) practitioners to see exactly what processing was done, reproduce results and perform further analysis.

A number of new opportunities are emerging in the way MT processing and modelling is currently done. Many of the time series processing codes were written decades ago in an era before a need for parallelisation and HPC. With the size of the MT data and processing, these codes must now be modernised and adapted for HPC so that whole surveys could be processed or reprocessed in a matter of minutes. Inversion codes could be advanced to run more efficiently on supercomputers which could lead to drastically reduced run times.

Time series datasets must also be revamped into HPD formats, allowing for much better scalability and performance and lowering the barrier for interoperability between different scientific disciplines. Processing pipelines and workflows could be developed and referenced in a communal environment, allowing for much more transparency in the way MT processing and modelling is conducted and ultimately leading to reproducible science. By reducing the computational barrier, HPC can lead to a step change in modelling sophistication, including better representation of uncertainty (Bryan, 2013).

HPC can allow us to do more interesting modelling with our MT data. For example, there is growing interest in probabilistic modelling of MT data which can provide interesting ways of modelling time-lapse data (Rosas-Carbajal et al., 2015), exploration of the solution space (Rosas-Carbajal et al., 2013; Conway et al., 2018; Mandolesi et al., 2018), and the ability to more accurately model the errors in the MT response (Sielie and Visser, 2018). However, such techniques are computationally burdensome and could greatly benefit from HPC. In addition, there is also some interest in applying machine learning techniques, particularly deep learning, to MT modelling (Conway et al., 2019), making the modelling process more efficient. This will require a large amount of high quality samples to create good training data that can then be used to provide accurate results. Using modern HPC deep learning libraries, the processing of these training samples is simple to parallelise and generate the inferencing model from these datasets.

All these opportunities could lead to a transformative change in the way MT data analysis is routinely conducted and such a change has the capacity to create new ways of doing collaborative and transparent MT analysis. However, to make MT datasets more interoperable and reusable, particularly on HPC, requires greater international agreement on data formats and consensus on what metadata attributes are to be collected both at the time of acquisition and during processing. To be really useful, data must be accompanied with information about how they are captured, processed, analysed and validated and other information that enables interpretation and use (Hills et al., 2015). The Australian proposal for metadata attributes (Kirkby et al., 2019) needs to be socialised with the international community. The international MT community already acknowledges that there is a need for change to modern formats (e.g. Kelbert et al., 2018) and although they have yet to agree on a preferred format for either a time series archival storage format or on a replacement for the EDI format, discussions are currently underway:
is inverted with the goal of rapid 3D inversions: we would like to go from 3D inversions taking days to hours to minutes. We have seen this progress happen with 1D and 2D inversions, so there is no reason to think that 3D inversion will not undergo the comparable performance improvements.

ACKNOWLEDGEMENTS

The Geosciences DeVL project was funded by ARDC and AusScope with in kind support from NCI, CSIRO and The University of Adelaide. The authors also acknowledge support from the Australian Government Department of Education, through the National Collaboration Research Infrastructure Strategy (NCRIS), and the National Computational Infrastructure (NCI) including its partners (ANU, CSIRO, BoM and GA) that contributed to building the data and compute infrastructures which enabled HPC techniques to be implemented on MT data at NCI. The DeVL project builds on trials with MT on HPC from 2012 to 2016 between GA and NCI.

REFERENCES


Robertson, K., Thiel, S., and Heinson, G., 2018a, Evolving 3D lithospheric resistivity models across southern Australia derived from AusLAMP MT: ASEG Extended Abstracts, 1, 1-5.

Robertson, K., Thiel, S., and Meqbel, N., 2018b, An investigation into modeling parameters with the ModEM3DMT inversion code: 24th EM Induction Workshop, Helsingør, Denmark.


http://dx.doi.org/10.1038/sdata.2016.18.

Figure 2. Comparison of the XY (left) and YX (right) mode apparent resistivity and phase curves for an example site from the legacy Renmark processing (blue curve) and the additional information retrieved from the same reprocessed time series data (red curve) available at NCI.
Table 1. The different stages of MT data processing based on the National Aeronautics and Space Administration (NASA) and The National Research Council Committee on Data Management and Computation (CODMAC) Processing Levels for Science Datasets (Neumann, 2016).

<table>
<thead>
<tr>
<th>Processing levels</th>
<th>Name</th>
<th>Description</th>
<th>Collected / Processed by</th>
<th>Typical Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packed Raw Data</td>
<td>Raw Time Series</td>
<td>Telemetry data streamed from site loggers.</td>
<td>Single researcher or Research Team</td>
<td>GBs to TBs</td>
</tr>
<tr>
<td>Level 0</td>
<td>Edited Time Series</td>
<td>Time ordered instrument recorded data (e.g. raw voltages, counts) at full resolution.</td>
<td>Single researcher or Research Team</td>
<td>GBs to TBs</td>
</tr>
<tr>
<td>Level 1A</td>
<td>Calibrated Time Series</td>
<td>Level 0 data that have been calibrated in a reversible manner and packaged with associated calibration equations.</td>
<td>Single researcher or Research Team</td>
<td>GBs to TBs</td>
</tr>
<tr>
<td>Level 1B</td>
<td>Resampled Time Series</td>
<td>Level 0 or 1A data that have been irreversibly transformed (e.g. resampled, noisy data removed, filters applied).</td>
<td>Can be processed by anyone with access to L1A</td>
<td>GBs to TBs</td>
</tr>
<tr>
<td>Level 2</td>
<td>Derived frequency domain processed data (e.g. EDI)</td>
<td>Geophysical parameters (e.g. impedance tensors) derived from frequency domain time series processing of Level 1A or 1B data.</td>
<td>Can be processed by anyone with access to L1A or L1B</td>
<td>MBs</td>
</tr>
<tr>
<td>Level 3A</td>
<td>Derived modelling inputs</td>
<td>Level 2 parameters converted into input files for modelling and inversion algorithms.</td>
<td>Can be processed by anyone with access to L2</td>
<td>MBs</td>
</tr>
<tr>
<td>Level 3B</td>
<td>Derived modelling outputs</td>
<td>Level 2 parameters mapped onto space-time grids.</td>
<td>Can be processed by anyone with access to L2 or L3A</td>
<td>MBs</td>
</tr>
</tbody>
</table>