




# Aurora: An open-source python implementation of the EMTF package for magnetotelluric data processing using MTH5 and mt\_metadata

Karl N. Kappler<sup>5</sup>, Jared R. Peacock<sup>1</sup>, Gary D. Egbert<sup>2</sup>, Andrew Frassetto<sup>3</sup>, Lindsey Heagy<sup>4</sup>, Anna Kelbert<sup>1</sup>, Laura Keyson<sup>3</sup>, Douglas Oldenburg<sup>4</sup>, Timothy Ronan<sup>3</sup>, and Justin Sweet<sup>3</sup>

<sup>1</sup> United States Geological Survey, USA <sup>2</sup> Oregon State University, USA <sup>3</sup> Earthscope, USA <sup>4</sup> University of British Columbia, USA <sup>5</sup> Independent Researcher, USA

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

## Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

The Aurora software package robustly estimates single station and remote reference electromagnetic transfer functions (TFs) from magnetotelluric (MT) time series. Aurora is part of an open-source processing workflow that leverages the self-describing data container MTH5, which in turn leverages the general mt\_metadata framework to manage metadata. These pre-existing tools greatly simplify the processing interface, reducing requirements for specialized domain knowledge in time series analysis, or data structures management and generating transfer functions with a few lines of code. The processing depends on two inputs – a table specifying the data to use for TF estimation, and a json file specifying the processing parameters, both of which are generated automatically, and can be modified if desired. Output TFs are returned as mt\_metadata objects, and can be exported to a variety of common formats for plotting, modelling and inversion.

## Introduction

Magnetotellurics (MT) is a geophysical technique for probing the electrical conductivity structure of the subsurface using co-located electric and magnetic field measurements. After data collection, standard practice is to estimate the time invariant (frequency domain) transfer function (TF) between electric and magnetic channels before proceeding to interpretation and modelling. When all channels are orthogonal the TF is equivalent to the 2x2 Impedance Tensor (Z) electrical impedance tensor (Z) (Vozoff, 1991).

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

where  $(E_x, E_y)$ ,  $(H_x, H_y)$  denote orthogonal electric and magnetic fields respectively. TF estimation involves management of metadata (locations, orientations, timestamps,) versatile data containers (for linear algebra, slicing, plotting, etc.) and uses a broad collection of signal processing and statistical techniques (Egbert (1997) and references therein). MTH5 supplies time series as xarray objects for efficient, lazy access to data and easy application of linear algebra and statistics libraries available in the python.

## Statement of Need

Uncompiled FORTRAN processing codes have been available for years (Chave (1989), Egbert et al. (2017)) but do not offer the readability of a high-level language and modifications are not often attempted (Egbert et al., 2017). Recently several python versions of MT processing codes have been released by the open source community, including Shah et al. (2019), Smaï & Wawrzyniak (2020), Ajithabh & Patro (2023), and Friedrichs (2022). Aurora adds to this canon of options but differs by leveraging the MTH5 and mt\_metadata packages eliminating a need for internal development of time series or metadata containers. By providing an example of workflows which employ mt\_metadata and mth5 as interfaces we hope that other developers will benefit from following this model, allowing researchers interested in signal-and-noise separation in MT to spend more time exploring and testing algorithms to improve TF estimates, and less time (re)-developing formats and management tools for data and metadata. As a python representation of Egbert's EMTF Remote Reference processing software, Aurora also provides a sort of continuity in the code space as the languages evolve.

This manuscript describes the high-level concepts of the software – for information about MT data processing Ajithabh & Patro (2023) provides a concise summary, and more in-depth details can be found in Vozoff (1991), Egbert (2002) and references therein.

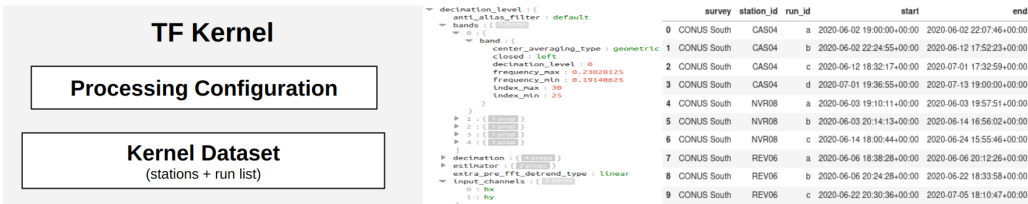
## Key Features

Simplifies:

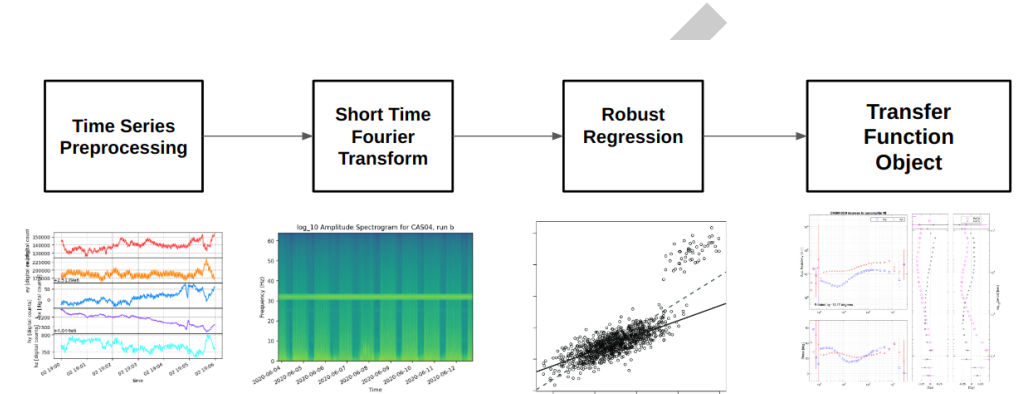
- Data indexing and management (via run summary) is tabular (pandas data frames), allowing simple programatic editing, or editing in a spreadsheet or text file.
- Processing parameters configuration interface (JSON) can be edited in a file or programmatically with JSON represented as a dictionary

Central to Aurora's process flow is the fact that an instance of a TF depends on two key prior decisions: a) The data input to the TF computation algorithm, b) the algorithm itself including the specific values of the various processing parameters. These concepts are formalized as classes (KernelDataset and Processing, respectively), and a third class TransferFunctionKernel (TFK Figure 1), a composition of the Processing, and KernelDataset provides a place for logic validating the data selection and processing parameters. TFK specifies all the information needed to make the calculation of a TF reproducible, thus supporting the R in FAIRly archived TFs.

Generation of robust TFs can be done in only a few lines starting from an MTH5 archive (Figure 3). Simplicity of workflow is due to the MTH5 container already storing comprehensive metadata, including a channel summary table, which describes all the time series stored in the archive including start/end times and sample rates. With this information already available, users can easily view a tabular summary of available data and select station pairs to process. Once a station – and optionally a remote reference station – are defined, the simultaneous time intervals of data coverage at both stations can be identified automatically, providing the Kernel Dataset. Reasonable starting processing parameters can be automatically generated for a given Kernel Dataset, and can be edited both programmatically or via a JSON file. Once the TFK is defined, the rest of the MT processing workflow automatically follows the flow of Figure 2.



**Figure 1:** TF Kernel concept diagram: A box representing the TF Kernel with two inlay boxes representing the processing config (JSON) and dataset (pandas DataFrame).



**Figure 2:** Cartoon depicting the main interfaces of aurora TF processing. Example time series from mth5 archive in linked notebook (using MTH5 built-in time series plotting capability), spectrogram from FC data structure, cartoon from Hand (2018) and TF from SPUD.

**Examples**

Here an example of the aurora data processing flow is given, using data from Earthscope. This section refers to several Jupyter notebooks that are intended as a companion to this paper. A relatively general notebook about accessing Earthscope data with mth5 can be found in the link from row 1 of Table 1.

**Table 1:** Referenced jupyter notebooks with links.

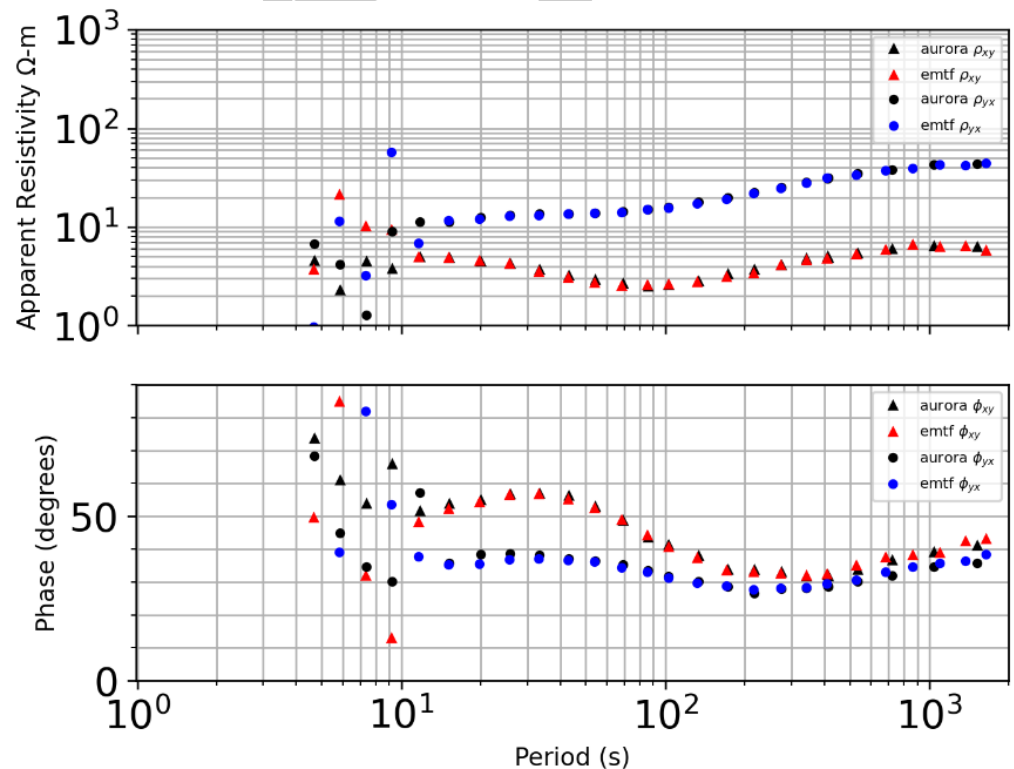
ID	Link
1	<a href="#">earthscope_magnetic_data_tutorial</a>
2	<a href="#">make_mth5_driver_v0.2.0</a>
3	<a href="#">process_cas04_multiple_station</a>

The MTH5 dataset can be built by executing the example notebook in the mth5 github repository in row 2 of Table 1. The data processing can be executed by following the tutorial in row 3 of Table 1. In that tutorial data from a station (CAS04) archived at Earthscope is processed with descriptive text, which can be condensed to the snippet in in Figure 3. The resultant apparent resistivities are plotted in Figure 4 along with the results hosted at Earthscope from Egbert’s FORTRAN EMTF code.

```
from aurora.config.config_creator import ConfigCreator
from aurora.pipelines.process_mth5 import process_mth5
from aurora.pipelines.run_summary import RunSummary
from aurora.transfer_function.kernel_dataset import KernelDataset
```

```
run_summary = RunSummary()
run_summary.from_mth5s(["8P_CAS04_NVR08.h5",])
kernel_dataset = KernelDataset()
kernel_dataset.from_run_summary(run_summary, "CAS04", "NVR08")
cc = ConfigCreator()
config = cc.create_from_kernel_dataset(kernel_dataset)
tf = process_mth5(config, kernel_dataset)
tf.write(fn="CAS04_rrNVR08.edi", file_type="edi")
```

**Figure 3:** Code snippet with steps to generate a TF from an MTH5 (generated by row 1 of Table 1). With MTH5 in present working directory, a table of available contiguous blocks of multichannel time series is generated a “RunSummary”, then station(s) to process are selected (by inspection of the RunSummary dataframe) to generate a KernelDataset. The KernelDataset identifies simultaneous data at the local and reference site, and generates processing parameters, which can be edited before passing them to process\_mth5 – the core processing method, and finally exporting TF to a standardized output file, in this case edi.



**Figure 4:** Comparison for a randomly selected station between Aurora and the EMTF Fortran code. While both curves exhibit some scatter in the low SNR MT “dead band” between 1-10s, most of the curves are very similar.

## Testing

The Aurora package uses continuous integration (Duvall et al., 2007) and implements both unit tests as well as integrated tests with code currently at 77% coverage as measured by CodeCov, and the core dependencies `mt_metadata` and `MTH5` have 84% and 60% of code covered by continuous integration testing respectively. Improvement of test coverage is ongoing. For integrated tests Aurora uses a small synthetic MT dataset (originally from EMTF) for TF estimation. A few processing configurations with manually validated results are stored in the repository. Deviation from these results causes tests to fail, alerting developers that a code change has resulted in an unexpected baseline processing result. In the summer of 2023, widescale testing on Earthscope data archives was performed and showed that the TF results of aurora are similar to those from the EMTF fortran codes, in this case for hundreds of real stations rather than a few synthetic ones. Before PyPI, and conda forge releases, example Jupyter notebooks are also run via github actions.

## Future Work

Aurora uses github issues to track tasks and planned improvements. In the near future we want to add noise suppression techniques, for example coherence and polarization sorting and Mahalanobis distance (e.g. Ajithabh & Patro (2023), Platz & Weckmann (2019)). We would also like to develop, or plug into a graphical data selection/rejection interface with time series plotting. Besides these improvements to TF quality, we also would like to embed the TFKernel information into both the `MTH5` and the output `EMTF_XML` (Kelbert (2020)). Unit and integrated testing should be expanded, with a larger dataset included in the tests and test coverage on data from audio frequency band (most test data is sampled at 1Hz). This work will continue to codevelop with `mt_metadata`, `MTH5` and `MTPy` to maintain the ability to provide outputs for inversion and modelling. Ideally the community can participate in a comparative analysis of the opensource codes available to build a recipe book for handling noise from various datasets, ideally using open-archived datasets.

## Conclusion

Aurora provides an open-source Python implementation of the EMTF package for magnetotelluric data processing. Aurora is a prototype worked example of how to plug processing into an existing opensource data and metadata ecosystem (`MTH5`, `mt_metadata`, & `MTPy`), and we hope that other open source MT processing authors will follow suit to also provide interfaces to these packages. It is hoped that these tools will contribute to workflows which can focus more on geoscience analysis, and less on the nuances of data management.

## Appendix

### Installation Instructions

The package is installable via the python Package Index (pip) as well as via conda forge. The installation in pip: `pip install aurora` And via conda forge: `conda install aurora`

### Documentation

Documentation is hosted by SimPEG Cockett et al. (2015) and can be found at this [link](#)

### Licence

Aurora is distributed under the [MIT](#) open-source licence.

## Acknowledgments

The authors would like to thank IRIS (now Earthscope) for supporting the development of Aurora.

TODO:

- ☐ Update links to ipynb to release branches after mth5/aurora releases.
- ☐ remove draft watermark
- ☐ Link these issues to discussion in future work? <https://github.com/kujaku11/mth5/issues/179>, [https://github.com/kujaku11/mt\\_metadata/issues/195](https://github.com/kujaku11/mt_metadata/issues/195)

Ajithabh, K., & Patro, P. K. (2023). SigMT: An open-source python package for magnetotelluric data processing. *Computers & Geosciences*, 171, 105270.

Chave, A. D. (1989). BIRRP: Bounded influence, remote reference processing. *Journal of Geophysical Research*, 94(B10), 14–215.

Cockett, R., Kang, S., Heagy, L. J., Pidlisecky, A., & Oldenburg, D. W. (2015). SimPEG: An open source framework for simulation and gradient based parameter estimation in geophysical applications. *Computers & Geosciences*.

Duvall, P. M., Matyas, S., & Glover, A. (2007). *Continuous integration: Improving software quality and reducing risk*. Pearson Education.

Egbert, G. D. (1997). Robust multiple-station magnetotelluric data processing. *Geophysical Journal International*, 130(2), 475–496.

Egbert, G. D. (2002). Processing and interpretation of electromagnetic induction array data. *Surveys in Geophysics*, 23(2-3), 207–249.

Egbert, G. D., Kelbert, A., & Meqbel, N. M. (2017). Mod3DMT and EMTF: Free software for MT data processing and inversion. *AGU Fall Meeting Abstracts*, 2017, NS44A–04.

Friedrichs, B. (2022). MTHotel. In *GitHub repository*. GitHub. <https://github.com/bfrmtx/MTHotel>

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3), 555–605.

Kelbert, A. (2020). EMTF XML: New data interchange format and conversion tools for electromagnetic transfer functions. *Geophysics*, 85(1), F1–F17.

Platz, A., & Weckmann, U. (2019). An automated new pre-selection tool for noisy magnetotelluric data using the mahalanobis distance and magnetic field constraints. *Geophysical Journal International*, 218(3), 1853–1872.

Shah, N., Samrock, F., & Saar, M. O. (2019). Resistics: A versatile native python 3 package for processing of magnetotelluric data. 28. *Schmucker-Weidelt-Kolloquium für Elektromagnetische Tiefenforschung*.

Smaï, F., & Wawrzyniak, P. (2020). Razorback, an open source python library for robust processing of magnetotelluric data. *Frontiers in Earth Science*, 8, 296.

Vozoff, K. (1991). *The Magnetotelluric Method*. <https://pubs.geoscienceworld.org/seg/books/book/2087/chapter-abstract/114406941/THE-MAGNETOTELLURIC-METHOD?redirectedFrom=fulltext>