Quality estimation of magnetotelluric impedance tensors using neural networks

Naoto Imamura[1] and Adam Schultz[2]

https://doi.org/10.1190/tle39100306.1

Abstract

While much of the workflow for geophysical data processing and analysis is automated, final quality assurance typically requires the decisions of skilled human analysts and interpreters. The quality and reliability of geophysical inverse models depends directly on the effectiveness of input data. The influence of spurious data on derived data products used in the inversion has been reduced by this effectiveness. Failure to identify and mitigate bias in data products can lead to costly errors. By applying supervised machine learning (ML), a neural network can be trained to recognize features in data that a human domain expert would identify as characteristic of poor data quality. In this study, we use magnetotelluric (MT) data as an example of a geophysical data set appropriate for such a training exercise. While MT data are used to estimate the resistivity structure of the subsurface, the concepts we discuss are universal to seismic, potential fields, and other geophysical data sets. We train a neural network, pyMAGIQ (Python-based magnetotelluric impedance qualifier), with multiple hidden layers and demonstrate that it successfully generates the nonlinear mapping function required to assess the quality of MT data. The training set is a large database of frequency-domain MT impedance tensors from the National Science Foundation-funded EarthScope MT project. A human-assigned quality index is associated with each impedance. We apply pyMAGIQ to unrated MT data from the United States and Canada and confirm that the ML-assigned quality factors are consistent with those assigned by trained human operators. We also apply sensitivity analysis to the trained neural network. This reveals that the human- and ML-assigned data quality index depends on the magnitude of the confidence limits on (1) the phases and (2) the continuity of the apparent resistivities and phases with respect to frequency.

Introduction

Solving inverse problems is a critical step in geophysical workflows. However, the inverse problem is typically ill posed, nonunique, and nonlinear (Tarantola, 2005). The uncertainty (confidence limits) on the parameters recovered from an inverse solution and their localization in space (resolving power) often scales nonlinearly with the size of the statistical confidence limits on the input data set in inverse modeling and with unmediated bias in the data. The bias here may be attributed to invalid assumptions made in representing the signal source. We draw on an example from magnetotellurics (MT), where over the past three decades, a variety of methods beyond classical least-squares estimation have been used to calculate the frequency-domain MT impedance tensors and induction vectors (i.e., the MT response functions that are inverted to produce resistivity images of the subsurface). The MT response functions can be defined in the frequency domain for a harmonic signal as:

$$\begin{bmatrix} E_x(\omega) \\ E_y(\omega) \end{bmatrix} = \begin{bmatrix} Z_{xx}(\omega) & Z_{xx}(\omega) \\ Z_{xx}(\omega) & Z_{xx}(\omega) \end{bmatrix} \begin{bmatrix} B_x(\omega) \\ B_y(\omega) \end{bmatrix} \tag{1}$$

$$B_z = \begin{bmatrix} T_x & T_y \end{bmatrix} \begin{bmatrix} B_x(\omega) \\ B_y(\omega) \end{bmatrix}, \tag{2}$$

where E is the Fourier coefficient of the electric field vector, Z is the impedance tensor, B is the Fourier coefficient of the magnetic field vector, and T is the induction vector. All are defined at radian frequency ω, while x, y, and z is a right-handed coordinate system representing north, east, and positive downward, respectively. The MT response functions are typically calculated within a given frequency band through the following relations (Vozoff, 1972). The relations represent a least-squares solution to minimize the misfit between electric and magnetic field components (for the impedance tensor) and between the vertical and two horizontal magnetic field components (for the induction vector) in the presence of incoherent noise:

$$\begin{aligned} Z_{xx} &= \frac{\langle E_x B_x^* \rangle \langle B_y B_y^* \rangle - \langle E_x B_y^* \rangle \langle B_y B_x^* \rangle}{\langle B_x B_x^* \rangle \langle B_y B_y^* \rangle - \langle B_x B_y^* \rangle \langle B_y B_x^* \rangle} \qquad Z_{xy} = \frac{\langle E_x B_x^* \rangle \langle B_x B_y^* \rangle - \langle E_x B_y^* \rangle \langle B_x B_x^* \rangle}{\langle B_y B_x^* \rangle \langle B_x B_y^* \rangle - \langle B_y B_y^* \rangle \langle B_x B_x^* \rangle} \\ Z_{yx} &= \frac{\langle E_y B_x^* \rangle \langle B_y B_y^* \rangle - \langle E_y B_y^* \rangle \langle B_y B_x^* \rangle}{\langle B_x B_x^* \rangle \langle B_y B_y^* \rangle - \langle B_x B_y^* \rangle \langle B_y B_x^* \rangle} \qquad Z_{yy} = \frac{\langle E_y B_x^* \rangle \langle B_x B_y^* \rangle - \langle E_y B_y^* \rangle \langle B_x B_x^* \rangle}{\langle B_y B_x^* \rangle \langle B_x B_y^* \rangle - \langle B_y B_y^* \rangle \langle B_x B_x^* \rangle}, \\ T_x &= \frac{\langle B_z B_x^* \rangle \langle B_y B_y^* \rangle - \langle B_z B_y^* \rangle \langle B_y B_x^* \rangle}{\langle B_x B_x^* \rangle \langle B_y B_y^* \rangle - \langle B_x B_y^* \rangle \langle B_y B_x^* \rangle} \qquad T_y = \frac{\langle B_x B_x^* \rangle \langle B_z B_y^* \rangle - \langle B_x B_y^* \rangle \langle B_z B_x^* \rangle}{\langle B_x B_x^* \rangle \langle B_y B_y^* \rangle - \langle B_x B_y^* \rangle \langle B_y B_x^* \rangle} \end{aligned} \tag{3}$$

[1]Formerly Oregon State University; presently Raithing Inc., Tokyo, Japan. E-mail: naoto.imamura@raithing.io.
[2]Oregon State University, Corvallis, Oregon, USA. E-mail: adam.schultz@oregonstate.edu.

where * represents the complex conjugate operator. The angle brackets represent crosspowers obtained through band average (the average of a set of complex Fourier coefficients summed over a set of adjacent frequencies), section average (the average of complex Fourier coefficients at a given frequency summed over a set of independent estimates obtained from different time series sections), or a combination of both. Generally, the larger the number of degrees of freedom in the averaging function, the smaller the confidence limits in the resulting response function. There is a trade-off between minimizing the variance on the estimate and the ability to resolve finer-scale structure if the averaging function smooths over too wide of a frequency band.

A fundamental limitation of the classical least-squares solution is that while it is statistically efficient, it is subject to potentially unbound bias in the presence of spurious data. One widely adopted data processing branch to mitigate this weakness relies on methods from robust statistics (Hampel et al., 1986). The methods seek to down weight the influence of data drawn from statistical populations that deviate from the main population of observed data. Generally, such methods do not impose an a priori assumption of a particular probability density function to describe the main population. Rather, they identify the presence of outliers (data or sections of data with statistical behavior that deviates significantly from the main population). The influence of outlying data on the calculation of the MT response functions is minimized (e.g., through Huber weighting [Huber, 1981] or similar methods). Then, a modified least-squares estimate obtains the robust solution (Egbert, 1997; Chave and Thomson, 2004).

Underlying this statistical approach is the assumption that within an observed MT data set, the dominant feature is the signature of the induced response of the subsurface resistivity structure to signal sources (magnetic fields due to lightning and/or ionospheric electric current systems). It is posited that data from other physical sources within the electromagnetic spectrum can be identified as outliers, distinct from the main induction source, which typically is assumed to have a plane wave structure. While this isn't a rigidly parametric approach, it is an assumption grounded in statistics rather than in a physical model of the signal source. A well-known shortcoming of this approach is the failure of statistically robust response function estimators to shield bias due to violations of assumed source field structure. Although, in practice such estimators provide substantial advances over nonrobust methods.

A second commonly used method in MT is coherence sorting and weighting (Jones and Jödicke, 1984). The external time-varying magnetic field, which serves as the MT signal source, induces both magnetic and electric fields that are measured at ground level. Absent of any noise sources, the inducing source magnetic fields and the induced electric fields are coherent. Data sections with low coherence between observed electric and magnetic fields are taken to represent episodes of poor signal-to-noise level. They are disregarded or their influence on the estimation of MT response functions is down weighted.

While the summarized methods led to remarkable improvement in the overall quality of MT response functions, substantial problems persist. Data sections of high coherence between electric and magnetic fields and/or sections where the influence of outlying values has been minimized not infrequently produce MT responses to be identified as biased, contaminated, or poor quality by experienced human operators. This may be evident to the trained operator (a subjective judgement usually based on years of experience). Or it may be revealed after attempts to invert seemingly high-quality response functions return unusable results, even when carefully constructed 3D inverse modeling attempts have been made.

We have explored the use of supervised machine learning (ML), employing a trained neural network to replace the human operator and to automate the process of assessing the quality of MT response functions. ML is a data-driven approach to extract features in data by using one or more ways to solve problems: supervised, unsupervised, and reinforcement learning. Artificial neural networks are one of the most commonly used supervised learning methods for classification and regression. An advantage of ML for geophysical data quality classification is that it can be implemented without the requirement of an underlying physical model of the structure of the signal source. It can also be implemented without knowledge of the earth response or of the multiplicity of noise sources.

In MT surveys, the complex-valued impedance tensor versus frequency and the induction vector serve as input to the inversion. Experienced human operators can often identify degraded response functions attributable to low signal-to-noise ratios, errors in recording due to sensor and timing failures, errors in metadata used to determine sensor configurations and gain/filter settings, and errors from the selection of suboptimal parameters governing the signal analysis stage. Human operators typically examine the apparent resistivity ρ and phase Φ curves versus frequency, which are real-valued quantities derived from complex-valued impedance tensors:

$$\rho_{kl}(\omega) = \mu_0 \left| Z_{kl}(\omega) \right|^2 \tag{4}$$

$$\phi_{kl}(\omega) = tan^{-1}\left(\frac{Im\left(Z_{kl}(\omega)\right)}{Re\left(Z_{kl}(\omega)\right)} \right). \tag{5}$$

These quantities assess the continuity or smoothness of the functions as well as the uncertainty as expressed through their calculated confidence limits (μ_0 is the magnetic permeability of free space). Apparent resistivities and phases indicate the integrated resistivity structure from the ground surface to a certain depth. Hence, they are expected to be smooth and continuous in the frequency domain.

In this paper, we propose a novel approach to qualify MT data based on an ML method. The approach is implemented in Python using ML libraries. We named this software "pyMAGIQ" (Python-based magnetotelluric impedance tensor qualifier). The software casts qualification of impedance tensors as a supervised classification problem. The artificial neural networks in pyMAGIQ are built using the TensorFlow framework (Abadi et al., 2016). They are trained on large data sets of MT apparent resistivities and phases with human-operator-provided data quality assessment

information acquired during the EarthScope MT project (Schultz, 2009). The effort to employ ML for the problem of automated MT data quality assessment requires the development of techniques to interpret and explain what the neural network has learned. Such an effort is a key component of the data validation procedure (Guidotti et al., 2018). We first evaluate the performance and limitations of our algorithm by using convergence history and a confusion matrix. Then, we apply it to a set of unrated MT impedance tensors obtained across the United States and Canada. In order to understand the decision-making process of pyMAGIQ, we perform a sensitivity analysis by analyzing the trained neural network.

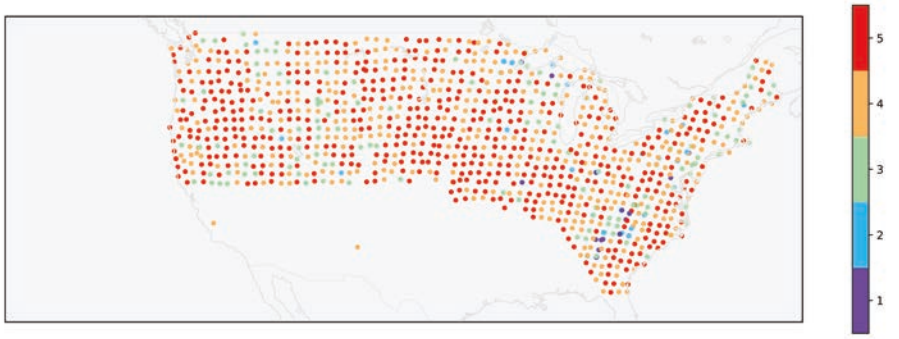

Figure 1. Location and data quality of MT impedance tensors in the USArray project. A higher rate indicates that the data quality of MT impedance tensors is better.

EarthScope MT data

We used MT response function data archived in the Incorporated Research Institutions for Seismology (IRIS) Searchable Product Depository (SPUD) database (Schultz, 2019). At the time of writing this paper, the database contained 1115 MT impedance tensors acquired during the EarthScope MT project. This is an effort overseen by Oregon State University that continues under NASA and USGS funding as we complete the systematic mapping of the electrical resistivity structure of the conterminous United States (Figure 1). In this database, an index of data quality for MT impedance tensors is provided on a scale from 0 to 5, based on the following criteria:

- An MT site with a quality index of 5 displays apparent resistivity and phase curves for the two principal (off-diagonal) elements of the impedance tensor. These vary smoothly with frequency and with confidence limits that are small (typically less than 5% standard error in apparent resistivity and less than 2° in phase across the frequency band spanning from 10^{-4} to 10^{-1} Hz).
- A rating of 4 indicates slightly larger confidence limits across part or all of the frequency band in comparison to a rating of 5. However, the MT impedance tensor can be safely used for inversion.
- A rating of 3 has significantly larger confidence limits than those with higher ratings, but the data still provide usable constraints to an inversion.
- A rating of 2 indicates that one or both principal elements of the impedance tensor were not usable for a range of periods.
- A rating of 1 means that the data were unreliable.
- A rating of 0 is a placeholder, indicating that the data were not yet assigned a rating.

Such a rating system based on the experience of human operators is clearly subjective and prone to inconsistency. This is because different human operators examine different data sets. The ability of supervised ML to assimilate a subjective training set and to produce a system capable of accurately attributing data quality assessments is a notable outcome of the present work.

To train the neural network, we employed TensorFlow (Abadi et al., 2016) and used 1101 MT impedance functions from the IRIS SPUD database. Each impedance function had a human-assigned rating from 1 to 5 and spanned a wide enough range of frequencies to substantially cover the band identified in the first bullet point. In this training set, there were 521 sites rated 5, 385 rated 4, 143 rated 3, 34 rated 2, and 18 rated 1. At the end of 2019, the SPUD database also had 1909 sets of unrated MT impedance data from locations in Canada and the United States. After training our neural network, we applied pyMAGIQ to these unrated MT impedance tensors to assess their data quality.

In an ML study, the distribution of classifications is ideally similar between different categories in a training data set. A biased training data set can result in a biased prediction. As shown earlier, the distribution in the training data sets is biased toward data ranked as higher quality. Approximately 82% of the data are rated as 4 or 5. We applied data augmentation to increase the ratio of the lower-rated data. New training data sets were created by horizontally rotating the original MT impedance tensors away from their cardinal coordinate system by angles between –15° and 15° in one-degree increments. These augmented data have the same rating as the original unrotated data. This is under the assumption that the rotation by a small angle does not change its perceived quality. After creating augmented data, we randomly selected 550 MT impedances for each of the data quality ratings, yielding a training set of 2750 impedance tensors distributed equally across the rating categories.

Method

The MT impedance tensor has four complex-valued elements at each frequency. In order to simplify the training, we only examine the off-diagonal impedance elements Z_{xy} and Z_{yx}. These are considered the principal elements of the impedance, and for a resistivity structure that varies in 1D or 2D, they are the only nonzero elements. The concept developed for pyMAGIQ could easily be extended to all four impedance tensor elements. Although, the quality of the diagonal elements is far more variable than the off-diagonal elements and less representative of overall data quality in most cases. We scale the off-diagonal impedance elements and their associated confidence limits into

their real-valued equivalent forms of apparent resistivity and phase, following equations 4 and 5.

Our ML software, pyMAGIQ, takes the apparent resistivities, phases, confidence limits, and continuity versus frequency of these quantities in the measurement reference frame as input parameters of the neural network. Continuity is calculated by taking the derivative, with respect to frequency, of each of these parameters. MT response function data in the EarthScope MT data sets are defined at 30 frequencies, which are spaced logarithmically from 7.3 to 18,724 s period. The total number of parameters for one MT site is 480, which is the product of 30 frequencies. At each frequency, there are 8° of freedom (apparent resistivities, phase, confidence limits, and continuity) for two off-diagonal impedance elements. The output of this neural network is a data quality rating assigned to each impedance that ranges from 1 to 5. This input and output information is used to train the weight parameters between nodes in the neural network.

The hidden core layers of our neural network are eight dense layers connecting all nodes between each layer (Figure 2). Dense layers are the rectified linear unit activation function for the first seven layers and the softmax activation function for the last layer. The parameter of the network is optimized to minimize discrepancy between the predicted rating and the given rating on the training set by using the cross-entropy loss function. The loss function is defined as:

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} y_{i,j}\log\left(p_{i,j}\right), \tag{6}$$

where $y_{i,j}$ denotes the true value (i.e., 1 if sample i belongs to rating j and 0 otherwise), and $p_{i,j}$ denotes the probability predicted by the model of sample i belonging to rating j. The number of nodes in each layer is set to 50. The total number of unknown weighting parameters is 42,155. Hyperparameters, including the number of layers, nodes, and regularization parameters in this study, are selected based on Optuna (Akiba et al., 2019), optimizing the validation accuracy. Using these parameters, we optimized the weight parameters with the Adamax algorithm with the default optimization parameters in the Keras library. We trained our weight parameters for 500 iterations, which took approximately 2 minutes on a 2.9 GHz Intel Core i5 processor. Once we train the weight parameters on the neural networks, a rating is estimated in a few seconds for any input.

After the neural network has been trained, we interpret the obtained neural network with a sensitivity analysis (Zurada et al., 1994; Sung, 1998; Khan et al., 2001). This is based on the model's locally evaluated gradient according to the following equation:

$$R_i(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_i}\right)^2, \tag{7}$$

where the gradient is evaluated at data point $\boldsymbol{x}$, and x_i denotes a parameter in the data. This gradient gives the rating change of each parameter x_i for $\boldsymbol{x}$. Sensitivity $R_i(\boldsymbol{x})$ will tell us how f will behave in response to infinitesimal perturbations. If $R_i(\boldsymbol{x})$ takes on a large magnitude, then f is sensitive to parameter x_i. We change every parameter of $\boldsymbol{x}$, a total of 480 parameters, to see its sensitivity. We use the predicted rating f for data analysis in this study. The gradient is discretized and calculated based on the central difference method with 4% perturbations.

Results and discussion

Before processing unrated data sets, we tested the accuracy of the trained neural network through the convergence and confusion matrix. The confusion matrix visualizes classification accuracy and errors made by the ML model. Each row of the matrix represents the given rating by human experts, while each column represents the predicted rating. In our evaluation, 90% of data sets randomly chosen are used for actual training from rated data sets. The remaining 10% of data sets are used for the validation of the trained neural network. These validation data sets are randomly selected from the human-rated MT database. However, their assigned ratings are ignored, and they are not used as part of the training data set. The training data sets are used for the actual training of the model. Validation data sets are not used for training, but are used to test the neural network for the nontraining data sets during the iteration. The model accuracy and loss for the trained neural network along with epoch are shown in Figure 3. After 500 iterations, there was no further improvement in the accuracy of the validation data sets. We obtain 98% accuracy for matching the predicted data quality ratings against those previously assigned to the actual training data sets. We matched the human-provided data quality rating of 91% for the validation data sets after 500 iterations. The high rating of concurrence with expert human data quality ratings indicates that the neural network is trained well. A review of the ML-guided

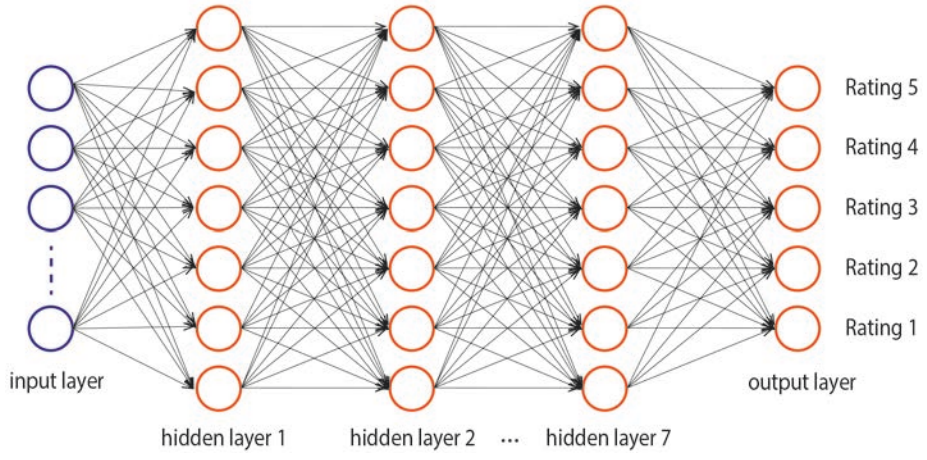

Figure 2. Overview of the neural network used in this study.

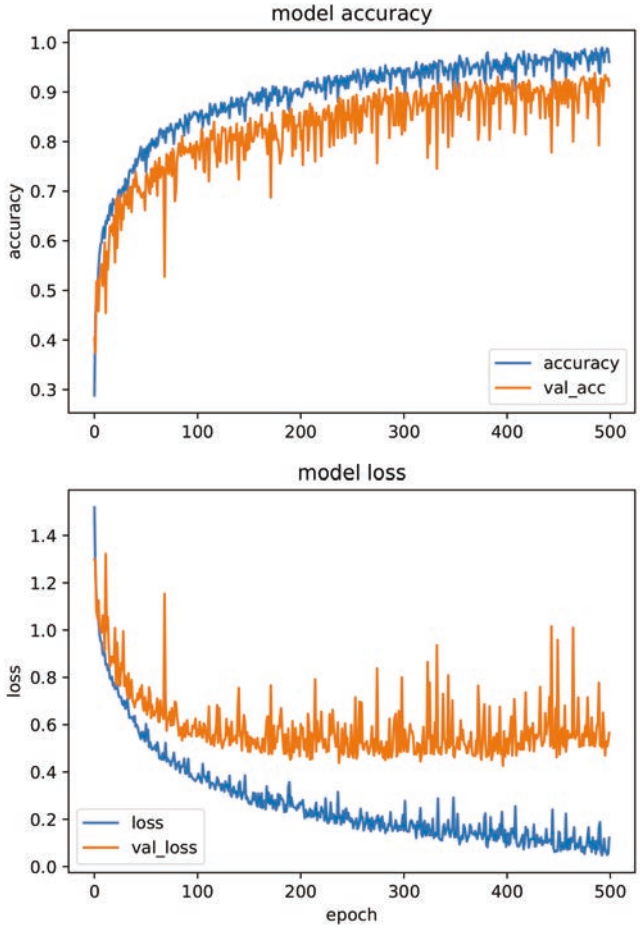

Figure 3. History of model accuracy and loss for the training and validation data sets. Blue lines are results from training data sets. Orange lines are results from validation data sets.

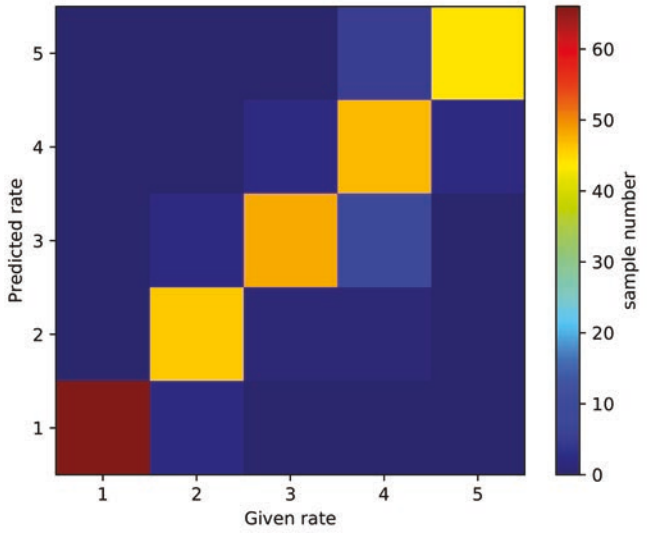

Figure 4. Confusion matrix for the validation data sets. Contour colors show the number of samples. Each row of the matrix represents the rating by human experts. Each column represents the predicted rating.

data quality ratings found discrepancies in some of the human-assigned data quality ratings. It was concluded (subjectively) that the neural network performed at least as well, if not better, than the trained human expert at assigning ratings that were consistent across different data sets. The loss function indicates that the validation loss hits bottom after 200 epochs. The loss function is used to predict rating probabilities. The accuracy line increases when the neural network's output is close to 1 for the right rating and close to 0 for other ratings. However, the loss function reaches a plateau after 200 epochs, while the validation accuracy continues to increase. This suggests that the neural network is trained to select the right rating. Although, the neural network returns high probabilities for multiple ratings rather than selecting one rating as the prediction with very high value. This happens because the neighbor rating in this study has similar characteristics. Therefore, it is hard to determine the prediction with high probabilities when the neural network is trained enough. Since validation accuracy increases after 200 epochs, we chose the neural network trained after 500 iterations for the following discussion.

We show in Figure 4 the confusion matrix between the predicted and given data quality ratings to identify erroneous ratings in the validation data sets. Here, the predicted rating is given by pyMAGIQ. The given rating is assigned by human experts, as reflected in the EarthScope MT database. Although the confusion matrix indicates that the predicted and given ratings match in most cases, there are some misclassified ratings that under- or overestimate the data quality. This is usually by a single integer rating level. This is because MT data-assigned neighboring ratings have similar characteristics, as described previously. Figure 5 shows examples of apparent resistivities and phases from these data sets. These data sets are displayed in measurement (i.e., geomagnetic coordinates). Apparent resistivities and phases show that the pyMAGIQ-assigned data quality rating conforms to the definitions provided previously in all examples. One can see that results in Figures 5c, 5e, and 5h are misclassified. However, it is difficult to judge correctly using only visual means and experience. In Figure 5c, the apparent resistivities and phases vary smoothly with frequency, and their confidence limits are relatively small. This example indicates that the predicted rating of 5 subjectively appears to be more reasonable than the human-assigned rating of 4. In the validation data sets, data from a few sites were assigned quality ratings by pyMAGIQ that appear to be more reasonable rather than those from corresponding human expert results (Figure 4). There are similar data sets potentially misrated in the training data sets, but these are not dominant. Because the weights in the neural network are optimized by the total data sets, these minor data sets do not significantly bias the weights during neural network training.

Following training of the neural network, we applied pyMAGIQ to unrated MT data from sites in Canada and the United States. Because the frequency bins of the EarthScope MT response function data and that of the unrated MT data can differ on a site-by-site basis, we interpolated the apparent resistivities, phases, and their confidence limits of the unrated data sets and resampled them using spline interpolation to fit the frequencies

at which the EarthScope MT program's data sets were defined. In general, apparent resistivities and phases at longer periods have larger confidence limits because there are fewer degrees of freedom associated with such data. While the EarthScope MT training data set includes long-period MT data, a significant fraction of the test data set from Canada comprises wideband MT that does not extend past a 1000 s period. Whereas, the EarthScope MT data extend well beyond a 10,000 s period. We exclude unrated data sets that do not extend beyond a 1000 s period to avoid the need to extrapolate wideband data well outside the range of their validity. There are 626 data sets satisfying this condition in the unrated data sets. The distribution and ML-predicted ratings of the unrated MT data are shown in Figure 6. In these predicted ratings, there are 231 data sets assigned a rating of 5, 200 rated as 4, 142 as 3, 20 as 2, and 33 as 1. By using these quality ratings, one can decide which sites have data appropriate for inversion and interpretation. One can also use the ML-guided data quality assessment to identify sites that may require relocation or reinstallation of MT equipment in order to address shortcomings in data quality. Unrated data sets and ML-generated data quality ratings are shown in Figure 7. The predicted data quality ratings appear to be reasonable in terms of the characteristics of the apparent resistivity and phase curves.

Although pyMAGIQ achieves a high congruence with expert human data quality assessment, we do not know precisely how pyMAGIQ's neural network arrives at a rating decision. We have explored which factors may have the most influence on neural network decision making. Figures 8a and 8b show a sensitivity analysis from Figures 5a and 5b. The *y*-axis in Figure 8 shows the sensitivity of the final data quality rating to the change of

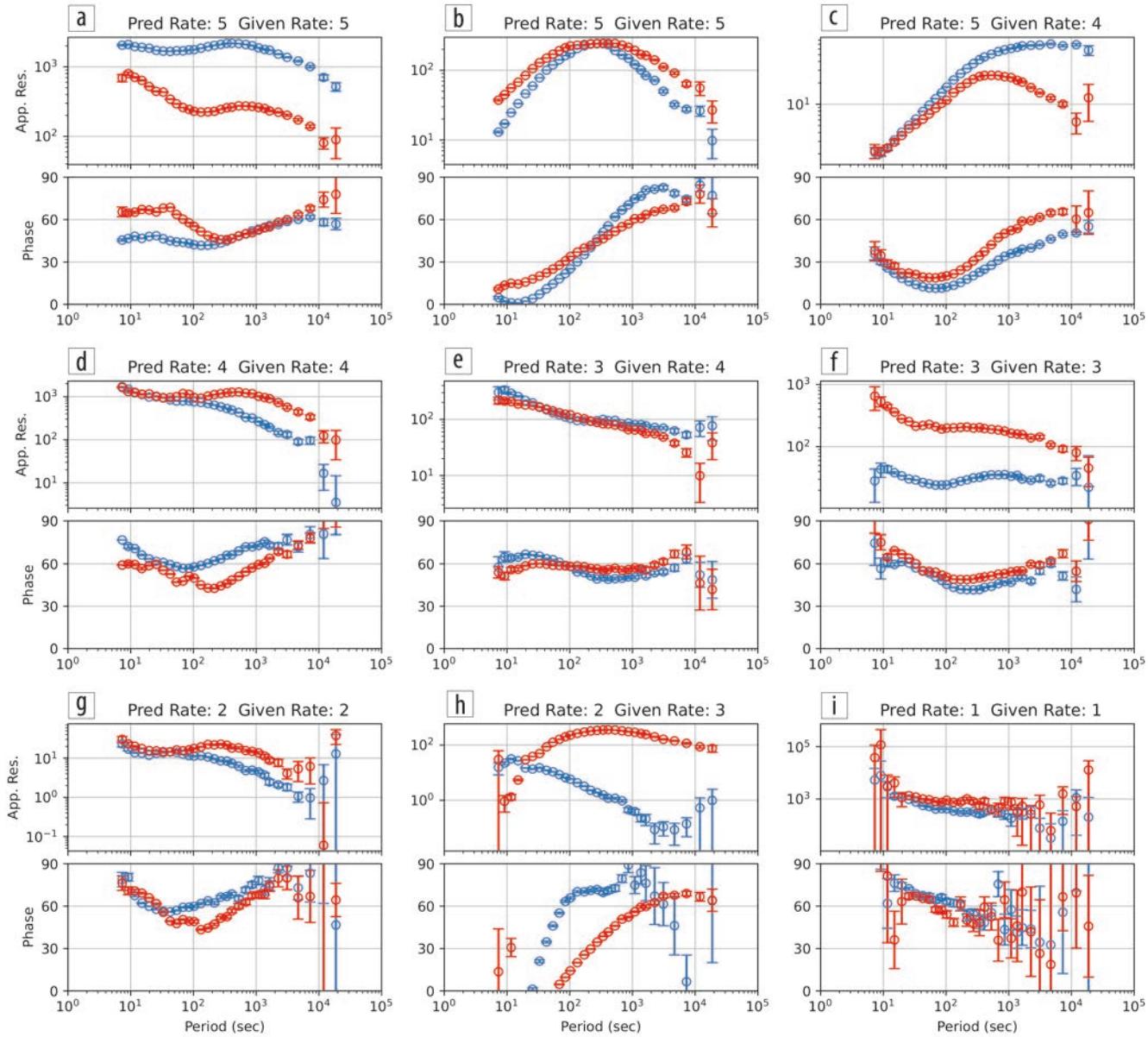

Figure 5. Predicted and given rates for apparent resistivities and phases for the validation data sets. Blue lines are *x-y* components. Red lines are *y-x* components. Predicted rating is for results of pyMAGIQ. Given rating is for the original rating in the USArray.

parameters. One can see that the continuity of apparent resistivity and phase versus frequency are relatively more significant factors than the other parameters. This means that these parameters are dominant for the rating decision. This corresponds with the traditional qualification, which focuses on smoothness of the parameters. The magnitude of the confidence limits on the phase and the continuity of the confidence limits on the apparent resistivity and phase follow in significance. Examination of the pattern of confidence limits reveals that there are often

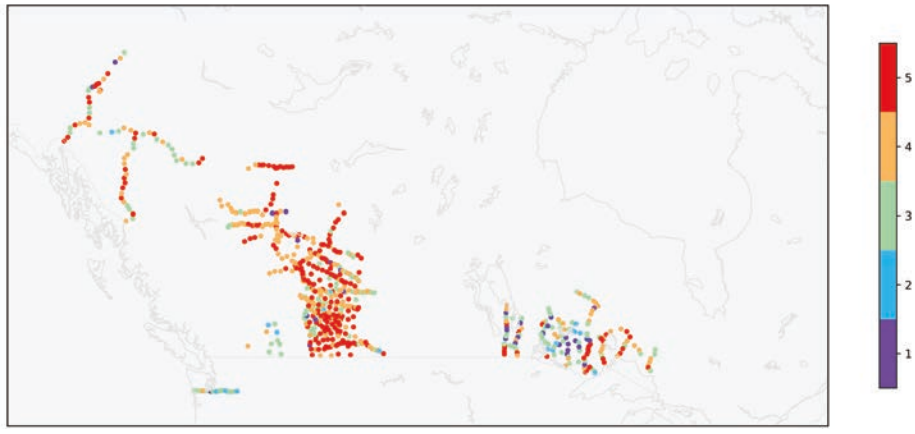

Figure 6. Location and predicted ratings for unrated data sets in North America.

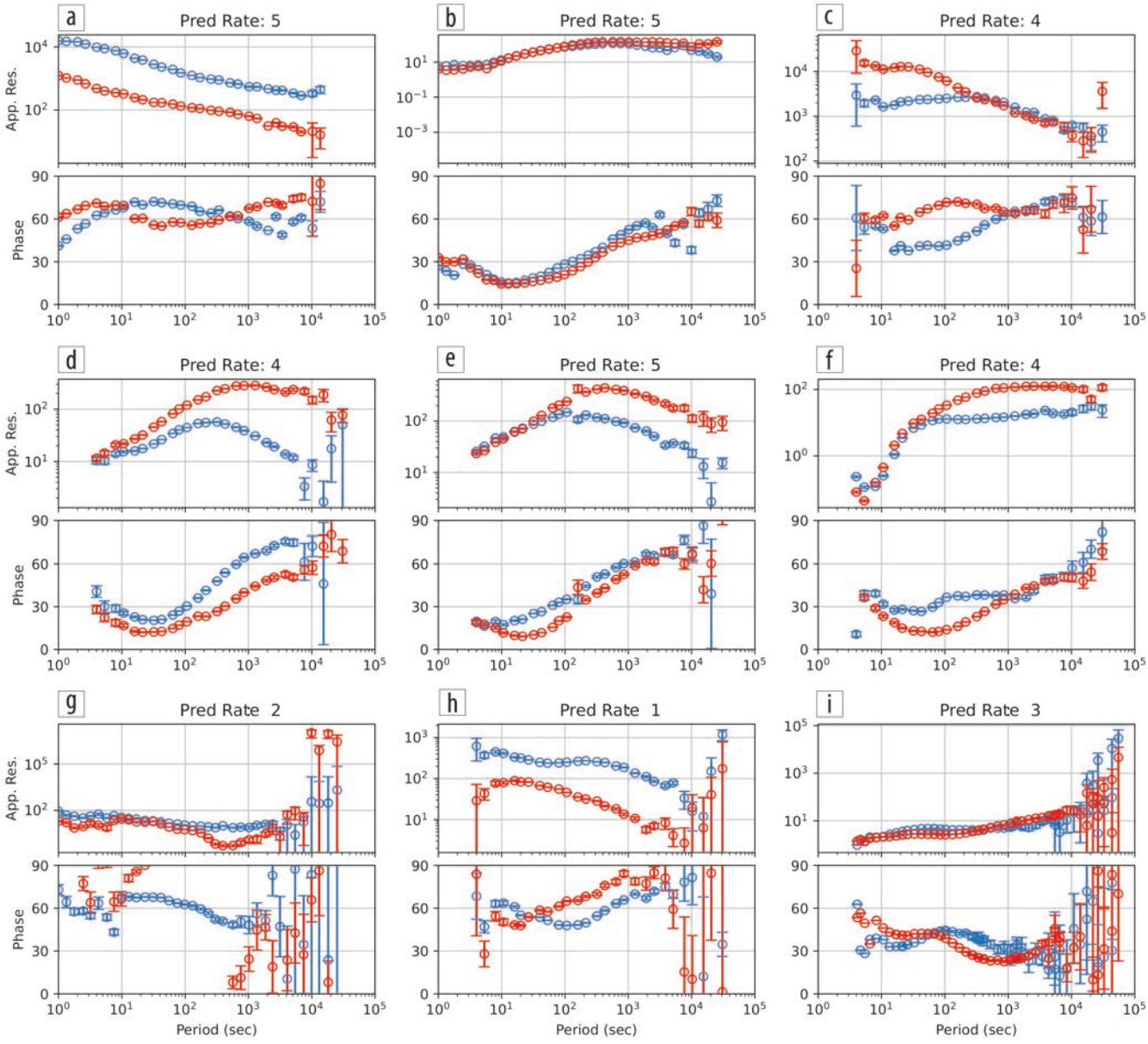

Figure 7. Predicted ratings for apparent resistivities and phases for unrated data sets in North America. Blue lines are *x-y* components. Red lines are *y-x* components. Predicted ratings are for results of pyMAGIQ.

two bands of larger confidence limits on the phase at periods of approximately 10^2 to 10^4 s and 10^1 s. The magnitude of confidence limits on phase generally becomes larger at the extreme short- and long-period ends of the spectrum for MT data of lower quality. The sensitivity analysis indicates that the ML-guided data quality rating is particularly sensitive to uncertainty in the phase at these periods. Previous research did not identify that the uncertainty in the apparent resistivity is a secondary factor to uncertainty in the phase as a driver of overall data quality.

Conclusion

We show that pyMAGIQ is capable of automatically ranking the quality of MT data as effectively as trained human operators. A deep neural network model trained with EarthScope MT project data sets was used to assess the data quality of unrated data sets throughout the United States and Canada. The ML-assigned data quality rating matches the expert human-assigned rating of 98% for the actual training data sets and 91% for the validation data sets. The confusion matrix suggests that when pyMAGIQ deviates from human-assigned data quality ratings, it does so by assigning them to the neighboring rating integer. This is a classification that indicates similar properties in apparent resistivity and phase to the neighboring classification. The apparent resistivities and phases of the validation data sets indicate that the rating predicted by pyMAGIQ in some sites is more reasonable than the previous expert human-assigned assessment rating. We applied pyMAGIQ to the unrated MT data and assigned data quality rankings to the data for the first time.

Sensitivity analysis of deep neural networks suggests that uncertainty in phase and the continuity versus frequency of apparent resistivity, phase, and confidence limits dominate the neural network's assessment of data quality. This is consistent with the standard visual approach for qualification, which focuses on smoothness of parameters and size of the associated confidence limits. On the

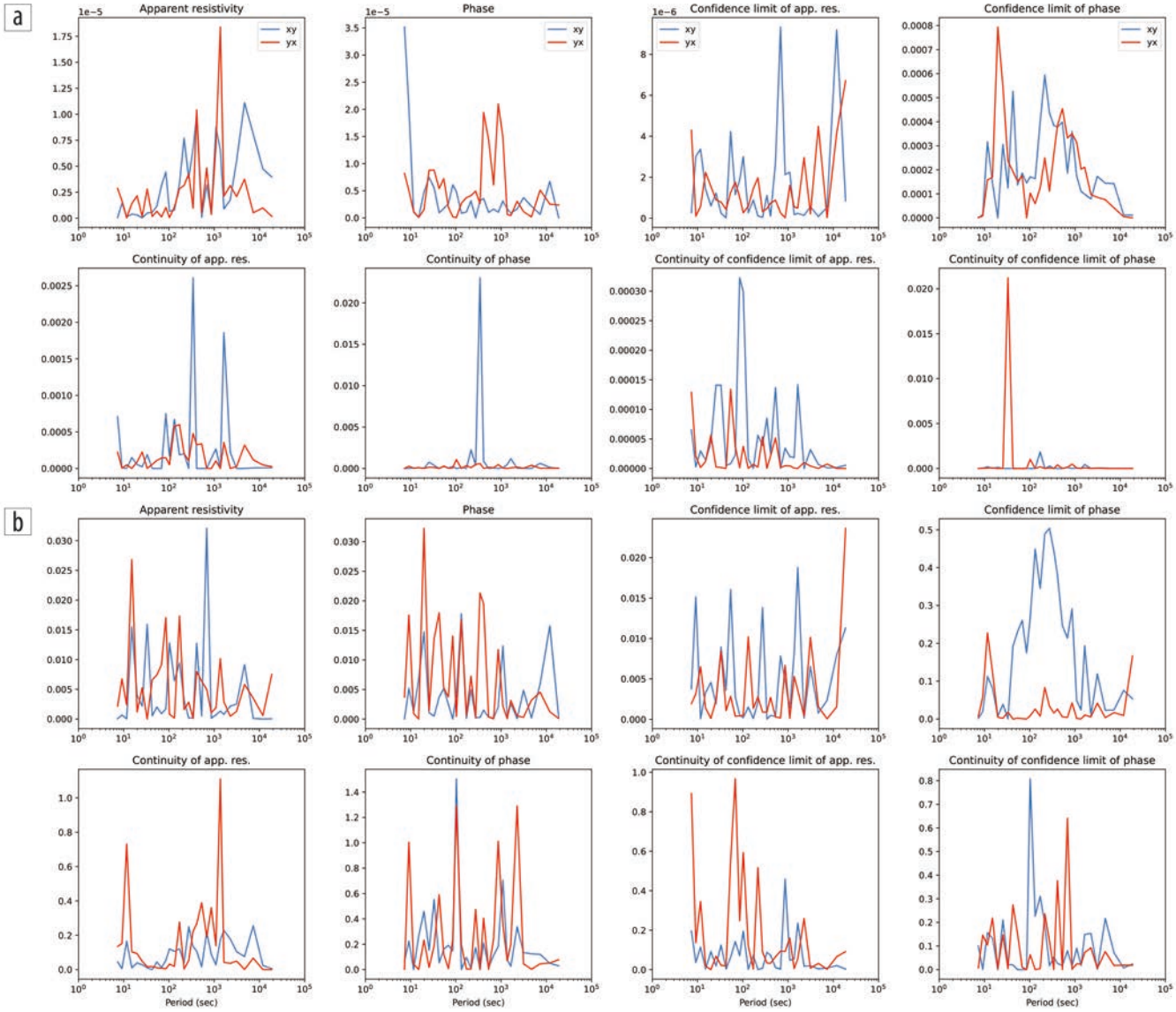

Figure 8. Sensitivity analysis of apparent resistivities and phases for three data sets. (a) Figure 5a. (b) Figure 5b. Blue lines are *x-y* components. Red lines are *y-x* components. The *y*-axis shows sensitivity to the change of parameters.

other hand, sensitivity of the data quality assessment to the confidence limits on the apparent resistivity, and the values of apparent resistivity and phase, are negligible compared to other parameters. This sensitivity analysis suggests that one should focus foremost on uncertainty in the phase and continuity of parameters.

In this study, the amount of MT data was limited. However, as that amount increases, more accurate prediction would be possible because the MT array program continues to map the 3D electrical structure across the United States. The computational requirements for pyMAGIQ to assess the quality of a data set are extremely modest. Hence, in the near future, pyMAGIQ could be used by field crews in real time to assess the quality of field data. As pyMAGIQ provides an arguably less subjective evaluation of the quality of MT impedance data, it could also be used to optimize the methods and parameters of data processing stages. Current research is looking into the practicality of using the trained neural network to randomly select from the large set of crosspowers that are used to calculate MT response functions. This aims to select the subset of crosspowers that yield responses of the highest quality rating. These could potentially serve as pilot estimates against which robust processing methods would refine to yield response functions less prone to biasing effects that escape purely statistical-based methods. Finally, while the examples we have drawn from are for MT data, the principles explored here are applicable across a wide range of geophysical and related data sets. TLE

Acknowledgments

The authors would like to thank the Oregon State University MT team, contractors, lab, and field personnel for assistance with data collection, QC, processing, and archiving. They thank Anna Kelbert for her creation of the original data quality rating index system. They also thank numerous districts of the U.S. Forest Service, Bureau of Land Management, U.S. National Parks, state land offices, and many private landowners who permitted access to acquire the MT transportable array data. The EarthScope MT array was funded through subawards to Oregon State University from IRIS under National Science Foundation Earth Science Division grants and/or cooperative agreements 0323311, 0350030, 0323309, 0733069, and 1261681.

Data and materials availability

Data associated with this research are available and can be accessed via https://doi.org/10.17611/DP/EMTF/USARRAY/TA.

Corresponding author: naoto.imamura@raithing.io

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, et al., 2016, TensorFlow: Large-scale machine learning on heterogeneous distributed systems: arXiv 1603.04467.

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama, 2019, Optuna: A next-generation hyperparameter optimization framework: Presented at the International Conference on Knowledge Discovery and Data Mining.

Chave, A. D., and D. J. Thomson, 2004, Bounded influence magnetotelluric response function estimation: Geophysical Journal International, **157**, no. 3, 988–1006, https://doi.org/10.1111/j.1365-246X.2004.02203.x.

Egbert, G. D., 1997, Robust multiple-station magnetotelluric data processing: Geophysical Journal International, **130**, no. 2, 475–496, https://doi.org/10.1111/j.1365-246X.1997.tb05663.x.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, 2018, A survey of methods for explaining black box models: arXiv 1802.01933.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, 1986, Robust statistics: The approach based on influence functions: John Wiley and Sons.

Huber, P. J., 1981, Robust statistics: John Wiley and Sons.

Jones, A. G., and H. Jödicke, 1984, Magnetotelluric transfer function estimation improvement by a coherence-based rejection technique: 54th Annual International Meeting, SEG, Expanded Abstracts, 51–55, https://doi.org/10.1190/1.1894081.

Khan, J., J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, et al., 2001, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks: Nature Medicine, **7**, no. 6, 673–679.

Schultz, A., 2009, EMScope: A continental scale magnetotelluric observatory and data discovery resource: Data Science Journal, **8**, PIGY6–IGY20, https://doi.org/10.2481/dsj.SS_IGY-009.

Schultz, A., 2019, The evolution of a continent: Thirteen years of EarthScope magnetotelluric three-dimensional imaging of the United States: Presented at the International Symposium on Deep Earth Exploration and Practices.

Sung, A., 1998, Ranking importance of input parameters of neural networks: Expert Systems with Applications, **15**, no. 3-4, 405–411, https://doi.org/10.1016/S0957-4174(98)00041-4.

Tarantola, A., 2005, Inverse problem theory and methods for model parameter estimation: Society for Industrial and Applied Mathematics, https://doi.org/10.1137/1.9780898717921.

Vozoff, K., 1972, The magnetotelluric method in the exploration of sedimentary basins: Geophysics, **37**, no. 1, 98–141, https://doi.org/10.1190/1.1440255.

Zurada, J. M., A. Malinowski, and I. Cloete, 1994, Sensitivity analysis for minimization of input data dimension for feedforward neural network: Proceedings of IEEE International Symposium on Circuits and Systems, 447–450, https://doi.org/10.1109/ISCAS.1994.409622.