Comparing Measurement Response and Inverted Results of Electrical Resistivity Tomography Instruments

Andrew D. Parsekian[1,2], Niels Claes[3], Kamini Singha[4], Burke J. Minsley[5], Bradley Carr[1], Emily Voytek[4], Ryan Harmon[4], Andy Kass[5], Austin Carey[3], Drew Thayer[1] and Brady Flinchum[1]
[1]Department of Geology & Geophysics, University of Wyoming, Laramie, WY
[2]Department of Civil & Architectural Engineering, University of Wyoming, Laramie, WY
[3]Department of Ecosystem Science & Management, University of Wyoming, Laramie, WY
[4]Hydrologic Science and Engineering Program, Colorado School of Mines, Golden, CO
[5]Crustal Geophysics & Geochemistry Science Center, US Geological Survey, Denver, CO

ABSTRACT

In this investigation, we compare the results of electrical resistivity measurements made by six commercially available instruments on the same line of electrodes to determine if there are differences in the measured data or inverted results. These comparisons are important to determine whether measurements made between different instruments are consistent. We also degraded contact resistance on one quarter of the electrodes to study how each instrument responds to different electrical connection with the ground. We find that each instrument produced statistically similar apparent resistivity results, and that any conservative assessment of the final inverted resistivity models would result in a similar interpretation for each. We also note that inversions, as expected, are affected by measurement error weights. Increased measurement errors were most closely associated with degraded contact resistance in this set of experiments. In a separate test we recorded the full measured waveform for a single four-electrode array to show how poor electrode contact and instrument-specific recording settings can lead to systematic measurement errors. We find that it would be acceptable to use more than one instrument during an investigation with the expectation that the results would be comparable assuming contact resistance remained consistent.

Introduction

With nearly a century of development, electrical resistivity is perhaps one of the most mature geophysical techniques. In recent decades, tomographic imaging capabilities have grown with increased computational resources, leading electrical resistivity tomography (ERT) to be a powerful tool for characterizing the subsurface in space and over time with widespread applications to geology, hydrology, the Cryosphere, and critical zone science, among other fields. Thanks to a rich history of development from numerous commercial instrumentation and software companies, high-quality ERT measurements can be acquired by non-geophysicists with only a modest amount of training. However, even with such extensive use of ERT instrumentation throughout the geosciences, comparisons of the reproducibility of measured data under controlled field conditions are absent from the literature. The objective of this investigation is to compare a group of ERT instruments, quantify measurement repeatability and errors, and to identify any disagreement among the results.

One of the most important reasons to compare instrument results is reproducibility. The scientific method dictates that experiments should be repeatable. When we attempt to recreate results of peer-reviewed research, it should be possible to obtain similar results regardless of instrument manufacturer. A second reason to justify a comparison is related to time-lapse measurements, where successive datasets are differenced to isolate a changing environmental variable. Due to the expense of ERT instrumentation and the variety of available instruments, it is not always possible to revisit a site with the same instrument each time, particularly for long-term observations. Demonstrating that changes in the electrical properties of the ground are solely due to environmental variation and not the instrumentation is

 DOI: #10.2113/JEEG22.3.249

essential for accurate interpretation. A third justification is related to concurrent-time-and-space measurements. For example, in comparative hydrology experiments where multiple watersheds are simultaneously instrumented with time-lapse ERT, identical instruments may not be available for each concurrent deployment. Regardless of the instruments deployed in different watersheds, knowledge that measurements are comparable and are not a function of the instrument is important.

Most comparative investigations are undertaken with the purpose of identifying the appropriate method to use for a particular experiment, or to evaluate errors and enable more robust interpretation. Examples of such comparative studies include using different geophysical techniques to image the same target (*e.g.*, Gallipoli *et al.*, 2000; Sass *et al.*, 2008; Yaramanci *et al.*, 1999), or relating non-invasive geophysical measurements with direct physical property measurements (Sudduth *et al.*, 2003; Skinner and Heinson, 2004). Comparing the same class of instrument and its associated physics is less common; however, there are examples such as Perri *et al.* (2012), who tested borehole ERT alongside surface ERT and found that comparing the resolution between the two types of measurements was difficult. Comparing the measurements from two instruments is also uncommon, but it has been done for electromagnetic instruments (Abdu *et al.*, 2007) and distributed temperature sensors (Tyler *et al.*, 2009). There are also examples of comparing elements of an ERT survey, such as array type, while using the same instrument (Seaton and Burbey, 2002; Dahlin and Zhou, 2004). In each of these cases there is only one ERT instrument used, working under the assumption that any other instrument will provide representative data that can be reproduced within measurement error.

Since ERT is commonly used in Earth science, especially in long-term studies where multiple instruments may be used, there is a value in validating the assumption that different ERT instruments will produce similar data. We motivate our study with the following questions: 1) how similar are measured apparent resistivity (ρ_a) values amongst all instruments? 2) how much of the measurement error can be attributed to systematic measurable sources (*i.e.*, contact resistance; R_c) and how much must be left to unknown sources? and 3) do all instruments record similar measurement errors—whether reciprocal or stacking errors—and how do these influence the final inversion results?

We undertake this experiment as objectively as possible with no bias towards personal preferences of instrument manufacturer. The choice of instruments was determined by availability, and we recognize that there are other manufacturers that could have been included. It is not our intention to pit different instruments against each other to identify a "best" or "worst", but rather to make general statements about the expected errors that could be encountered if using different instruments within the same experiment.

Methods

Electrical Resistivity Tomography Principles

The physical and mathematical principles governing ERT measurements are thoroughly covered in many book chapters and review papers (*e.g.*, Grant and West, 1965; Herman, 2001; Binley and Kemna, 2005; Binley 2015) and only a brief introduction will be provided here. The principle of resistivity is rooted in Ohm's law, which states that the current moving between two points in a conductive material is proportional to the potential difference between the two points:

$$R = \frac{V}{I} \tag{1}$$

where V is the measured electrical potential difference in volts, I is the applied current in amperes, and R is the resistance expressed in Ohms. Electrical potential is measured between two electrodes (often referred to as $P_1 - P_2$ or M - N) as a function of the current driven between two other electrodes ($C_1 - C_2$ or A - B); a set of four electrodes used in this manner may be termed a quadrupole. Accounting for quadrupole geometry, measured resistance can be translated to apparent resistivity, or ρ_a, in units of Ohm-m. A geometric factor (K) that relates resistance to apparent resistivity varies depending on the array type (*i.e.*, geometric location of current and potential electrodes on the Earth's surface) used for measurement, where C_1P_1 indicates the distance between the first transmitting electrodes and the first receiving electrode, P_1C_2 indicates the distance between the first potential electrode and the second transmitting electrode, and so forth:

$$\rho_a = RK = R \times 2\pi \left(\frac{1}{C_1P_1} - \frac{1}{P_1C_2} - \frac{1}{C_1P_2} + \frac{1}{P_2C_2} \right)^{-1} \tag{2}$$

Each array type is characterized by a different geometric quadrupole configuration of C_1C_2-P_1P_2 with characteristic geometric factors, sensitivities, and signal-to -noise ratios.

A 2D or 3D model with estimates of the true spatial distribution of resistivity values in the subsurface is

determined through inversion of the measured resistance data and estimates of error (*e.g.*, Oldenburg, 1978).

Reciprocal measurements compare differences associated with the directionality of the measurement by swapping the current and potential electrode pairs whereas stacking errors measure the repeatability of an individual measurement. Following the principle of reciprocity, swapping the current and potential electrode pairs should produce the same apparent resistivity measurement (Habberjam, 1967; Binley *et al*., 1995; Slater *et al*., 2000). Since the reciprocal measurements of the dipole-dipole array have the same electrode separation for the transmitting current pair as for the forward measurement transmitting pair (in comparison to Wenner or Schlumberger arrays where the reciprocal results in swapping of the transmitter electrode separation with the receiver electrode separation), values from forward and reverse should not have different geometry-related noise characteristics and can be used to yield an estimate of error (Zhou and Dahlin, 2003; Deceuster *et al*., 2013). In a nested array such as Wenner, the large transmit offset and smaller receiver offset serves to improve signal to noise, while an inverse Wenner would have a small current offset and large receiver offset that would enhance noise. Alternatively, it is common to repeat, or stack, each measurement several times – statistics may then be drawn from the group of stacked measurements to assess error, although stacking errors are thought to be less reliable for quantifying total measurement error (Deceuster *et al*., 2013). Collection of reciprocal measurements can greatly increase survey duration, beyond the time required to collect multiple stacks, which can be problematic when using time-lapse ERT to evaluate processes on a short time scale.

Sources of Measurement Error

Often, all sources of measurement errors are aggregated into a single value prior to data-quality assessment or being incorporated into a weighted inversion scheme. This is partially for simplicity because evaluating each source of error independently may be time consuming, but it is also because some sources of error are poorly understood, unquantifiable, or unknown. Measurement errors fall under three categories: 1) experimental design errors caused by inaccuracy of electrode placement; 2) quantifiable errors, such as the precision of the resistance measurement (as evaluated by stacking or reciprocals) and contact resistance (can be measured once or more by each instrument, though not directly used in data analysis), or differences in instrument engineering specification (such as the available transmit current amplitudes or the transmitted waveform); and 3) other unknown errors that are difficult to quantify, such as the accuracy of the resistance measurement, proprietary instrument design specifications, or electrode polarization. The effect of electrode polarization on measurement electrodes that were recently used to inject current is known (*e.g.*, Dahlin, 2000) but it is not independently measured by the instrument. In this study, common sources of error such as differences in electrode material (LaBrecque and Daily, 2008) or electrode positioning (Szalai *et al*., 2009; Oldenborger *et al*., 2005) are the same for each instrument used and are therefore not addressed in our analysis.

Field Site

Measurements were acquired on January 29, 2016 at the Red Buttes test site about 19 km south of Laramie, Wyoming. The geologic structure consists of three shallow-dipping, sub-horizontal layers in the near surface: an unconsolidated clay loam, the Satanka Formation, and the Epsilon Sandstone Member of the Casper Formation. The unsaturated 3-4 m thick clay loam is a low-resistivity layer that overlies the lower-resistivity silty-claystone of the Satanka Formation, found to 12 m depth. Finally, starting approximately 12 m below the surface is the consolidated Casper Formation Epsilon sandstone. The Epsilon sandstone is generally observed to have a much higher resistivity than the overburden. Although it was wintertime and contact resistance within the near surface soils was predicted to be higher than normal due to partially frozen soils, preceding warm weather had melted most of the snow and the ground was $<$50% snow covered; various patches of remaining snow were less than 20 cm deep. Grasses and forbs occupied ground that was not snow-covered. During the day, air temperature rose from a low of -3°C in the morning to 3°C in the afternoon.

Instruments

Four brands of commercially produced ERT instruments were used for the test: Iris Instruments SyscalPro (Orleans, France), Lippmann 4PL (Shaufling, Germany), Multi Phase Technologies (MPT) EIT-2000 (Sparks, Nevada, USA), and Advanced Geosciences Inc. SuperSting R8 (Austin, Texas, USA). Following the example set in previous instrument comparison research (Tyler *et al*., 2009), the instruments used in this study will be referred to only numerically (*e.g.*, Unit 1, Unit 2, etc.), although not in the order listed above, because the manufacturer is not relevant to the evaluation. The purpose of this work is to understand the character of how instruments may produce different results, not to

Table 1. Instruments' engineering parameters that cannot be directly changed by the user.

	Maximum output parameters					
	Current [A]	Peak-to-peak voltage [V]	Power [W]	Waveform	Rx waveform sampling freq. [Hz]	No. avail. channels
Iris SyscalPro	2.5	1,000	250	+ -	100	10
Lippmann 4PL	0.1	380	10	+ -	300	1
MPT EIT-2000	5.0	1,000	2,500	+ off - off	Unknown	2
AGI SuperSting R8	2.0	800	200	+ - - +	5,000	8

find the "best" one. There were three units of one brand available (Units 4, 5 and 6) to enable comparison between the same model of instrument, bringing the total number of instruments in the comparison to six. An effort was made to keep acquisition parameters as uniform as possible in an attempt to collect the most comparable data, but ultimately some variation was unavoidable for two reasons: 1) some parameters have predefined ranges of values and there is no equivalent among all instruments (*e.g.*, transmitter current injection time); and 2) to ensure that the measurements could be completed in a single day and therefore retain consistent Earth electrical properties, after multichannel instruments completed measurements, the transmit time for one- and two-channel instruments was reduced. We recognize that current injection time may influence errors in the data, however each instrument has a set of predefined current injection times such it is impossible to set them precisely equal to each other.

The design parameters and electrical specifications for each instrument unit that are not variable in the case of this experiment are listed in Table 1. Common acquisition parameters are listed in Table 2. A user error was made during programming of Unit 3 (see details in Discussion) where it was improperly set up to inject a lower transmit current, although those data have been retained for comparison as described.

Electrode Array and Measurement Sequence

Measurements were conducted on one 78 m-long array with 40 electrodes at 2 m electrode spacing. Although the terrain was relatively flat, the line location was measured by GPS and electrode topography was corrected using data from an airborne LiDAR survey collected in 2015 (vertical accuracy <2 cm RMS in comparison to calibration control points). Stainless-steel electrodes, 30 cm long, were deployed the day before and were not disturbed during the course of subsequent measurements. We ensured that all takeouts on the cables were off the ground to eliminate possible coupling, either by direct connection between the electrode and cable, or by draping the cable across the electrode when jumper cables were used. Electrodes 1-30 were inserted about 25 cm into the ground, while contact resistance was intentionally degraded for electrodes 31-40 by only inserting them about 10 cm into the ground.

A dipole-dipole sequence was used for all measurements with a maximum spacing between current electrodes (max C_1C_2 spacing) of 6 electrodes and maximum spacing between current and potential electrodes of 5 electrodes. A dipole-dipole array was chosen because the multi-channel instruments could exploit an acquisition speed advantage as well as the reciprocity benefits described previously. Reciprocal measurements (forward, followed by reverse where the current and potential electrodes were swapped) were taken for all quadrupoles after the full set of forward measurements. The measured values used in our analysis are only from those quadrupoles that are common to all datasets ($n = 582$).

Table 2. Measurement parameters. SuperSting stacks reported are ½ of what is measured since each 'cycle' consists of two positive and two negative phases.

	Stacks	Time before measurement [s]	Transmit time [s]	Transmit freq [Hz]	Max I [mA]	Min I [mA]	Duration [min]
Iris SyscalPro	5	0.58	1	1	415	2	50
Lippmann 4PL	5	0.40	0.4	2.5	50	15	77
MPT EIT-2000	5	0.10	0.167	6	3	0.5	Unknown
AGI SuperSting R8	2	0.12	1.2	0.83	304	69	36

A follow-up investigation was conducted in April 2017 at the same location with the same array geometry and dipole-dipole sequence to investigate how transmitted current amplitude may affect data quality. In this case, only one instrument was used (Unit 4) at a transmit time of 0.4 s with 3 stacks, and the maximum allowable transmit current was limited to: 1000 mA, 500 mA, 100 mA, 50 mA, 10 mA and 5 mA.

Calculating Contact Resistance

R_C is measured differently by each instrument, typically for the purposes of indicating how well electrodes are connected to the ground and to set the input current. The procedure for each instrument is as follows:

- Lippmann: The software performs an auto-calibration (detailed below) to record R_C for the first three electrodes on the line and then each of electrodes 4 - 40 is compared with the autocalibrated first electrode. A list of R_C with length equal to the number of electrodes is stored once at the beginning of the measurement sequence.
- SuperSting and MPT: The R_C is measured at every line of the acquisition sequence between the current (transmit) pair of electrodes for that command.
- Iris: The R_C is measured for every line of the acquisition sequence between the potential (Rx) pair of electrodes for that command.

Here we convert R_C measured between pairs of electrodes to the R_C associated with each individual electrode in order to have a common comparison between instruments. Contact resistance between two electrodes is the sum of R_C at each electrode (*e.g.*, C_1+C_2), plus the resistance of the Earth along the path between those two electrodes. To calculate the R_C of one electrode we need the resistance of at least one reference electrode.

Consider three electrodes where we want to know contact resistance: R_C,1 R_C,2 and R_C,3 from measurements R_C (1,2), R_C (2,3), and R_C (1,3). We can set this up as a linear system of equations, Ax = b:

$$\begin{bmatrix} 1 & 1 & 0 & \cdots \\ 0 & 1 & 1 & \cdots \\ 1 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} Rc,1 \\ Rc,2 \\ Rc,3 \\ Rc,N \end{bmatrix} = \begin{bmatrix} Rc(1,2) \\ Rc(2,3) \\ Rc(1,3) \\ \vdots \end{bmatrix} \quad (3)$$

Each row of A corresponds to a measured two-electrode R_C value and states that the measured R_C is the sum of the two single-electrode R_C values involved in the measurement. The columns of A that correspond to electrodes used in the R_C measurement are 1; the others are zero. This system of equations is easily built for any suite of R_C (a,b) measurements, possibly with many more rows than columns. We then solve using the method of least squares to get the single-electrode R_C values, x, where $x = inv(A'*A)*A'*b$. Estimates of R_C may be biased in the case where the earth path resistance is significant compared to the electrode contact resistances. However, this approach is useful in identifying the relative magnitudes and patterns of electrode contact resistance in many cases.

Inversion

We recognize that if the instruments measure the same data, then inverted result should also be the same. However, inversions that weight data based on instrument-derived data errors can yield different subsurface models even if the absolute R values are identical. To explore this issue, we invert a dataset from each tested instrument, along with the associated data error. To address the question of whether or not stacking errors and reciprocal errors impact resultant inversions differently, we perform separate inversions that incorporate the stacking error or reciprocals in the weighting process. All data were inverted using the Occam's type inversion R2 software (Binley and Kemna, 2005) with a triangular mesh. Quadrupoles with a reciprocal or stacking error higher than 10% were removed from the dataset prior to inversion. After this procedure, we construct an error vector for each collected set of quadrupoles after Koestel *et al*. (2008). The error function allows for the assignment of a weight to each quadrupole that is associated with its standard deviation. Measurements within each dataset are binned by R into groups of 25 to 30 and an error function is calculated as:

$$\varepsilon_{N/R} = a_{N/R} + b_{N/R}|R_{N/R}| \quad (4)$$

where ε is the standard deviation (Ohm) of each bin, R is the mean resistance value (Ohm), and the constants a and b are defined by a best linear fit. For every collected dataset, this error function is defined and the weights for each quadrupole are assigned based on Eq. 4. The regularized inversion of each dataset is carried out with a model norm consisting of the smallest deviation from a homogeneous starting model. This starting model has a resistivity value of 80 ohm-m, which is within 4 ohm-m of the average apparent resistivity of each collected dataset. The desired misfit for each inversion is an RMS error of 1.0 or less, defined by the misfit between observed and modeled data scaled by their respective weights.

The sensitivity of the inverted model is assessed with two methods. The first assessment follows the depth-of-investigation (DOI) method as described by Oldenburg and Li (1999), where different homogeneous reference models are used to determine where the data inform the model. In our case, the first reference model has a resistivity of 50 Ohm-m and the second has a resistivity of 500 Ohm-m. The difference in inversion results is used as an indicator of sensitivity, S, according to:

$$S(x,z) = \frac{m_{500}(x,z) - m_{50}(x,z)}{\log10(500) - \log10(50)} \quad (5)$$

where m_{500} and m_{50} are the inverted resistivities in $\log_{10}$. Values of S close to 0, indicate a higher sensitivity to the data, while higher values indicate the inversion results are affected by the choice of starting model. We use these values of S to mask our inversion results when presenting them. Oldenburg and Li (1999) suggest that S values between 0 and 0.2 are sensitive to the data and presented without transparency. Values between 0.2 and 0.4 indicate reduced sensitivity and are presented with semi-transparency. Areas with sensitivity values greater than 0.4 are not shown.

A second assessment of sensitivity is provided based on a sensitivity matrix as described in detail in Binley and Kemna (2005). Instead of using a full resolution matrix, which can be computationally expensive, this method uses a sensitivity matrix comprising of a data weighing matrix and the Jacobian matrix evaluated for the final model. Higher values indicate areas with higher sensitivity, while lower values indicate the opposite (Binley and Kemna, 2005). We present this sensitivity assessment for every inverted measurement set.

Effects of Electrode Contact and Instrument-Specific Recording Settings

A separate experiment was conducted in Denver, Colorado in September 2016 to investigate in more detail some of the specific sources of systematic measurement error. Here, Unit 5 was deployed in a dry sandy soil, with a four-electrode array and 1 m between each electrode. Forward and reciprocal measurements were made using a dipole-dipole geometry, and for three electrode coupling scenarios: (1) all four electrodes fully planted (approximately 25 cm) into the ground; (2) two fully planted electrodes and two only partially inserted into the ground (approximately 10 cm); and (3) all four electrodes only partly inserted into the ground. During each measurement, the receiving electrode pair was monitored with a Picotech 5444b digital oscilloscope (Pico Technology, Tyler, Texas) using a high sampling frequency to capture the entire measured waveform. This is a 1 M Ohm impedance instrument that is high enough as to not cause significant measurement error in this case, however if R_C is very high a larger impedance oscilloscope would be required. Finally, we tested the importance of the instrument delay time before measurement (Table 2), with 10 repeat measurements and the fully planted electrode scenario (#1) for each of three different delay times: 60 ms, 100 ms, and 200 ms.

Results

Apparent Resistivity and Resistance Pseudosections

Given that each instrument transmits a different amount of current for each quadrupole and resistances can vary by orders of magnitude, the most raw form of data for easy comparison is apparent resistivity, ρ_a. Figure 1(a) shows an apparent resistivity pseudosection for each instrument; qualitatively, the structure of all of these results appears similar. The mean and standard deviation of resistance at each quadrupole collected across the six instruments was calculated and plotted as a single pseudosection with color values representing the ratio of these statistics (coefficient of variation, or standard deviation divided by the mean) as shown in Fig. 2. From this figure, we can see that the errors are distributed throughout the dataset with a possible pattern of higher errors clustered around 60 – 78 m along the line, where R_C was intentionally degraded, and for quadrupoles with large pseudo depth.

Error Comparisons

Figure 1(b) shows resistance pseudosections with the difference between forward (R_F) and reverse (R_R) reciprocal measurements normalized by the average for each instrument. Unit 1 (Fig. 1(b.1)) and Unit 2 (Fig. 1(b.2)) show similar spatial patterns in reciprocal error, with larger reciprocals observed towards the 60 – 78 m section of the line with degraded R_C. Both of these datasets also have larger reciprocal errors at larger pseudo depths. In both datasets, we observe that nearly all of the measurements have $(R_F\text{-}R_R)/\text{mean}(R) < 0.1$. The reciprocal errors from the Unit 3 do not follow an obvious spatial pattern, and overall the errors are larger than all other instruments. Each of Units 4, 5 and 6 (Fig. 1(b.4),1(b.5),1(b.6)) has a similar spatial pattern, with the largest errors again observed between 60 – 78 m along the line. Overall, Units 4, 5 and 6 have reciprocals that are the smallest of all instruments.

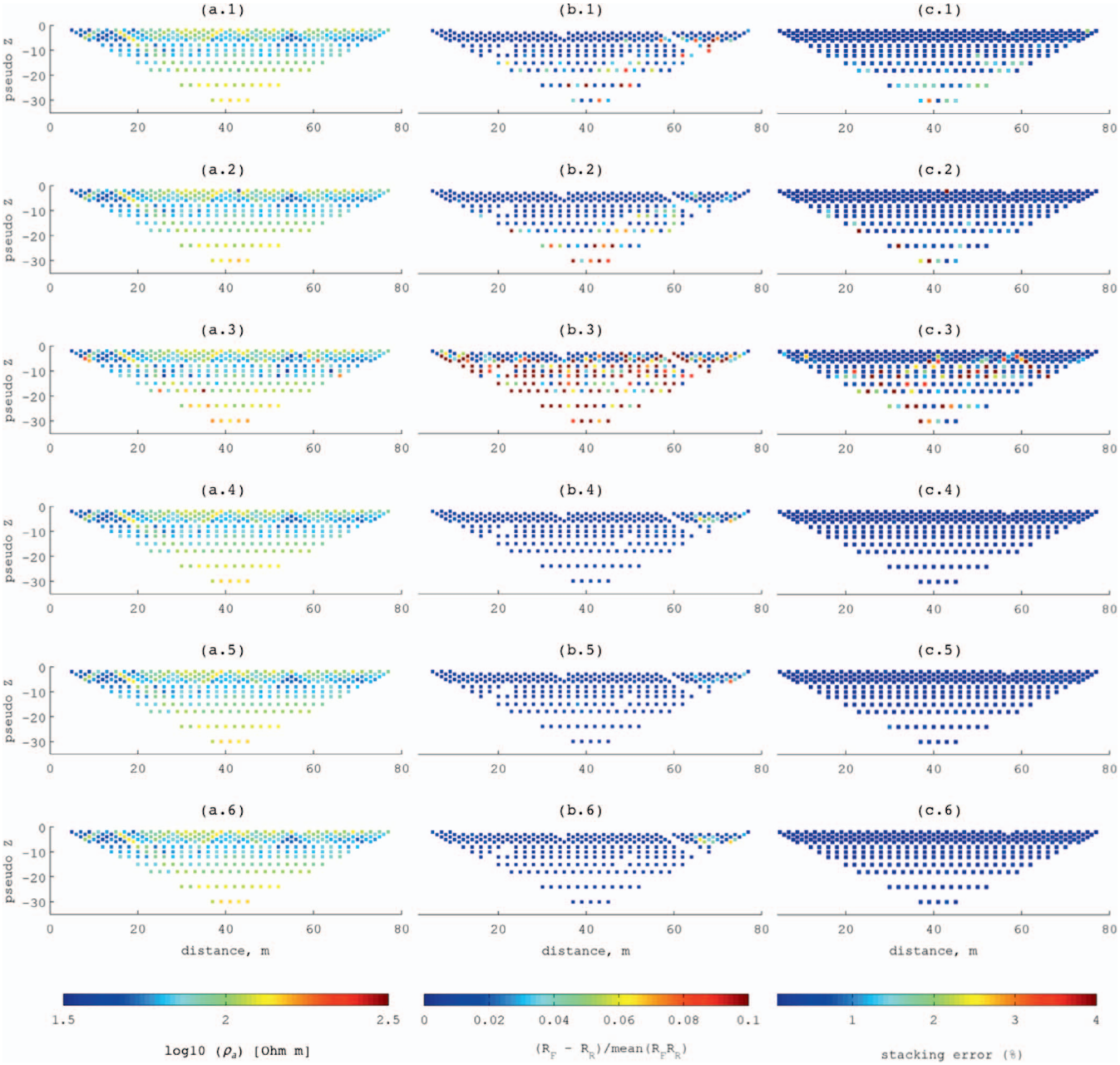

Figure 1. (a) Pseudosections with apparent resistivity plotted for each instrument using all datapoints recorded. (b) The reciprocals within each separate dataset and plotted the difference in percent as a pseudosection. (c) Error reported by the instrument as an average of stacks. The colorscale is equal for all panels in each column of this figure.

Stacking errors are automatically calculated by each instrument, and reported as coefficient of variation shown in Fig. 1(c). Unit 1 recorded generally low stacking errors (Fig. 1(c.1)), with 90% of the values less than 1% error. Several points with stacking errors in the range of 2% - 4% occurred on the 60 – 78 m portion of the line, and at larger pseudo depths. Unit 2 also recorded low stacking errors (Fig. 1(c.2)), with 94% of the values below 1% error, except for several measurements around 5% error near the left edge of the pseudosection, and again in the 60 – 78 m portion of the line. Although Unit 3 has the most points with stacking error >1% among all of the instruments (Fig. 1(c.3)), 73% are below 1%. In this case, the points with elevated errors (*i.e.*, >1%) are distributed approximately randomly throughout the pseudosection. Units 4, 5 and 6 (Fig. 1(c.4), Fig. 1(c.5), Fig. 1(c.6)) show uniformly low

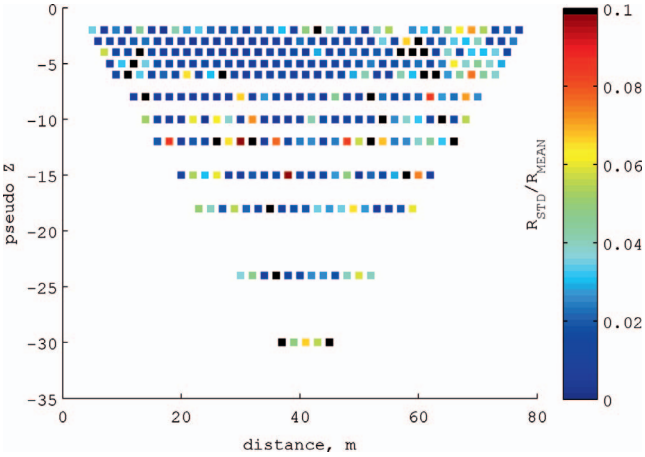

Figure 2. Coefficient of variation resistance pseudosection for points common to all instruments. Measurements with highest coefficient of variation (>0.10) account for 9.2% of the dataset and are masked in black. Of the remaining data, 52% have a coefficient of variation <0.02, and 79% have coefficient of variation <0.05.

stacking errors <1%, with slightly elevated errors in the 60 – 78 m portion of the pseudosection.

Figure 3 compares the reciprocal error with stacking errors at the same quadrupoles using a cross plot. However, we observe that there is a bias towards the points falling preferentially below the one-to-one line, indicating that the reciprocal measurements are generally larger than stacking errors. It is important to note, however, that stacking was done up to five times while a reciprocal only consists of two measurements.

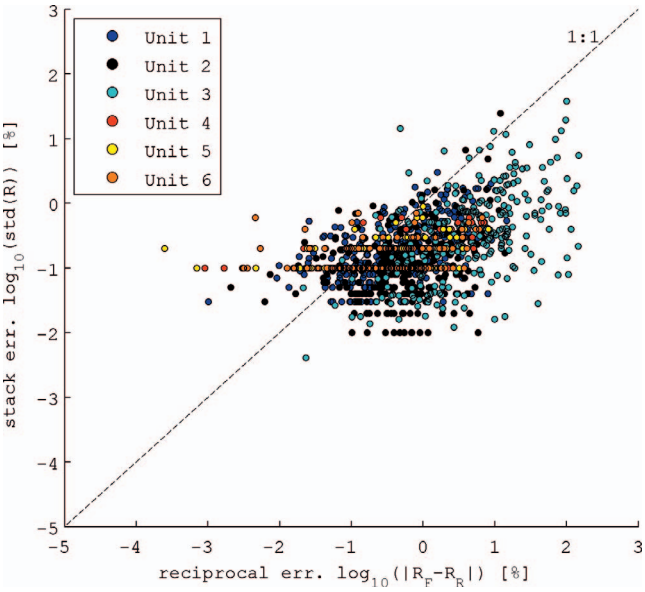

Figure 3. Cross plot of the stacking error vs. the reciprocal error.

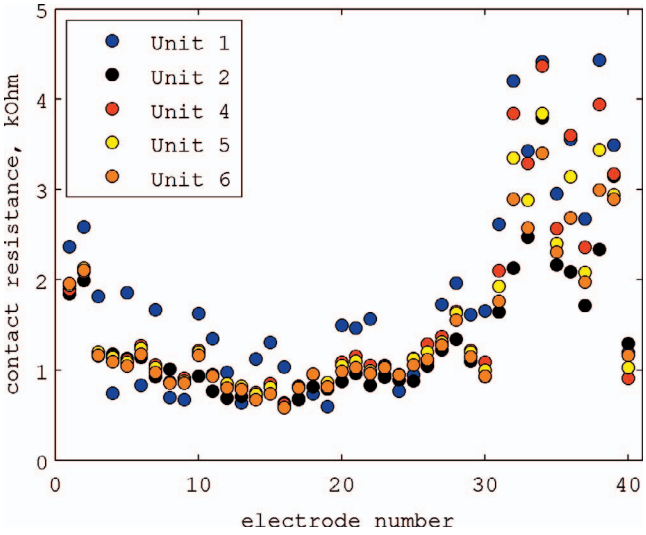

Figure 4. Contact resistance (R_C) for each electrode measured by each instrument.

Contact Resistance

Contact resistance is determined by the electrical connection between the ground and the electrode, and therefore the following observations are related to the set of electrodes with degraded R_C and the overall conditions of R_C during the test. As expected, electrodes 31 – 40 had higher R_C (mean $R_C = 2.6$ kOhm , std = 0.99 kOhm; Fig. 4) than the rest of the electrodes (mean $R_C =$ 1.1 kOhm, std = 0.38 kOhm). Electrodes 1 and 2 were fully inserted into the ground, but ended up having slightly higher R_C than electrodes 3- 30. For electrodes 1 – 30, Unit 2, 4, 5, and 6 recorded very similar R_C, within 0.1 kOhm – 0.2 kOhm. About half of Unit 1's measurements are in the same range as the other instruments, while the other half are ~0.5 kOhm larger. In the zone of degraded R_C the only discernable pattern is that values recorded by Unit 2 are generally the lowest. Otherwise, the variability in R_C measurement is larger, ~2 kOhm. Unit 3 did not record R_C.

To visualize the relationship between measurement errors and R_C, the data were sorted to find each quadrupole that used a particular electrode in each dataset, then we extracted the stacking error for all measurements that included that electrode and took the mean (standard deviation error bars are not shown to preserve legibility of the figure). Stacking error is plotted against R_C in Fig. 5(a), and a similar comparison for reciprocal errors is shown in Fig. 5(b). A linear fit was calculated for each dataset in Fig. 5 although lines are not plotted in order to retain clarity of the figure. The slope and coefficient of determination for each fit line are presented in Table 3. Slopes for stacking errors ranged from 0.012 to 0.050 % $k\Omega^{-1}$ and r^2 ranged from

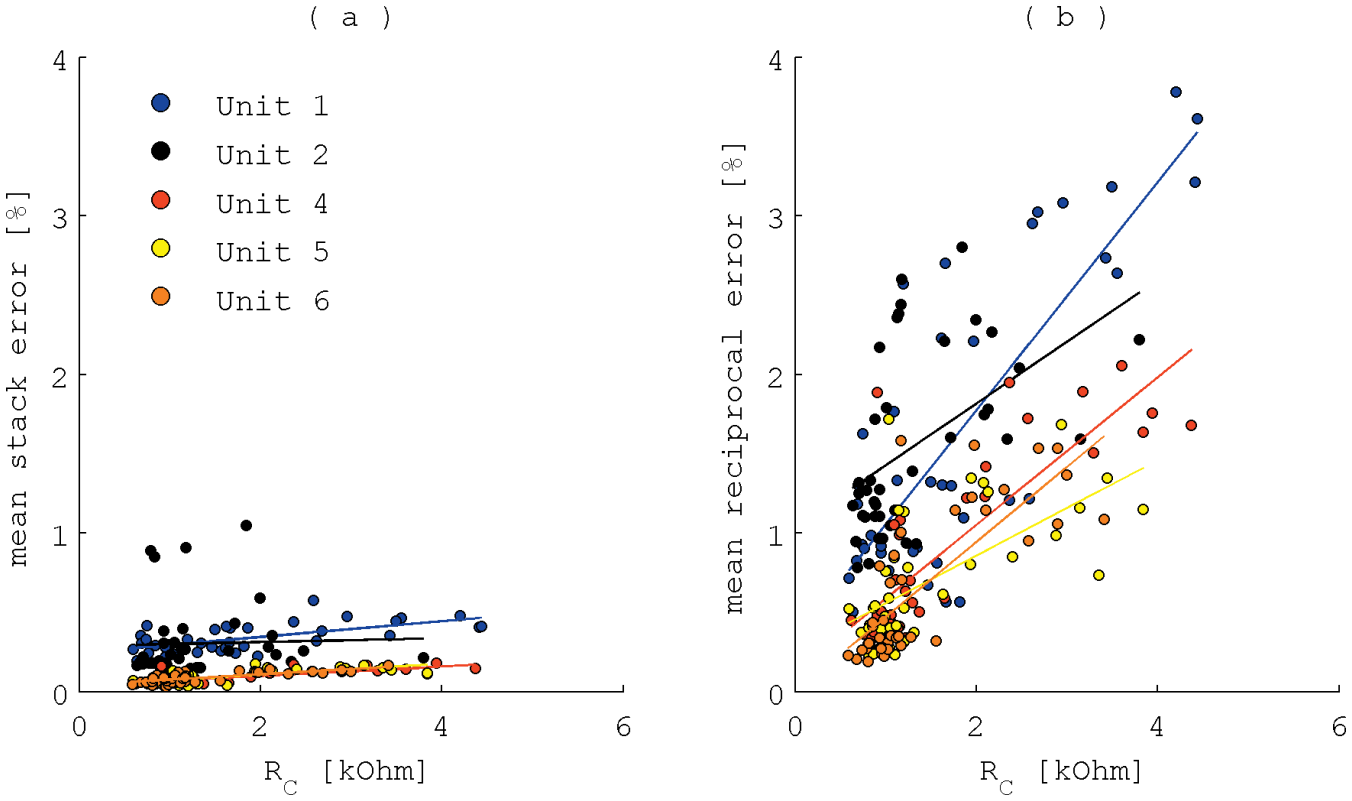

Figure 5. (a) The relationship between contact resistance and stacking error. (b) The relationship between contact resistance and reciprocal error. Lines of best fit are shown: associated slopes, coefficients of determination, and *p*-values are included in Table 3.

<0.01 to 0.59. Slopes for reciprocal errors ranged from 0.30 to 0.72 % $k\Omega^{-1}$ with associated r^2 of 0.24 to 0.66.

Statistical Comparisons of Measured Data

We investigated differences among several parameters in our dataset using the Kruskal-Wallis test (Breslow, 1970), a non-parametric one-way Analysis of Variance (ANOVA) on ranks that compares whether data originate from the same distribution. All statistical analysis scripts are in the standard 2012b version of the MATLAB (Mathworks, Inc.) Statistics Toolbox. All statistical comparisons shown are relative to the 95% confidence interval, with a critical $p = 0.05$. The results of the statistical comparisons (Fig. 6) are shown as box plots (McGill *et al.*, 1978) where the center horizontal line of each box indicates the median, the top and bottom edges indicate the upper and lower quartile, whiskers are 1.5x the interquartile range and the crosses are outliers. Overlapping notches in adjacent boxes indicate that there is not a significant difference between the median values. For the apparent resistivities (sample size, n =582) corresponding to all coincident quadrupoles (Fig. 6(a)), we fail to reject the null hypothesis that these instruments measured the same apparent resistivity (p=0.74), indicating that all instruments produce apparent resistivity values that are not significantly different. Considering R_C (Fig. 6(b)), we fail to reject the null hypothesis that R_C was constant during the day of the test (p=0.31). Unit 3 is not included in this analysis because the R_C was not recorded in the field. With regards to reciprocal (Fig. 6(c)) and stacking (Fig. 6(d)) errors, we reject the null hypothesis that the instruments

Table 3. Linear fit slopes and coefficients of determination for data shown in Fig. 5. *P*-values show that the line fits are significant at the 95% confidence interval, except where underlined.

	Stacking			Reciprocal		
	slope [% $k\Omega^{-1}$]	r^2	*p*-value	slope [% $k\Omega^{-1}$]	r^2	*p*-value
Unit 1	0.050	0.39	$1.5\text{x}10^{-5}$	0.719	0.65	$<10^{-6}$
Unit 2	0.012	<0.01	$\underline{8.2\text{x}10^{-1}}$	0.389	0.24	$1.3\ \text{x}10^{-3}$
Unit 4	0.029	0.53	$<10^{-6}$	0.466	0.66	$<10^{-6}$
Unit 5	0.034	0.51	$<10^{-6}$	0.302	0.39	$1.8\ \text{x}10^{-5}$
Unit 6	0.033	0.59	$<10^{-6}$	0.474	0.59	$<10^{-6}$

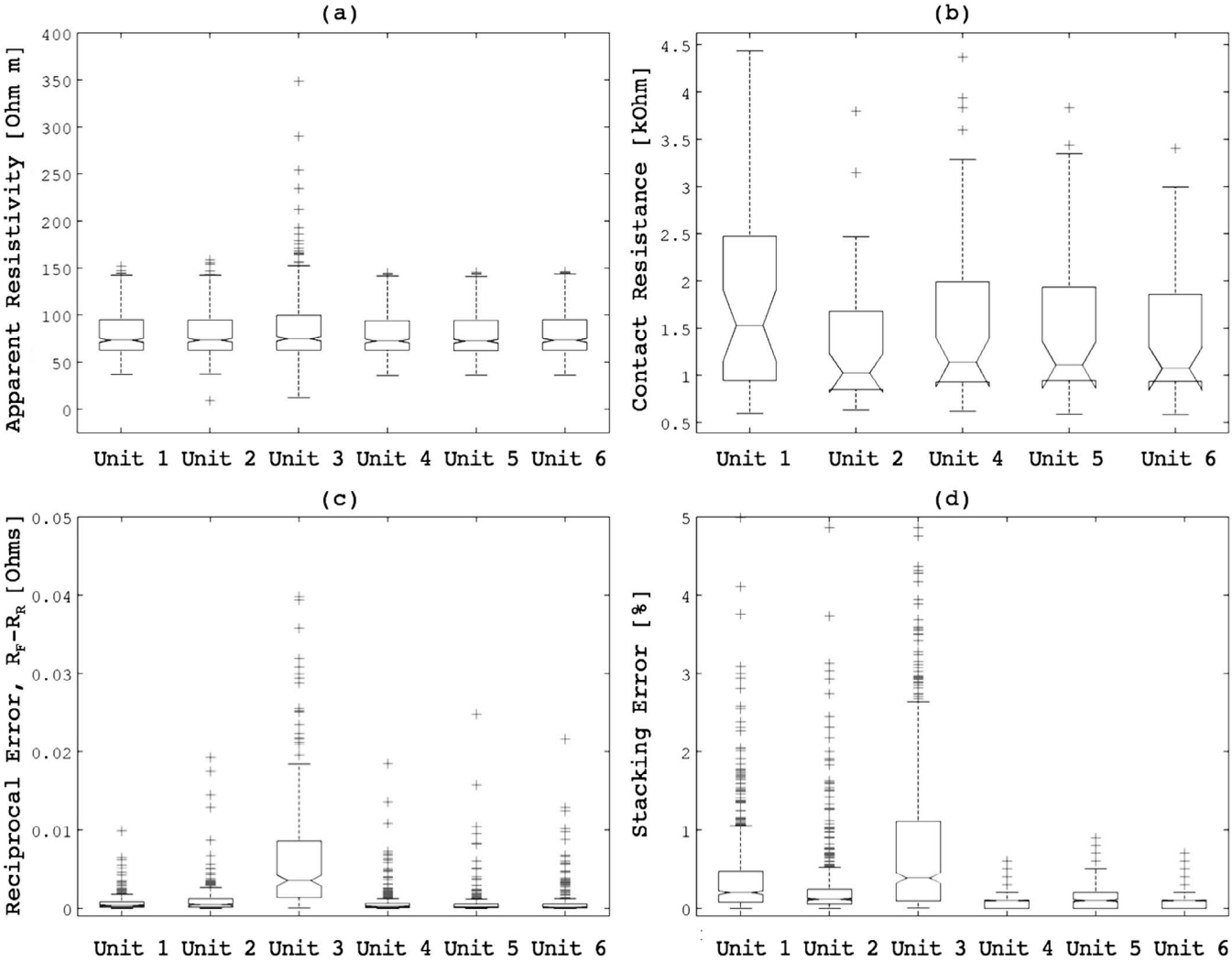

Figure 6. Distributions used in statistical analysis for (a) apparent resistivity, (b) contact resistance, (c) reciprocal errors, and (d) stacking errors.

have the same errors ($p<10^{-6}$ in both cases) indicating that the errors associated with each instrument are not statistically similar. We also reject the null hypothesis even when omitting the visually obvious outlier, Unit 3 (Fig. 6(c) and 6(d)) from the statistical analysis.

Inverted Results

All statistics related to inversion results in the following paragraphs are summarized in Table 4. Using Kruskal-Wallis, we compare the whole group of six inverted resistivity results using reciprocals for weighting (Fig. 7(a)) and found that these populations are not statistically similar ($p<10^{-6}$), and we similarly compared the whole group of six results using stacking errors for weighting (Fig. 7(b)) and found that these populations are also not statistically similar ($p<10^{-6}$). To ensure that Unit 3 was not an outlier biasing the results, we ran the same two tests omitting Unit 3; however, the outcome of the statistical test did not change (reciprocal weighted $p=10^{-5}$, stacking weighted $p<10^{-6}$).

Kolmogorov-Smirnov (Massey, 1951), a non-parametric test that compares several distributions, was used to compare the stacking and reciprocal inversion results from each instrument. The results showed that Unit 1 and Unit 3 had significantly different results when stacking errors were used as weights versus reciprocal weighting, while the rest of the units showed no statistical difference between stacking and reciprocal weighted inversion results.

Finally, for each instrument, we found the difference between the stacking-weighted and reciprocal-weighted inversion results by quadrupole, and compared those six sets of data using a Kruskal-Wallis test and found that the distributions of differences are not

Table 4. Summary of statistics comparing inversion results. All tests are judged relative to the 95% confidence interval. Underlined numbers indicate comparison is not significantly different. Values for n refer to the size of the population for comparisons of two populations; values n_{pop} indicate the number of populations compared when more than two populations are compared.

Do stacking and reciprocal data weights produce the same inverted resistivities for each instrument?

Test	Kolmogorov-Smirnov					
Name	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6
p	$<10^{-6}$	0.50	$<10^{-6}$	0.22	0.79	0.33
n	3,094	3,223	2,291	3,393	3,333	3,395

Are the distributions of inverted resistivities the same across instruments?

Test	Kruskal-Wallis	
Inversion weight	Reciprocal	Stacking
p	$<10^{-6}$	$<10^{-6}$
n_{pop}	6	6

Are the distribution of differences between inversions ($\rho_{W,stack}$ - $\rho_{W,recip}$) the same across instruments?

Test	Kruskal-Wallis
p	$<10^{-6}$
n_{pop}	6

equal ($p<10^{-6}$), which is not surprising given the obvious visual differences as seen in Fig. 7(c). We also ran the same test using only Units 4, 5 and 6 that appear to be very similar and also found that the differences are still not equal ($p = 4\text{x}10^{-4}$). It is important to note that in this case of differenced data, each population is of different size because only values not masked by sensitivity analysis are used, and the sensitive area is slightly different for each instrument (see Appendix A for details). We recognize that treating the datasets as populations in this way is a very coarse comparison that neglects the spatial component of the inverted resistivity section; however, a detailed image processing analysis is outside the scope of this work. Inverted tomograms are presented in Appendix A.

Effects of Electrode Contact and Instrument-Specific Recording Settings

Measured receiver waveforms for the separate four-electrode dipole-dipole experiment are shown in Fig. 8. The waveforms are plotted in units of resistance (Ohms), which are calculated by dividing the recorded voltage curves by the instrument-reported injected current (Eq. 1). This normalization is used because differences in electrode coupling lead to differences in the amplitude of injected current and measured voltage (Table 5), but their ratio remains relatively constant. Forward measurements, where A – B is the transmitter pair and M – N is the receiver pair, are shown in blue, with reciprocal measurements shown in red. Each curve captures two cycles, where each cycle comprises a (+ - - +) transmitter waveform. The grey shaded regions indicate the delay time (100 ms in this instance) before measurements begin after each change in polarity (Table 2).

In the separate test of the influence of the delay time before measurement, 10 repeat measurements of two cycles each were made with delay times 60 ms, 100 ms, and 200 ms. All measurements used the same four-electrode array with fully coupled electrodes. The mean and standard deviation of apparent resistivity reported by the instrument for each set of 10 measurements was: 60 ms delay, 40.74±0.04 Ohm-m; 100 ms delay, 41.77±0.13 Ohm-m; and 200 ms delay, 42.89±0.01 Ohm-m.

Effects of Varying Transmit Current

While the instrument was set to a predetermined range of maximum allowable transmit current, the actual average transmit currents observed were: 327 mA, 316 mA, 106 mA, 56 mA, 10 mA, and 6 mA, indicating that the measurement was voltage limited at the highest two settings. The average contact resistance during the set of measurements was 1.12 kOhm. The resulting distributions of stacking errors are shown in Fig. 9(a), while the associated distributions of apparent resistivity values are

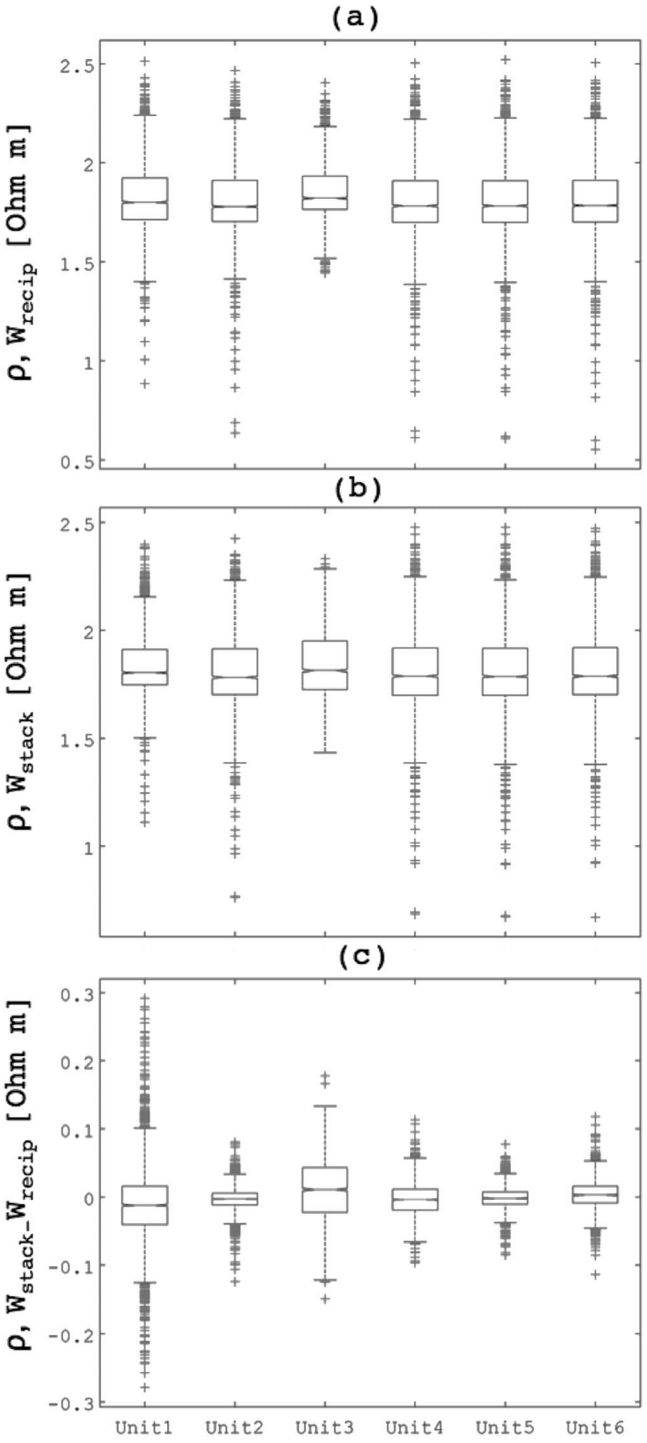

Figure 7. (a) Distributions of inverted resistivity values using reciprocals for data weights (W_{recip}), (b) distributions of inverted resistivity values using stacking errors for data weights (W_{stack}), and (c) distribution of the differences between inverted resistivity values using the two weighting options.

shown in Fig. 9(b). The Kruskal-Wallis test tells us that the distributions of stacking errors are significantly different ($p<10^{-6}$), while the apparent resistivity values were not significantly different (p=0.063).

Discussion

Unit 3 Operation Error

Here we provide a brief explanation of the apparently poor performance of Unit 3, and justifications for our interpretation that this was user error and not related to instrument design. Unit 3 performs an automatic calibration for each receiver channel relative to internal reference resistors prior to measurement. The calibration is performed at log spaced intervals between 0.05 Ohm and 500 Ohm; for this experiment, the average deviation from the known reference resistance was 1.7%. Given this small calibration error, we interpret the cause of the relatively higher stacking and reciprocal errors that were observed in Unit 3 (Fig. 3) not to be an instrument malfunction or a problem with hardware, but instead due to the small input current that needs to overcome R_C, resulting in less signal in the ground. Due to the error in setting up Unit 3, transmitter current was accidently limited to a maximum of only 3 mA (Table 2). Even though the raw data showed large stacking and reciprocal errors for many quadrupoles, filtering before inversion removed the bad measurements. The data errors that were observed for Unit 3 were statistically different from other units (Table 4).

Measured Data Comparison

We find that the measured apparent resistivity and R_C were statistically similar across all instruments. If we refer to the average observed R of all instruments for each quadrupole in Fig. 2, the percentage of data points with a coefficient of variation <0.02 is 52% and the percentage of data points with a coefficient of variation <0.10 is 90%. If we ignore Unit 3 data, the percentage of data points with a coefficient of variation <0.02 grows from 52% to 77% and the percentage of data points with a coefficient of variation <0.10 grows from 90% to 99%. This answers a key question: as long as the instrument is operated correctly, the tested resistivity meters should measure R within a coefficient of variation <0.10 at 99% of the quadrupoles. Additionally, the area of elevated coefficient of variation near the surface around 60 - 78 m corresponds with elevated R_C (Fig. 4), indicating that all instruments suffered from reduced data quality for quadrupoles where at least one electrode had higher R_C. Although we expect the degraded R_C

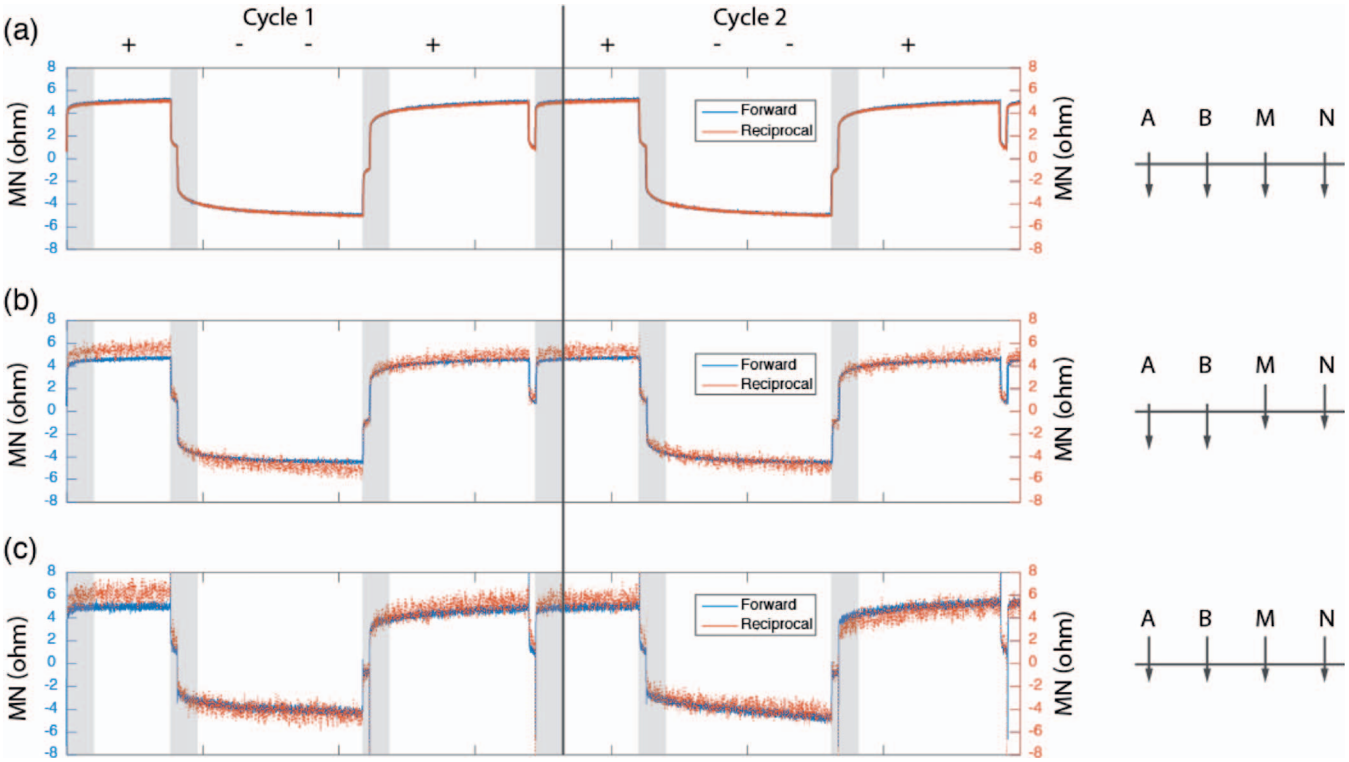

Figure 8. Measured forward and reciprocal waveforms for three different electrode-coupling scenarios. (a) All four electrodes fully planted in the ground; (b) one electrode pair fully planted in the ground, and one only partly in the ground; and (c) all four electrodes only partly planted in the ground. Greyed-out areas represent the measurement the delay time before measurements begin after each change in polarity (Table 2). Measurements are plotted in units of resistance (Ohms), which is calculated by dividing each measured voltage curve by the instrument-reported injected current.

impacts the inverted result, we were not able to precisely determine the effect because we do not have a corresponding dataset with electrodes 31 – 40 inserted 25 cm into the ground (*i.e.*, non-degraded R_C). Finally, we note that the differences in transmit time may have contributed to the variability in measured errors presented in Fig. 5.

Table 5. Injected current, measured voltage, and calculated apparent resistivity for the four-electrode experiment shown in Fig. 8.

	$\text{Current}_{A,B}$ [A]	$\text{Potential}_{M,N}$ [V]	ρ_a [Ω m]
All electrodes fully coupled (Fig. 9(a))			
Forward	0.49	2.32	89.22
Reciprocal	0.454	2.15	89.36
Two electrodes fully coupled, two partly coupled (Fig. 9(b))			
Forward	0.487	2.12	82.22
Reciprocal	0.019	0.89	88.23
All electrodes partly coupled (Fig. 9(c))			
Forward	0.058	0.25	82.69
Reciprocal	0.0143	0.065	85.92

Cause of Measurement Errors

For all instruments, increased measurement errors were most closely associated with degraded R_C. We evaluated this spatially (Fig. 1) and empirically (Fig. 5, Table 3), and in both cases the pattern is clear that larger errors are associated with the 60 – 78 m portion of the line where R_C ranged between 2 and 4.5 kOhm. The exception is Unit 3 that had randomly distributed high-error results. However, this is likely attributed to low input current (Table 2, and detailed above). This is supported by our follow-up measurements varying transmitter current where the 5 mA data had higher stacking errors (Fig. 9(a)). Nonetheless, many measurements made using electrodes in the degraded R_C zone were not highly corrupted, so it is certainly possible to acquire data with low errors in zones with sub-optimal R_C. We also note that the degraded R_C conditions in this study are not as extreme as sometimes encountered in the field where very dry, frozen, or hard-rock conditions exist, which may lead to greater measurement error. This result indicates that variability between ERT measurements repeated with two different instruments is most

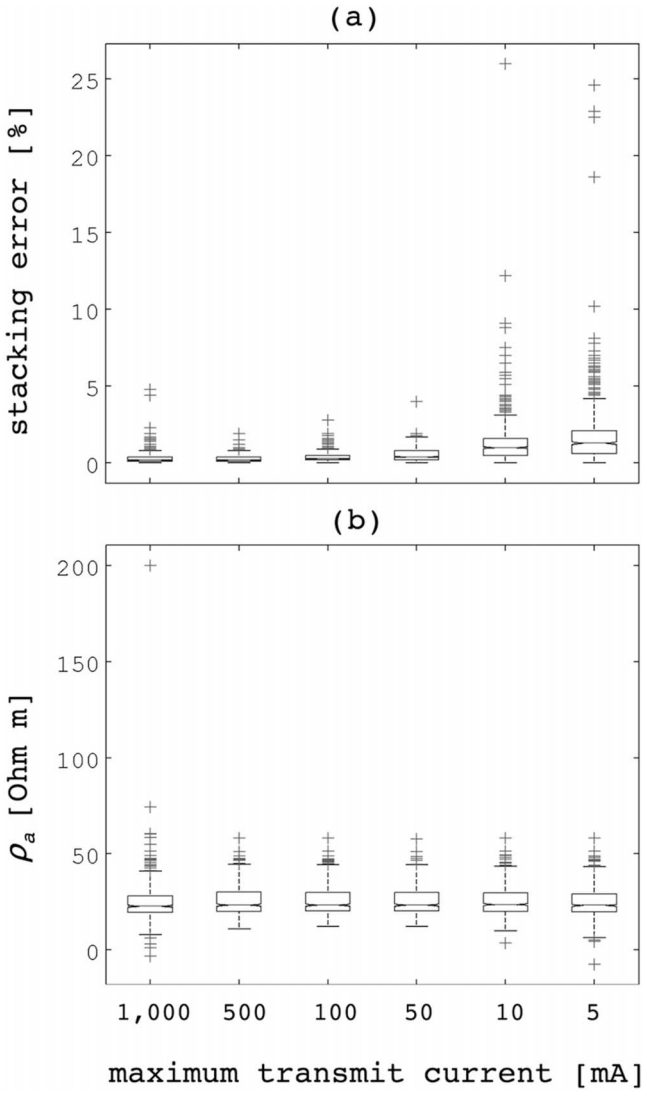

Figure 9. (a) Distributions of stacking error and (b) apparent resistivity as a function of mean transmitted current amplitude.

likely to be attributed to how the instrument is set up by the user (*e.g.*, how much effort the operator puts into achieving low R_C, using maximum available transmitter current), rather than the instrument itself. There do, however, appear to be some differences between instruments. For example, the spatial pattern of reciprocal errors between the Unit 1 (Fig. 1(c.1)) and Units 4, 5 and 6 (Fig. 1(c.4), 1(c.5), 1(c.6)) shows that Unit 1 resulted in more errors at large electrode offsets, while Units 4, 5 and 6 only had noticeably elevated errors around the electrodes with degraded R_C. Unit 4 transmitted less current than Unit 1, and similarly Unit 4 measured smaller received voltages than Unit 1, after normalizing V and I for electrode spacing.

We acknowledge that this was not a perfect comparison since the acquisition parameters were not precisely equal among instruments (Table 2), although in many cases some acquisition parameters are outside user control (*e.g.*, transmitter current setting within the instrument; transmitter time setting; limitations of the experiment duration to make the measurements under as consistent environmental conditions as possible). Nonetheless, some of the errors observed in Fig. 1 and Fig. 2 must be attributed to differences in acquisition. For example, the transmitter current for Units 4, 5 and 6 never dropped below 69 mA and these instruments measured all mean reciprocal errors below 2% (Fig. 5(b)). Unit 1 and Unit 2 had some or all measurements transmitted at current lower than Units 4, 5 and 6 and these two instruments observed mean reciprocal errors up to 4% or 6% (Unit 1 and Unit 2, respectively).

When comparing the effect of electrode coupling on reciprocal measurements in our separate four-electrode test (Fig. 8), we observe just 0.1% measured reciprocal error in forward and reciprocal apparent resistivity (Table 5) when both transmitter and receiver electrode pairs are well coupled (Fig. 8(a)), 7.3% reciprocal measurement repeatability when one pair of electrodes is well coupled to the ground and the other is poorly coupled (Fig. 8(b)), and 3.9% reciprocal measurement repeatability when both electrode pairs are poorly coupled (Fig. 8(c)).

It is a somewhat surprising result that the measured reciprocal error is smaller when all four electrodes were poorly coupled (3.9%) compared with the case with only one poorly coupled electrode pair (7.3%). We suggest that this occurs because of significant differences in electrode complex (quadrature) impedance for the well coupled and poorly coupled electrode pairs, which leads to non-reciprocal measurements. In the case where all four electrodes are poorly coupled, the electrode impedances are similar to one-another and a smaller reciprocal error is measured. In the fully coupled case, which we expect to be the most accurate measurement, both forward and reciprocal measurements recorded an apparent resistivity of approximately 89 Ohm-m. For the mixed coupled-uncoupled case, apparent resistivity values of 82 and 88 Ohm-m were recorded for the forward and reciprocal measurements. This is a large reciprocal error, but only the forward measurement is biased from the expected value of 89 Ohm-m. For the fully uncoupled case, both forward and reciprocal apparent resistivity values (83 and 86 Ohm-m) deviate from the expected value of 89 Ohm-m, but with smaller reciprocal error. This result suggests that measures of reciprocal error may misrepresent and potentially under-

estimate true measurement error. Forward and reciprocal measurements may be similar – but both equally biased – when all four electrodes are poorly coupled. Further experiments should explore this observation, and will be the focus of future studies.

In all three coupling scenarios there is a signature of electrode polarization during measurements, with the resistance increasing in magnitude throughout the duration of the measurement cycle. Most of this increase in amplitude occurs immediately after each change in polarity and is therefore not sampled because of the delay time before measurement. However, the remaining electrode polarization during the measurement cycle is a potential source of error. In the well-coupled scenario (Fig. 8(a)), the resistance curves are very similar and therefore have low reciprocal error. In the poorly coupled scenarios (Figs. 8(b)-8(c)), differences in electrode contact lead to polarization curves that are offset from one-another, and therefore result in high reciprocal error. The poorly coupled scenarios also exhibit greater levels of random noise due to the lower signal levels for these cases. Although averaging over the measurement window will reduce the influence of random noise, a significant difference remains in the shape of the forward and reciprocal curves.

Because electrode polarization causes increasing resistance amplitudes throughout the measurement cycle, instrument-measured apparent resistivity values depend on the size and position of the averaging window. In particular, the delay time before measurement can be an important parameter because measured voltages tend to increase most rapidly at early times after each change in polarity. We observed a systematic increase in apparent resistivity of approximately 2.5% when the delay time before measurement was increased from 60 ms to 100 ms, and another 2.7% increase when the delay time was set at 200 ms. This increase in apparent resistivity is attributed to the fact that larger delay times result in averaging a portion of the received waveform with larger measured values. Differences in delay time and averaging window length for different instruments (Tables 1–2) may therefore be an important cause for the different responses measured by different systems.

Stacking vs. Reciprocals for Error Quantification

Reciprocals are often used for assessing error in ERT data (Zhou and Dahlin, 2003; Slater *et al*., 2000; Binley *et al*., 1995), even though nested arrays with different signal-to-noise ratios for forward and reverse measurements may produce estimates of error that are larger than "true" (but known) measurement errors (Deceuster *et al*., 2013). Many time-lapse studies with multichannel tools do not have the requisite time to collect reciprocals due to the speed of the environmental processes being investigated (*e.g*., hyporheic exchange, Ward *et al*., 2012). Interestingly, it has also been suggested that using measured reciprocals directly in the inversion may underestimate data errors (Robinson *et al*., 2012; Koestel *et al*., 2008), although it is not clear how this could be judged given that the "true" error is not known.

There is a conundrum here, however, given that reciprocals only consist of differencing two resistance measurements, while for stacking, a measurement may be repeated an arbitrary number of times. Two stacks (or reciprocals) pulled from a random distribution are likely not representative of the mean, let alone the standard deviation. Furthermore, modeling errors may be larger than measurement errors in some cases, suggesting that inversions weighted by modeling errors would be more important than concern over whether reciprocal or stacking data errors should be used for weights (Singha *et al*., 2014). Similarly, electrode-positioning errors are comparable to the signal strength and often neglected, occasionally exceeding 10% in certain measurement geometries (Oldenborger *et al*., 2005). As predicted, we observed reciprocal errors are most frequently larger than stacking errors for all instruments (Fig. 3), and because reciprocals intrinsically account for the difference in R_C when the transmitting and receiving dipoles are swapped, we interpret the bias to mean that stacking is underestimating error relative to reciprocals. In the four-electrode coupling test (Fig. 8, Table 5) both stacking error and reciprocal error were small (0.1% - 0.2 %) for the well-coupled scenario (Fig. 8(a)). However, in the poorly coupled scenarios (Figs. 8(b)-8(c)) stacking errors (0.1 %) were smaller than reciprocal errors (3.9% - 7.3 %). Figure 5 indicates that stacking and reciprocal errors are more consistent with one another when both are small (well-coupled, low R_C scenarios), but that reciprocal errors are systematically greater than stacking errors when both are large (poorly coupled, high R_C scenarios). We also find that the distributions of measured errors in the field experiment are not statistically similar among all instruments based on the Kruskal-Wallis analysis (Figs. 6(c) and 6(d)).

Inversion Result Comparison

In four cases the inverted results are not statistically different given the stacking versus reciprocal errors (Table 4). It is not surprising that Unit 3 was found to have a significant difference in inversions between weighting methods because of the setup error; however,

Unit 1 was operated correctly and therefore would be expected to fall in line with the other four instruments that were operated correctly. To explain this, we refer to the measured errors, where we can see that Unit 1 did have the second largest average of stacking errors (Fig. 6(d)), despite having one of the smallest average reciprocal errors (Fig. 6(c)). Given that the measured errors by Unit 2 had the largest difference amongst all instruments, it is not surprising that the inversion results weighted by stacking or reciprocal error would be different. Although the statistical evaluation indicates that the inverted results are not equal across all instruments, we acknowledge that, in some cases, only qualitative interpretations of the results are needed. In those cases, we believe that it would be acceptable to use more than one instrument during an investigation at a site with the expectation that the results would be interpreted similarly in terms of electrical structure assuming R_C could be kept consistent and low.

Environmental Factors During the Test

At the time of the experiment, our reasoning for constraining the total time period was that completing all of the measurements on the same day would minimize the impact of environmental factors: primarily changes in R_C and changes to the ground moisture conditions (though we also anticipate that very different ambient air temperature may have a small effect on the instrument electronics). Based on our results, this is at least partly justified – maintaining consistent, low R_C is critical to measurement quality - though, of course, contact resistance is not determined by the instrument. In essence, there was no perfect way to do the experiment given the limitation of relatively short daylight in winter. Air temperature started below freezing and rose above freezing around 9:00 am, when Unit 1 (first to measure) was acquiring data. Therefore, the ground surface began to thaw possibly impacting R_C, and possibly explaining why Unit 1 has higher R_C in Fig. 4. We found that 77% of electrodes experienced a decrease in R_C during the day as measured by different instruments, suggesting that the increase in liquid water at the surface improved contact resistance. For electrodes that had a decrease in R_C larger than one standard deviation, the observed average rate of decrease was 0.19 kOhms per hour. Since the total change in R_C over the course of the day averaged only 0.62 kOhm, similar to the standard deviation of the non-degraded R_C electrodes (mean R_C = 1.1 kOhm, std = 0.38 kOhm), we predict that the change in environmental conditions did not have a negative impact on our interpretations.

Conclusions

This investigation demonstrates that the measured apparent resistivity and R_C data across six instruments on the same transect with partially degraded R_C were statistically similar; however, the measured reciprocal and stacking errors across these instruments were not. The weighted inversion resistivity results based on those data were visually similar, and four of the six instruments produced resistivity tomograms that were not significantly different, suggesting that in some cases the choice of error assessment is not important to the final result. Although the inverted results were not identical across all instruments, we judge that any conservative assessment of the final inverted resistivity models would result in a similar geological interpretation for each. Based on the results of statistical analysis that show no discernable difference in apparent resistivity between the instrument, we conclude that it would be acceptable to use more than one instrument during an investigation at a site with the expectation that the results would be similar assuming R_C could be kept consistent and a continuing record of R_C should be measured throughout the experiment to ensure compatible data quality.

Acknowledgments

We thank E. Kempema, W.S. Holbrook, and C. Moulton for participation in the field measurements. We are greatly appreciative of R. Van Dam, L. Slater, L. Ball, and one anonymous reviewer for their constructive comments that greatly improved the quality of this work. We also acknowledge the Wyoming Center for Environmental Hydrology and Geophysics (WyCEHG) EPSCoR Research Infrastructure Improvement (RII) Track-1 project, funded by NSF (Award #1208909) that provided three of the ERT instruments. We also thank R. Knight (Stanford U.) for loan of one of the ERT instruments. We thank A. Binley (Lancaster U.) for his freely available R2 electrical resistivity tomography inversion software used for all inversions in this investigation. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Abdu, H., Robinson, D.A., and Jones, S.B., 2007, Comparing bulk soil electrical conductivity determination using the DUAL-EM-1S and EM38-DD electromagnetic induction instruments: Soil Science Society of America Journal, **71**, 189–196.

Binley, A., Ramirez, A., and Daily, W., 1995, Regularised image reconstruction of noisy electrical resistance tomography data: Process Tomography—1995, Proceedings of the 4th

Workshop of the European Concerted Action on Process Tomography, Bergen.

Binley, A., and Kemna, A., 2005, DC resistivity and induced polarization methods: Hydrogeophysics, **50**, 129–156.

Binley, A., 2015, DC electrical methods: *in* G. Schubert (ed.), Treatise on Geophysics, 2nd ed. **11**, 233–259.

Breslow, N., 1970, A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship: Biometrika, **57**, 579–594.

Dahlin, T., 2000, Short note on electrode charge-up effects in DC resistivity data acquisition using multi-electrode arrays: Geophysical Prospecting, **48**, 181–187.

Dahlin, T., and Zhou, B., 2004, A numerical comparison of 2D resistivity imaging with 10 electrode arrays: Geophysical Prospecting, **52**, 379–398.

Deceuster, J., Kaufmann, O., and Van Camp, M., 2013, Automated identification of changes in electrode contact properties for long-term permanent ERT monitoring experiments: Geophysics, **78**, E79–E94.

Gallipoli, M., Lapenna, V.I., Lorenzo, P.I., Mucciarelli, M., Perrone, A., Piscitelli, S., and Sdao, F., 2000, Comparison of geological and geophysical prospecting techniques in the study of a landslide in southern Italy: European Journal of Environmental and Engineering Geophysics, **4**, 117–128.

Grant, F.S., and West, G.F., 1965, Interpretation theory in applied geophysics: McGraw-Hill Book Company, 583 pp.

Habberjam, G.M., 1967, On the application of the reciprocity theorem in resistivity prospecting: Geophysics, **32**, 918–919.

Herman, R., 2001, An introduction to electrical resistivity in geophysics: American Journal of Physics, **69**, 943–952.

Koestel, J., Kemna, A., Javaux, M., Binley, A., and Vereecken, H., 2008, Quantitative imaging of solute transport in an unsaturated and undisturbed soil monolith with 3D ERT and TDR: Water Resources Research, **44**, W12411.

LaBrecque, D., and Daily, W., 2008, Assessment of measurement errors for galvanic-resistivity electrodes of different composition: Geophysics, **73**, F55–F64.

Massey, F.J.J., 1951, The Kolmogorov-Smirnov test for goodness of fit: Journal of the American Statistical Association, **46**, 68–78.

McGill, R., Tukey, J.W., and Larsen, W.A., 1978, Variations of boxplots: The American Statistician, **32**, 12–16.

Oldenborger, G., Partha, A., Routh, S., and Knoll, M.D., 2005, Sensitivity of electrical resistivity tomography data to electrode position errors: Geophysical Journal International, **163**, 1–9.

Oldenburg, D.W., 1978, The interpretation of direct current resistivity measurements: Geophysics, **43**, 610–625.

Oldenburg, D.W., and Li, Y., 1999, Estimating depth of investigation in dc resistivity and IP surveys: Geophysics, **64**, 403.

Perri, M.T., Cassiani, G., Gervasio, I., Deiana, R., and Binley, A., 2012, A saline tracer test monitored via both surface and cross-borehole electrical resistivity tomography: Comparison of time-lapse results: Journal of Applied Geophysics, **79**, 6–16.

Robinson, J.L., Slater, L.D., and Schäfer, K.V., 2012, Evidence for spatial variability in hydraulic redistribution within an oak–pine forest from resistivity imaging: Journal of Hydrology, **430**, 69–79.

Szalai, S., Novák, A., and Szarka, L., 2009, Depth of investigation and vertical resolution of surface geoelectric arrays: Journal of Environmental and Engineering Geophysics, **14**, 15–23.

Sass, O., Bell, R., and Glade, T., 2008, Comparison of GPR, 2D-resistivity and traditional techniques for the subsurface exploration of the Öschingen landslide, Swabian Alb (Germany): Geomorphology, **93**, 89–103.

Seaton, W.J., and Burbey, T.J., 2002, Evaluation of two-dimensional resistivity methods in a fractured crystalline-rock terrane: Journal of Applied Geophysics, **51**, 21–41.

Singha, K., Day-Lewis, F.D., Johnson, T., and Slater, L.D., 2014, Advances in interpretation of subsurface processes with time-lapse electrical imaging: Hydrological Processes, **29**, 1549–1576.

Skinner, D., and Heinson, G., 2004. A comparison of electrical and electromagnetic methods for the detection of hydraulic pathways in a fractured rock aquifer, Clare Valley, South Australia. Hydrogeology J., **12**(5), 576–590.

Slater, L., Binley, A.M., Daily, W., and Johnson, R., 2000, Cross-hole electrical imaging of a controlled saline tracer injection. Journal of Applied Geophysics, **44**, 85–102.

Sudduth, K.A., Kitchen, N.R., Bollero, G.A., Bullock, D.G., and Wiebold, W J., 2003, Comparison of electromagnetic induction and direct sensing of soil electrical conductivity: Agronomy Journal, **95**, 472–482.

Tyler, S.W., Selker, J.S., Hausner, M.B., Hatch, C.E., Torgersen, T., Thodal, C.E., and Schladow, S.G., 2009, Environmental temperature sensing using Raman spectra DTS fiber-optic methods: Water Resources Research, **45**, W00D23.

Ward, A.S., Gooseff, M.N., and Singha, K., 2012, How does subsurface characterization affect predictions of hyporheic exchange?: Ground Water, **51**, 14–28.

Yaramanci, U., Lange, G., and Knödel, K., 1999, Surface NMR within a geophysical study of an aquifer at Haldensleben (Germany): Geophysical Prospecting, **47**, 923–943.

Zhou, B., and Dahlin, T., 2003, Properties and effects of measurement errors on 2D resistivity imaging surveying: Near Surface Geophysics, **1**, 105–117.

APPENDIX A

Tomograms inverted with reciprocals as weights using identical inversion parameters are shown in Figs. A1(a) through A1(f) (similar inversions were done with stacking weights, however these results are visually very similar and therefore omitted for brevity). As described earlier, we have plotted the results with low sensitivity as defined by the DOI as translucent. More specifically, a value below 0.2 has no transparency, a value between 0.2 and 0.4 is semi-transparent and a value above 0.4 is fully transparent. Among all of the results, the resistivity structure was found to be around 200-250 Ohm-m within the top 3 m, underlain by a zone of about 70 Ohm-m that is 3-m thick. The deepest

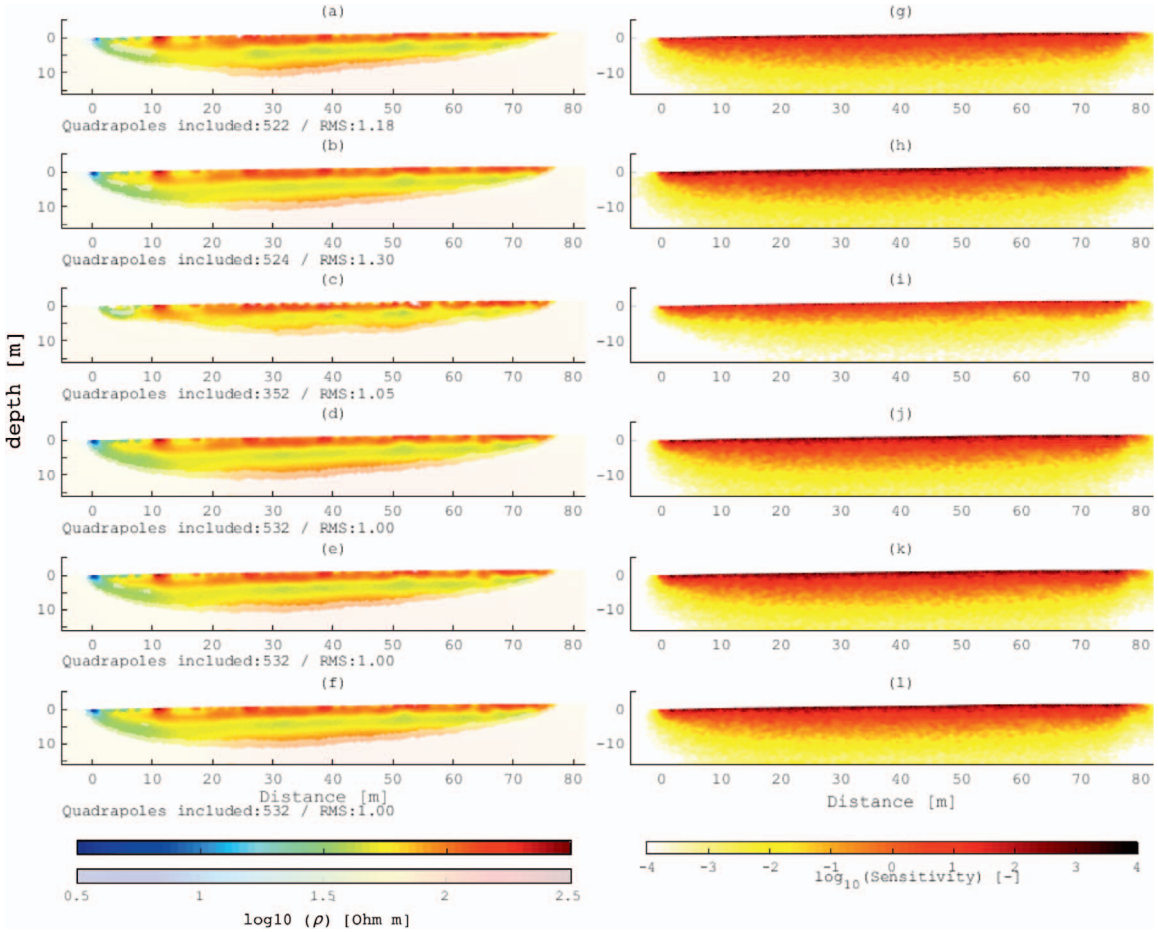

Figure A-1. (a-f) Inversion results using reciprocal weighting for each instrument and (g-i) sensitivity from those inversion.

observed layer has higher resistivity starting at around 6 – 8 m depth. Also similar to all instruments is that the DOI is slightly greater on the left side of the line in comparison with the right side. The sensitivity section associated with each resistivity model is shown in Figs. A1(g) through A1(l). Here, the most similar results are Units 1, 4, 5 and 6. Both Units 2 and 3 have slightly shallower, lower sensitivity.

We found that by inverting each dataset with the same inversion parameters, instrument-specific data error weights, and the same criteria for rejecting measurements with high error, the resulting inverted sections are visually similar (Figs. A1(a) through A1(f)), with exception to Unit 3 (Fig. A1(c)), which differs slightly from the others with a shallower depth of investigation, and the high resistivity anomaly at the ground surface 40 m along the line is not well-resolved. The two instruments with lowest transmitting current – Unit 2 and Unit 3 – had difficulty resolving some portions of the top 1 m. The sensitivity is similar between all instruments up to a depth of 10 m (Figs. A1(g) through A1(l)) with only the low-current Unit 3 measurements differing in a visible way from the rest of the instruments.