A Data Variance Technique for Automated Despiking of Magnetotelluric Data with a Remote Reference

Karl N. Kappler*

*Lawrence Berkeley National Laboratory,

Earth Sciences Division

1 Cyclotron Road, Berkeley, CA, 94720

February 28, 2011

ABSTRACT

The magnetotelluric method employs co-located surface measurements of electric and magnetic fields to infer the local electrical structure of the earth. The frequency-dependent “apparent resistivity” curves can be inaccurate at long periods if input data are contaminated—even when robust remote reference techniques are employed. Data despiking prior to processing can result in significantly more reliable estimates of long period apparent resistivities. This paper outlines a two-step method of automatic identification and replacement for spike-like contamination of magnetotelluric data; based on the simultaneity of natural electric and magnetic field variations at distant sites. This simultaneity is exploited both to identify windows in time when the array data are compromised, and to generate synthetic data that replace observed transient noise spikes. In the first step, windows in data time series containing spikes are identified via intersite

comparison of channel 'activity' — such as the variance of differenced data within each window. In the second step, plausible data for replacement of flagged windows is calculated by Wiener filtering coincident data in clean channels. The Wiener filters — which express the time-domain relationship between various array channels — are computed using an uncontaminated segment of array training data. Examples are shown where the algorithm is applied to artificially contaminated data, and to real field data. In both cases all spikes are successfully identified. In the case of implanted artificial noise, the synthetic replacement time series are very similar to the original recording. In all cases, apparent resistivity and phase curves obtained by processing the despiked data are much improved over curves obtained from raw data.

INTRODUCTION

Raw magnetotelluric (MT) data are records of natural variations of the earth's magnetic and electric fields at fixed locations. It is well known that these data carry information about the conductivity structure of the earth near the location of the recordings. Most MT methods are based on the assumption of vertically incident plane wave sources. Superposed on the plane-wave signal are other electric and magnetic variations due to localized noise phenomena. These noise spikes appear as anomalous features in the raw time series. They are typically of short duration and not recorded at all sites in an array when intersite spacings are large (tens to hundreds of km). Disturbances are often present simultaneously on all channels at a particular site, or in groups of channels measuring the same field type (electric or magnetic) at a single site. This can be due to a malfunctioning component of the data acquisition system common to the spiking channels or a local noise source. Single channel spikes are less common, but do occur in electrodes, possibly in association with poorly understood chemical reactions in the near surface, (Perrier and Morat (2000)), and are occasionally observed in single channel magnetics as well. Time series often return to pre-disturbance 'baseline' levels following the contamination (spikes), but sometimes settle to a new baseline vastly different from that before the disturbance (steps). Szarka (1987) provides a discussion of noise characteristics and sources. No attempt is made here to speculate on the sources of the transients in the data — for processing purposes it is sufficient to identify and remove these transients. Regardless of the reason for the noise, intersite data variance ratios are an effective tool for automatic identification of artificial transients in MT data.

There are a variety of techniques dedicated to 'cleaning' MT data. These typically fall

into categories of time or frequency domain. The frequency domain category partitions into methods dealing explicitly with Fourier coefficients of windowed time series (e.g. Chave and Thomson (1989), Egbert (1997), and Oettinger et al. (2001)), and those which operate on impedance tensor estimates (e.g. Tzanis and Beamish (1989) and Weckmann et al. (2005)). Many techniques require, or at least significantly benefit from the inclusion of a remote reference (RR) station (e.g. Egbert (1997) and Hattingh (1989)). Robust remote reference processing in frequency domain significantly stabilizes MT data processing results, but at longer periods the performance of those methods decays as shown in the results of this paper. Junge (1996) provides a review of many of these frequency domain (and some time domain) methods. Here the discussion will be restricted to time domain methods; specifically those which treat transient noise. For removal of harmonic noise in time domain delay-line filters are described by Fontes et al. (1988).

Several techniques for time-domain spike identification and repair have been previously discussed in the literature. Short spikes/gaps on the order of a few points were repaired using simple interpolation schemes, for example the median filter method of Jones et al. (1989). Often, larger gaps were removed from analysis altogether. Egbert et al. (1997) filled gaps in time domain electric field data by dictating probable data values based on a multivariate impulse response operator (IRO) convolved with coincident magnetic field data. That method of IRO calculation however requires preliminary estimates of the MT impedance tensor at various frequencies, and an optimized inversion routine discussed in Egbert (1992). For continuously present noise of broad spectral content, the correlated adaptive noise cancellation (CANC) approach of Hattingh (1989) is applicable. CANC however has the drawback that a user selected convergence factor needs to be adapted is it not demonstrated to be effective when the amplitude of the transient noise is orders of

magnitude larger than the underlying signal. Trad and Travassos (2000) employ the wavelet transform, and then downweight all wavelet coefficients which are larger than some threshold. This technique effectively removes many large spikes, and works when data from only one site is available. In practice however, as the user lowers the threshold to identify and remove less ostentatious outliers, progressively smaller spikes are removed at the cost of many false positives — flagging natural variations as noise — because many sudden variations in signal amplitude are in fact part of the natural MT signal. Both Edwards and Hastie (1997) and Leśniak et al. (2009) explore use of the Kalman filter to replace contaminated MT data, but neither paper discusses the process used to flag data for replacement. The recursive Kalman filter is also a valid way to replace data, but here it is shown that simpler Wiener filtering yields good results. Edwards and Hastie (1997) acknowledge that the accuracy of the Kalman procedure tends to decrease as the contaminated window increases in length, which does not seem to be an issue with the method developed here. Neither of the Kalman method papers discuss how the replaced data are finally spliced back into the observed data.

The method described here is applicable to transient noise observed in time domain, requires a remote reference, and deals with spikes and steps easily even if they have very large amplitudes. It will be abbreviated IARWR for Intersite Activity Ratio Wiener-filter Replacement. IARWR retains the strengths of previous time domain methods while improving on them in several ways. IARWR handles considerably larger missing or contaminated segments than interpolation methods. Like Egbert et al. (1997) IARWR calculates a multivariate IRO for convolution, but implementation is purely in the time domain. Residuals obtained by subtracting dictated estimates from observations suggest the technique here is at least as effective as that in Egbert (1992). The 'fill-in' methods cited however do not

address step-offsets — a simple data replacement will only translate a step to the edge of the replacement window — whereas IARWR removes such steps effectively. Furthermore, previous studies provide no assurance (beyond manual inspection of the data) that IROs are calculated using uncontaminated data, nor that time series used for replacement are calculated with uncontaminated channels. IARWR calculation of the IRO (and replacement data) is done in a way which automatically guards against use of corrupt data. Also a match-filter is employed at the edges of replaced windows to mitigate boundary artifacts. Moreover, IARWR differentiates natural variations from variations of local origin by comparing channels at different sites, and flagging sharp variations which are not present arraywide, thus allowing natural spikes (signal) to inhabit the data, but flagging artificial spikes. The algorithm can be adapted to compare intersite wavelet coefficients after Trad and Travassos (2000), or a user customized measure of site activity, but for simplicity a more crude statistic — the variance of short time series of time-differentiated raw data — is compared from site to site. The Theory section of the paper mathematically describes the method in the context of matrix algebra and time series. The Application section shows results from applying the technique to contaminated field data — first with synthetic noise and then with actual recorded spikes. The Discussion section covers method applicability and limitations and also includes some practical considerations when programming the method.

THEORY

In the absence of overwhelming local noise, MT sensors at different sites are strongly correlated, with nearly identical time variations in magnetic fields, and electric fields differing by a scale factor dependent on local conductivity structure. Figure 1 shows orthogonal

electric and magnetic field data sampled at two MT sites (PKD and SAO) separated by 120km. Note the similarity in field variations between channels of same orientation and field type as well as identical scaling of the magnetic components, compared to the scale factor difference in the electric channels. This similarity extends from periods of hours down to less than a second. The inherent simultaneity in the natural fields at different sites can be exploited both to identify windows in time when the array data are compromised as well as to generate synthetic data which replaces observed transient noise spikes.

IARWR comprises two principal steps: identification and replacement. In the identification step, simultaneous data from two separated locations are broken into short, overlapping windows, and the variance of the differenced time series in each window is calculated. This generates a condensed 'variance' time series. Windows containing spikes are identified by scanning intersite ratios of these window variances. The user specifies window length, and a parameter that controls an adaptive identification threshold. In the replacement step flagged windows are replaced by synthetic data determined by convolution of clean data — coincident with the flagged window — with Wiener filters. A list of individual smaller processing steps is shown in Figure 2 and these steps are described in detail in the following two subsections.

IDENTIFICATION

Identification starts with simultaneous time series from at least two MT sites, scaled in units of instrument counts. Data are windowed, and a measure of 'activity' of observed

fields is calculated for each window. Time series of activity ratios between distant channels, recording the same field type at the same cardinal orientation, are typically stationary about some mean value. In the case of magnetic fields, this value is near 1, and in the case of electric fields the average activity ratio will reflect site effects such as differences in conductivity or local distortion of electric fields. The criteria used to identify spikes is independent of this median value, and independent of scaling site data by a constant. Outliers in the activity ratio distributions for particular channel pairs correspond to spikes, and are identified according to a user-selected threshold.

The steps of identification are:

1. **Window the time series:** All time series are divided into windows of length L and overlap V. Each N-point time series is thus transformed into $W = floor(N/(L-V))$ windows. Window length (L) should be sufficiently narrow that moderate amplitude spikes significantly affect activity. If the window is chosen too wide, then only very large spikes will be caught, as smaller spikes will not significantly impact activity measures. If the time series do not neatly partition into windows of length L and overlap V, take the last L points of each time series as an activity window.

2. **Measure Activity of windows:** Define a measure of window 'activity'. Apply this measure to each window, resulting in a time series of activity for each channel.

3. **Calculate Activity Ratios:** For distant channels of the same field type and same cardinal orientation, compute the ratio of activity for simultaneous windows. Activity ratios reflect the difference in energy seen by similar channels at distant sites. Before calculating statistics on the activity ratios it is helpful to take their logarithm making their distribution nearly normal.

4. **Select Thresholds:** (α, n_{std}) Spike infected windows are identified as outliers of the log activity ratio distributions. Those windows lying more than n_{std} standard deviations from the median are flagged for removal. The standard deviation is calculated using only log-ratios in the α to 100-α percentiles. The so-called α-trimmed standard deviation prevents a few very large spikes from driving the standard deviation, and thus the threshold so high that smaller spikes are not caught. A lower bound constraint (l_{std}) is applied to the threshold to the effect of reducing the number of false positives on days when variance ratios are particularly stable.

5. **Identify Offending Channels:** Of the two channels making a ratio deemed an outlier, the channel having a larger variance is replaced. Because the log scale assigns negative values to ratios less than 1, and positive to values greater than one, outliers to the left of the distribution are associated with spikes in the denominator channel, and outliers to the right with spikes in the numerator channel.

At the end of the identification stage, the user has a catalog of spikes, each spike specified by two parameters: the channel registering the disturbance (c), and the time window (w) during which the spike occurred.

SPIKE REPLACEMENT

The spike replacement starts with the list of channel-window pairs from the identification routine. Windows containing spikes can be considered as gaps in the data stream. These gaps, together with any actual gaps in acquired data are replaced with time series of physically plausible data. Replacement data are calculated by convolution of clean channels with Wiener filters which relate the contaminated channel to the clean channels. The filter

coefficients themselves are calculated over some uncontaminated window in time. The steps of replacement for a particular channel c which has a spike during time window w are listed below.

1. **Identify the set of non-flagged channels:** Identify all channels in the array which were not flagged during w. Call this set $\mathcal{T}$ for 'training' data. The number of channels in $\mathcal{T}$ are denoted by the integer K.

2. **Select time window for training Wiener filters:** Identify a window in time when c as well as all channels in $\mathcal{T}$ are unflagged. Training data for electric field replacements should have length on the order of the gap. This is because raw electric field data contain high amplitude long period energy, thus using too long a segment of training data to calculate a least-squares filter will preferentially fit long periods at the cost of poorly fitting short periods. This is not the case with raw magnetic data. Since instrument counts are not corrected for search coil instrument transfer functions, long period energy is significantly diminished in the raw magnetic time series. The data from channel c is denoted $\mathbf{d}$, and the training data channels are denoted $\mathbf{m}_1 ... \mathbf{m}_K$

3. **Calculate Wiener filters:** Using clean data from step 2, solve for the Wiener filter coefficients which best predict channel c when convolved with the training data. That is, solve:

$$\mathbf{d} = \sum_{k=1}^{K} \Psi_k \star \mathbf{m}_k \tag{1}$$

where $\star$ denotes convolution and the Wiener filters are denoted $\mathbf{\Psi}_k$. Training channels $\mathbf{m}_k$ are mean-subtracted prior to all calculations, and so $\mathbf{d}(\mathrm{t})$ is plausible in its

variations, but offset by a shift (accounted for in step 5). Equation 1 has a matrix representation:

$$\mathbf{d} = \mathbf{M}\mathbf{\Psi} \tag{2}$$

where $\mathbf{d}$ is a column vector of data observed in channel c, $\mathbf{M}$ is a multichannel convolution matrix, and $\mathbf{\Psi}$ is the concatenation of all the Wiener filters. If the training data have T observations, and each Wiener filter is of length Q, then Equation 2 has the following expanded form:

$$\begin{bmatrix} d_{1+\frac{Q-1}{2}} \\ d_{2+\frac{Q-1}{2}} \\ \vdots \\ d_{T-\frac{Q-1}{2}} \end{bmatrix} = \left[\begin{array}{ccc|c|ccc} m_{1,1} & \dots & m_{1,Q} & \dots & m_{K,1} & \dots & m_{K,Q} \\ m_{1,2} & \dots & m_{1,Q+1} & \dots & m_{K,2} & \dots & m_{K,Q+1} \\ & \vdots & & \vdots & & \vdots & \\ m_{1,T-2Q+1} & \dots & m_{1,T-Q} & \dots & m_{K,T-2Q+1} & \dots & m_{K,T-Q} \end{array}\right] \left[\begin{array}{c} \psi_{1,1} \\ \vdots \\ \psi_{1,Q} \\ \hline \\ \vdots \\ \\ \hline \\ \psi_{K,1} \\ \vdots \\ \psi_{K,Q} \end{array}\right] \tag{3}$$

$\mathbf{M}$ is a concatenation of K convolution matrices (one for each training channel). Vertical lines partition $\mathbf{M}$ into submatrices, each submatrix performing convolution with a single training channel. Subscripts i, j, on the elements of $\mathbf{M}$ denote the j^{th} element of the i^{th} training channel in $\mathcal{T}$. Subscripts i, j on $\psi_{i,j}$ denote the j^{th} filter

coefficient for i^{th} training channel. Horizontal lines partition the column vector $\mathbf{\Psi}$ between individual training channels. Forming $\mathbf{d}$ and $\mathbf{M}$ from clean training data, Equation 3 is solved for the $K \times Q$ Wiener filter coefficients $\mathbf{\Psi}$. Note when $\mathbf{M}$ is constructed of input time series having length T we do not recover T points of $\mathbf{d}$, since $(Q-1)/2$ points are lost on either side as the filter is 'entering or leaving' $\mathbf{M}$. Because the training data are unflagged (spike-free), ordinary least squares is sufficient to solve the problem.

4. **Calculate Replacement data:** With an appropriate $\mathbf{\Psi}$ obtained by solving Equation 3, $\mathbf{d}(t)$ is dictated by again solving Equation 3, but this time with $\mathbf{\Psi}$ as a known input, and the $\mathbf{M}$ matrix replaced by the observed data in the time window w. The resulting $\mathbf{d}_{plaus}$ is the plausible data used to fill the gap. Enough padding is applied to the $\mathbf{m}_j$ so that the resultant $\mathbf{d}_{plaus}$ is wider than the gap. A width of $1.1 \times T$, (5% overlap of predicted data to either side) is used in this paper. The gap and the overlap to either side of the gap are then replaced following the scheme described in the next step.

5. **Replace gaps:** The noisy data window is replaced by the synthetic data calculated in the previous step. Match filtering is applied at the edges of the window to smooth the transition between observed and synthetic data. Five disjoint regions in time are considered separately: $\mathcal{R}_i$, $i = 1..5$ increasing from left to right. $\mathcal{R}_3$ is defined by the bounds of the contiguous flagged window(s). $\mathcal{R}_2$ and $\mathcal{R}_4$ extend 5% of the width of $\mathcal{R}_3$ to the left and right of $\mathcal{R}_3$, respectively. $\mathcal{R}_1$ and $\mathcal{R}_5$ extend to the ends of the datastream left and right of $\mathcal{R}_2$ and $\mathcal{R}_4$, respectively. These bounds are shown, partitioned by vertical lines, in Figure 3. In each region, data are adjusted to some

combination of the predicted and observed data as described in Equations 4 through 8.

$$\mathbf{p}_1(t) = \mathbf{d}_{obs}(t) \quad \forall t \in \mathcal{R}_1 \tag{4}$$

$$\mathbf{p}_2(t) = [(1 - \mathbf{T}_{down})(\mathbf{d}_{obs} - \bar{\mathbf{d}}_{obs}) + \mathbf{T}_{up}\mathbf{d}_{plaus}] \forall t \in \mathcal{R}_2 \tag{5}$$

$$\mathbf{p}_3(t) = \mathbf{d}_{plaus} \quad \forall t \in \mathcal{R}_3 \tag{6}$$

$$\mathbf{p}_4(t) = [\mathbf{T}_{up}(\mathbf{d}_{obs} - \bar{\mathbf{d}}_{obs}) + (1 - \mathbf{T}_{down})\mathbf{d}_{plaus}] \quad \forall t \in \mathcal{R}_4 \tag{7}$$

$$\mathbf{p}_5(t) = \mathbf{d}_{obs}(t) \quad \forall t \in \mathcal{R}_5 \tag{8}$$

The diagonal matrices $\mathbf{T}_{down}$ and $\mathbf{T}_{up}$ are tapering matrices (Equation 9) which 'splice' the predicted data to the observed data. The result of applying these matrices is that predicted time series are gradually upweighted and observed data correspondingly downweighted as the distance to the flagged window boundary decreased.

$$\mathbf{T}_{up(i,i)} = cos(\frac{-\pi(i - Q)}{(2Q)}) \qquad \mathbf{T}_{down} = \mathbf{I} - \mathbf{T}_{up} \tag{9}$$

An option to add a trendline on top of the replaced data can be invoked if the gap corresponds to a clipped spike as mentioned in the discussion. Finally, the time series in the five regions are shifted to be free from sharp offsets at the boundaries, and concatenated into a cleaned data vector $\mathbf{d}_{new}$. To each of regions 2, 3, and 4 a shift s_{12} is added representing the difference in the median values at the joining edges of $\mathbf{p}_1$ and $\mathbf{p}_2$. To $\mathbf{p}_5$ a shift (s_{45}) is added representing the difference in the median values of $\mathbf{p}_4$ and $\mathbf{p}_5$, where only a few points (n_{med}) nearest the boundary are considered in the median calculation. Thus, a whole day's data channel is finally replaced by $\mathbf{d}_{new}$, defined as:

$$\mathbf{d}_{new}(t) = [\mathbf{p_1} \mid \mathbf{p_2} + s_{12} \mid \mathbf{p_3} + s_{12} \mid \mathbf{p_4} + s_{12} \mid \mathbf{p_5} + s_{45}] \tag{10}$$

where

$$s_{1,2} = median\{\mathbf{p}_1(t)\ s.t.\mid t_{1,2} - \frac{t}{dt} < n_{med}\mid\} - median\{\mathbf{p}_2(t)\ s.t.\mid t_{1,2} - \frac{t}{dt} < n_{med}\mid\} \tag{11}$$

$$s_{4,5} = median\{\mathbf{p}_4(t)\ s.t.\mid t_{4,5} - \frac{t}{dt} < n_{med}\mid\} - median\{\mathbf{p}_5(t)\ s.t.\mid t_{4,5} - \frac{t}{dt} < n_{med}\mid\} \tag{12}$$

where dt is sampling rate, and $t_{i,j}$ is the time corresponding to the boundary between $\mathcal{R}_i$ and $\mathcal{R}_j$. Setting $n_{med} = 1$ corresponds to forcing the two time series segments to assume the same value at the points immediately to either side of the boundary.

Trains of contiguous flagged windows are automatically grouped together and replaced en masse. Some replacement examples are shown in Figure 3. Note the shift between $\mathcal{R}_4$ and $\mathcal{R}_5$ is particularly important when there are steps in the data. In summary, the parameters that control the algorithm are the initial time series lengths (N), Window length (L), window overlap (V), the percent of data to reject when calculating thresholds (α), the threshold multiplier (M), and a lower bound flagging threshold. Values for each of these parameters are described in the following section.

APPLICATION

Two ultra-low frequency (1mHz-1Hz) electromagnetic observatories located at Parkfield (PKD) and Hollister (SAO) California are the sources of the magnetotelluric data in the

examples. The sites are separated by a distance of approximately 120km. At each site are two orthogonal 100m electrodes and two orthogonal BF4 induction coils.

In this section, IARWR is applied to synthetically contaminated data as well as genuinely contaminated field data. The method improves sounding curves at both sites but for brevity, results are plotted for Parkfield only, with Hollister acting as remote reference site. In all examples, data are sampled at 1 Hz. $L = 256$ point windows and $V = 64$ point overlap are used for activity windows. The variance of the first-differenced data in each window is defined as the 'activity'. For practical threshold choices, $n_{std} = 5$ worked well for magnetic fields and $n_{std} = 6$ for electrics, that is, all spikes were identified and few false positives were detected while testing. A lower bound threshold constraint of $l_{std} = 0.4$ is used with Wiener filter lengths set at $Q = 13$. Five points are used to calculate the shifts s_{12} and s_{45}, ($n_{med} = 5$). Thirty minute segments of training data are used to calculate filters for gaps in magnetic data, whereas for electric field data, the training data is the same length as the gap. In the case of severe synthetic contamination — 45-50% of windows with spikes implanted — I employed α=0.85, whereas when less than 10% data are contaminated, α=0.03 is sufficient. Initial choice of α is typically done by visual inspection of activity ratio time series.

Two examples of synthetically contaminated data are now described, where spikes were implanted in otherwise clean time series at Parkfield, and then IARWR was applied to these data. Implanted spikes randomly alternated between sinc functions, chirp functions, and scaled uniform random zero-mean noise. The raw time series (Figure 1), span one day, and show good signal to noise ratios across all bands, ranging from 30-40 dB from 100-1000s, 20-30dB from 10-100s, and around 10dB in the dead band. Windowing these time series

results in 450 time windows. In the first synthetic contamination experiment, 225 of the 450 time windows were selected at random for implanted spikes. For each selected window, one channel at PKD was selected at random for contamination. Figure 4 shows the synthetically contaminated PKD time series. Spike identification was performed by analysis of intersite activity ratios between PKD and the simultaneous (uncontaminated) data at SAO. Wiener filters were trained on time series drawn from a clean 30 minute segment of data recorded the following day, and replacement data were calculated by convolution of these filters and clean channels simultaneous with the spikes. Application of the IARWR algorithm yielded the cleaned time series shown in Figure 5. The original, contaminated, and cleaned time series were all processed using robust remote-reference code (Egbert (1997)), and resultant sounding curves are shown in Figure 6. The sounding curves can readily be seen to show significant improvement in low frequency estimates after IARWR cleaning.

A second synthetically contaminated dataset is shown in Figure 7(A) with 45% of windows contaminated. In this case all channels at PKD are implanted with spikes, rather than just one channel. The IARWR method was applied just as in the previous case. Even given this severely contaminated data, IARWR yielded significantly stabilized processing results shown in Figure 7(B).

IARWR was also applied to data acquired at Parkfield which was contaminated in the raw record, i.e. no synthetic contamination was applied to this data. Signal to noise ratio in the dead band during the acquisition of this data at both PKD and the remote site was approximately 0dB. Contamination was present at both sites, rather than only at one site as was the case for synthetic data. Processing results calculated with raw and IARWR-treated time series are shown in Figure 8. In this case, the longest period processing results

improved significantly, but the poor results in the dead band remain after treatment. This is because the effective SNR in the dead band cannot be improved when inputs to the Wiener filter contain no signal in that band.

DISCUSSION

Despite modern robust remote reference (RRR) processing techniques, noise spikes in raw MT data can severely corrupt apparent resistivity curves at low frequencies, as shown in Figures 6, 7, and 8. MT data processing typically involves decimating time series and windowing the data (similar to the windowing applied here to obtain variance ratios). Statistical processing is applied — at each frequency — to the set of Fourier coefficients (FCs) from all windows. The robust method of Egbert (1997) relies on a tendency for median of the FC amplitude distribution to be stable, even when noise is present. A finite dataset however is limited in the number of FCs it can offer at each harmonic, and in general, fewer FCs are available as the period of interest increases. Thus at long periods, in the presence of noise, the sample median may be unrepresentative of the median which would be observed in a noise free environment. In these cases, robust methods which migrate outliers towards the sample median fail. IARWR seems to stabilize these median estimates at long periods. At short periods, the method will be less effective, since there are many more estimates of short period transfer functions (TFs) from a given length time series, so repairing even a large number of windows may not significantly perturb the median TF.

All synthetic spikes implanted in the data were correctly identified for replacement, and no windows were flagged which did not contain spikes. IARWR is not very prone to false positives as observed (Kappler et al. (2010)) when the method was applied to a four-year

dataset. Additional insurance against the use of contaminated training data is obtained by calling a final loop in the code which compares the variance of each flagged window before replacement to after. If no reduction in variance occurs the window is not replaced — and is presumably flagged for supervised despiking. If multiple sites are available, replacement data could be calculated using each of several sites independently. With only two sites available in this dataset, no experiments were performed to determine the effectiveness of this variant on the algorithm.

The method is limited in its application to removing disturbances of a transient nature. One cannot fairly replace the entirety of a many hour record based solely on a half hour of clean training data since the information regarding the E to H ratios at periods longer than the training data is simply not present in so short a segment of data. Because of this, continuous noise, present over the duration of acquisition cannot be removed by this method in theory. Choice of thresholds for spike identification is important, but by examining a few distributions of variance ratio, the user can select an appropriate threshold easily. In a worst case, identification can be done in a supervised way. It is essential that some uncontaminated time series be available simultaneously at both sites in order to accurately calculate Wiener filter coefficients. In this paper, selection of training data was done by simply selecting (at random), unflagged time series, but training data could also be selected manually.

Clearly some artifacts will be introduced into the data at the region boundaries where the data are spliced together. These artifacts appeared to be well within the capabilities of robust RR processing to handle however if the user encounters a case where errors are suspected to be sourced by these splices, or by the trendline in the replacement data, it may

be profitable simply place the synthetic data directly into $\mathcal{R}_\ni$ without a trendline or match filter ($n_{med} = 1$), and then pass the cleaned data with the CANC filter of Hattingh (1989). While improved quality of fitting synthetic to true data (in the case of artificial spikes) was observed when a trend linking the left and right edges of the deleted window is added to the synthetic time series, a significant degradation of apparent resistivity stability was not observed when the trend line was omitted. Again this suggests that RRR processing can handle the artifacts relating to such a trend, thus the user may consider omitting this trend. Also, it is clearly inappropriate to apply a trend in the case of large step offsets. If the user wishes to include the trend for spikes, but not for steps, they can supervise the algorithm, sorting spikes from steps, or can apply an algorithm which discriminates steps from other transient disturbances. One way to do this, is as follows: Take the deleted window of length L, together with L points left and right of the window, and detrend it; call this time series x. Calculate the time series of partial sums $p(x)$. In the case of a spike, $p(x)$ will have the character of noise and a near-zero mean, but in the case of a step disturbance, $p(x)$ will have a distinct "V-shape" and fitting $p(x)$ with a two-segment piecewise linear fit will yield a signature of two lines of nontrivial slope with opposing signs, with the critical point lying in the deleted window.

A considerable amount of the time programming the IARWR algorithm is spent in book-keeping during the spike replacement phase. Physical limitations of the machine-memory for executing the algorithm render the natural datastream too large to read at once; therefore the data must be stored in sections, such as the day long sections used in the synthetic examples. A challenge occurs when boundaries occur in data storage. Specifically, when a spike or a sequence of spikes occurs in the first or last window of a day, reading the sequence

of data in for replacement requires special handling. Similarly the 5% taper edge segments of data need to be read from a separate file. Coding around these cases requires care and should be addressed early on.

Variance is considered as the activity measure here, but a different intersite comparative measure can be used such as coherence, or wavelet coefficient amplitude (Trad and Travassos (2000)), eigenvalues (Kappler et al. (2010)), or others. The key is to select a measure which makes windows of length L which containing spikes stand out as statistically significant from those windows which do not contain spikes.

Conclusions

A method of automated spike identification and replacement has been described. It has a particular application in impedance estimation in magnetotellurics where only a few spikes or steps can severely corrupt cross-power averaging. The activity ratios are generally applicable for anomaly detection in arrays which typically record nearly stationary time series. The identification is based on identifying windows of data where the ratio of variance between distant instruments recording the same field type is anomalous. Once an outlier (corrupt window) was identified, Wiener filtering was used to replace the corrupt data with synthetic data. The Wiener filter coefficients are determined on relatively clean data, and as such it is critical to have at least some uncontaminated data at multiple sites to implement data replacement. Examples of synthetically contaminated data were treated with the IARWR method. All synthetic spikes implanted in the data were correctly identified for replacement, and no clean windows were incorrectly flagged. In the event that a spike

is misidentified, the algorithm typically fills the data in with a dictated time series that is nearly indiscernible from the recorded time series. The IARWR algorithm outlined here significantly reduced the effect of transient time domain noise contamination on apparent resistivity curves in the cases of synthetic contamination. IARWR was also applied to real data where transient noise of unknown origin was recorded in the field, and yielded significant improvements to processing results.

ACKNOWLEDGMENTS

Data acquisition was supported by USGS Grant Numbers 05HQGR0077 and 05HQGR0079 and the efforts of Sierra Boyd. The Berkeley Seismological Laboratory acted to warehouse the data presented here. I am indebted to Frank Morrison for valuable discussions, and to Ute Weckman, Erika Gasperikova, and two anonymous reviewers for critical reading of this manuscript which resulted in significant improvement to the work. The author also acknowledges DOE-LBNL contract number DE-AC02-05CH11231.

REFERENCES

Chave, A., and D. Thomson, 1989, Some comments on magnetotelluric response function estimation: J. geophys. Res, **94**, 215–14.

Edwards, K., and L. Hastie, 1997, Processing magnetotelluric data with modern statistical and numerical techniques: Exploration Geophysics, **28**, 43–47.

Egbert, G., 1992, Noncausality of the discrete-time magnetotelluric impulse response: Geophysics, **57**, 1354.

———, 1997, Robust multiple station magnetotelluric data processing: Geophysical Journal International, **130**, 475–496.

Egbert, G., J. Booker, and A. Schultz, 1997, Very long period magnetotellurics at Tucson Observatory: Estimation of impedances: Journal of Geophysical Research-Solid Earth, **97**.

Fontes, S., T. Harinarayana, G. Dawes, and V. Hutton, 1988, Processing of noisy magnetotelluric data using digital filters and additional data selection criteria* 1: Physics of the Earth and Planetary Interiors, **52**, 30–40.

Hattingh, M., 1989, The use of data-adaptive filtering for noise removal on magnetotelluric data: Physics of the Earth and Planetary Interiors, **53**, 239–254.

Jones, A., A. Chave, G. Egbert, D. Auld, and K. Bahr, 1989, A comparison of techniques for magnetotelluric response function estimation: J. geophys. Res, **94**, 14–201.

Junge, A., 1996, Characterization of and correction for cultural noise: Surveys in Geophysics, **17**, 361–391.

Kappler, K., H. Morrison, and G. Egbert, 2010, Long-term monitoring of ulf electromagnetic fields at parkfield ca: J. geophys. Res.

Leśniak, A., T. Danek, and M. Wojdyła, 2009, Application of Kalman Filter to Noise

Reduction in Multichannel Data: Schedae Informaticae, **17**, 63–73.

Oettinger, G., V. Haak, and J. Larsen, 2001, Noise reduction in magnetotelluric time-series with a new signal–noise separation method and its application to a field experiment in the Saxonian Granulite Massif: Geophysical Journal International, **146**, 659–669.

Perrier, F., and P. Morat, 2000, Characterization of electrical daily variations induced by capillary flow in the non-saturated zone: Pure and Applied Geophysics, **157**, 785–810.

Szarka, L., 1987, Geophysical aspects of man-made electromagnetic noise in the eartha review: Surveys in Geophysics, **9**, 287–318.

Trad, D., and J. Travassos, 2000, Wavelet filtering of magnetotelluric data: Geophysics, **65**, 482–491.

Tzanis, A., and D. Beamish, 1989, A High-Resolution Spectral Study of Audiomagnetotelluric Data and Noise Interactions: Geophysical Journal International, **97**, 557–572.

Weckmann, U., A. Magunia, and O. Ritter, 2005, Effective noise separation for magnetotelluric single site data processing using a frequency domain selection scheme: Geophysical Journal International, **161**, 635–652.

LIST OF FIGURES

sounding curves at long periods are clearly much more reasonable than those generated from the contaminated data.

7 (A) "Severe" synthetic contamination of data from Figure 1. Rather than implanting spikes in one channel for each randomly selected window (Fig. 4), here a spike is planted in all Parkfield channels. Y-axes values are counts. (B) Apparent resistivity curves obtained from original (red), corrupt (Blue), and repaired (Black) datasets. IARWR data show dramatic improvement over the contaminated data.

8 Four-year median apparent resistivity calculated at PKD (error bars are smaller than symbol size) shown in red, ρ_a calculated from field data from a six hour time series from Julian day 217, 2003 (blue), and ρ_a calculated from field data treated with IARWR (black). The IARWR data show dramatic improvement over the contaminated data at long periods, but virtually no improvement at short periods as discussed in the text.

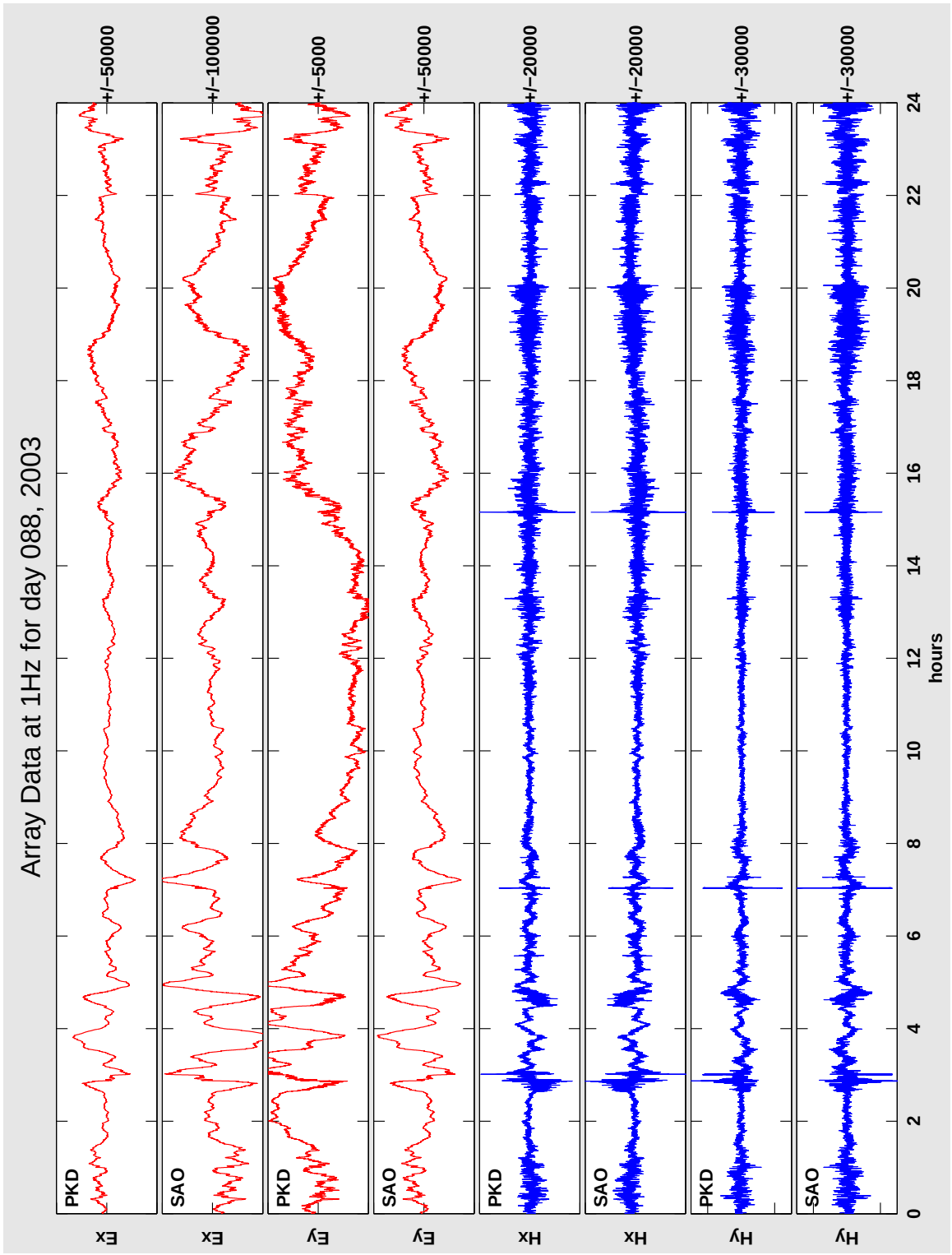

Figure 1: Mean-subtracted array data spanning 24 hours. Electric fields in red and magnetics in blue. Plots alternate between sites at each field polarity. Y values are counts (axis limits shown on right). Note the similar character of the recorded waveforms, especially the simultaneous variations in co-linear channels at distant sites.

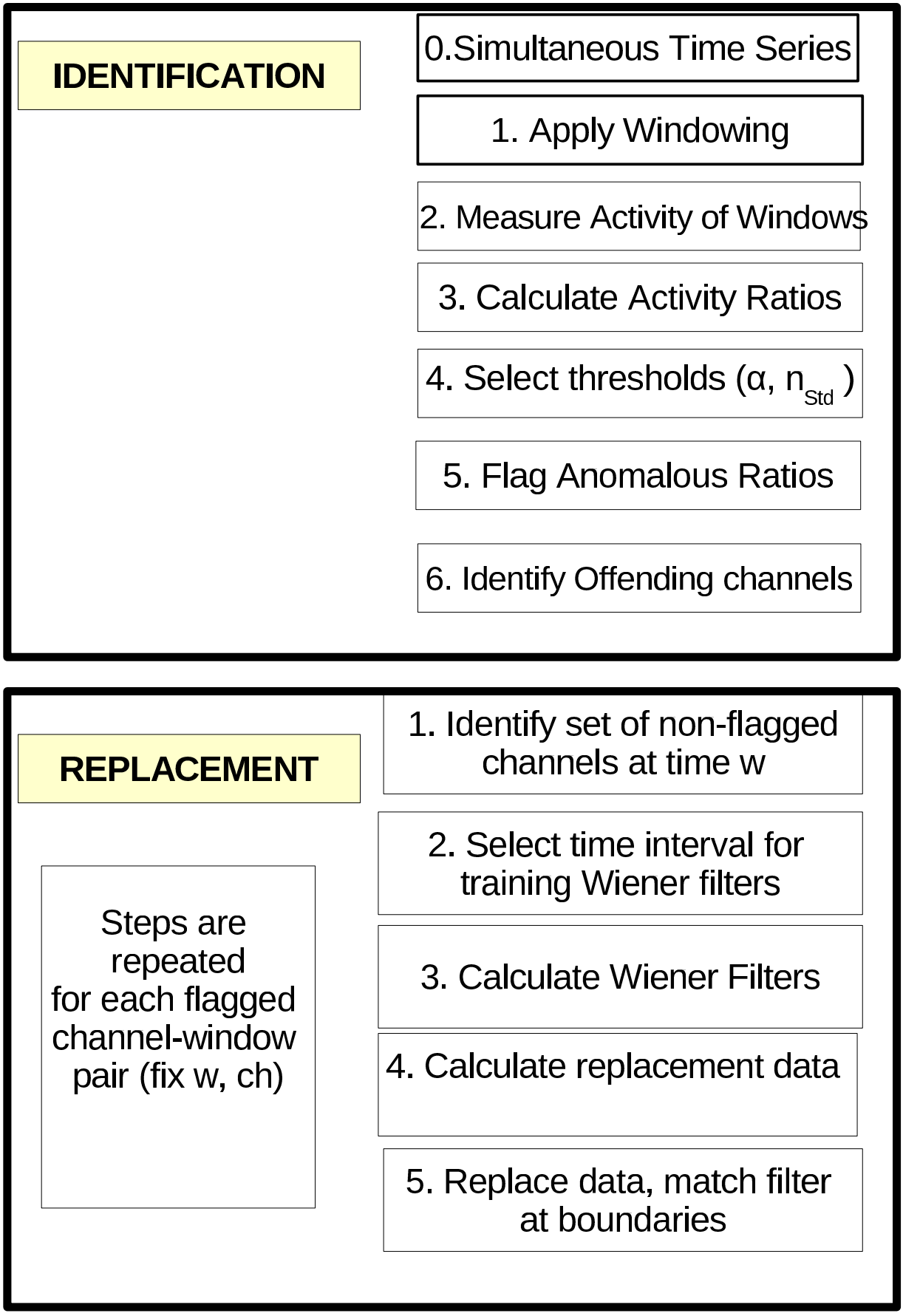

Figure 2: Chart detailing the primary stages in the IARWR method. The method is broken into two primary algorithms; identification and replacement. Each subtask listed is described in the text.

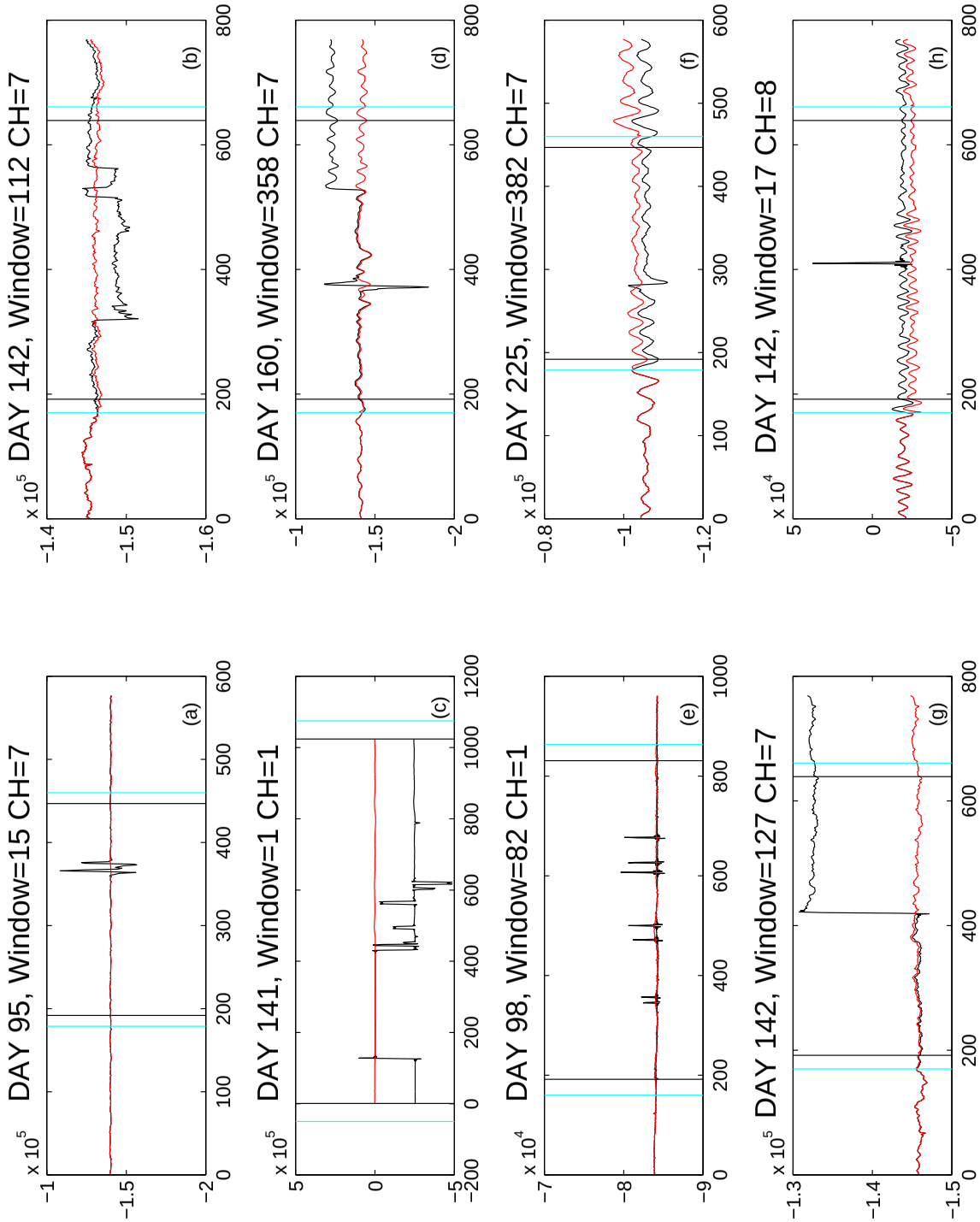

Figure 3: Some example despiking results. Recorded data is black, and plausible data calculated by the method outlined in the text is red. Black vertical lines bound the clipped section ($\mathcal{R}_3$) and the regions bounded between the cyan and black vertical lines are $\mathcal{R}_2$ and $\mathcal{R}_4$ (the splice regions). All y-axes are in units of data logger machine counts, and all x-axes are time in seconds.

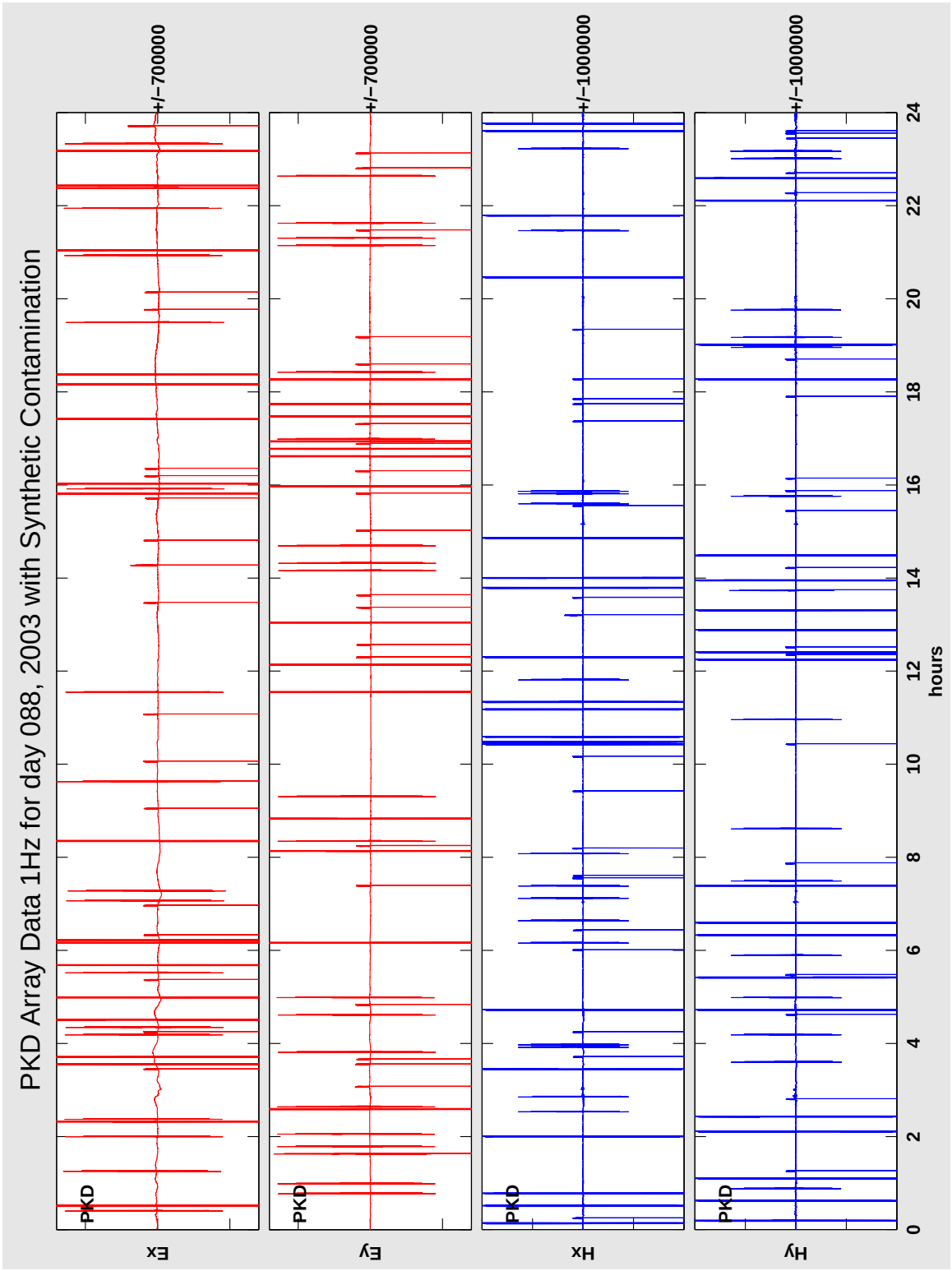

Figure 4: Synthetic contamination: The Parkfield electric and magnetic time series shown in Figure 1, but with half of the time windows have been randomly selected for contamination. Spikes are distributed evenly amongst the four channels. All y-axes are in units of data logger machine counts (axis limits shown on right).

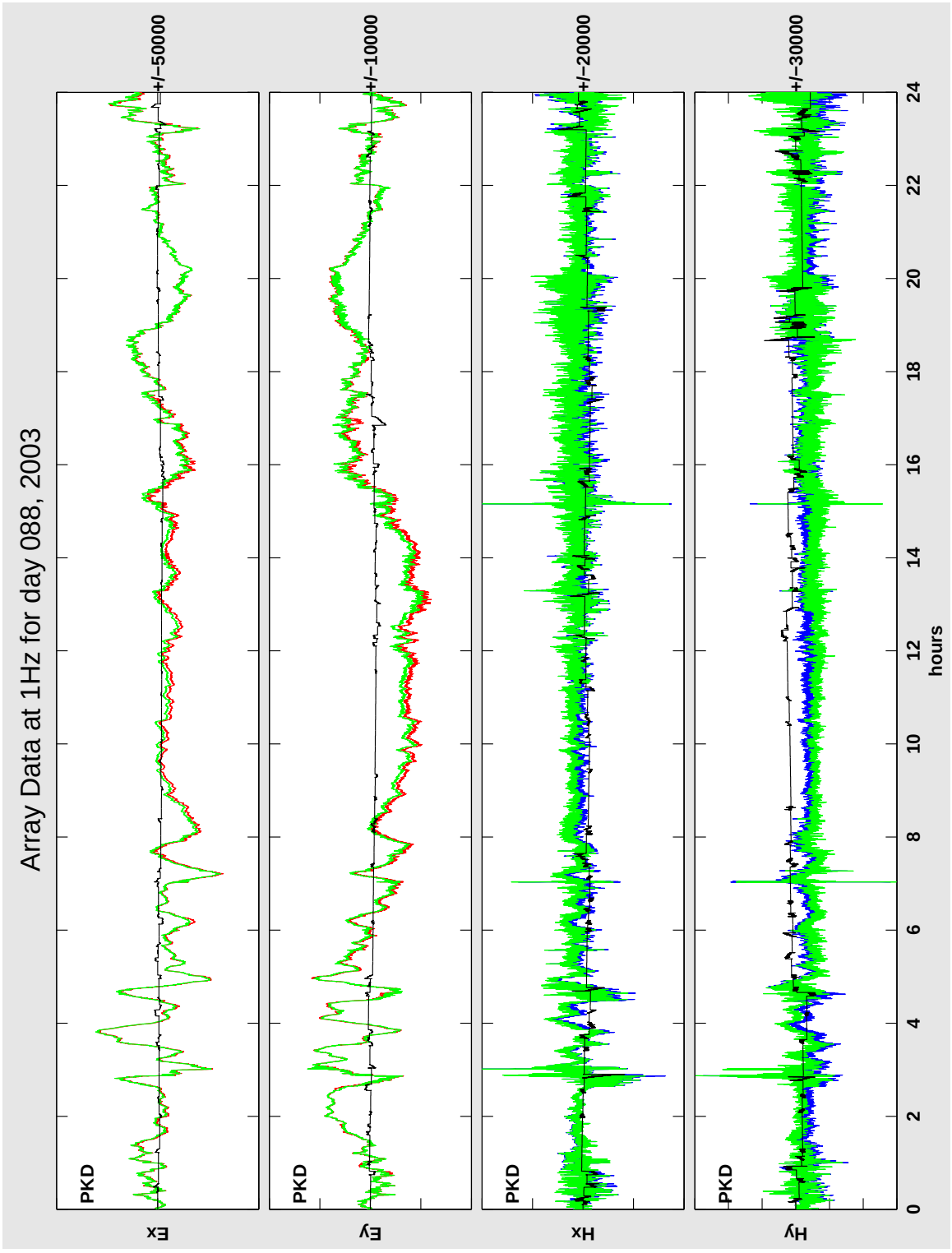

Figure 5: Artificially contaminated time series from Figure 4 after IARWR. Original data in red (electric) and blue (magnetic). Replacement data are green, and residuals are black. All y-axes units are datalogger counts. Residuals clearly deviate from zero but artifacts in the residual time series are less detrimental to MT processing than spikes.

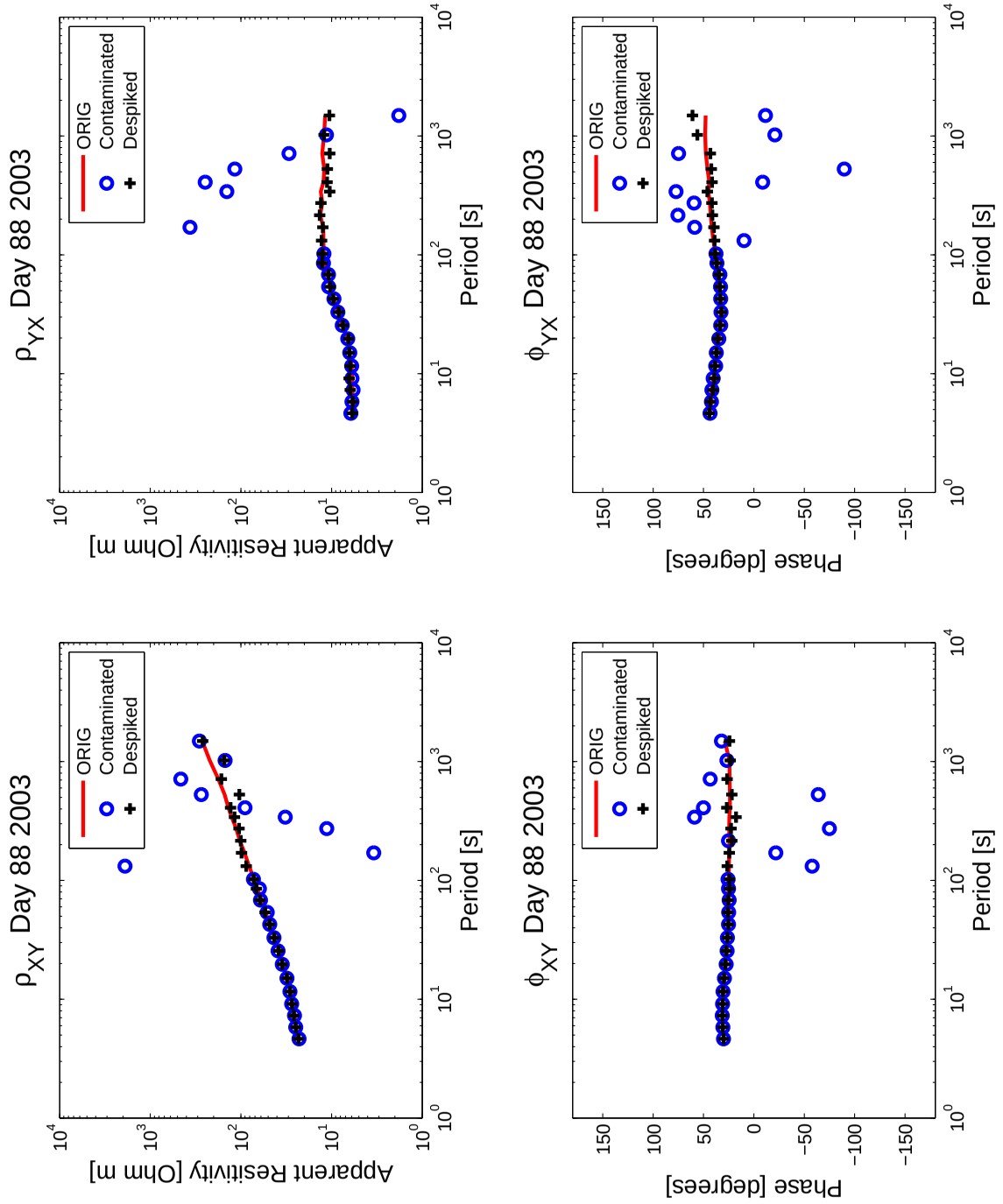

Figure 6: Sounding curves generated by (a) the actual field data of Figure 1 in red, (b) the artificially contaminated data — shown as blue circles, and (c) the IARWR data marked by black crosses. While the original ρ_a curves are not quite recovered, the IARWR generated sounding curves at long periods are clearly much more reasonable than those generated from the contaminated data.

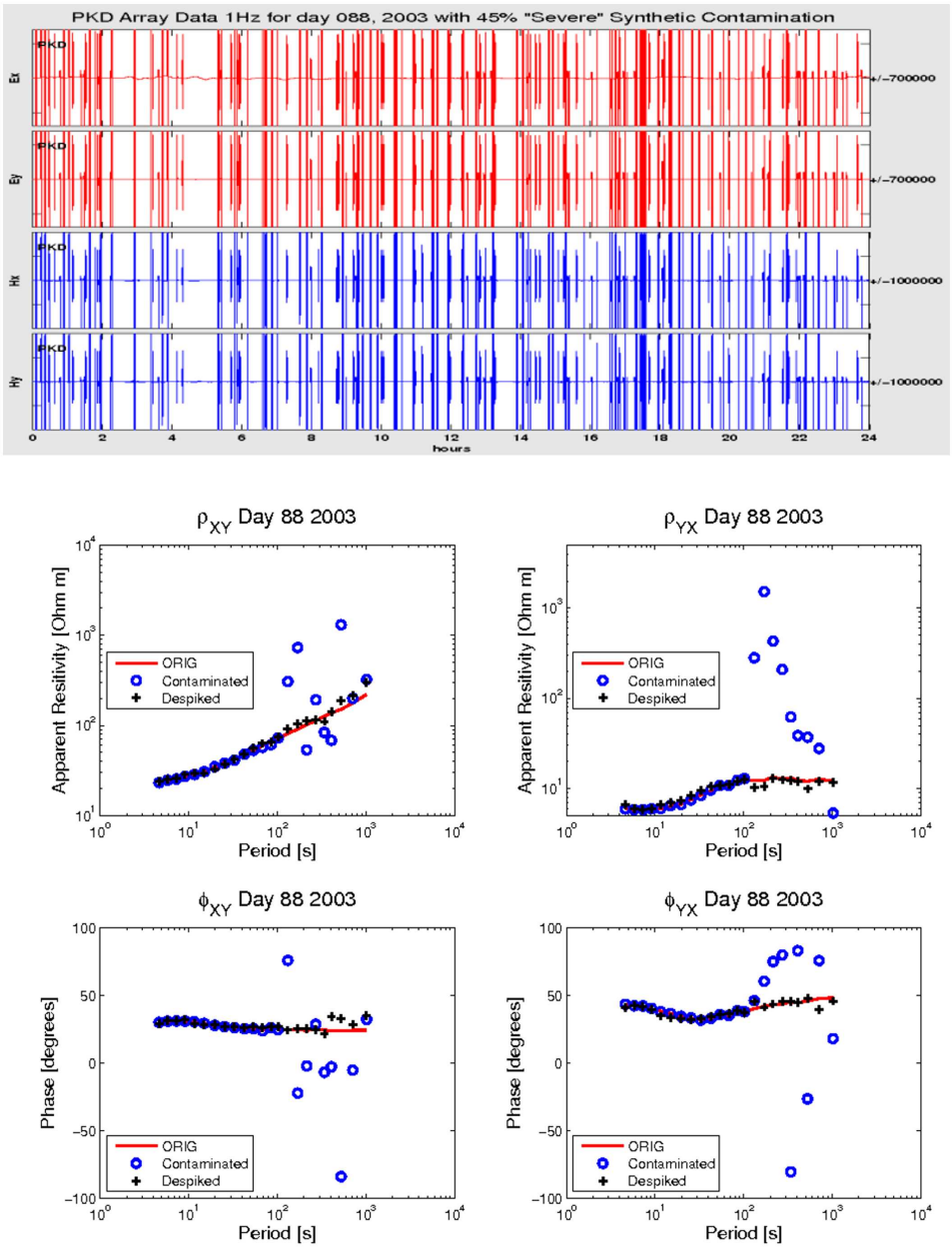

Figure 7: (A) "Severe" synthetic contamination of data from Figure 1. Rather than implanting spikes in one channel for each randomly selected window (Fig. 4), here a spike is planted in all Parkfield channels. Y-axes values are counts. (B) Apparent resistivity curves obtained from original (red), corrupt (Blue), and repaired (Black) datasets. IARWR data show dramatic improvement over the contaminated data.

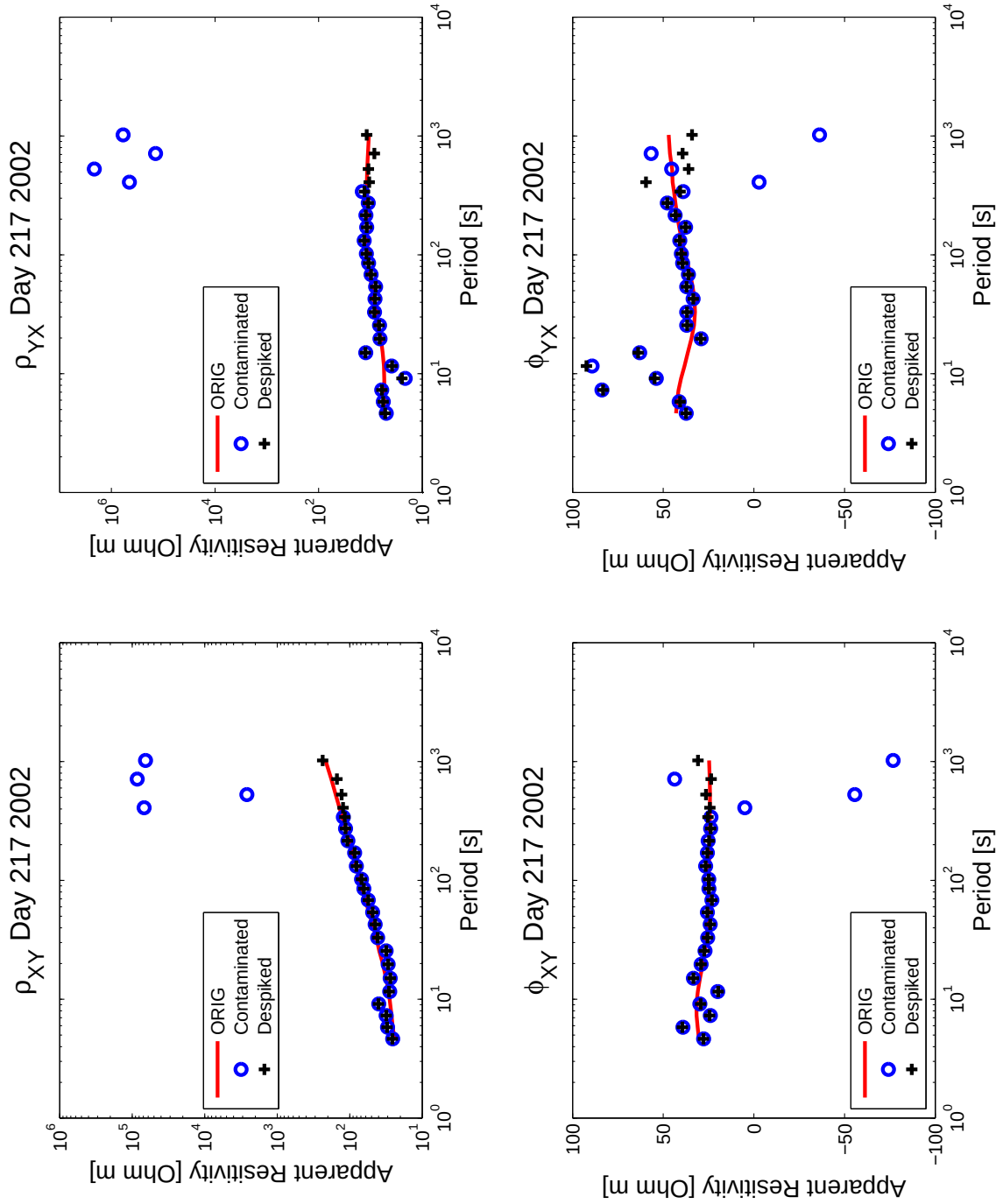

Figure 8: Four-year median apparent resistivity calculated at PKD (error bars are smaller than symbol size) shown in red, ρ_a calculated from field data from a six hour time series from Julian day 217, 2003 (blue), and ρ_a calculated from field data treated with IARWR (black). The IARWR data show dramatic improvement over the contaminated data at long periods, but virtually no improvement at short periods as discussed in the text.