Platz, A., Weckmann, U. (2019): An automated new pre-selection tool for noisy Magnetotelluric data using the Mahalanobis distance and magnetic field constraints. - *Geophysical Journal International*, *218*, 3, pp. 1853—1872.

DOI: http://doi.org/10.1093/gji/ggz197

Geophysical Journal International

Geophys. J. Int. (2019) **218,** 1853–1872
Advance Access publication 2019 April 30
GJI Marine Geosciences and Applied Geophysics

doi: 10.1093/gji/ggz197

An automated new pre-selection tool for noisy Magnetotelluric data using the Mahalanobis distance and magnetic field constraints

A. Platz[1,2] and U. Weckmann[1,2]

[1] *GFZ German Research Centre for Geosciences, Potsdam, Germany. E-mail: aplatz@gfz-potsdam.de*
[2] *University of Potsdam, Institute of Earth and Environmental Science, Potsdam, Germany*

Accepted 2019 April 29. Received 2019 April 15; in original form 2018 September 14

SUMMARY

In Magnetotellurics (MT) natural electromagnetic field variations are recorded to study the electrical conductivity structure of the subsurface. Thereby long time-series of electromagnetic data are subdivided into smaller segments, which are Fourier transformed and typically averaged in a statistically robust manner to obtain MT transfer functions. Unfortunately, nowadays the presence of man-made electromagnetic noise sources often deteriorates a significant fraction of the recorded time-series by overprinting the desired natural field variations. Available approaches to obtain undisturbed and high quality MT results include, for example robust statistics, remote reference or multi-station analyses which aim at the removal of outliers or uncorrelated noise. However, we have observed that intermittent noise often affects a certain time span resulting in a second cluster of transfer functions in addition to the expected true MT distribution. In this paper, we present a novel criterion for the detection and pre-selection of EM noise in form of outliers or additional clusters based on a distance measure of each data segment with regard to the centre of the data distribution. For this purpose, we utilize the Mahalanobis distance (MD) which computes the distance between two multivariate points considering the covariance matrix of the data that quantifies the shape and the size of multivariate data distributions. As the MD considers the covariance matrix, it corrects not only for different variances but also for any correlation between the data. The computation of both, the mean value and covariance matrix, is susceptible to ouliers (e.g. noise) and requires a statistically robust estimation. We tested several robust estimators, for example median absolute deviation or minimum covariance determinant algorithm and finally implemented an automatic criterion using a deterministic minimum covariance determinant algorithm. We will present results using MT data from various field experiments all over the world, which illustrate successfull data improvement. This approach is able to remove scattered data points as well as to reject complete data cluster originating from noise sources. However, like all purely statistical algorithms the criterion is limited to cases where the majority of the recorded data is well-behaved, that is noise content is below 50 per cent. If the majority of data points originates from noise sources, the new criterion will fail if used in an automatic way. In these cases, additional input by the user either manually or in an automated fashion can be utilized. We therefore suggest to use an add-on criterion to back the MD selection and subsequent robust stacking in form of a physically motivated constraint based on the magnetic incidence direction. This property indicates whether the magnetic field originates from various sources in the far field or from a strong and well defined source in the near field.

Key words: Magnetotellurics; Statistical methods; Time-series analysis.

1 INTRODUCTION

The Magnetotelluric (MT) method senses the electrical conductivity structure of the subsurface through measurements of orthogonal components of natural magnetic and electric field variations at the Earth's surface. In the frequency domain, we can mathematically describe the linear relationship between horizontal magnetic and

electric field components by the impedance tensor **Z**:

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{pmatrix} \cdot \begin{pmatrix} B_x \\ B_y \end{pmatrix} \quad (1)$$

with **E** being the electric field [V m^{-1}], **B** the magnetic field [*T*] and Z_{ij} $(i; j = x; y)$ the components of the impedance tensor **Z** [m s^{-1}]. The complex valued impedance tensor carries the information about the Earth's electrical conductivity structure and if not used in a monitoring application, this quantity is treated as time independent. However, if we use eq. (1) within a statistical model, we have to consider an additional error term. Within the course of estimating the components of **Z**, time-series are often divided into shorter time segments that are treated as individual and independent 'measurements' and subsequently transformed into frequency domain. Spectral values of these different segments are used to calculate the impedance tensor elements. To solve eq. (1), least-squares (LSQ) methods can be applied and give unbiased results if noise affects only the output channels of the equation system (E_x, E_y for the impedance in eq. 1): In addition, they are statistically optimal and/or unbiased if noise is independent and Gaussian distributed (Weckmann *et al.* 2005). Unfortunately, in many regions we observe a growing number of industrial sites and advanced electric infrastructure which results in strong man-made electromagnetic (EM) noise. This noise contribution is superimposed on the desired natural MT signal and needs to be addressed by advanced data processing approaches as it has long been known that simple single-station ordinary LSQ methods cannot be used to calculate a meaningful impedance **Z** (e.g. Sims *et al.* 1971) as a significant portion of the time segments distorts the calculated mean value. The introduction of robust statistics and the remote reference method significantly improved the estimation of the transfer functions. Nowadays several robust statistical algorithms (e.g. Egbert & Booker 1986; Chave *et al.* 1987; Chave & Thomson 1989, 2004 ; Ritter *et al.* 1998; Smirnov 2003), mainly relying on data-adaptive weighting schemes, are in use to decrease the influence of outliers. The remote reference method involves simultaneously recorded EM fields from at least two sites that are composed of highly correlated signal and uncorrelated noise (Goubau *et al.* 1978; Gamble *et al.* 1979); in addition, accurate time is a prerequisite. The most promising results are achieved with a combination of both methods: a robust remote reference processing (e.g Larsen 1989; Oettinger *et al.* 2001; Chave & Thomson 2004). In comparison to a single-station approach, remote reference processing uses an errors-in-variables model and therefore accommodates for noise in all channels (Chave & Jones 2012); as a consequence remote reference results have commonly been accepted as superior to the simple single-site approach (e.g. Jones *et al.* 1989; Larsen *et al.* 1996; Egbert 1997). Although, field experiments are almost always designed to record one or several remote stations simultaneously, it is often difficult to identify a suitable reference site as cultural noise signals can be widespread and coherent over large areas. If both local and reference sites are affected by the same noise, the remote reference method can give misleading processing results (e.g. Pedersen *et al.* 1992; Ritter *et al.* 1998). Using data from a local station array, Ritter *et al.* (1998) showed that in some frequency bands up to 99 per cent of the measured time-series were contaminated with correlated noise.

The multi-station processing presented by Egbert (1997) is based on a multivariate model, which uses all available data from many simultaneously recording stations to improve the signal-to-noise (S/N) ratio. Egbert's approach (1997) is a robust version of an underlying principal component analysis; he proposed to weight the different channels according to their uncorrelated noise. Although it is desirable to use simultaneous recordings from different MT stations with some of the above mentioned processing approaches, we often face the problem of correlated EM noise over large distances or sometimes insufficient time accuracy. As a result, remote reference and multi-station approaches cannot be used for data processing and the practitioner is set back to single-site processing.

While outliers are in general satisfactorily handled through the robust statistic approach, intermittent EM noise contributions, for example originating from near-field noise sources and/or the violation of the plane-wave assumption, often form a own cluster of transfer functions overlying the MT signal distribution. In this case, robust processing schemes can be succoured through a frequency domain pre-selection approach that reduces the amount of EM noise to a level robust statistics can deal with. Travassos & Beamish (1988) and Weckmann *et al.* (2005) suggested interactive selection algorithms that are based on physical criteria to eliminate disturbed parts of the MT recordings. For their data pre-selection, Weckmann *et al.* (2005) examined physical properties such as spectral power densities, coherences or polarization directions of MT data and observed that these quantities calculated for each subsequent event exhibit temporal variations within the recording time. Such a pattern indicates noise sources that are switched on temporarily or are coupled with production periods. In heavily industrialized or populated areas, night-time recordings are often regarded as less noise affected as during daytime. To overcome this obstacle, noisy parts of the time-series can be truncated based purely on time (e.g. only night-time recordings). But the practitioner might pay dearly for omitting all daytime recordings with an excessive reduction of the length of the time-series. Furthermore, these data pre-selection approaches are often very tedious and time consuming and require experienced users.

Therefore, we introduce an automated approach to improve the S/N ratio prior to the final and robust stacking of the MT transfer functions. On one hand side, we try to classify the distribution of MT transfer functions of sequenced events through their distance to the mean value of the distribution and their variance. Distances are traditionally computed using the Euclidean distance, but for MT data (either cross- and autospectra or impedance tensor components) the usage of the Mahalanobis distance (MD) might be more expedient.

The MD is used in multivariate statistics for outlier detection (e.g. de Maesschalck *et al.* 2000; Filzmoser *et al.* 2005; Srinivasaraghavan & Allada 2006; Friebel *et al.* 2010; Brereton 2015) or in discriminant analysis (e.g. Kleinschmidt *et al.* 1994; Wu *et al.* 1997; Hayashi *et al.* 2001; Srinivasaraghavan & Allada 2006) in a wide spectrum of fields reaching from biology and chemistry to lean manufacturing; however, it has not been applied as a selection criterion prior to robust stacking. Furthermore, the squared Mahalanobis distance is part of the density function of the multivariate Gaussian distribution. Since outliers have a strong influence on the empirically estimated mean value and covariance matrix, a robust estimation of the data centre and the covariance matrix is essential for an effective MD calculation. Several methods have been used for this purpose, for example the median absolute deviation from the median (Gnanadesikan & Kettenring 1972; Huber 1981; Falk 1997; Friebel *et al.* 2010) or more complex algorithms like, for example the minimum volume ellipsoid or the minimum covariance determinant method (Rousseeuw 1984, 1985). We tested different robust approaches with MT data and decided upon a deterministic

minimum covariance determinant (MCD) algorithm, which also allows that nearly half of the data can either be outliers or belong to a second data cluster. We implemented the approach as a confinement and pre-selection criterion into the robust EMERALD processing (Ritter *et al.* 1998; Weckmann *et al.* 2005; Krings 2007). We will show improved processing results from various MT stations (South Africa, Germany, Tajikistan and Venezuela, respectively; Muñoz *et al.* 2010; Korolevski *et al.* 2014; Schmitz *et al.* 2013; Platz 2018). However, like all purely statistical algorithms this first criterion is limited to cases where the majority of the recorded data is well-behaved, that is noise content is below 50 per cent. If the majority of data points originates from noise sources, this criterion will fail if used in an automatic way. In these cases, several options can be applied: (i) usage of a remote station, (ii) additional input by the user either manually or (iii) in an automated fashion. We therefore suggest to use an add-on criterion to back the MD selection and subsequent robust stacking in form of a physically motivated constraint based on the magnetic incidence direction. This second criterion was originally designed as an add-on for the MD criterion, but it can also be successfully applied without the MD criterion and has the advantage that it does not require any user input. Although we will mainly show advances and limitations of two novel criteria for single-site processing to estimate the impedance tensor, these tools can also be applied to the vertical magnetic transfer and inter-station transfer functions, as well as for remote reference or multi-station processing.

2 ELECTROMAGNETIC NOISE IN MT DATA

We will first focus on the characterization of intermittent EM noise and its removal, as it is not the purpose of this paper to compare different MT processing approaches.

Although the MT results obtained from the EMERALD processing (Fig. 1a) for an exemplary station from South Africa allows to perceive the true curve of apparent resistivity and phase, some scatter can be observed in the period range >1 s. A slightly improved result is obtained by using the processing approach based on Egbert & Booker (1986, fig. 1a). Both processing algorithms exhibit some inherent differences and since the main work flow of the EMERALD software package is not widely known, we will briefly summarize it below. Eq. (1) or similar equations for remote reference processing or the vertical magnetic or inter-station transfer functions are solved by a bivariate linear regression, thereby:

(i) Bandpass filtered time-series are divided into short, contiguous time windows or segments of fixed length, for example 128 samples.

(ii) Tapered segments are Fourier transformed and corrected for instruments responses (= events).

(iii) For each target period and event, auto- and crossspectra estimates are computed from the calibrated Fourier coefficients (e.g. $[E_x B_x^*]_i$) and averaged into logarithmically distributed target periods (*cf.* Schmucker & Weidelt 1975) which subsequently contribute to the calculation of the impedance tensor components (e.g. $Z_{xy} = \frac{\langle E_x B_y^* \rangle \langle B_x B_x^* \rangle - \langle E_x B_x^* \rangle \langle B_x B_y^* \rangle}{\langle B_x B_x^* \rangle \langle B_y B_y^* \rangle - \langle B_x B_y^* \rangle \langle B_y B_x^* \rangle}$).

The bracketed terms represent stacked auto- and cross-spectra, with for example, $\langle E_x B_x^* \rangle = \sum_{i=1}^{N} w_i \left[E_x B_x^* \right]_i$ with the asterisk denoting the complex conjugate. The weights w_i are calculated by stacking the smoothed spectra from many events with an iterative robust weighting algorithm described in the appendix of Ritter *et al.* (1998). The robust algorithm to address the regression problem (Junge 1990, 1992, 1994) combines two main parts: the χ^2 and the consistency criterion. The χ^2 criterion examines whether a single event spectrum fits into the majority of all data. Subsequently, the influence of this single event spectrum is increased or decreased by a robust weighting scheme using a combination of Huber and Tukey weights. This means that single event spectra data are declared as outliers, if they have large errors based on single event transfer functions. The consistency criterion reduces in a second step non-stationary contributions in the transfer functions by iteratively replacing a certain amount of noisy data with predicted data. In view of recent papers by Chave (2014) and Chave (2017), the underlying Gaussian model within EMERALD might not be adequate for MT data and might require modifications; however, in this paper we focus on two pre-stacking selection tools whereby at least the second tool is independent on the assumed underlying Gaussian distribution.

In addition, optionally a coherence threshold can be applied prior to the robust stacking to remove single events. However, if MT stations are affected by a high amount of EM noise, the MT transfer function estimates from this automatic robust data processing scheme might still be insufficient (Fig. 1a). A comparison with processing algorithms based on Egbert & Booker (1986) and Egbert (1997) reveals partly improved apparent resistivity and phase curves for periods <2 s, however, for longer periods scattered data points indicate that the influence of EM noise becomes more severe and beyond of being manageable by robust statistics. The multivariate processing is applied to single-site data with coherent noise in the channels; however since we only use five channels of one station, significantly improved results cannot be expected as this approach can play to its potential primarily by using one or more additional stations.

To apply additional filters or approaches to reduce the scatter observed in the MT curves, it is necessary to look at scatterplots or histograms of real and imaginary parts of the single event transfer functions from which the final estimates are calculated in a statistically robust manner (Figs 1c and d). These plots illustrate that the single event transfer functions are separated into two different distributions; one caused by the natural signal and the second one presumably by noise. This distribution combined of two clusters is a nightmare for each statistical approach.

Often quantile–quantile (q–q) plots are suggested to check whether the assumed model for the robust approach is valid (Chave 2014). The residual q–q plot for $T = 1\,s$ (Fig. 1b) indicates that the residuals are systematically long tailed compared to an assumed Gaussian distribution which would result in a straight line. However, they do not give additional insight into the actual problem of two clusters.

Since this sort of EM noise is often present in MT data and available processing algorithms that (partly) rely on statistics are at their limits, we suggest to use additional automatic confinement criteria to remove both, outliers and additional and unwanted distributions within the MT recordings.

3 DATA CONFINEMENT BASED ON THE MAHALANOBIS DISTANCE

The basic idea of this criterion is to confine the data to an ideally noise-free or noise reduced subset that is subsequently used in the regression problem. Outliers and events belonging to noise, for example in the tail of the distribution (as in Fig. 1d) or forming a

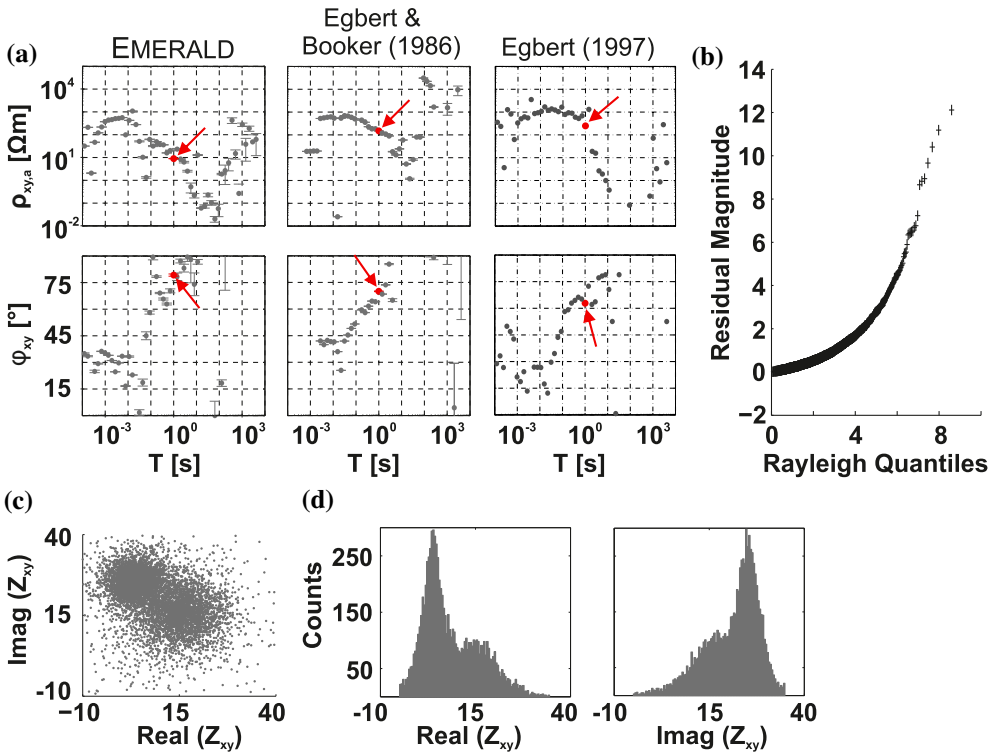

Figure 1. (a) Processing results of station SA-4 for the Z_{xy} component using the robust EMERALD (Ritter *et al.* 1998; Weckmann *et al.* 2005) as well as processing algorithms after Egbert & Booker (1986) and Egbert (1997) showing scattered apparent resistivity and phase curves for very short periods and periods >1 s due to EM noise. The multivariate processing is applied here only to single-site data with coherent noise in the channels; therefore significantly improved results cannot be expected as this approach is aimed primarily for the use of multiple stations. Whereas the robust processing after Egbert & Booker (1986) results in mostly smooth curves for periods <2 s, severely disturbed results exist for longer periods. (b) The corresponding q-q plot of the residual magnitude versus the Rayleigh distribution quantiles for the period of $T = 1$ s indicates that the residuals are systematically long tailed; the existence of two distributions as seen in (c) and (d) is concealed. (c) The Argand diagram of the real and imaginary part of the Z_{xy} component of each event of the entire time-series reveals two distributions. (d) Histograms of the real and imaginary part of all single event transfer functions Z_{xy} for $T = 1\,s$ indicate that one larger cluster and one smaller cluster are merged.

separate noise distribution in the data space, are assumed to have a larger distance to the desired MT data distribution. Therefore, only events with a distance value smaller or equal to a critical distance are considered further. The statistically motivated confinement criterion will be applied after a coherence sorting, but prior to the actual regression (see Fig. 2).

The Mahalanobis distance (MD; Mahalanobis 1936) is an important distance measure for multivariate data. It represents a generalization of the well-known Euclidean distance (ED) and allows for correlated data by taking the inverse of the covariance matrix $\mathbf{C}_x^{-1}$ into account. The distance MD_i between the ith observation of a multivariate measurement x_i (expressed in a row vector) and the data centre μ of a distribution is calculated by

$$MD_i = \sqrt{(x_i - \mu)\mathbf{C}_x^{-1}(x_i - \mu)^T} \qquad (2)$$

(Mahalanobis 1936; Lehmann 2012; Lohninger 2012; Brereton 2015). For our purpose, the row vectors x_i can be summarized into a $n \times p$ data matrix $\mathbf{X}$ with n and p as the number of observations and variables, respectively. The MD describes the distance between a multivariate point x_i and the data centre of the distribution in terms of multiples of standard deviations. Furthermore, the MD is affine invariant and unitless. The MD is commonly used for outlier detection in a wide range of science and technology applications or the production and quality control in manufactures (e.g. de Maesschalck *et al.* 2000; Dickhaus 2003; Filzmoser *et al.* 2005; Srinivasaraghavan & Allada 2006; Friebel *et al.* 2010; Brereton 2015). In all of these cases, the examined variables have different metrics/units, and therefore show different variability. In MT, the Fourier coefficients and the cross- and autospectra also exhibit such a behaviour as they originate from electric and magnetic fields, which have different units. Furthermore, it is possible that the real and imaginary parts of the used variables (either cross- and autospectra or transfer functions) have different standard deviations. As the ED does not account for this, the variable with the largest range would dominate the results. Therefore the variables have to be scaled before calculating the ED or an alternative measure as the MD have to be used. We prefer the MD, as it also corrects for any correlations between the variables, which is not possible with the ED.

The idea of using a covariance matrix in MT data processing is not new. Modern processing approaches based on multivariate regression (e.g. Egbert 1997; Smirnov & Egbert 2012) calculate the sample covariance matrix (called spectral density matrix) as part of their robust estimation scheme. In contrast to these flavours of MT data processing, the EMERALD processing suite (Ritter *et al.* 1998; Weckmann *et al.* 2005) is based on a bivariate regression. Weckmann *et al.* (2005) showed that solving each electric field component (row of the equation system 1) individually often results in smoother transfer functions in case of EM noise affecting only one component. For our criterion, we use the covariance matrix to explicitly calculate a distance value for each single event

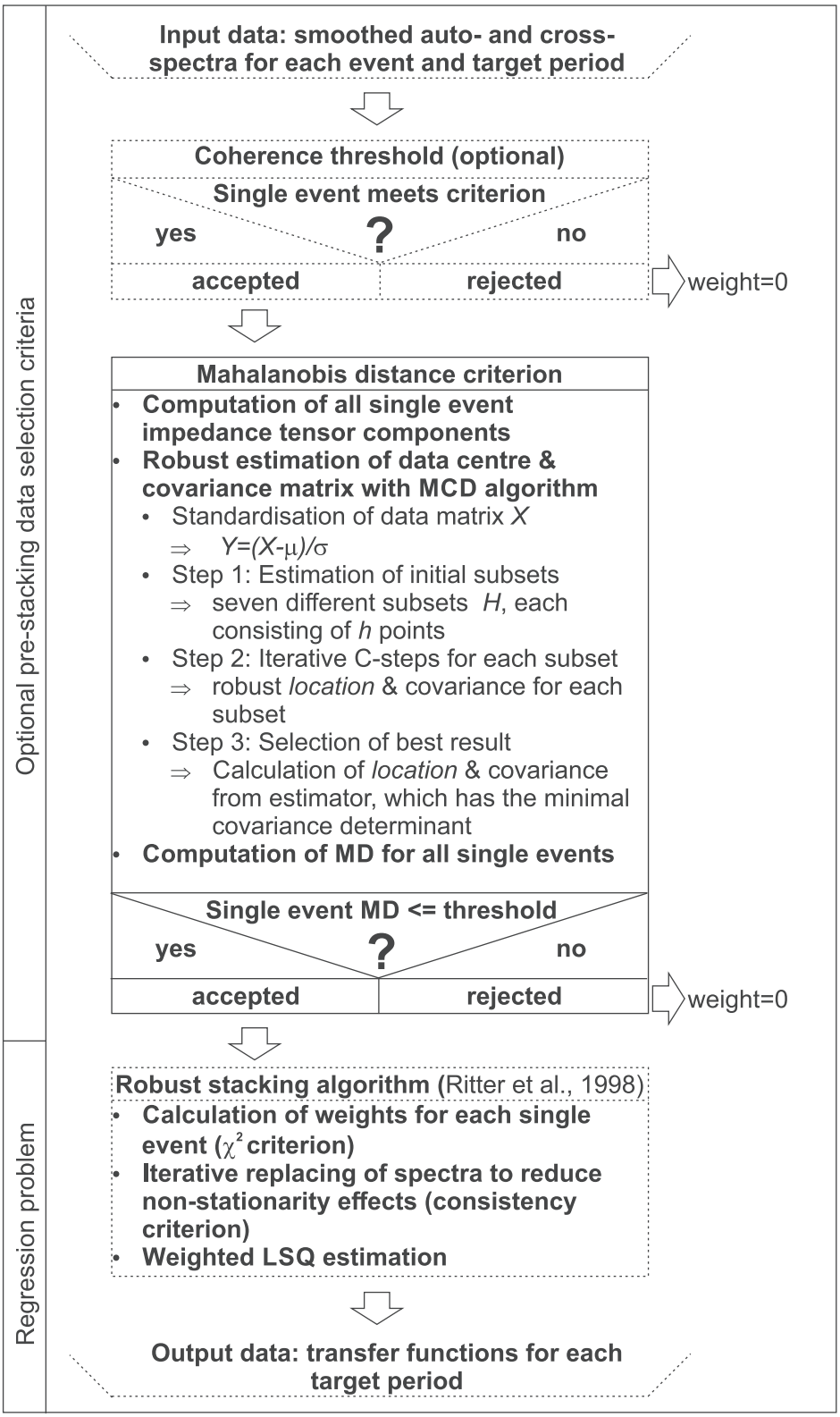

Figure 2. Work flow of the final stacking routine within the EMERALD processing (see Section 2) using smoothed auto- and cross-spectra for each event as input data. Optional data selection criteria, for example using a coherence threshold or statistical approaches as the MD criterion can be applied prior to the actual robust stacking algorithm.

and to confine data for the subsequent regression problem to remove outliers and noise clusters prior to the actual robust stacking algorithm.

We expect that single event transfer functions caused by the natural signal form one cluster in the complex plane (Argand diagram). Therefore, we use the real and imaginary parts of all single event transfer functions calculated from the smoothed auto- and cross-spectra for each event as input data for the data matrix **X**; when using the EMERALD processing to solve eq. (1) for each electric field component individually, this results in $p = 4$ variables for our data matrix (e.g. real and imaginary parts of Z_{xx} and Z_{xy}). The number of observations n is given by the number of events for the examined period and varies for broadband MT data between several tens (for the longest periods) and several hundred thousand of events (for the shortest periods).

3.1 Robust estimation of the Mahalanobis distance

Unfortunately, the computation of the MD in eq. (2) has at least two shortcomings:

(i) The calculation of the covariance matrix requires that $n > p$, that is that we need more samples or events than variables (de Maesschalck *et al.* 2000; Brereton 2015). Since we use the real and imaginary parts of the single event transfer functions for each direction of the electric field **E** (rows in eq. 1) as variables, we need at least five events per period. Due to the large number of events for most of the target periods, this restriction is negligible.

(ii) More important is the robust calculation of the MD_i values, that is the data centre μ and the covariance matrix $\mathbf{C_x}$. A straightforward calculation of these two quantities by the arithmetic mean and the sample covariance matrix will strongly depend on the outliers within the data set and is therefore not advisable, as the MD_i values will be distorted (Rousseeuw & van Driessen 1999; Filzmoser *et al.* 2005; Hubert & Debruyne 2010; Lehmann 2012).

Several robust estimators of location and scale can be found in relevant literature, for example the data centre and the covariance. The simplest class of these estimators is based on the median and the median absolute deviation (from the median, e.g. Huber 1981; Falk 1997; Friebel *et al.* 2010). However, the obtained covariance matrices do not necessarily have to be positive (semi-) definite resulting in negative squared MD values. Our tests on a variety of stations with different noise contaminations suggest that these estimators worked quite well for many stations and a broad period range, however, we occasionally observe non-positive (semi-) definite covariance matrices for long period data with very few events and for severely disturbed periods, for example. around the fundamental frequency of power grids. Therefore, we refrain from using these estimators with MT data as we aim to have positive (semi-) definite covariance matrices to ensure positive squared MD values.

We finally opt for a minimum covariance determinant (MCD) algorithm (Rousseeuw 1984, 1985; Rousseeuw & van Driessen 1999). The basic idea of the MCD algorithm is to use only an ideally noise-free subset of the entire data set to calculate data centre and covariance. The qualification of a chosen subset is assessed through the determinant of the covariance matrix. Broadly speaking this measure describes the volume of a distribution (Basu & Ho 2006), that is the absolute value of the determinant is the volume of a parallelepiped spanned by a set of real vectors. The larger the determinant, the more dispersed the data points. The application of the MCD algorithm has been rapidly increased over the last decades, particularly since Rousseeuw & van Driessen (1999) published a computationally fast algorithm. Smirnov & Egbert (2012) mentioned the MCD algorithm as a promising alternative to the affinely invariant covariance estimator.

The robust MD calculation for our data confinement criterion is mainly based on the deterministic MCD algorithm from Hubert *et al.* (2012) which is a slight modification of the original MCD algorithm from Rousseeuw & van Driessen (1999). It was originally implemented in Matlab and is part of LIBRA, the Matlab Library for Robust Analysis (Verboven & Hubert 2010). We programmed the main parts of this routine into our own software package EMERALD written in C++ and adopted the algorithm to better serve our needs by adding a seventh initial estimator.

3.1.1 *The modified deterministic MCD algorithm*

The MD algorithm implemented in our processing requires the data matrix **X** that contains n rows representing the observations (i.e. number of events for a given period) and p columns representing the variables (i.e. real and imaginary parts of the transfer functions within chosen bivariate equation). The deterministic MCD algorithm first applies a column-wise standardization (studentization) of the data matrix **X**:

$$\mathbf{Y} = \frac{\mathbf{X} - \mu}{\sigma} \tag{3}$$

with the coordinate-wise median μ and the robust scale estimator σ. Hubert *et al.* (2012) used two different robust scale estimators depending on the size of the data matrix: the Q_n scale estimator of Rousseeuw & Croux (1993) for $n < 1000$ or the τ-scale of Yohai & Zamar (1988) for $n \geq 1000$.

The next step of the MCD algorithm is the selection of various initial subsets H of size h, with $h = \left[\frac{n+p+1}{2}, n\right]$. The sensible selection of the initial subsets is important, because otherwise the final result of the MCD algorithm might not represent the global minimum. While the original MCD algorithm from Rousseeuw & van Driessen (1999) solved this problem by taking many (by default 500) randomly chosen subsets, the deterministic MCD algorithm from Hubert *et al.* (2012) only utilizes six well-chosen subsets determined by different estimators. We used the same six estimators for our modified deterministic MCD algorithm, which are listed in appendix A. Each of these six estimators calculates a preliminary estimate of the correlation matrix $\mathbf{S_k}$ of **Y** with $k = 1, ..., 6$. The following three steps are applied to each of the six correlation matrices $\mathbf{S}_k$ individually to calculate initial estimates $\hat{\mu}_k(\mathbf{Y})$ and $\hat{\mathbf{C}}_k(\mathbf{Y})$ for data centre and covariance matrix:

(i) Computation of the matrix **E** containing the eigenvectors of $\mathbf{S}_k$ and calculation of $\mathbf{B} = \mathbf{YE}$.

(ii) Estimation of the covariance of **Y** by $\hat{\mathbf{C}}_k(\mathbf{Y}) = \mathbf{ELE}^T$ with $\mathbf{L} = diag(\sigma^2(\mathbf{B}_1), \ldots, \sigma^2(\mathbf{B}_p))$.

(iii) Calculation of the data centre estimate $\hat{\mu}_k(\mathbf{Y}) = \hat{\mathbf{C}}_k^{1/2}(median(\mathbf{Y}\hat{\mathbf{C}}_k^{-1/2}))$.

Each of the six estimates $(\hat{\mu}_k(\mathbf{Y}), \hat{\mathbf{C}}_k(\mathbf{Y}))$ are then used to compute a distance MD_{ik} for each event i and to form the k initial subsets by using the h observations with the smallest distance.

Tests of the starting values for data centre and covariance estimates obtained by these estimators revealed that they are suitable for MT data; however, the initial data centre estimates are often close to the final estimated value and therefore exhibit only small variability at start. Since we might fail to find the global minimum of the MCD objective function, it is important to start from initial, significantly

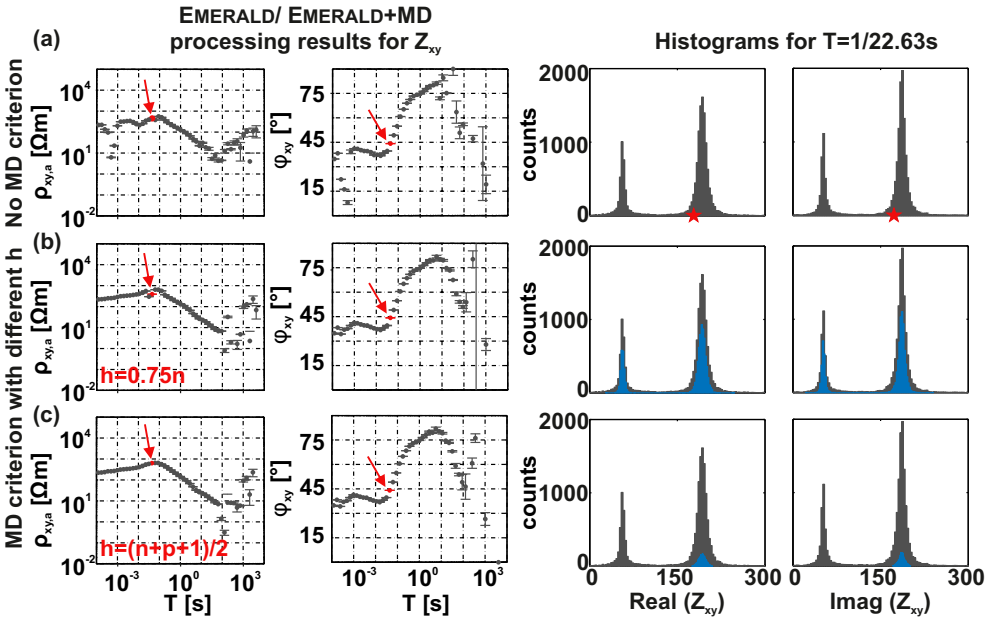

Figure 3. Processing results of station SA-1 for different processing parameters using an additional coherence threshold of 0.9. Columns from left- to right-hand side: (1) Apparent resistivity and (2) phase curves of Z_{xy} over period; histograms of the (3) real and (4) imaginary part of Z_{xy} for all events for $1/22.63\,s$. (a) To illustrate the effect of an inappropriate choice of the size of the subset, we focus on the apparent resistivity curve based on the standard EMERALD processing, which exhibits minor wobbles for periods around $1/22.63\,s$. Histograms reveal that the data distribution is separated into two clusters. Based on results calculated for the adjacent period of $1/16\,s$ (red asterisks) we can classify the left cluster as noise-related. (b) The EMERALD+MD processing with h is $0.75n$ which is the default value of the LIBRA routines cannot improve the processing result for the selected period , as the underlying assumption of the total noise content is violated. The single event distribution (blue) in the histograms show the remaining events used for subsequent robust stacking. (c) Using a size of the subset h of $\frac{n+p+1}{2}$, representing the smallest possible subset H leads to the desired smooth apparent resistivity curve. Please note that for better readability extreme outliers are not displayed.

Table 1. *α-quantiles* for a χ^2-distribution with four degrees of freedom from Morrison (1967) and the derived thresholds MD_{crit} (square root from x_α) for standard single-station processing with four variables representing real and imaginary parts of the two transfer functions components in each row of eq. (1).

α	0.500	0.750	0.900	0.950	0.9750	0.990	0.995
x_α	3.36	5.39	7.78	9.49	11.14	13.28	14.86
$MD_{crit} \triangleq \sqrt{x_\alpha}$	1.83	2.32	2.79	3.08	3.33	3.64	3.85

different subsets. Therefore, we included a seventh estimator. Our supplementary estimator uses the final data centre and covariance matrix estimates μ_{T-1} and $\mathbf{C}_{T-1}$ from a previously processed, adjacent period in eq. (2) to calculate a MD_i value for all events of the currently examined period and selects the h events with the smallest MD value to form the seventh initial subset H_7. The reasoning is based on the physics of induction processes by which MT transfer functions vary only smoothly with period, as the induction space of adjacent periods increases only marginally. This supplementary estimator is (i) significantly different from the other six (statistical motivated) estimators and therefore yields significantly different data centre and covariance estimates, (ii) warrants to be composed of a different subset of data and (iii) utilizes a physical relationship inherent to MT data. If the result of the adjacent period was biased, this seventh estimator does not automatically result in a wrong result for the currently processed period. The remaining six estimators are able to prevail against the seventh starting subset, if the majority of the data points of the current examined period do not confirm the previously derived solution. Therefore, the algorithm will get rid of false initial estimates during the iterative process. However, thorough tests confirmed that if undisturbed results of an adjacent period exist, our novel seventh estimator often represents the best solution of all seven initial subsets. If no previously processed, adjacent period is available (e.g. for the first period processed), only the six estimators of the original deterministic MCD algorithm will be used.

After the selection of the initial subsets, the kernel routine of the MCD algorithm is iteratively applied to each subset individually (concentration steps / C-step) to obtain an improved approximation to the MCD. Rousseeuw & van Driessen (1999) proved the convergence of the C-steps in a finite number of iterations.

Each C-step j is divided into several steps:

(i) Computation of the distances MD_i for $i = 1, \ldots, n$ using $\hat{\mu}_{j-1}$ and $\hat{\mathbf{C}}_{j-1}$ from the previous iteration step in eq. (2). For the first iteration ($j = 1$), $\hat{\mu}_0$ and $\hat{\mathbf{C}}_0$ are calculated from the initial subsets.

(ii) Sorting of the distances MD_i in ascending order and yielding a permutation π for which $MD(\pi(1)) \leq MD(\pi(2)) \leq \ldots \leq MD(\pi(n))$.

(iii) Formation of a new subset H with $H = \{\pi(1), \pi(2), \ldots, \pi(h)\}$.

(iv) Computation of new data centre and covariance estimates with $\hat{\mu}_j = \frac{1}{h}\sum_{i\in H} y_i$ and $\hat{\mathbf{C}}_j = \frac{1}{h-1}\sum_{i\in H}(y_i - \hat{\mu}_j)(y_i - \hat{\mu}_j)^T$.

The C-steps are repeated until convergence is reached. Finally, the subset with the smallest determinant of the covariance matrix is selected and the corresponding $\hat{\mu}_{raw}$ and $\hat{\mathbf{C}}_{raw}$ are called raw solution of the deterministic MCD algorithm calculated by all observations within this subset. The final $\hat{\mu}$ and $\hat{\mathbf{C}}_x$ estimates are, in

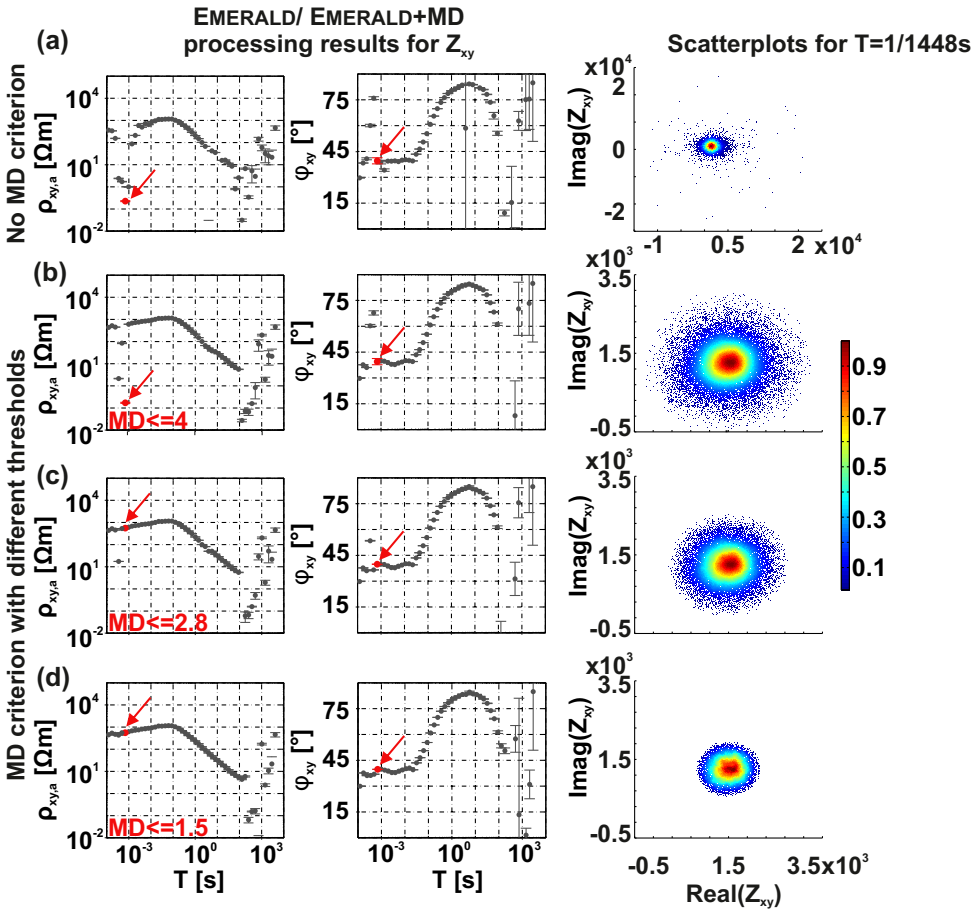

Figure 4. Processing results of station SA-2 with different thresholds for the MD criterion. Columns from left- to right-hand side: (1) Apparent resistivity and (2) phase curves for the Z_{xy} component; (3) Scatterplot for a period of $^1/_{1448}\,s$ (red arrow), each single event is colour-coded with its smoothed likelihood (red $\hat{=}$ high, blue $\hat{=}$ low likelihood). All processing results use an additional coherence threshold of 0.9. (a) Standard EMERALD processing reveals good results for a wide period range, however, short period data show scatter. The scatterplot indicates that a significant fraction of the single events scatter around the distribution. EMERALD+MD processing results with a threshold of (b) 4, (c) 2.8 and (d) 1.5 and their corresponding scatterplots exhibit lesser outliers with decreased threshold values.

contrast to the raw solution, then obtained as weighted mean and covariance matrix with the weights

$$w_i = \begin{cases} 1 & \text{if } MD_i \leq \sqrt{\chi^2_{p,0.975}} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

following the deterministic MCD approach from Hubert *et al.* (2012) and the original fast MCD algorithm from Rousseeuw & van Driessen (1999). We compared processing results obtained by using the raw- and reweighted solutions for many stations with different noise content. In general, the processing results using the reweighted MCD solution are slightly superior, and result in smoother apparent resistivity and phase curves. However, in many cases the difference is negligible.

3.2 Important parameters for the MD criterion

The MCD algorithm requires an initial choice of the key parameters h, which is the size of the subset H, with $\frac{n+p+1}{2} \leq h \leq n$, and the maximum number of allowed C-steps; furthermore, we have to define a threshold or critical distance of a single event to the centre. The first two parameters are hard-wired programmed to ensure a high breakdown point and a fast computation. A sensible threshold depends on the EM noise characteristics in the MT data and therefore has to be chosen by the user. For a more detailed description of the deterministic MCD algorithm, we refer to Hubert *et al.* (2012).

3.2.1 Size of the subset H

The smallest possible size of a subset $h = \frac{n+p+1}{2}$ describes the case where the algorithm has its highest possible breakdown value. The breakdown point is defined as the limiting fraction of outliers a robust algorithm can handle. It normally cannot exceed 50 per cent; meaning that the majority of the data has to be well-behaved (Huber 1981; Hampel 1986). For $h = n$, the results of the MCD algorithm correspond to the normal arithmetic mean and the sample covariance matrix. The default value of h in the routines DetMCD (Hubert *et al.* 2012) and the FASTMCD (Rousseeuw & van Driessen 1999) of the Matlab library LIBRA (Verboven & Hubert 2010) is set to $h = 0.75n$ as this value represents a compromise between retaining a high breakdown point and having a good statistical efficiency (Rousseeuw & van Driessen 1999; Hubert *et al.* 2008, 2012; Hubert & Debruyne 2010; Verboven & Hubert 2010). However, with $h = 0.75n$ we assume that the data are contaminated by less than 25 per cent noise. We tested three different h values for a variety of MT stations with $h = (\frac{n+p+1}{2},\ 0.6n,\ 0.75n)$. For stations with

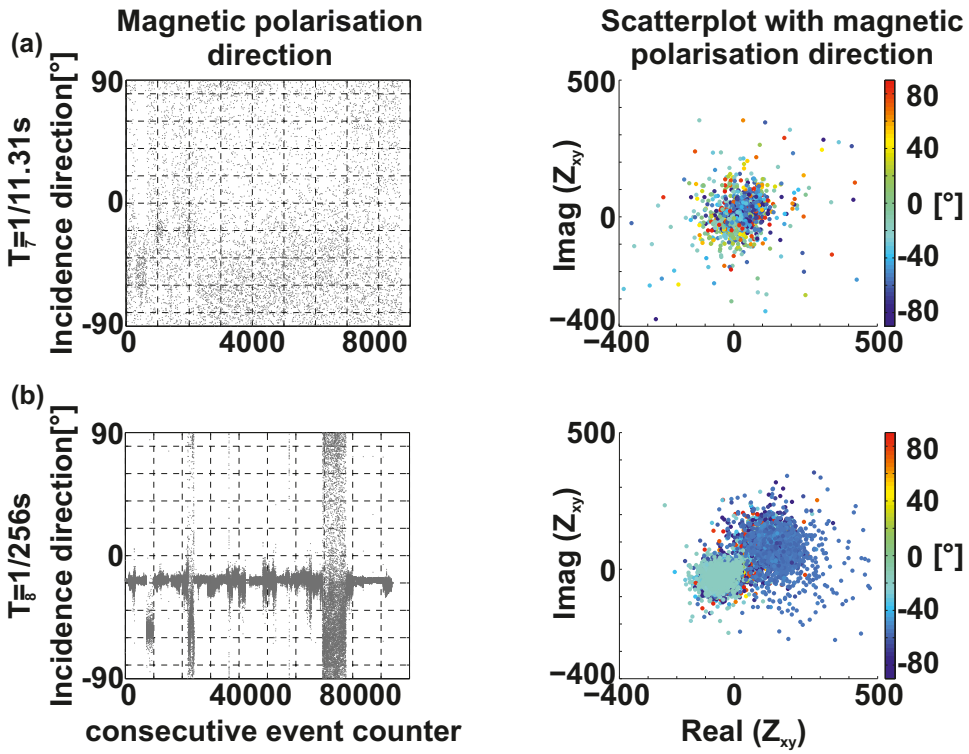

Figure 5. Distribution of polarization direction of the magnetic field (see Weckmann *et al.* 2005) and scatterplots for station V-2 (*cf.* Fig. 8e) for periods of (a) $T_7 = 1/11.31\,s$ and (b) $T_8 = 1/256\,s$. (a) Left-hand panel: this period is not affected by coherent noise; the polarization directions of the magnetic field for single events are equally distributed between −90 and 90°. Right-hand panel: each single event is colour-coded with its polarization angle for the magnetic field showing a rather random scatter. (b) Left-hand panel: graph of the polarization directions of the magnetic field exhibit two dominant polarization directions around −20° and −50° to −60°; they are interrupted by two shorter time segments with a broad range of incidence directions. Right-hand panel: colour-coded single events indicate that two noise sources from these different directions are active.

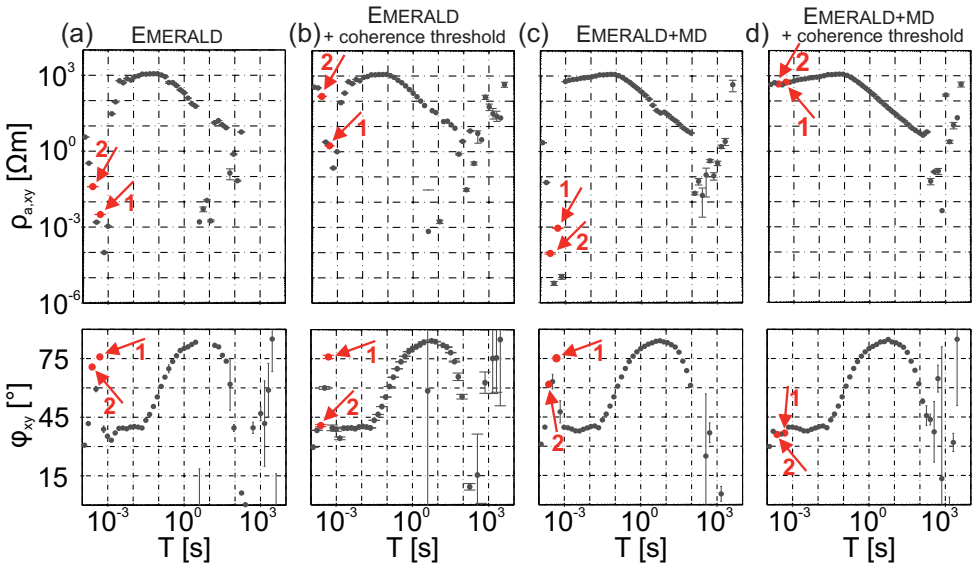

Figure 6. Apparent resistivity and phase curves of Z_{xy} for station SA-2. (a) Processing result from the robust EMERALD processing (a) without any additional data selection criteria. (b) using an additional coherence threshold of 0.9, (c) using the MD approach and (d) with a combination of a coherence threshold and the MD approach. Although we observe improved apparent resistivity and phase values for several periods, reasonable transfer functions for very short periods $< 1/1000\,s$ can only be obtained by a combination of coherence sorting and MD criterion.

medium to acceptable data quality, the differences are usually small, but for those with a large amount of EM noise we observe significant differences. The best results in these cases were observed for the lower limit of h due to the higher robustness of the data centre and covariance matrix estimates against outliers and additional distributions. Therefore, we fixed h to $\frac{n+p+1}{2}$ to ensure a high breakdown point. To illustrate the effect of different h-values we have chosen a station with acceptable apparent resistivity and phase values for

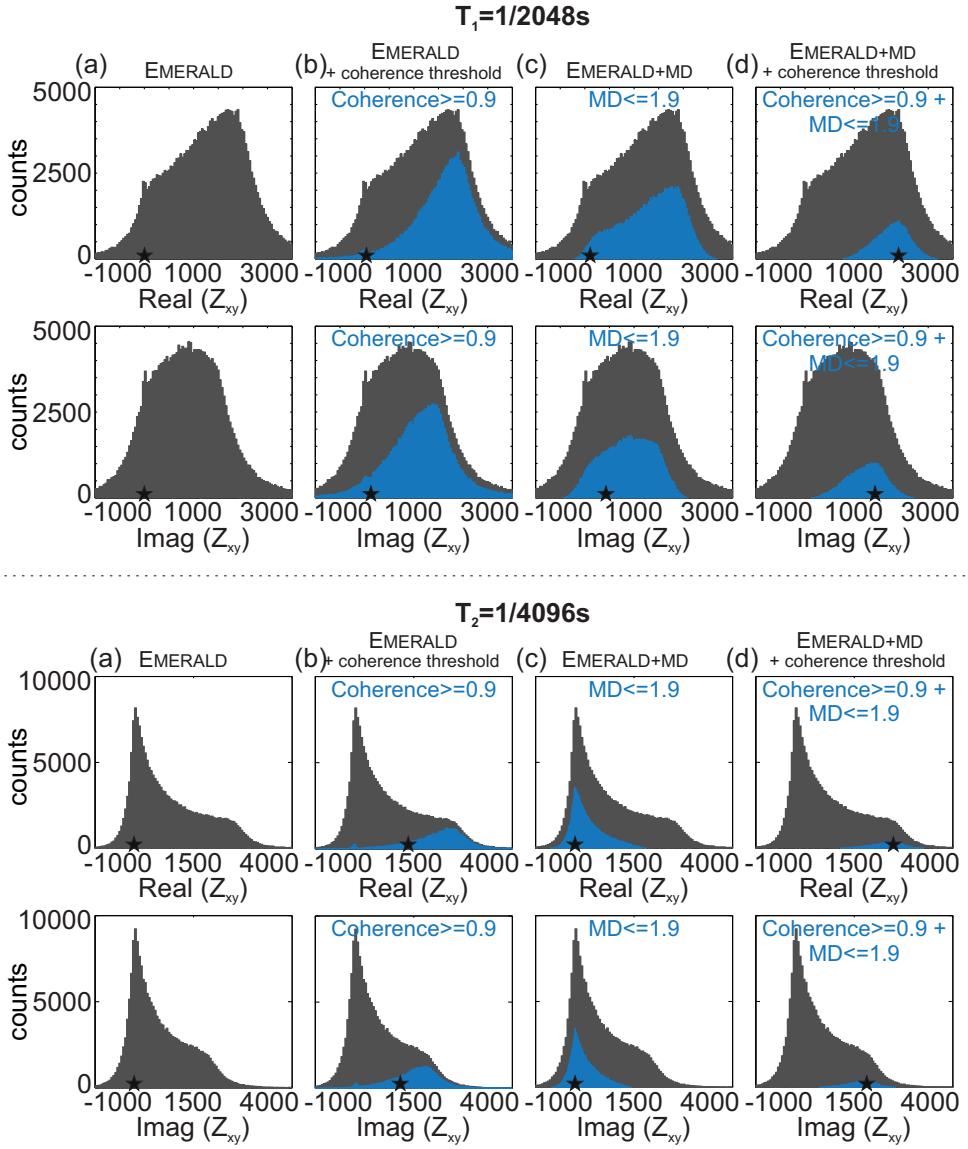

Figure 7. Histograms of the real and imaginary parts of Z_{xy} (station SA-2) obtained by different processing approaches with and without coherence sorting for $T_1 = {}^1/_{2048}\,s$ (upper panel) and $T_2 = {}^1/_{4096}\,s$ (lower panel). The black asterisks represent the final processing result. Due to the extremely long tails (not shown here, as the histograms focus on the main part of the distribution) the black asterisks are located far away from the data centre. (a) The grey-coloured distributions represent single events that are used to calculate the final transfer function using the robust EMERALD processing. Both values of real and imaginary parts lie close to −1000. (b) The grey-coloured distributions of (a) are overlaid with the distribution that in addition fulfil a coherence threshold of 0.9 (blue colour). (c) EMERALD+MD processing selects a subset of data with MD values ≤ 1.9 (blue) which resembles the shape of the original distribution and thus leads to values for real and imaginary parts that are comparable to solution (a). (d) A combination of Emerald+MD and a coherence threshold of 0.9 yield values of real and imaginary parts that result in a smooth transfer function (see Fig. 6d). The combination of both criteria removes about 80 per cent and 90 per cent of all events for T_1 and T_2, respectively. However, for both periods we still have more than 45,000 and 15,000 events for the subsequent stacking process which exhibit a high coherence value as well as a small distance to the assumed desired MT distribution due to pre-selection.

higher frequencies (Fig. 3). The highlighted period of $T = {}^1/_{22.63}\,s$ obtained through standard EMERALD processing (Fig. 3a) with an additional coherence threshold of 0.9 shows only small deviations from an ideally smooth apparent resistivity curve and almost no sign of noise contamination in the phases. Obviously, our processing together with a coherence sorting reveals two different clusters of transfer functions which differ in size. Based on the robustly estimated results for the adjacent period of $T = {}^1/_{16}\,s$ (red asterisk in Fig. 3a) we conclude that the cluster to the right belongs to natural MT excitation. Luckily the largest cluster belongs to the desired signal so that the applied robust statistics is almost able to deal with it. However, the MD criterion prior to the EMERALD robust statistics, modifies the data set in such way that the amount of events of both clusters is almost equal (blue histograms in Fig. 3b). Due to the underlying, obviously wrong assumption that only 25 per cent of the data set contains noise, the effect is counterproductive, as now the robust stacking finds disadvantageous conditions and thus results in worse estimates of apparent resistivities. By choosing a much higher breakdown value (Fig. 3c), almost no events are taken from the noise cluster and finally even the small wobbles in the apparent resistivities are gone.

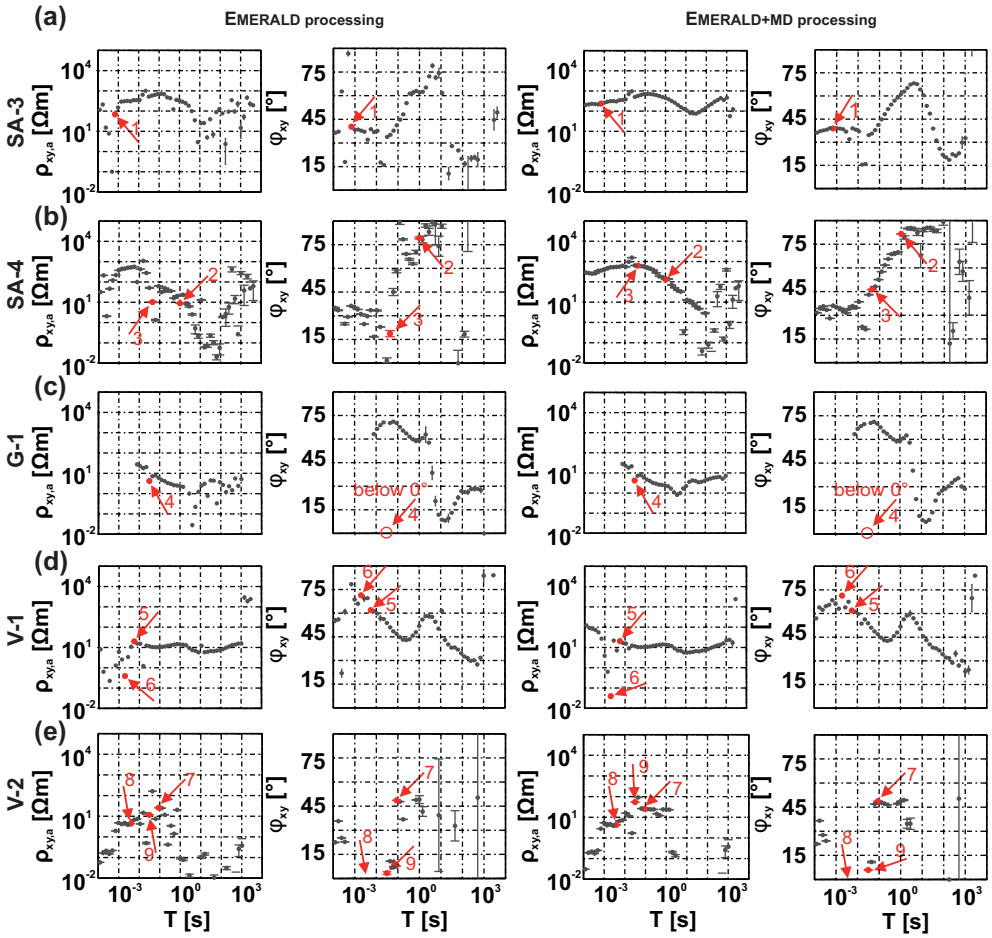

Figure 8. Apparent resistivities and phases of one impedance tensor component of five different MT stations after using standard EMERALD (two columns to the left) and EMERALD+MD (two columns to the right) processing: Please note that we did not apply any additional notch or delay line filter to remove, for example the $^1/_{50}\,s$ signal from the power grid. However, a coherence threshold of 0.9 was applied to all processing runs. In almost all cases we can achieve a significant improvement for apparent resistivities and phases. The particular data distribution and characteristics of each station and the highlighted periods $T_1 - T_9$ will be discussed in the subsequent figures. Often, the EMERALD+MD fails to improve the data significantly for, for example periods in the so-called dead band (e.g. b and e) or at long periods due to the small amount of available events.

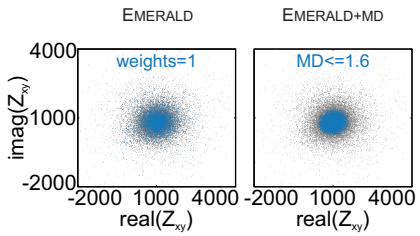

Figure 9. Scatterplots of station SA-3 for $T_1 = {}^1/_{1448}\,s$ comparing accepted events for standard EMERALD and EMERALD+MD processing (Fig. 8a). The distribution of all events (grey dots) in the complex plane is overlain by those events with the highest possible weight ($w = 1$) in the robust stacking process of the standard EMERALD processing (blue) and those events below a certain MD value (EMERALD+MD), respectively. Application of the MD criterion leads to a more focused subset of data and rejection of scattered data points. Please note that for the final stacking result within Emerald we use weights between 0 and 1 and thus much more events can be integrated into the final result with weights slightly lower than 1.

3.2.2 Maximum number of C-steps

Preferably, C-steps are applied until convergence is reached. In the interest of computational time and in particular for large data sets, Rousseeuw & van Driessen (1999) proposed to use a limited number of C-steps. Hubert *et al.* (2012) fixed this maximum number of C-steps in their algorithm to 100. Tests with different MT stations and data qualities suggest that the MCD algorithm always needs fewer C-steps so that this limit seems to be more than sufficient. A moderate increase of applied C-steps is observed for short periods in comparison to long periods, which is caused by the larger number of events.

3.2.3 Critical distance

Noise affected data in form of outliers in the tail of a distribution or additional clusters from a separate noise distribution in the data space usually have a certain, but larger distance to the data centre. A critical distance (threshold) MD_{crit} for the largest allowed MD_i value is often defined by a certain quantile of the χ^2-distribution, because for multivariate normal distributed data the squared MDs are approximately central χ^2-distributed with p degrees of freedom. A value x_α is called α-quantile of a probability distribution P, when it divides the distribution into two intervals so that

$$P\left((-\infty, x_\alpha]\right) \geq \alpha \;\; and \;\; P\left([x_\alpha, +\infty)\right) \geq 1 - \alpha \qquad (5)$$

holds with $\alpha \in (0, 1)$. Typical values for the α-quantiles and the derived critical distances, that is the square root of x_α, are given in Table 1 for a χ^2-distribution with four degrees of freedom.

If we assume that our data are noise-free, that is with no outliers or second distributions, the squared MD_i values with a probability α lie in the interval $[0, x_\alpha]$. This means that only events with $MD_i \leq MD_{crit}$ are accepted and used in the subsequent robust stacking algorithm. As the EMERALD processing assumes an underlying Gaussian model, we use quantiles of the χ^2 distribution to determine a suitable range of thresholds. Typically, we use thresholds $MD_{crit} < 4$ following the values in Table 1.

Apparent resistivity and phase curves for periods $< 10^{-3}\,s$ obtained through standard EMERALD processing (Fig. 4a) show scattered data points. Scatterplots of the affected period $^1/_{1448}\,s$ exhibit several large outliers around the desired distribution which hamper the estimation of the transfer functions. In Figs 4(b)–(d) the MD criterion prior to standard EMERALD processing is applied with decreasing thresholds. Outliers are successively removed by the MD criterion. Transfer functions calculated from these confined clusters result in data improvement. The smaller the MD threshold, the more events are rejected. However, the most effective threshold differs for each data set and depends on data quality. If the chosen threshold is too small, the transfer function becomes unstable due to an insufficient amount of accepted events. On the other side, if the threshold is chosen too large, the MD criterion is not able to remove an adequate amount of outliers and therefore cannot decrease the effect of EM noise on the obtained transfer functions. Here the user has to find a good compromise between these two end members.

4 THE MAGNETIC POLARIZATION DIRECTION (MPD)

Station V-2 (*cf.* Fig. 8e) is a typical example for which the MD criterion fails - at least if applied in an automated way. Many periods are heavily affected by coherent noise sources so that only a minor part of the data originates from the natural MT signal. To improve the processing results interactive selection algorithms can be used, which provide additional information on e.g power spectra, polarization directions of electric and magnetic field or coherence values. A promising property to distinguish between near-field noise and MT excitation is the magnetic field polarization direction (Fowler *et al.* 1967) which was already discussed in Weckmann *et al.* (2005). Given the different sources of the natural magnetic field (e.g. lightning and ionospheric current systems) we expect to see the full range of incidence directions ($[0^\circ - 360^\circ]$; *cf.* Fig. 5a, left-hand panel). This assumption is supported by the colour-coded scatterplots of real and imaginary parts of all events (Fig. 5a, right-hand panel). However, if the magnetic field is generated by a nearby noise source we observe one or several distinct polarization directions (*cf.* Fig. 5b, left-hand panel). However, there is a small part around event number 8000 which shows the desired evenly distributed polarization directions for the magnetic field. Colour-coded single events indicate that two noise sources from two different directions are active (*cf.* Fig. 5b, right-hand panel). Therefore, we implemented a physically motivated data selection criterion based on the polarization (incidence) direction of the magnetic field (MPD) as an add-on to the statistically based MD criterion. In contrast to Weckmann *et al.* (2005), the MPD criterion presented in this paper works in a fully automated manner and does not require any user input. Although we mainly applied the MPD criterion together with the MD approach, it can be used separately.

4.1 Implementation of the MPD criterion

The MPD criterion is based on the detection of confined magnetic polarization directions, which means that within a certain time span an exceptionally large number of events have the same incidence direction. For this purpose, the incidence direction (or magnetic polarization direction angle) α_B is calculated for all events of the considered target period using the smoothed single event cross- and autospectra:

$$\alpha_{B,i} = arctan \frac{2 * Real([B_x B_y^*]_i)}{[B_x B_x^*]_i - [B_y B_y^*]_i} \tag{6}$$

Depending on the definition of the arctan routine, obtained values of $\alpha_{B,i}$ reside in the codomain of $[-90^\circ, 90^\circ]$ or $[0^\circ, 180^\circ]$. To identify events with incidence angles accumulating at certain angles, all incidence directions α_B of the current period are organized into a histogram of 180 bins with a bin width of 1°. To decide which amount of events per class is considered as exceptionally large, the actual number of events in a class is compared with a uniform distribution. Thereby the expected value $E_k = \frac{Number\ of\ events}{180}$ is the same for each bin. A suitable threshold within the MPD criterion was found empirically after testing many stations with different polarization patterns and the total number of events is essential (cf. Table B1 in the appendix). The chosen limits are selected in a conservative manner to assure that only events corresponding to a distinct polarization direction are removed. For stations that do not show any preferred polarization direction for a particular period, these conservative thresholds ensure that all events are accepted. As a consequence of these conservative limits, the automatic MPD criterion sometimes does not remove all events which would be identified by visual inspection and removed subsequently, in particular for stations which suffer from strong and complicated noise sources and thus show complex polarization pattern.

5 APPLICATION OF THE MD AND THE MPD CRITERIA TO DIFFERENT DATA SETS

We tested the MD criterion on several MT data sets collected worldwide which include a broad range of different noise contaminations. During the respective field experiments, we were also able to obtain a good overview on the specific noise sources active in the area. In this section, we show advances of the MD data confinement criterion and at the same time assess and discuss limitations of this approach. Furthermore, we apply the MPD criterion as a physically based constraint. In combination, our suggested approach is able to deal with noise contents far more than 50 per cent as long as the noise exhibits distinct magnetic polarization directions.

5.1 Combination of the MD criterion with the coherence criterion

The MD criterion as a purely statistical criterion cannot distinguish between physically reasonable MT data and near-field EM noise contributions. Instead, it will always focus on the cluster/distribution which consists of the majority of events. Therefore it is often advisable to use coherence sorting as a physically based data selection criterion prior to the application of any MT processing algorithm and in particular the MD criterion to remove data points with a small coherence. Thereby, the user makes sure that the amount of events caused by incoherent or noisy data is reduced so that the robust processing can work effectively. It often improves the transfer

function estimation significantly (see e.g. comparison in Figs 6a and b), however, subsequent standard robust processing will fail since near-field EM noise is also characterized by high coherence values. Therefore, the use of the coherence sorting has to be carefully examined for each station separately and it should not be applied systematically without any verification. The coherence and the MD thresholds presented in this paper were carefully selected for each station by testing many different values and evaluating the result by visual inspection of the transfer functions. The grey-coloured graphs in Fig. 7 represent the distributions of the real and imaginary parts of all single events corresponding to the highlighted periods $T_1 = {}^1/_{2048}\,s$ and $T_2 = {}^1/_{4096}\,s$ in Fig. 6. Although no second cluster originating from, for example coherent noise sources is visible in the histograms, the depicted distributions cover a wide range. The blue-coloured graphs in Fig. 7(b) mark events, which have a bivariate quadratic coherence value greater than 0.9. Especially for T_2, the majority of the original events is removed by the coherence criterion.

The MD criterion can be applied independent of the coherence criterion (see Fig. 6c). It improves the processing results for periods where the majority of all events are caused by the natural signal. But it fails if the majority of all events is caused by noise, for example events with a low coherence value (Fig. 7c). The MD criterion as a purely statistical approach will concentrate on the majority of data without recognizing that these points do not represent physically meaningful MT data. In the worst case, the data points originating from natural MT signal can be almost completely removed (Fig. 7c for T_2). The combination of both criteria (Fig. 6d) leads to the best results in this case. It can be recognized that removing some of these points that do not behave according to our linear relationship of electric and magnetic fields (i.e. data with low bivariate coherence values) in a first step, can be essential to achieve good results. Histogram plots (Figs 7c and d) substantiate this finding, as most single events in Fig. 7(c) with small coherences yield poor results in Fig. 6(c). Single events (Fig. 7d) with at least a coherence value of 0.9 lead to removal of the majority of low quality data. This example demonstrates impressively that the MD criterion as a purely statistical method will always focus on distributions with the majority of the data regardless of which points represent physically reasonable MT data. Therefore we recommend to use the MD criterion in combination with physically based data selection criteria.

5.2 Application of the MD criterion to data sets with different noise contaminations

The MT stations discussed here suffer from different noise contaminations and are suitable to illustrate advances and limitations of the new criterion as well as to demonstrate under which circumstances improved results can be obtained for selected periods. We compare apparent resistivities and phases obtained by the standard EMERALD processing (left-hand columns in Fig. 8) with apparent resistivities and phases achieved by the EMERALD+MD processing (right-hand columns). For all these examples, we applied a coherence threshold of 0.9.

These examples illustrate that the application of the MD criterion can in general improve MT impedances independent of the period range; however, for longer periods it is more difficult because of the small number of available events.

A detailed analysis will be helpful to understand under which circumstances and for which type of distorted data the MD criterion will work well.

5.2.1 *One distribution plus scattered noise*

Although station SA-3 (South Africa) has an overall good data quality, standard EMERALD processing results are only acceptable for some periods in the ranges between $10^{-3} - 10^{-2}\,s$ and ${}^1/_{32} - 1\,s$ (left-hand columns in Fig. 8a). The application of the MD criterion significantly improves the processing results and yields much smoother apparent resistivity and phase curves (right-hand columns in Fig. 8a). For the highlighted period ($T_1 = {}^1/_{1448}\,s$), scatterplots of both processing approaches indicate that the events are arranged in one cluster/distribution (Fig. 9). However, a minor fraction of data scatters around this cluster. Some of these scattered events have the highest possible weight of 1 in the standard EMERALD processing and therefore have a large influence on the final processing result which hampers the transfer function estimation. The application of the MD criterion prior to the stacking procedure removes these events as they have a larger distance to the actual data centre. This results in more focused clusters and consequently in smoother transfer functions.

This example demonstrates that the MD criterion is capable of improving processing results of stations affected by noise which is expressed as scatter around the true MT distribution by confining the data to a focused subset.

5.2.2 *Two data clusters*

Histograms of the real and imaginary part as well as scatterplots of station SA-4 (South Africa) for a) $T_2 = 1\,s$ and b) $T_3 = {}^1/_{22.63}\,s$ are shown in Fig. 10.

Obviously, the data of SA-4 are distributed into two clusters for $T_2 = 1\,s$ (upper panel in Fig. 10a). To identify the cluster originating from the natural MT signal, we use the processing results of an undisturbed adjacent period, which constitutes one of the initial subsets during the robust calculation of the MD. The real and imaginary parts of this estimate are indicated by the red asterisks. A scatterplot of imaginary versus real parts shows much clearer that we deal with two different distributions (upper panel in Fig. 10a). In this case the smaller cluster (consisting of less events) to the right is caused by coherent noise sources and the majority of the events originates from the natural MT signal. Again the red ellipse indicates the distribution of the undisturbed adjacent period and allows to differentiate noise and data cluster. Since the majority of events belongs to the desired MT signal, the essential criterion for a successful application of the MD approach is met and all events of the noise distribution are removed. In contrast, the standard EMERALD processing fails for this period as events from both clusters achieve the highest possible weight. For the period $T_3 = {}^1/_{22.63}\,s$ we observe only one larger although elongated cluster (Fig. 10b). Close inspection suggests that there a second distribution is merged with the main distribution (tail towards higher values of the real part of Z_{xy}). Again, from the adjacent period we know where the cluster of the "true" MT transfer function is located (red asterisk and ellipse in Fig. 10). Application of the MD criterion removes events with a large distance to the majority of all points and therefore uses only points which are much more focused towards the left, that is smaller values of the real part of Z_{xy}. The subsequent EMERALD processing is able to remove outliers with regard to the confined subset of data. In comparison, the standard EMERALD processing (rightmost column) of the entire data set is not able to completely down-weight events, which originate from this tail.

These examples underline that the MD criterion can (completely) remove a separate noise distribution in the data space in two cases:

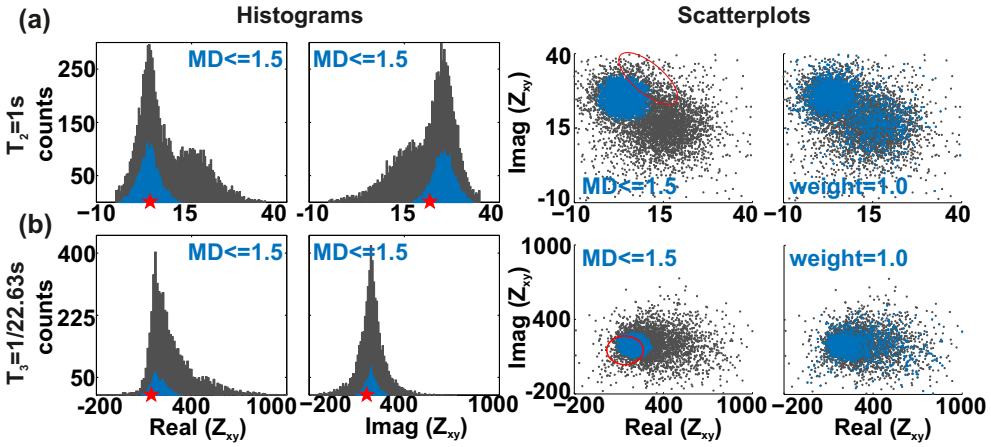

Figure 10. Histograms and scatterplots for the Z_{xy} component (station SA-4 in Fig. 8b) for a period of (a) $T_2 = 1\,s$ and (b) $T_3 = {}^1/_{22.63}\,s$. Grey bars or dots indicate the data distribution which was used for the standard EMERALD processing, blue colour highlights those events after application of our confinement approach. The red asterisks and ellipses (covariance matrix) indicate the result of an undisturbed adjacent period. (a) For a period of $T_2 = 1\,s$ the real and imaginary part of Z_{xy} show two overlapping distributions. The MD criterion is able to focus on the blue data subset. The third column depicts the same distribution in an Argand plot with two neighbouring clusters of Z_{xy}. The covariance of the undistorted, adjacent period characterizes the cluster to the right as EM noise. The fourth column shows those events which obtain the highest weight within the standard robust algorithm of EMERALD. As events from both clusters are used for computation of the final stacked transfer function, the robust processing still uses events from the cluster associated with EM noise and yields distorted transfer functions. (b) Analogue figures to (a) now for a period of $T_3 = {}^1/_{22.63}\,s$. Here, two distributions have a significant overlap so that they almost merge into one cluster (see e.g. tail to the right in the histogram of the real part). The MD criterion focuses on the left part of the distribution, whereas the standard robust processing uses events from the entire cluster including distorted ones.

(i) The noise distribution consists of the minority of all events and is spatially separated from the desired distribution of natural signal (Fig. 10a); (ii) The noise distribution is merged with the desired MT distribution, but consists of significantly less events (Fig. 10b).

5.2.3 Application to remote reference processing and inter-station transfer functions

Several tests with different stations suggest that the robust single-site EMERALD+MD processing can lead to similarly good or even better results than using the robust remote reference processing as long as the noise content represents the minority of all data. This is a remarkable result as often a reference station either does not have sufficiently clean data or has experienced some kind of technical problems so that we finally cannot rely on the robust remote reference processing to obtain acceptable data quality. However, the existence of a true remote reference station is still preferable and essential in many cases to improve processing results as shown in the following example. Although the MD criterion was designed for single-site processing, especially for cases without an adequate remote station, it can also be applied as an add-on prior to the robust remote reference processing. The comparison (Fig. 11) using an exemplary MT station from Tajikistan shows results from robust single-site and robust remote reference processing (separation 33 km) with and without the application of the MD criterion. The period range around 10 s exhibits the influence of EM noise in the dead band (see red ellipses in Fig. 11a) that can be slightly improved by the remote reference processing (Fig. 11c); in contrast, a more pronounced improvement is observed for the EMERALD+MD processing (Fig. 11b). The outliers for $T = 8\,s$ and $T = 16\,s$ for the remote reference processing remain independent of the chosen coherence threshold. The best result for this period range is obtained by using the MD criterion as add-on prior to the remote reference processing. Furthermore, additional scatter in the data is discernible for periods $> {}^1/_{500}s$. The period $T_1 = {}^1/_{1024}\,s$ was chosen to illustrate under which circumstances EMERALD+MD single-site processing (Fig. 11b) fails to improve the transfer function in contrast to EMERALD remote reference processing (Fig. 11c). This happens (Figs 11b and 12a) if the amount of noise is relatively large or two distributions are merged to a large extent. Here, a true remote station is essential to distinguish between near-field noise and desired MT signal (Figs 11c and 12b). The MD criterion results in a subset of MT data which represents both data clusters (blue distribution in Fig. 12a) and as a consequence is not able to obtain undisturbed transfer functions. In contrast, the remote reference processing can distinguish between desired MT signal and near-field noise as indicated by the unimodal distributions of all events (Fig. 12b). The period of $T_2 = 8\,s$ represents an example for which the EMERALD+MD single-site processing (Fig. 11b) is superior to the simple remote reference processing (Fig. 11c), although the best result for this period (as for the entire period range) is obtained by using a combination of remote reference processing and the MD criterion (Fig. 11d). For both processing approaches, single-site and remote reference processing, the events scatter broadly (upper scatterplots in Figs 12c and d). Due to the small number of available events, the robust weighting scheme fails for these scattered distributions. In contrast, the MD criterion focus only of a subset of events in the centre of the distributions (~26 per cent of all single-site events using a threshold of 1.4 and ~19 per cent of all remote reference events using a threshold of 5.3), which finally lead to an improved transfer function. For remote reference processing, a higher MD threshold is sufficient, indicating that the data quality is already improved by the use of a remote station in contrast to the single-site processing. In general, this example illustrates that remote reference is an essential tool to improve the transfer function estimation, but in some cases it can still be further improved by using it in combination with the MD criterion.

Although we mainly focus on the estimation of MT impedances in this paper, the MD criterion can be used in the same way for the

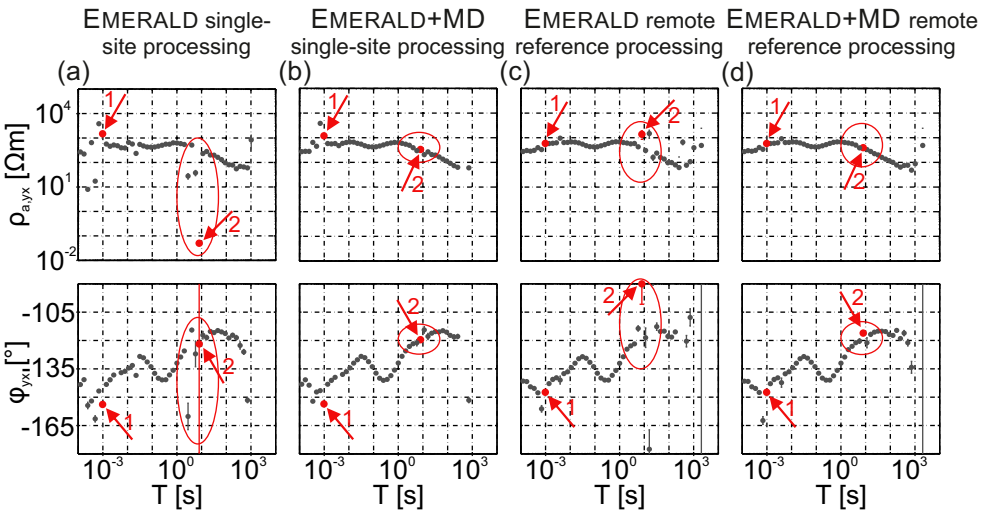

Figure 11. Apparent resistivity and phase curves of the Z_{yx} component of station T-420 for (a and b) single-site and (c and d) remote reference processing comparing standard EMERALD and EMERALD+MD processing. Single-site as well as remote reference data are processed using a coherence threshold of 0.9. (a) Single-site processing exhibits problems around 10 s and for very short periods <0.001 s. (b) EMERALD+MD processing yields significant improvement around the period of 10 s. (c) Robust remote reference processing is (as expected) superior to single-site processing, but cannot recover smooth curves around the period of 10 s. (d) Using the EMERALD+MD approach prior to robust remote reference processing results in less disturbed data points.

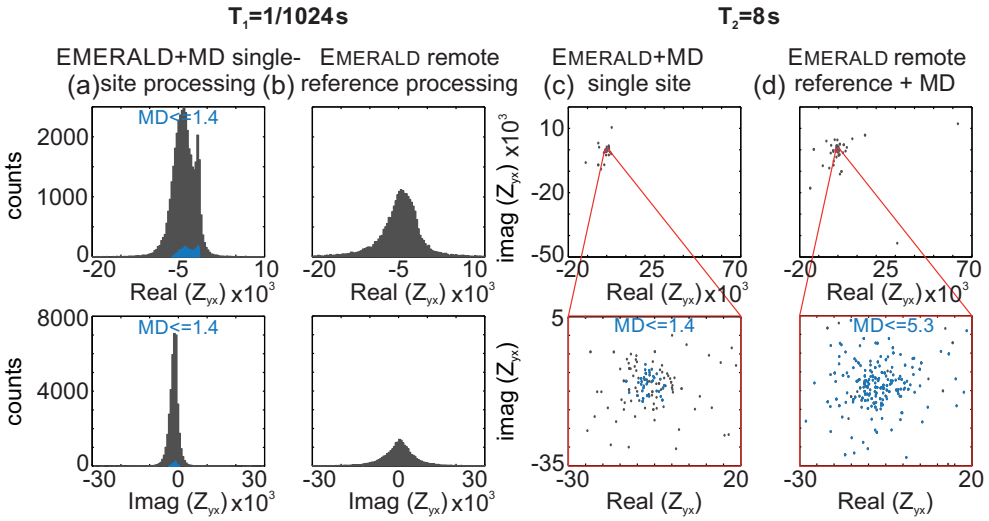

Figure 12. Histograms of the real and imaginary parts of Z_{yx} for the remote reference example in Fig. 11 for $T_1 = {}^1/_{1024}\,s$ (a + b) and scatterplots for $T_2 = 8\,s$ (c + d). (a) The grey-coloured distribution of the real and imaginary parts of all events used in the single-site processing indicate two distributions that are merged to a large extent. In this case, the application of the MD criterion fails as events of both distributions are selected (blue-coloured subset). (b) In contrast, the remote reference processing can distinguish between desired MT signal and near-field noise as indicated by the unimodal distributions of all events. (c and d) The scatterplots show that all events (blue and grey colours) for single-site (see Fig. 11a) as well as for remote reference processing (see Fig. 11c) scatter broadly. Therefore both processing approaches fail. In contrast, the subsequent application of the MD criterion selects only events in the centre of these distributions (blue colours in lower panel of c and d) for the robust stacking and results in improved estimates.

estimation of other transfer functions such as inter-station (Fig. 13) or the vertical magnetic transfer functions.

5.2.4 Limitations in presence of two data clusters

Station G-1 from Germany exhibits only small improvements after the application of the MD criterion (Fig. 8c). A scatterplot for $T_4 = {}^1/_{32}\,s$ reveals the existence of two spatially separated clusters (Fig. 14). As we often observe that noise only affects a certain, relatively narrow period range, we select an adjacent period which shows reasonable processing results (red asterisks for the undisturbed adjacent period of $T = {}^1/_{22.63}\,s$). Under the assumption that the processing results of this period are correct, the smaller upper cluster can be identified as desired MT signal. Within the algorithm the result of the adjacent period is used as the seventh MCD estimator. In this example, the EM noise cluster consists of the majority of data, therefore the MD criterion as well as the original robust stacking algorithm fail, although the seventh estimator hints at the correct solution. The entire desired MT signal is removed by the MD criterion and only events from noise sources are accepted. This example represents the worst case: Although we have nicely separated clusters, the majority of all events emanate from EM noise and consequently an automated statistical confinement criterion focuses on the majority of data. In this case, interactive selection

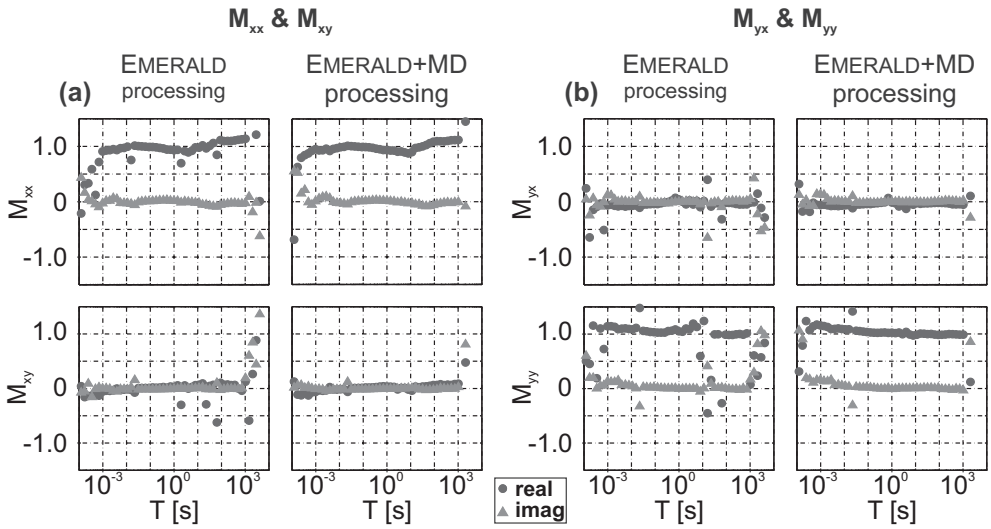

Figure 13. Inter-station transfer functions of station SA-6 with SA-7 comparing standard EMERALD and EMERALD+MD processing for (a) M_{xx} and M_{xy} and (b) M_{yx} and M_{yy} shown as real (dots) and imaginary (triangles) parts. The results of the MD criterion are superior and lead to smoother transfer functions for most periods.

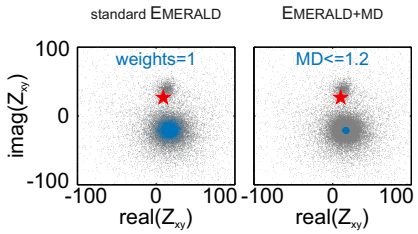

Figure 14. The scatterplots of station G-1 for a period of $T_4 = {}^1/_{32}\,s$ show two spatially separated distributions. Obviously, EMERALD as well as theEMERALD+MD will fail if the majority of all events is caused by noise. The red asterisks represent the processing results of the undisturbed adjacent period of $T = {}^1/_{22.63}\,s$ indicating the distributions of the desired MT signal, which unfortunately consist of the minority of all events (smaller cluster). As the majority of all events (larger clusters) is caused by EM noise, the MD criterion consequently fails although 74 per cent of the entire events are removed. This chunk however also included the desired MT signal.

algorithms (e.g. from Weckmann *et al.* 2005) or information of an undisturbed adjacent period have to be used to manually remove events corresponding to the EM noise cluster.

Similarly challenging is the existence of a second cluster (originating from noise) with a moderate number of events that overlays the true distribution to a large extent. In this case, the ellipsoids derived from the MD criterion are distorted so that the noise distribution remains unnoticed and data points originating from the noise cluster do not necessarily have a high MD value with regards to our desired MT data centre. In Fig. 15(a), histograms of real and imaginary part as well as the corresponding scatterplot are shown for $T_5 = {}^1/_{181}\,s$ of the exemplary station V1 (Venezuela). This period is affected only by a small amount of EM noise that forms a smaller second cluster visible as a long tail in the histograms - similar to station SA-4 (Fig. 10b). Events of this noise cluster can be completely removed by the MD criterion analogue to the example from station SA-4. However, a nearby period of $T_6 = {}^1/_{512}\,s$ reveals a different distribution, where the EM noise cluster dominates and thus the application of EMERALD and EMERALD +MD fails.

5.3 Application example of the MPD criterion to remove distinct polarization bands

Because the amount of noise within the data of station SA-5 (Fig. 16a) exceeds 50 per cent, the standard EMERALD processing results in poor MT curves especially in the period range between $0.1\,s$ and $1\,s$. The application of the MD criterion (Fig. 16b) was able to improve the short periods, but fails in the period range between $0.4\,s$ and $1\,s$, which seems to suffer from noise with preferred magnetic polarization directions (see exemplary periods in Fig. 17). For the period of $T_{10} = 0.5\,s$, a polarization band between -72° and -64° (left-hand column in Fig. 17a) can be observed. In contrast, the second exemplary period of $T_{11} = 1.4\,s$ exhibits two separate polarization bands between -72° and -64° and between -84° and -76° (left-hand column in 17b). In both cases, the bands are not continuous over the entire recording time, but are interrupted by some undistorted time spans. The additional use of the MPD criterion prior to the MD criterion removes most of the events in the disturbed segments (right-hand columns in Fig. 17) and further improves the result (Fig. 16c) to finally obtain almost completely smooth apparent resistivity and phase curves. A quantitative analysis exhibits that the MPD criterion removed more than 50 per cent of the events.

5.3.1 Application of the MPD criterion to complex polarization pattern

As shown in Fig. 5(b), station V-2 is highly affected by EM noise that shows a distinct polarization direction. The additional application of the MPD criterion can improve the exemplary period $T_8 = {}^1/_{256}\,s$ (see Fig. 18a) by removing large parts of the polarization band (Fig. 18b) which results in rejection of approximately 80 per cent of all events. This high amount of noise explains why a statistical approach will never succeed. However, in contrast to the depicted examples that are characterized by more or less distinct polarization bands, EM noise can also exhibit a complex polarization pattern (Fig. 18c). Such complicated polarization patterns are difficult to remove with an automatic criterion and require manual editing by an

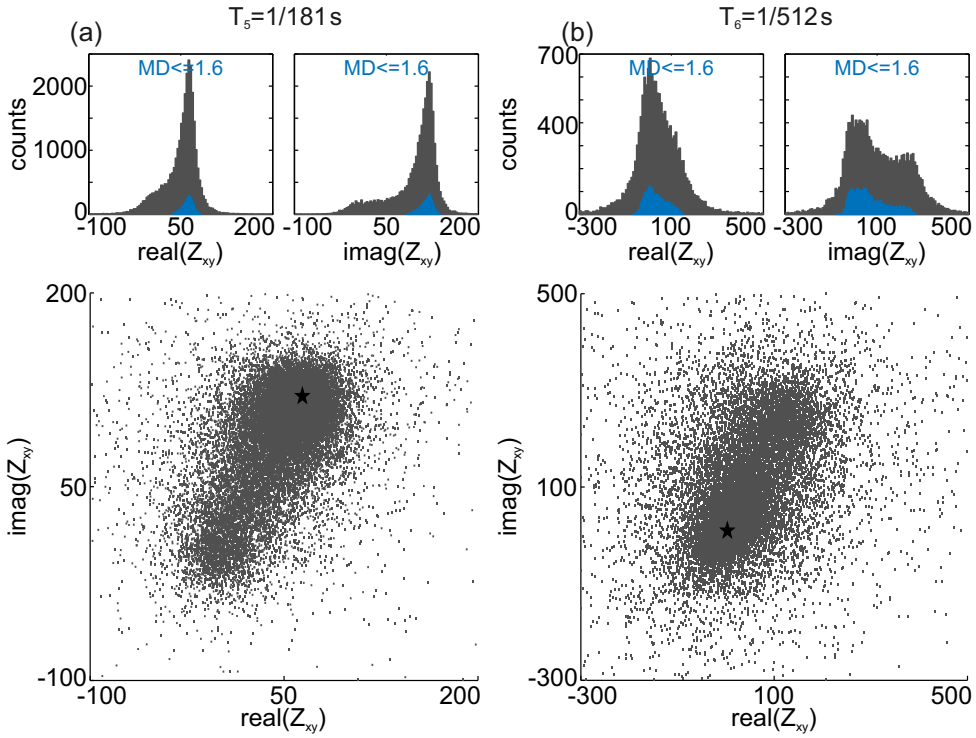

Figure 15. Histograms and scatterplots (Argand diagrams) of station V-1 for (a) $T_5 = 1/181\,s$ and (b) $T_6 = 1/512\,s$. (a) The distribution of all events (grey colour in histograms) is long tailed, best visible in the histogram of the imaginary part and in the corresponding scatterplot. The MD criterion completely removes events corresponding to this tail. (b) Two merged distributions can be depicted for this period, whereby the majority of data now belongs to EM noise; therefore, the MD criterion fails. The black asterisks represent the processing results for the current periods.

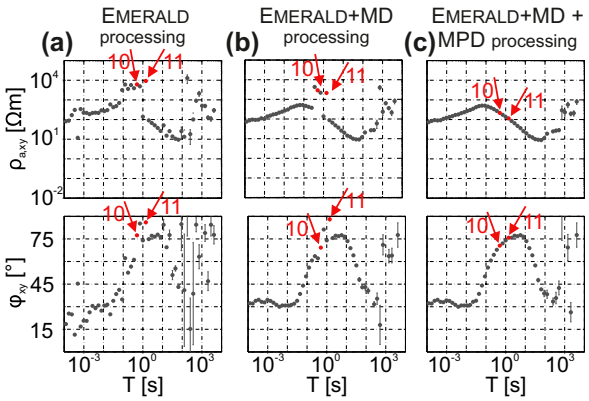

Figure 16. Apparent resistivities and phases of one impedance tensor component of station SA-5. (a) Processing results of the standard EMERALD processing using a coherence threshold of 0.9. The additional application of the MD criterion in (b) improves the result for several periods. However, some periods suffer from noise, which exhibits preferred magnetic polarization directions and can only be improved by using the MPD criterion together with the MD criterion (c). Two of these periods ($T_{10} = 0.5\,s$ and $T_{11} = 1.4\,s$) are highlighted and their incidence directions over the entire recording time are shown in Fig. 17.

experienced user. The success of the MPD criterion in these cases can therefore not be guaranteed. However, sometimes the MPD criterion is able to remove enough distorted events of these patterns to obtain improved processing results as for $T_9 = 1/32\,s$ for station V-2 (see Figs 18a and c). In this case more than 85 per cent of all events were removed by the MPD criterion.

Although the MPD criterion is presented here as a physically based add-on for the MD criterion, it can also be applied without the MD criterion. However, best results are normally obtained by using the combination of both criteria.

6 CONCLUSIONS

To obtain the MT impedance tensor, we use smoothed cross- and autospectra of contiguous time windows, called events, which are averaged by a statistically robust approach. Especially in populated and industrialized areas, robust processing approaches often fail to estimate physically meaningful MT results as the desired natural MT signal is superimposed by man-made EM noise signals. While remote reference and multi-station processing are powerful tools to tackle this problem, we often face the problem that remote stations still suffer from correlated EM noise over large distances or sometimes insufficient time accuracy. In these cases, the practitioner is set back to single-site processing. We observe that intermittent EM noise contributions often form their own distribution of transfer functions overlying the desired MT signal distribution. Therefore we introduce a pre-stack data confinement criterion based on the Mahalanobis distance to classify the distribution of MT transfer functions through their distance. This criterion can be used for single-site as well as an add-on for the remote reference processing. Outliers and events belonging to an intermittent EM noise distribution are assumed to have a larger distance to the desired MT data distribution than events caused by this distribution of natural signal. The basic idea of this criterion is to confine the data to an ideally noise-free or noise reduced subset, to improve the conditions for the subsequently applied robust statistics in the regression problem. The MD is a commonly used measure to detect outliers in many fields

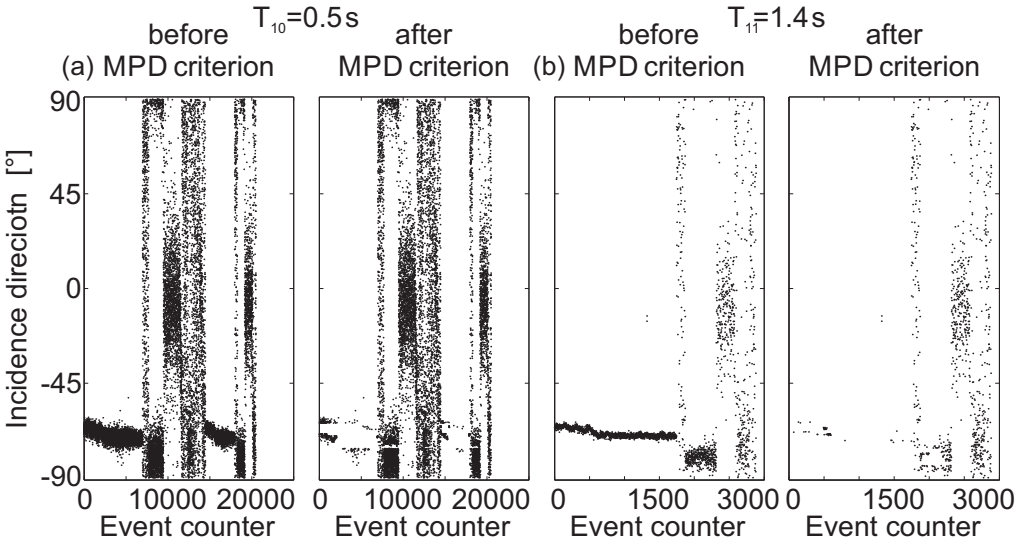

Figure 17. Plots of the incidence directions of the magnetic wave field α_B of all events before and after the application of the MPD criterion for (a) $T_{10} = 0.5\,s$ and (b) $T_{11} = 1.4\,s$ in an interval of $[-90°, 90°]$. Most of the events belonging to a distinct polarization band are removed by the MPD criterion. In both cases more than 50 per cent of all events were removed.

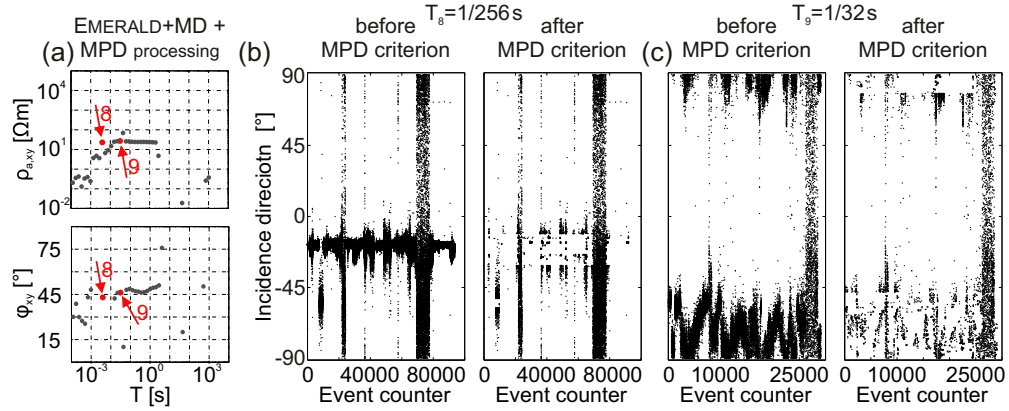

Figure 18. Application of the MPD criterion as add-on for the statistical MD criterion for station V-2. (a) Apparent resistivity and phase curve using MD and MPD criterion in addition to the standard EMERALD processing. Both exemplary periods can be improved (see also Fig. 8e). Incidence directions of the magnetic wave field α_B before and after the application of the MPD criterion for (b) $T_8 = 1/256\,s$ and (b) $T_9 = 1/32\,s$ show that large parts of the polarization band as well as the complex polarization pattern are removed.

of science and economy. Here, we used the MD as an additional measure in MT data processing to classify potential outliers and additional noise clusters and to improve the signal-to-noise ratio prior to the robust stacking process. To calculate the MD for each single event, we need the data centre and covariance matrix. Although the MD seems superior at first glance, its properties are also very much dependent on a robust estimation of data centre and covariance. A deterministic approach is required that results in various initial subsets of data and computation of different correlation matrices for the estimation of MD properties. Since all of these different estimators are based on statistical approaches, we have added a 'physical' estimator to this set which reflects a direct consequence of the inductive processes in MT. We therefore assume that induction spheres for EM field variations of adjacent periods do not change abruptly resulting in smoothly varying apparent resistivity and phase curves. Accordingly we consider the data centre and covariance of a previously estimated adjacent period as an additional estimator. To implement the MD as a measure to detect and remove outliers into the data processing suite EMERALD, we used real and imaginary parts of the MT impedance tensor components as input data. Single events which have a high distance to the estimated data centre are rejected from the further processing. Events that have passed the MD criterion are subsequently stacked in a robust manner within the standard EMERALD processing scheme. This pre-stack criterion was tested for various MT data sets from different regions and suffering from different noise contaminations. A comparison of processing results with and without the MD criterion reveals that for stations with less than 50 per cent noise contamination, data quality of apparent resistivity and phase curves could be improved over the entire period range, even in the so-called dead band. Since EM noise often forms a completely independent cluster of transfer functions, the MD criterion is able to remove such clusters as well as to reduce scatter around the desired cluster of MT transfer functions. A necessary prerequisite is that events of the desired natural MT data dominate over the amount of events contaminated by EM noise. Many MT stations fulfil the requirement that noise does not outweigh natural MT signal and thus this criterion is a useful measure to improve the estimation of transfer functions in an automated way, in particular

when no other methods like remote reference can be applied. However, the MD is a purely statistical measure and therefore cannot distinguish between physically reasonable MT data and EM noise signal. For stations affected by a high amount of noise (i.e. higher than 50 per cent) or in cases where the transfer function distribution of EM noise overlaps with the MT transfer function distribution, the automated MD criterion can result in either totally misleading transfer functions or does not show any improvement. In these cases, an adequate remote station or manual rejection of noisy events, for example by physically based data selection criteria or other *a priori* information are necessary to ensure that the majority of all events is well-behaved for the estimation of correct and undisturbed MT transfer functions. For such cases, we also introduced a physically motivated data selection criterion based on the magnetic polarization (incidence) direction and showed its successful application as an add-on for the statistical MD criterion. We observe that based on the removal of strongly polarized magnetic fields, even more than 80 per cent of the entire events creates optimal conditions for the successful application of the MD criterion or if used separately without MD for the subsequent robust stacking.

ACKNOWLEDGEMENTS

We would like to thank Jose Cruses for providing MT data examples from Venezuela. Furthermore, we thank Michael Becken, Matthew Comeau and Dominik Harpering for providing us the processing results of SA-4 using the approach after Egbert & Booker (1986). The presented MT data were recorded with instruments from the Geophysical Instrument Pool Potsdam (GIPP), which is highly appreciated. We also appreciate comments and suggestions by Gary Egbert, Alan Chave, Anne Neska, Elena Sokolova and one anonymous reviewer that helped to improve the manuscript.

REFERENCES

Basu, M. & Ho, T.K., 2006. *Data Complexity in Pattern Recognition,* Springer London.

Billor, N., Hadi, A.S. & Velleman, P.F., 2000. BACON: Blocked adaptive computationally efficient outlier nominators, *Comput. Stat. Data Anal.,* **34**(3), 279–298.

Brereton, R.G., 2015. The Mahalanobis distance and its relationship to principal component scores, *J. Chemom.,* **29**(3), 143–145.

Chave, A.D., 2014. Magnetotelluric data, stable distributions and impropriety: an existential combination, *Geophys. J. Int.,* **198**(1), 622–636.

Chave, A.D., 2017. Estimation of the magnetotelluric response function: the path from robust estimation to a stable maximum likelihood Estimator, *Surv. Geophys.,* **38**(5), 837–867.

Chave, A.D. & Jones, A.J., 2012. *The Magnetotelluric Method,* Cambridge University Press.

Chave, A.D. & Thomson, D.J., 1989. Some comments on magnetotelluric response function estimation, *J. geophys. Res.: Solid Earth,* **94**(B10), 14 215–14 225.

Chave, A.D. & Thomson, D.J., 2004. Bounded influence magnetotelluric response function estimation, *Geophys. J. Int.,* **157**(3), 988–1006.

Chave, A.D., Thomson, D.J. & Ander, M.E., 1987. On the robust estimation of power spectra, coherences, and transfer functions, *J. geophys. Res.,* **92**(B1), 633.

de Maesschalck, R., Jouan-Rimbaud, D. & Massart, D.L., 2000. The Mahalanobis distance, *Chemom. Intell. Lab. Syst.,* **50**(1), 1–18.

Dickhaus, T., 2003, Statistische Verfahren fuer das Data Mining in der pharmazeutischen Forschung. *Diploma thesis,* Fachhochschule Aaachen, Juelich.

Egbert, G., 1997. Robust multiple-station magnetotelluric data processing, *Geophys. J. Int.,* **130,** 475–496.

Egbert, G.D. & Booker, J.R., 1986. Robust estimation of geomagnetic transfer functions, *Geophys. J. R. astr. Soc.,* **87**(1), 173–194.

Falk, M., 1997. On mad and comedians, *Ann. Inst. Stat. Math.,* **49**(4), 615–644.

Filzmoser, P., Garrett, R.G. & Reimann, C., 2005. Multivariate outlier detection in exploration geochemistry, *Comput. Geosci.,* **31**(5), 579–587.

Fowler, R.A., Kotick, B.J. & Elliot, R.D., 1967. Polarization analysis of naturally and artificially geomagnetic micropulsations, *J. geophys. Res.,* **72,** 2871–2883.

Friebel, T., Stockmann, M. & Haber, R., 2010. Sensorueberwachung mit einer robusten zweidimensionalen Regelkarte, in *AALE 2010,* pp. 71–76.

Gamble, T.D., Goubau, W.M. & Clarke, J., 1979. Magnetotellurics with a remote magnetic reference, *Geophysics,* **44**(1), 53–68.

Gnanadesikan, R. & Kettenring, J.R., 1972. Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data, *Biometrics,* **28**(1), 81.

Goubau, W.M., Gamble, T.D. & Clarke, J., 1978. Magnetotelluric data analysis: Removal of bias, *Geophysics,* **43**(6), 1157–1166.

Hampel, F.R., 1986. , *Robust Statistics: The Approach Based on Influence Functions. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics,* Wiley.

Hayashi, S., Tanaka, Y. & Kodama, E., 2001. A new manufacturing control system using Mahalanobis distance for maximising productivity, in *Conference Proceedings / 2001 IEEE International Symposium on Semiconductor Manufacturing,* pp. 59–62, IEEE Operations Center, Piscataway, NJ.

Huber, P.J., 1981. *Robust Statistics: Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics,* John Wiley.

Hubert, M. & Debruyne, M., 2010. Minimum covariance determinant, *Comput. Stat.,* **2**(1), 36–43.

Hubert, M., Rousseeuw, P.J. & van Aelst, S., 2008. High-breakdown robust multivariate methods, *Stat. Sci.,* **23**(1), 92–119.

Hubert, M., Rousseeuw, P.J. & Verdonck, T., 2012. A deterministic algorithm for robust location and scatter, *J. Comput. Graph. Stat.,* **21**(3), 618–637.

Jones, A.G., Chave, A.D., Egbert, G., Auld, D. & Bahr, K., 1989. A comparison of techniques for magnetotelluric response function estimation, *J. geophys. Res.: Solid Earth,* **94**(B10), 14201–14213.

Junge, A., 1990. Robust estimation of bivariate transfer functions (in German), in *Protokol Kolloquium Elektromagnetische Tiefenforschung,* Hornburg.

Junge, A., 1992. Zur Schätzung der effektiven Anzahl der Freiheitsgrade bei der Bestimmung magnetotellurischer Übertragungsfunktionen, in *Protokol Kolloquium Elektromagnetische Tiefenforschung,* Borkheide.

Junge, A., 1994. Induced telluric fields - new observations in North Germany and the Bramwald (in German), *Habilitation thesis,* Faculty of Physics, University of Göttingen.

Kleinschmidt, P., Mitterreiter, I. & Piper, J., 1994. Improved chromosome classification using monotonic functions of mahalanobis distance and the transportation method, *Math. Methods Operat. Res.,* **40**(3), 305–323.

Korolevski, W., Ritter, O., Weckmann, U., Rybin, A. & Matiukov, V., 2014. Magnetotelluric Study of the Southern Pamir, Tajikistan, in *AGU Fall Meeting Abstracts*.

Krings, T., 2007, The influence of robust statistics, remote reference, and horizontal magnetic transfer functions on data processing in magnetotellurics. *Diploma thesis,* University Muenster, Muenster.

Larsen, J.C., 1989. Transfer functions: smooth robust estimates by least-squares and remote reference methods, *Geophys. J. Int.,* **99,** 645–663.

Larsen, J.C., Mackie, R.L., Manzella, A., Fiordelisi, A. & Rieven, S., 1996. Robust smooth magnetotelluric transfer functions, *Geophys. J. Int.,* **124**(3), 801–819.

Lehmann, R., 2012, Der Einfluss statistischer Ausreisser auf die Schaetzung der natuerlichen Variabilitaet in Daten zu Biota. *PhD thesis,* RWTH Aachen University, Aachen.

Lohninger, H., 2012. *Fundamentals of Statistics,* Epina e-Book Team.

Mahalanobis, P.C., 1936. On the generalized distance in statistics, *Proc. Natl. Inst. Sci. India,* **2,** 49–55.

Maronna, R.A. & Zamar, R.H., 2002. Robust estimates of location and dispersion for high-dimensional datasets, *Technometrics,* **44**(4), 307–317.

Morrison, D.F., 1967. *Multivariate Statistical Methods,* McGraw-Hill.

Muñoz, G., Ritter, O. & Moeck, I., 2010. A target-oriented magnetotelluric inversion approach for characterizing the low enthalpy Groß Schönebeck geothermal reservoir, *Geophys. J. Int.,* **183**(3), 1199–1215.

Oettinger, G., Haak, V. & Larsen, J.C., 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, *Geophys. J. Int.,* **146**(3), 659–669.

Pedersen, L.B., Juhlin, C. & Rasmussen, T.M., 1992. Electric resistivity in the gravberg-1 Deep Weel, Sweden, *J. geophys. Res.,* **97**(B6), 9171.

Platz, A., 2018, Novel pre-stack data confinement and selection for magnetotelluric data processing and its application to data of the Eastern Karoo Basin, South Africa. *PhD thesis,* University of Potsdam, Potsdam.

Ritter, O., Junge, A. & Dawes, G., 1998. New equipment and processing for magnetotelluric remote reference observations, *Geophys. J. Int.,* **132**(3), 535–548.

Rousseeuw, P.J., 1984. Least median of squares regression, *J. Am. Stat. Assoc.,* **79**(388), 871–880.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point, in *Mathematical Statistics and Applications,* pp. 283–297, eds Grossmann, W., Pflug, G.C., Vincze, I. & Wertz, W., Springer Netherlands.

Rousseeuw, P.J. & Croux, C., 1993. Alternatives to the median absolute deviation, *J. Am. Stat.Assoc.,* **88**(424), 1273–1283.

Rousseeuw, P.J. & van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator, *Technometrics,* **41**(3), 212.

Schmitz, M., Orihuela, N.D., Klarica, S., Gil, E., Levander, A., Audemard, F.A., Mazuera, F. & Avila, J., 2013. Lithospheric scale model of Merida Andes, Venezuela (GIAME Project), in *AGU Spring Meeting Abstracts*.

, Schmucker, U. & Weidelt, P., 1975. *Electromagnetic Induction in the Earth,* Lecture notes, Aarhus.

Sims, W.E., Bostick, F.X. & Smith, H.W., 1971. The estimation of magnetotelluric impedance tensor elements from measured data, *Geophysics,* **36**(5), 938–942.

Smirnov, M.Y., 2003. Magnetotelluric data processing with a robust statistical procedure having a high breakdown point, *Geophys. J. Int.,* **152,** 1–7.

Smirnov, M.Y. & Egbert, G.D., 2012. Robust principal component analysis of electromagnetic arrays with missing data, *Geophys. J. Int.,* **190**(3), 1423–1438.

Srinivasaraghavan, J. & Allada, V., 2006. Application of mahalanobis distance as a lean assessment metric, *Int. J. Adv. Manufact. Technol.,* **29**(11-12), 1159–1168.

Travassos, J.M. & Beamish, D., 1988. Magnetotelluric data processing-a case study, *Geophys. J. Int.,* **93**(2), 377–391.

Verboven, S. & Hubert, M., 2010. MATLAB library LIBRA, *Comput. Stat.,* **2**(4), 509–515.

Weckmann, U., Magunia, A. & Ritter, O., 2005. Effective noise separation for magnetotelluric single site data processing using a frequency domain selection scheme, *Geophys. J. Int.,* **161**(3), 635–652.

Wu, T.-J., Burke, J.P. & Davison, D.B., 1997. A measure of DNA sequence dissimilarity based on mahalanobis distance between frequencies of words, *Biometrics,* **53**(4), 1431.

Yohai, V.J. & Zamar, R.H., 1988. High breakdown-point estimates of regression by means of the minimization of an efficient scale, *J. Am. Stat. Assoc.,* **83**(402), 406.

APPENDIX A: INITIAL ESTIMATORS FOR THE DETERMINISTIC MCD ALGORITHM

The first six estimators of our modified deterministic MCD algorithm are the same as in the original deterministic MCD algorithm from Hubert *et al.* (2012), which represent different types of correlation matrices:

(i) $\mathbf{S}_1 = corr(\mathbf{W})$ with $W_j = tanh(Y_j)$ for $j = 1, \ldots, p$
(ii) $\mathbf{S}_2 = corr(\mathbf{R})$ with R_j being the ranks of the column Y_j
(iii) $\mathbf{S}_3 = corr(\mathbf{T})$ with $T_j = \Phi^{-1}((R_j - 1/3)/(n + 1/3))$ and the normal cumulative distribution function Φ
(iv) $\mathbf{S}_4 = (1/n)\sum_{i=1}^{n} k_i k_i^T$ with $k_i = y_i/\|y_i\|$
(v) $\mathbf{S}_5$ is based on the first steps of the BACON algorithm (Billor *et al.* 2000)
(vi) $\mathbf{S}_6$ is based on the raw orthogonalized Gnanadesikan–Kettenring estimator from Maronna & Zamar (2002)

APPENDIX B: THRESHOLD FOR INCIDENCE ANGLES OF THE MAGNETIC FIELDS

With 180 bins for values of the incidence angle between $-90°$ and $90°$, the expected amount of events in one bin $E_k = \frac{Number\ of\ events}{180}$ depends on the total number of events. Empirical thresholds for treating an incidence angle as part of a strongly polarized magnetic fields are given in Table B1. All these values are found by trial and error after testing many stations with different polarization patterns. The chosen limits are selected in a conservative manner to assure that only events corresponding to a distinct polarization direction are removed and to ensure that all events are accepted for stations that do not show any preferred polarization direction.

Table B1. Threshold for flagging a bin as caused by strongly polarized magnetic field.

Period with no. of events	≤ 90	≤ 180	≤ 900	>900
No. of events in bin	$>10 \cdot E_k$	$>10 \cdot E_k$	$>11 \cdot E_k$	$>4 \cdot E_k$ & consecutive events