New MT impedance estimators in frequency-domain and time-domain using automatic preselection of time-series data

陳, 浩

出版情報：Kyushu University, 2022, 博士（工学）, 課程博士
バージョン：
権利関係：

New MT impedance estimators in frequency-domain and time-domain using automatic preselection of time-series data

Ph.D. Thesis

Exploration Geophysics Laboratory
Department of Earth Resources Engineering
Graduate School of Engineering

九州大学
KYUSHU UNIVERSITY

Author: Hao Chen (Student ID: 3TE19453P)
Supervisor: Hideki Mizunaga
Co-Supervisor: Toshiaki Tanaka

May 2022

Contents

Chapter 1: INTRODUCTION AND MOTIVATION

1.1 Introduction

The magnetotelluric (MT) method is a passive electromagnetic (EM) method used to infer the subsurface electrical conductivity. It was first proposed by Rikitake (1948), Cagniard (1953) and Tikhonov (1950). Fig. 1.1 shows the sketch map of the MT field campaign. The MT source is the natural EM signal coming from the Earth's magnetosphere and ionosphere or global lightning. Three magnetic components in three orthogonal directions and two horizontal electric components in the direction of N-S and E-W are measured simultaneously. The subsurface electrical conductivity distribution can be determined from the natural geomagnetic and geoelectric fields observed on Earth's surface.

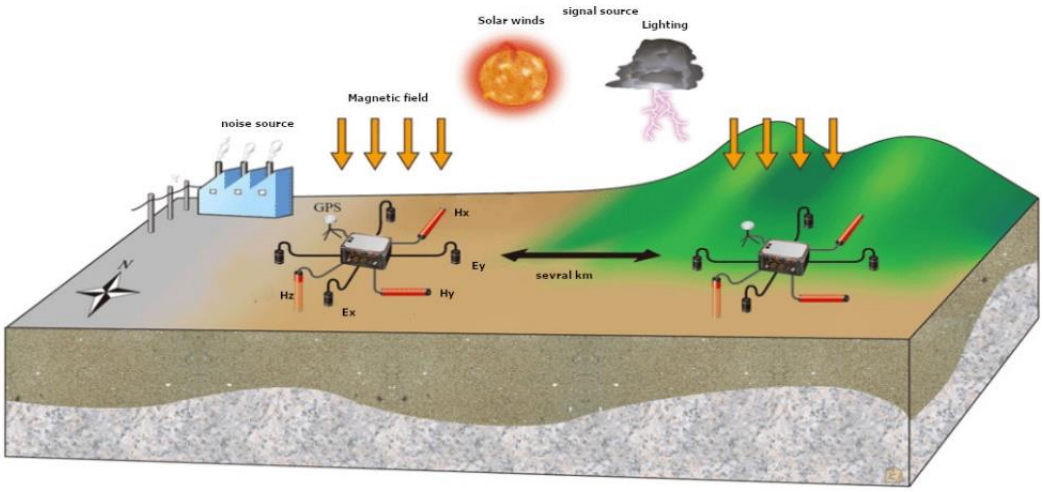

Fig. 1.1 The sketch map of the MT field campaign (download from: http://www.chikatansa.co.jp /tansa-tansa_03.html).

Conductivity is sensitive to mineral composition, salinity, temperature, and fluids. Fig. 1.2 shows the conductivity corresponding to the mineral composition in different conditions. MT method is widely used for economic, environmental, and engineering applications and hazard assessments in earthquake and volcano monitoring. Numerous case studies have been published, including mineral exploration, direct imaging of ore bodies, petroleum exploration, and geotherm.

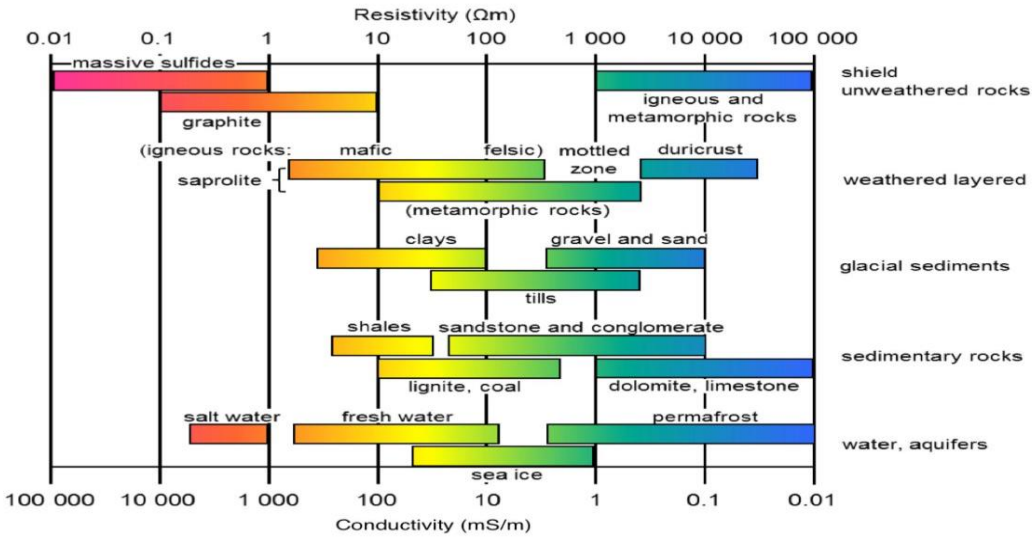

Fig. 1.2 The conductivity corresponding to the mineral composition in different conditions (taken from: https://www.eoas.ubc.ca/courses/eosc350/content/foundations/properties/resistivity.htm).

1.2 The Basic Theory of the MT method

Maxwell's equations can describe the electromagnetic fields within a material as follows:

$$\nabla \times \boldsymbol{E} = -\,\partial \boldsymbol{B}/\partial t, \text{ (Faraday's law)} \tag{1.1a}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{j} + \partial \boldsymbol{D}/\partial t, \text{ (Ampere's law)} \quad (1.1b)$$
$$\nabla \cdot \boldsymbol{B} = 0, \text{ (Gauss's law for magnetism)} \quad (1.1c)$$
$$\nabla \cdot \boldsymbol{D} = \rho, \text{ (Gauss's law)} \quad (1.1d)$$

where **E** (V /m) and **H** (A/m) are the electric and magnetic fields, **B**(T) is the magnetic induction, **D** (C/m^2) is the electric displacement, **j** and ∂**D**/∂t (A/m^2) are the current density and the displacement current, respectively. ρ (C/m^3) is the electric charge density owing to free charges.

The magnitudes in Maxwell's equations can be related through their constitutive relationships:

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}, \quad (1.2a)$$
$$\boldsymbol{B} = \mu \boldsymbol{H}, \quad (1.2b)$$
$$\boldsymbol{j} = \sigma \boldsymbol{E}, \quad (1.2c)$$

where σ, ε and μ describe the intrinsic properties of the materials. σ (S/m) is the electrical conductivity (its reciprocal is the electrical resistivity ρ:1/σ (Ω·m)), ε (F/m) is the dielectric permittivity, and μ (H/m) is the magnetic permeability. These magnitudes are scalar quantities in isotropic media. In anisotropic materials, they must be expressed in a tensorial form.

By the assumption that the magnetic field is quasi-homogeneous, the displacement currents (∂**D**/∂t) can be neglected relative to conductivity currents (**j**) (eq. 1.1b) for the period range 10^{-5} s to 10^{5} s and for not extremely low conductivity values. Moreover, in the absence of charges, the right term of Eq. 1.1 d vanishes and the electric and magnetic field solutions depend solely upon angular frequency (ω) and conductivity. The Maxwell function can be rewritten as follows:

$$\nabla \times \boldsymbol{E} = i\omega\mu \boldsymbol{H} \quad (1.3a)$$
$$\nabla \times \boldsymbol{H} = \sigma \boldsymbol{E} \quad (1.3b)$$
$$\nabla \cdot \boldsymbol{H} = 0 \quad (1.3c)$$
$$\nabla \cdot \boldsymbol{E} = 0 \quad (1.3d)$$

I take the curl of Eq. (1.3 a):

$$\nabla \times \nabla \times \boldsymbol{E} = i\omega\mu \ \nabla \times \boldsymbol{H} \quad (1.4)$$

The left side of the function (1.4):

$$\nabla \times \nabla \times \boldsymbol{E} = \nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E} = -\nabla^2 \boldsymbol{E} \quad (1.5)$$

And plug (1.3 b) into the right side of Eq. (1.4), I can get:

$$-\nabla^2 \boldsymbol{E} = i\omega\mu\sigma \boldsymbol{E} \quad (1.6)$$

Let's set k=$\sqrt{-i\omega\mu\sigma}$, then I can get:

$$\nabla^2 \boldsymbol{E} - k^2 \boldsymbol{E} = 0 \quad (1.7)$$

Similar processing to the ***H***, I can get:

$$\nabla^2 \boldsymbol{H} - k^2 \boldsymbol{H} = 0 \quad (1.8)$$

In the Cartesian coordinate system, the z-axis is downward, and the x and y axes are in the horizontal plane. In the homogeneous half-space, there is the following relationship as follows for the TE ($E_x - H_y$) mode:

$$\frac{\partial \boldsymbol{E}_x^2}{\partial Z^2} - k^2 \boldsymbol{E}_x = 0, \quad (1.9)$$
$$\frac{\partial \boldsymbol{E}_x}{\partial z} = i\omega\mu \boldsymbol{H}_y. \quad (1.10)$$

The general solution to the second-order differential equations (1.9) is:

$$\boldsymbol{E}_x = Ae^{-kz} + Be^{kz}. \quad (1.11)$$

When $\boldsymbol{Z} \rightarrow \infty$; $\boldsymbol{E}_x$=0; therefore, B=0;I can get:

$$\boldsymbol{E}_x = Ae^{-kz}. \quad (1.12)$$

Combine (1.12) and (1.10), I can get:

$$\boldsymbol{H}_y = \frac{1}{\sqrt{-i\,\omega\mu\rho}} \ Ae^{-kz}. \quad (1.13)$$

Therefore the impedance:

$$Z = Z_{TM} = \frac{E_x}{H_y} = \sqrt{-i\omega\mu\rho}. \quad (1.14)$$

According to the forward simulation, I can infer the impedance at the Earth's surface. Fig. 1.3 shows the apparent resistivity and phase curve on the surface of the Earth; it shows that even if the resistivity changes rapidly, the apparent resistivity and phase curve varies smoothly.

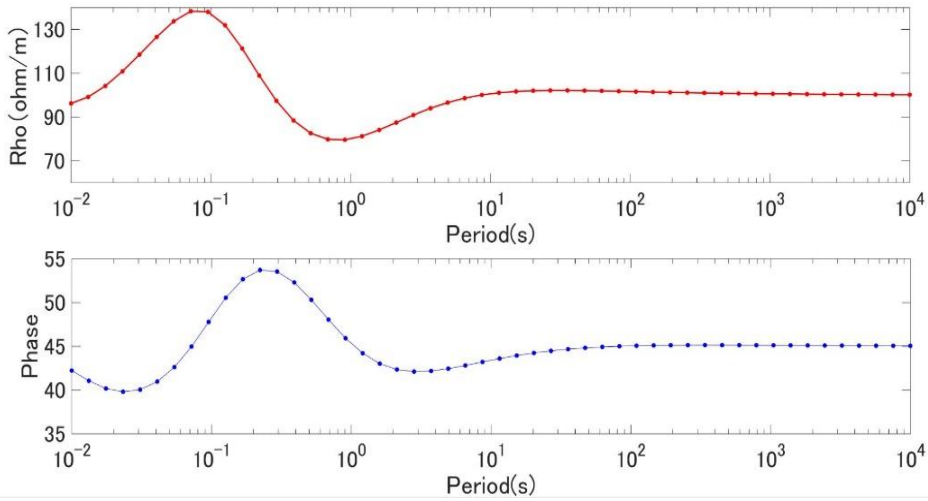

Fig 1.3 The apparent resistivity and phase curve on the surface of the earth, when the thickness and resistivity of each layer are as follows: h=[500,1000,750,500,1500,2000]; rho = [100, 200, 50, 20, 400, 500, 100].

Skin effect describes the influence of the EM wave attenuation with depth. Consider a 1-D medium; the horizontal electric and magnetic field only vary with depth. And the skin depth (δ) is defined as the depth at which the amplitude attenuates to be 1/e of that on the surface:

$$A(\delta)=A*exp^{-\boldsymbol{k}*\boldsymbol{\delta}}=A*exp^{-1}, \tag{1.15}$$

$$\delta=\sqrt{(2*\rho/\omega*\mu_0)}\approx 503*\sqrt{\rho*T}, \tag{1.16}$$

here δ in meter, ρ in ohm/m, T in seconds. Skin depth is expressed as the penetration depth of EM signals. The lower frequency and higher resistivity correspond to slower attenuation and deeper penetration. In the MT study, the period is usually acting as the depth.

1.3 Impedance Tensor

In the MT method, the magnetic field (***H***) and the electric field (***E***) have a linear relationship in the frequency domain. The impedance tensor at a specific frequency can be calculated in the frequency domain as follows:

$$\begin{pmatrix}\boldsymbol{E_x}(\omega)\\ \boldsymbol{E_y}(\omega)\end{pmatrix}=\begin{pmatrix}Z_{xx}(\omega)\ Z_{xy}(\omega)\\ Z_{yx}(\omega)\ Z_{yy}(\omega)\end{pmatrix}\begin{pmatrix}\boldsymbol{H_x}(\omega)\\ \boldsymbol{H_y}(\omega)\end{pmatrix}, \tag{1.17}$$

where ***E*** and ***H*** are the horizontal electric and magnetic fields at a specific frequency, respectively, ω denotes the angular frequency, and ***Z*** means the MT impedance. The suffix *x* denotes the north-south direction, and *y* denotes the east-west direction. I can obtain the impedance from the measured geoelectric and geomagnetic time-series data.

The components of ***Z*** are complex numbers. Two scalar magnitudes (The apparent resistivity and impedance phase) are defined from the modulus and phase of the ***Z*** in a specific frequency. The apparent resistivity, which is an average resistivity for the volume of Earth sounded at a particular period, is defined as follows:

$$\rho_a=|\boldsymbol{Z}|^2/(\omega\mu), \tag{1.18}$$

The impedance phase is the phase of the ***Z*** and is defined as follows:

$$\varphi=tan^{-1}(\frac{imag(\boldsymbol{Z})}{real(\boldsymbol{Z})}). \tag{1.19}$$

1.4 Motivation and Evolution

MT data interpretation is generally divided into two steps. The first step in MT data processing is to estimate the frequency-domain impedance tensor from the measured time-series data. The second step is to do the inversion based on the impedance and finally get the subsurface resistivity structure. All MT data interpretations are based on the MT impedance. Therefore, it is very important to obtain a reliable impedance. Any subsequent analysis and high-precision inversion will be meaningless without a high-quality and credible response function. Therefore, studying estimating the MT impedance tensor is the primary issue in improving the magnetotelluric method's effectiveness.

The use of natural sources is both an attraction and a weakness. The nature EM signal is extremely weak,

and the cultural noise easily influences it. Artificial disturbances to electromagnetic observations are becoming more serious with urban constructions. The near-field noises produced by human activities such as vehicles, subways, and high-speed railways can easily influence the MT field data. Estimating a reliable impedance can be challenging under the influence of cultural noise. This thesis focuses on researching the noise reduction method of MT time-series data and getting a reliable impedance.

We proposed a new method to estimate the data quality of the MT time-series data. By this method, we can extract the high signal-to-noise ratio (SNR) data automatically. We can get a reliable result assuming the field data is not contaminated by the continuous noise. We also used the data quality analysis method to investigate the influence of geomagnetic storms on the MT impedance calculation. We found that the geomagnetic storm is beneficial to data quality; the high SNR data may appear during a geomagnetic storm. And we can get a more reliable result using the data observed during the geomagnetic storm in a noisy site.

This thesis is organized as follows. Chapter 1 introduces the basic theory of the MT method. Chapter 2 introduces the signal and noise for the MT method. Chapter 3 introduces the conventional MT impedance estimation methods. Chapter 4 introduces a robust estimator based on the windowed FFT for data quality analysis and impedance calculation. Chapter 5 introduces a robust estimator based on the Hilbert-Huang transformation (HHT). Chapter 6 introduces the influence of geomagnetic storms on the MT impedance calculation. Chapter 7 summarizes the practical way to reduce the noise of MT time-series data.

Chapter 2: THE SIGNAL AND NOISE FOR THE MT METHOD

Magnetotelluric (MT) field data contain natural electromagnetic (EM) signals and artificial noise sources (instrumental, anthropogenic, etc.). The MT method assumes that the EM source is located at a large distance from the measurement area, and the signal propagates as plane waves. This condition is also known as the far-field assumption (Oettinger et al., 2001; Zonge and Hughes, 1987) and is dependent on the target period. The natural EM signal from the Earth's magnetosphere and ionosphere or global lighting is far from the observation site and can be treated as plane waves. On the contrary, the local anthropogenic sources violate the plane wave assumption and are regarded as noise in the MT method. I can subdivide the field data ($\boldsymbol{E}$ and $\boldsymbol{H}$) into three parts as follow:

$$\boldsymbol{E} = \boldsymbol{E}^{MT} + \boldsymbol{E}^{CN} + \boldsymbol{E}^{IN}, \quad (2.1\ a)$$

$$\boldsymbol{H} = \boldsymbol{H}^{MT} + \boldsymbol{H}^{CN} + \boldsymbol{H}^{IN}, \quad (2.1\ b)$$

where the *CN* denotes the coherent noise, the *IN* denotes incoherent noise, and the *MT* denotes electromagnetic signals (Oettinger et al., 2001).

2.1 Natural Electromagnetic Signal

Many works have focused on the Earth's EM environment (Constable, 2016; Constable and Constable, 2004; Garcia and Jones, 2002; Hennessy and Macnae, 2018; Mareschal, 1986; McPherron, 2005). Generally, the high-frequency signals (> 1 Hz) originate from worldwide thunderstorm activity. In comparison, the low-frequency signals (< 1 Hz) originate from the interaction between solar winds and the Earth's magnetosphere and ionosphere.

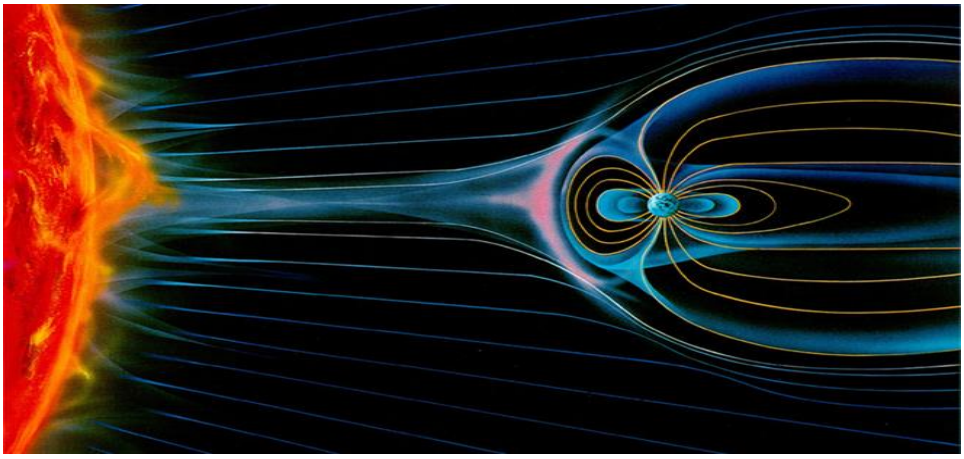

Fig. 2.1 Interaction between the magnetosphere and the solar wind (Taken from: https://en.wikipedia.org/wiki/Geomagnetic_storm).

The solar wind is a plasma stream ejected from the upper atmosphere of the Sun. It consists of high-energy electrons and protons. Earth has a strong internal magnetic field. The geomagnetic field can be roughly a dipole field with an axis tilted about 11 degrees from the spin axis without any external drivers. Forcing by the solar wind modifies the field, creating a cavity called the magnetosphere. The magnetosphere is formed by the interaction of the solar wind with Earth's magnetic field. Fig. 2.2 illustrates the shape and size of Earth's magnetic field. The plane wave assumption is generally acceptable at midlatitudes (Lezaeta et al., 2007; Viljanen et al., 1993). However, it is violated at high magnetic latitudes because the source field is nonuniform during geomagnetic storms (Mareschal, 1981; Viljanen et al., 1993; Garcia et al., 1997; Lezaeta et al., 2007). Possible biases in the MT transfer function due to the source effect are considered only at long periods (> 1000 s) and near the auroral or equatorial electrojets (Murphy and Egbert, 2018).

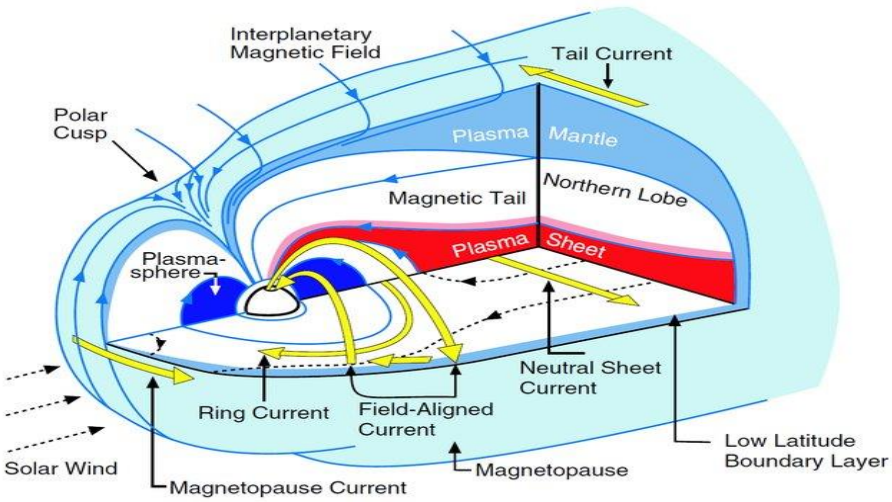

Fig. 2.2 Three-dimensional cutaway view of the magnetosphere. The light blue outer surface is the magnetopause. And its boundary layers are shown in darker blue. Magnetic field lines are shown in blue, and electric currents in yellow. The polar cusp is the polar region where the magnetic field lines converge (Figure taken from Chandorkar, 2019).

Geomagnetic pulsations result from the interactions between the solar wind and the magnetosphere, and these events will appear on the EM time series as short oscillations. The pulsation amplitude ranges from tenths of nT to hundreds of nT on the ground, and its period ranges from tenths of seconds to a few minutes. As shown in Table 2.1 and Fig. 2.3, pulsations are classified into continuous pulsations (Pc) and irregular pulsations (Pi); they are further divided according to their period within these two groups. This classification was proposed by IAGA (International Association for Geomagnetism and Aeronomy) in 1964.

Table 2.1 The Geomagnetic pulsations classification proposed by IAGA

Continuous pulsations		Irregular pulsations	
Notation	Period(s)	Notation	Period(s)
Pc1	0.2 - 5	Pi1	1 - 40
Pc2	5 - 10	Pi2	40 - 150
Pc3	10 - 45		
Pc4	45 - 150		
Pc5	150 - 600		

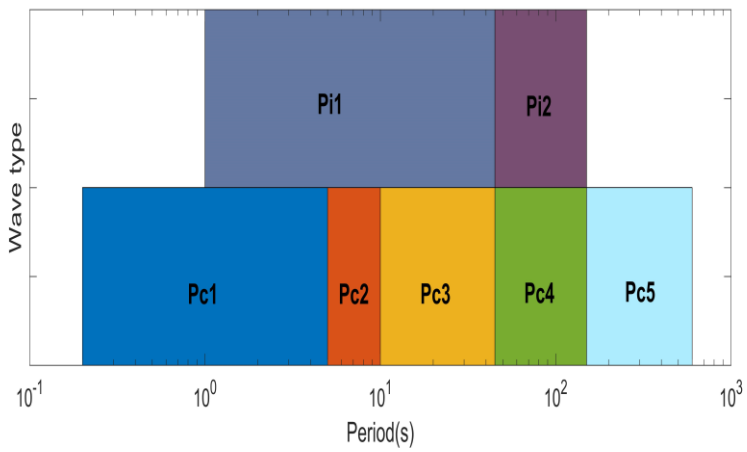

Fig. 2.3 The Geomagnetic pulsations classification proposed by IAGA

Fig. 2.4 shows the typical EM signal strength observed at mid-latitudes during periods of moderate

geomagnetic activity. It shows that the signal strength increases between 10^{-5}~10^{5} seconds, which means that the artificial noise influences the low period data more easily. Moreover, the spectrum has a minimum of around 1 s and 1,000 Hz, and it is most sensitive to culture noise, so they are called a "dead band".

On the other hand, there are spectrum peaks around 7.83 Hz (fundamental), 14.3, 20.8, 27.3 and 33.8 Hz in the extremely low frequency (ELF) portion of the Earth's electromagnetic field spectrum, which is named the Schumann resonances. Schumann resonances are global electromagnetic resonances generated and excited by lightning discharges in the cavity formed by the Earth's surface and the ionosphere.

Moreover, the natural EM signal strength will increase around 10^{1}~10^{4} seconds in the presence of a geomagnetic storm. Next, I will introduce the geomagnetic storm.

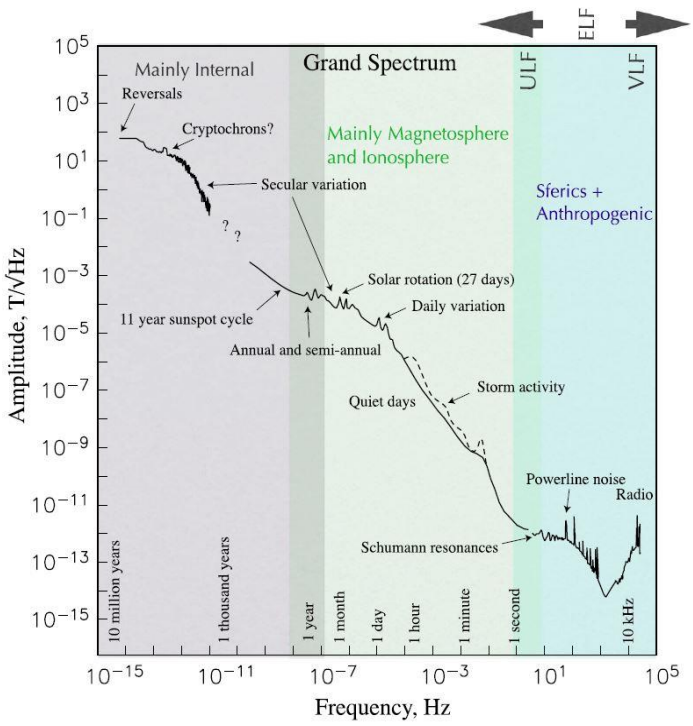

Fig. 2.4 Sketch of the amplitude spectrum of geomagnetic (Figure taken from Constable, 2016).

2.2 Geomagnetic Storm

The geomagnetic storm is a temporary disturbance of the Earth's magnetosphere caused by the solar wind's shock wave interacting with the Earth's magnetic field. Geomagnetic storms start when the enhanced energy of the solar wind transfer into the magnetosphere intensifies the magnetospheric ring current at a distance of 3-8 Earth radius (Chave and Jones, 2012; Tsurutani et al., 2006). It has been known that the horizontal geomagnetic field is depressed during the geomagnetic storm and gradually recovers to its average level. The disturbance storm time (*Dst*) is an index of magnetic activity derived from the four low-latitude horizontal magnetic fields and is defined as follows (see the homepage of WDC Kyoto Observatory):

$$Dst(T)= \sum_{i=1}^{4} \frac{D_i(T)}{cos(\varphi_i)}, \tag{2.2}$$

where φ is the dipole latitudes of the observatory, and T denotes the time in hours. $D(T)$ is the disturbance variation for each observatory and is defined by :

$$D(T)=\Delta H(T) - S_q(T), \tag{2.3}$$

where $\Delta H(T)$ denotes the deviations from the baseline for each observatory, and $S_q(T)$ denotes the solar quiet daily variation. *Dst* can indicate the occurrences of magnetic storms and their severity. When the *Dst* is less than -50 nT, it is categorized as a geomagnetic storm. When the *Dst* is less than -100 nT, it is categorized as a strong geomagnetic storm.

The Phoenix Geophysics system's broadband frequency 5-component MT time-series data were used to show the signal strength variation corresponding to the geomagnetic storm. The MT time-series data were sampled in three frequency bands. The high-frequency band (2,400 Hz) was sampled for 1 second at

intervals of 4 minutes from the beginning of the minute. The middle-frequency band (150 Hz) was sampled for 16 seconds at intervals of 4 minutes from the beginning of the minute. The low-frequency data (15 Hz) was sampled continuously. The data were observed from August 20 to August 28, 2018, at Sawauchi station, Japan. The geomagnetic storm occurred on August 26. The MT time-series data were stored in three files.

First, I analyzed the spectrum variation along with the *Dst* index. I first applied a set of Slepian tapers to obtain precise spectral information from these datasets and then used the fast Fourier transform to the time series (Garcia and Jones, 2002). Fig. 2.5 shows the time-frequency distribution against the *Dst* index and the Hx component time-series data. The sampling rate is 15 Hz, and the upper figure shows the spectrum variation from August 20 to August 28. The color denotes the value of 10·log10 (amp.), and the "amp" denotes the spectrum amplitude. The lower figure shows the Hx component time series and the *Dst* index. This figure shows that the amplitude between approximately 1 second and 1,000 seconds increases dramatically and is correlated with the geomagnetic storm around August 26. The high-frequency (< 1 Hz) amplitude does not correlate with the geomagnetic storm.

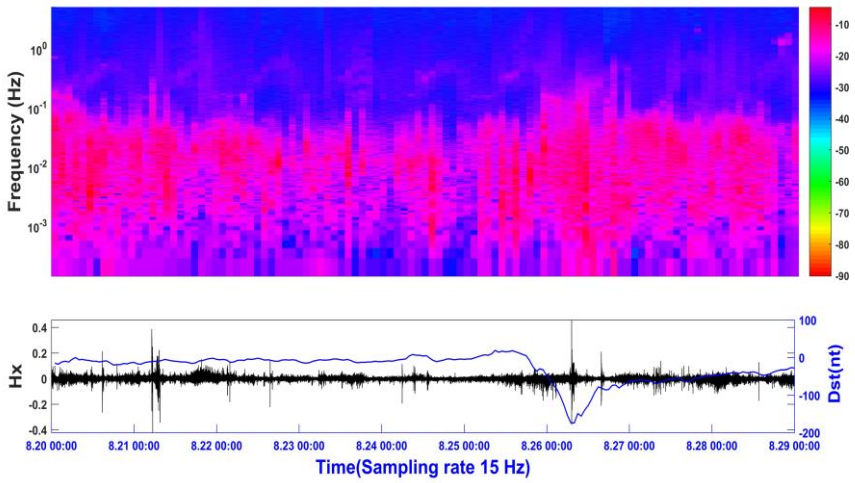

Fig. 2.5 The time-frequency distribution against the *Dst* index variation and the sampling rate is 15 Hz. The upper figure shows the time-frequency distribution from August 20 to 29. The color denotes the value of 10·log10 (amp.). The lower figure shows the Hx component's time variations and the *Dst* index. The unit of Hx is nT. The horizontal axis denotes the date.

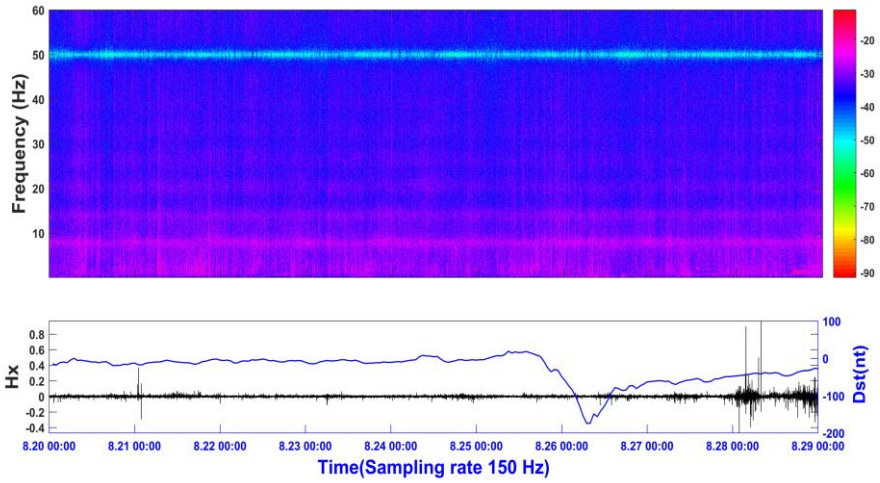

Fig. 2.6 The time-frequency distribution against the *Dst* index. The sampling rate is 150 Hz. The content is the same as Fig. 2.5.

Fig. 2.6 shows the time-frequency distribution against the *Dst* index in the mid-frequency band. This figure had no obvious changes in the intensity correlated with the storms. The result confirmed that the interaction between the solar wind and magnetosphere doesn't contribute to the high band frequency. The

signal strength is extremely low because the 50 Hz signal is filtered out during fieldwork. On the other hand, distinct peaks appeared at approximately 7.83, 14.3, 20.8 and 27.3 Hz. These frequencies correspond to the frequencies of Schumann's resonances (SRs).

Another example shows the signal strength variation corresponding to the geomagnetic storm. Fig. 2.7 shows the geomagnetic field's X (N-S) component during the storm and non-storm days at Japan's Kakioka (KAK) station. In 1973, the KAK Magnetic Observatory was designated as one of four facilities to calculate the disturbance storm time (*Dst*) index, representing the strength of the equatorial ring current encircling the Earth. The magnetic field intensity observed during the storm day can be almost two orders stronger than during a non-storm day.

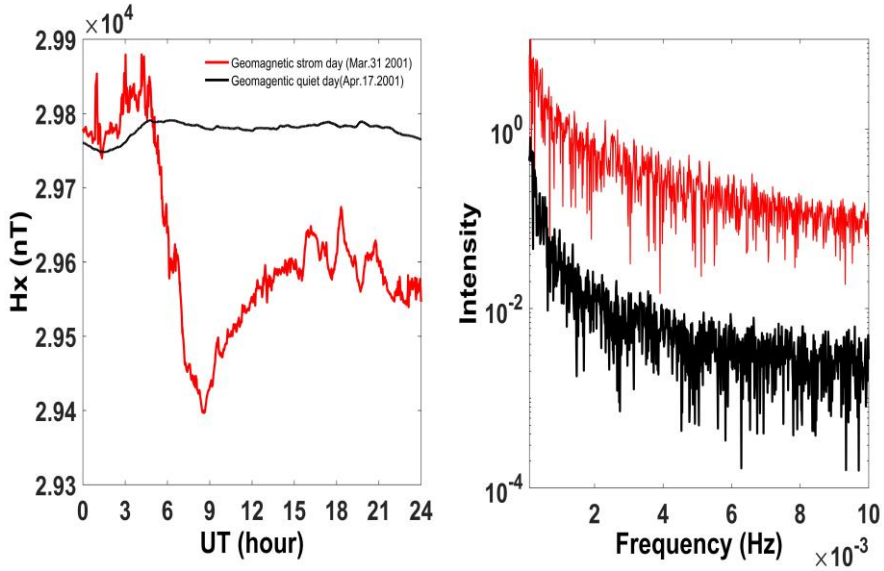

Fig. 2.7 The geomagnetic intensities along the N-S direction during the storm and non-storm days. The black lines denote the non-storm day's data, and the red lines denote the storm day's data. The left is a profile in the time domain, and the right is in the frequency domain.

Next, I introduce the statistical analysis of the geomagnetic storm based on the *Dst* index. Fig. 2.8 shows the distribution of the Dst index from 1957 to 2020; the orange line denotes the boundary of the geomagnetic storm (Dst <= -50 nT), and the light blue line denotes the boundary of the strong geomagnetic storm (*Dst* <= -100 nT). It shows that geomagnetic storms did not appear frequently, and the probability of a strong storm is less than 1% per day.

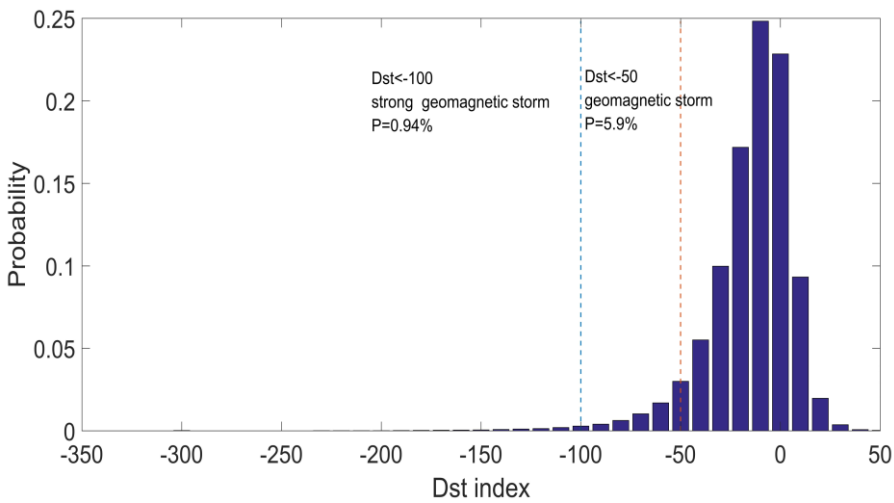

Fig. 2.8 The distribution of strong storms based on the *Dst* index between 1957 - 2020, the orange line denotes *Dst* (<= -50 nT), and the light blue line denotes *Dst* (<= -100 nT).

Fig. 2.9 shows the monthly cumulative count of strong geomagnetic storms. One hour was recorded as one count in this figure. For example, a 3-hour storm is counted as three storms. The high probability of a strong geomagnetic storm occurred around April and October.

Fig. 2.10 shows the yearly cumulative count of geomagnetic storms each year. Fig. 2.11 shows the FFT result of the yearly count of storms from 1957 to 2020. There is a 10.7-year peak, which corresponds to the 11-year solar cycle.

This section concludes that the geomagnetic storm has a seasonal and 11-year solar cycle. The strong geomagnetic storm doesn't happen frequently and causes significant EM field variations observed on the Earth's surface.

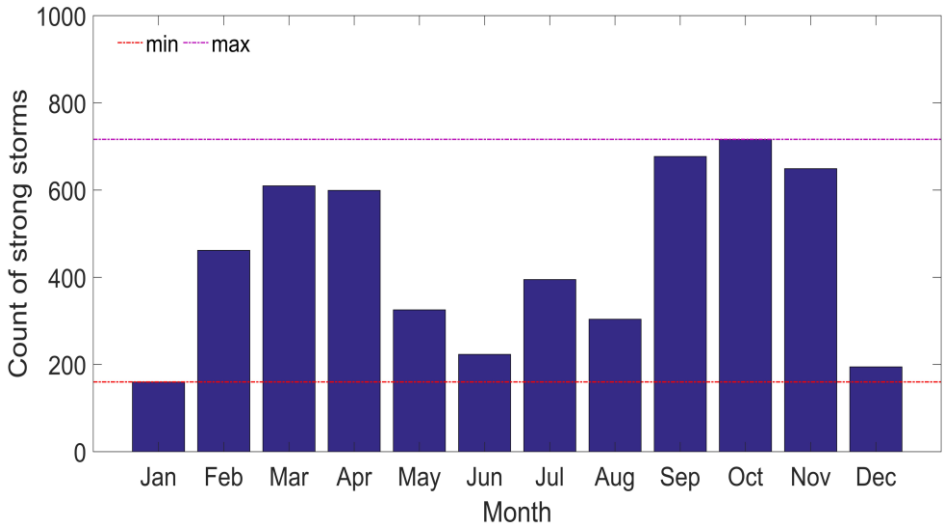

Fig. 2.9 The monthly cumulative count of strong geomagnetic storms based on the *Dst* index from 1957 to 2020.

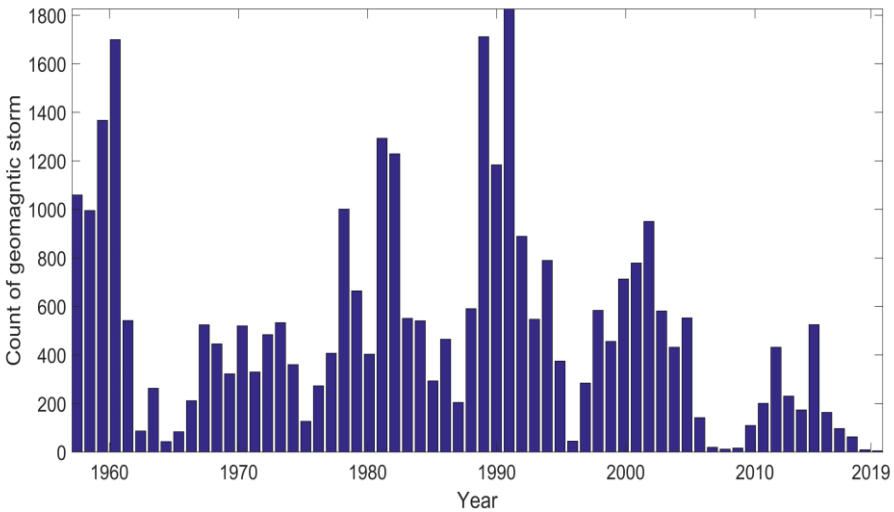

Fig. 2.10 The yearly cumulative count of geomagnetic storms based on the *Dst* index from 1957 to 2020.

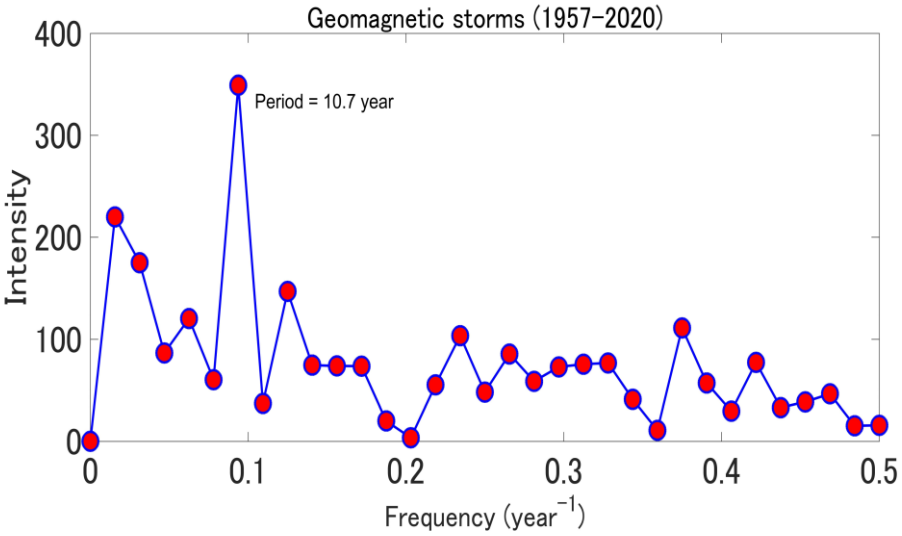

Fig. 2.11 The calculated periods by Fourier analysis using the yearly count of geomagnetic storms from 1957 to 2020.

2.3 Noise for MT method

Szarka (1988) and Junge (1996) summarized the active and passive noise sources observed in MT measurements. Most of the artificial noise is caused by earthing currents originating from electrical equipment. The most serious noise is electrified railways. Table 2.2 summarizes noise sources observed in MT measurements (Szarka,1988). The artificial EM noise can be categorized as active and passive, depending on its source. Active noise is created by sources that produce parasitic EM fields on the Earth, including power stations and power lines, electric trains, corrosion-protection systems in metal pipelines, electric fences, electric pumps, and electrical networks in mining areas. Passive noise can be attributed to technical constructions like pipelines or ditches, which are superficial resistivity inhomogeneities.

Table 2.2 Manmade EM noise in geophysical measurements, taken from Szarka (1988).

Passive	Active	Other local man-made effects
(affect as compact or elongated superficial resistivity inhomogeneities)	(produce regular or irregular parasitic EM field in the earth)	
• conductive: – power lines – pipelines – telephone lines – cable systems – fences, etc. • resistive: – roads – ditches, etc.	• electric power transmission lines • rectifiers • DC traffic substations • AC railway lines • arc furnaces • complicated mining electric networks • anti-corrosion systems • EM wave transmitters, etc.	• magnetic or EM effects of vehicles • EM induction due to vibrations of man-made origin • piezomagnetic anomaly due to artificial ground loading, etc.
• interaction with natural phenomena (magnetic storms, lightning, wind and ionisation of the air, etc.)		

Many noise signals can be attributed to the near-field of a grounded electric dipole. Zonge & Hughes (1987) describe the EM fields of a horizontal electric dipole grounded in a

homogeneous halfspace (Fig. 2.12). It can be divided into three parts depending on the separation distance *r* between transmitter and receiver: the near-field zone, the near-field/far-field transition zone, and the far-field zone.

When the separation distance r is smaller than the depth of penetration (r_A<<δ). Where the $\delta = \sqrt{2/\omega\mu_0\sigma}$, ω denotes angular velocity and μ_0 denotes free space magnetic permeability and σ denotes the electrical conductivity of the halfspace. The electric and magnetic field varies as $1/r^2$ and $1/r^3$ respectively, both independently of frequency. The apparent resistivity curves show as 45° rises in the log scale. The phase is zero for all periods.

In the near-field/far-field transition zone ($r_A \approx \delta$). The magnetic field varies between $1/r^2$ and $1/r^3$ and the apparent resistivity depends on geometry, frequency and resistivity.

In the far-field zone ($r_A >> \delta$). The electric and magnetic field decay as $1/r^3$. The ratio ***E/H*** is independent of the separation distance *r*. The far fields of industrial noise approach the MT assumption of quasi-homogeneous fields and can be regarded as part of the MT signal.

From these considerations of a grounded electric dipole, it is clear that the correlated noise that originates from the near-field and transition zone causes distortion and scatters the apparent resistivity and phase curves.

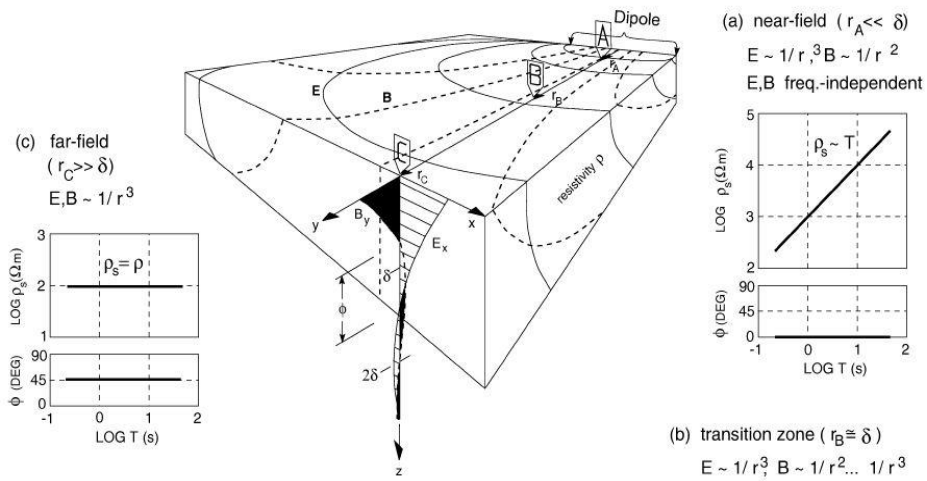

Fig. 2.12 Diagrammatic sketch of the EM field propagation in a homogeneous halfspace (modified after Zonge & Hughes 1987). MT measurement in three zones from an electric dipole, with δ= depth of penetration in a uniformly conducting halfspace: (a) Near-field (r_A<<δ); (b) Near-field/far-field transition zone ($r_A \approx \delta$); (c) Far-field (r_A >>δ).

Chapter 3: THE CONVENTIONAL MT IMPEDANCE ESTIMATION METHODS

The first step of MT data processing is to estimate the frequency-domain impedance tensor from the measured time-series data. There is a linear relationship between the geoelectric and geomagnetic fields in the frequency domain as follows:

$$\begin{pmatrix} E_x(\omega) \\ E_y(\omega) \end{pmatrix} = \begin{pmatrix} Z_{xx}(\omega) & Z_{xy}(\omega) \\ Z_{yx}(\omega) & Z_{yy}(\omega) \end{pmatrix} \begin{pmatrix} H_x(\omega) \\ H_y(\omega) \end{pmatrix}, \tag{3.1}$$

where $\boldsymbol{E}$ and $\boldsymbol{H}$ are the horizontal electric and magnetic field components at a specific frequency, respectively, ω denotes the angular frequency, and $\boldsymbol{Z}$ means the MT impedance. The term x denotes the north-south direction, and y denotes the east-west direction.

The MT impedance estimation is divided into two steps. The first step is to transform the time-series data into the frequency domain by windowed FFT. The second step is MT impedance estimation in the frequency domain. Fig. 3.1 show the procedure of conventional MT impedance estimators. The initial MT impedance estimator is based on the least-squares theory (Sims et al., 1971), and its result might be severely disturbed by cultural noise. Several MT impedance estimators have been proposed to reduce the influence of noise, as shown in Fig. 3.1. When most of the recording data is well behaved, those estimators can remove the noise. When the noise's content is over 50%, the robust methods may be biased, while the preselection strategy can obtain a reliable impedance.

This chapter introduces the conventional MT impedance estimation methods. Section 3.1 introduces the procedure of spectrum estimation by the windowed FFT. Section 3.2 introduces the different methods to estimate the impedance. Section 3.3 introduces the preselection strategy. Section 3.4 compares the performance of the single-site MT response function estimator. Finally, section 3.5 concludes and discusses the performance of the conventional impedance estimator.

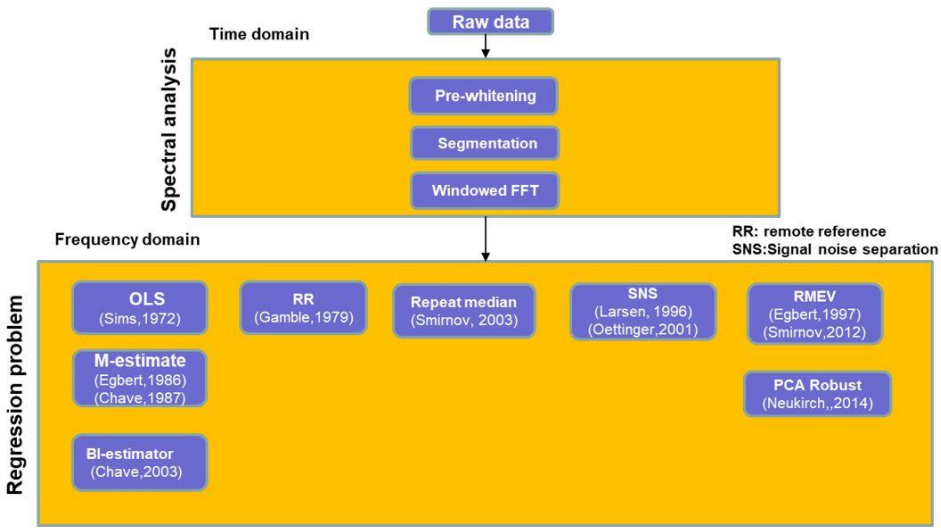

Fig. 3.1 The procedure of conventional MT impedance estimators. The time-series data is transformed into the frequency domain by windowed FFT. Several methods have been proposed for MT impedance estimation.

3.1 Spectrum Estimation

The cascade decimation proposed by Wight and Bostick (1980) was widely used to calculate the spectrum for different frequencies. This scheme has a significant advantage: it can compute the

spectrum fast and produce an evenly spaced spectrum on a log scale. When I perform the Fourier transform, there is

$$f=n*fs/N, \tag{3.2}$$

where the f is the frequency corresponding to the nth complex coefficient, n is the nth complex coefficient within each segment, the fs denotes the sampling rate, and the N denotes the segment's length. The cascade decimation method keeps the same segment's length N and changes the sampling rate (fs) to get the different frequencies' complex coefficients. It divides the time-series data into segments with a specific window length (e.g., 32, 128, 256 or 512). The 50% overlapping rate is usually used; I can consider the segments independent and get more samples. The nth (e.g., 5th and 8th) complex coefficients are grouped to calculate the impedance tensor. The original time-series data are then passed through a low pass filter and downsampled by two. The sampling frequency is reduced by two whenever the level raises one stage. The downsampled time-series data is divided into segments with the same window length, and the same n^{th} complex coefficient within each segment is extracted to calculate the impedance tensor; in this way, the frequency f is down by two times than that in the last level. This procedure is repeated to calculate the impedance in different frequencies.

The success of any MT response estimator depends on the fidelity of the spectra analysis. For the real MT time series, e_x, e_y, h_x, h_y, h_z, the electric field unit is transformed into mv/km; the magnetic field unit is transformed into nT. For each level, the procedure used to produce the accurate spectrum for each segment is as follows:

(1) Use the first difference filter to remove the trends and means; in the case of e_x, this process yields the following new series.

$$\widetilde{e_x}(k) = e_x(k+1) - e_x(k). \tag{3.3}$$

(2) Adopt a multitaper method to reduce the bias in the spectral estimation. A multitaper process is an extension of single-taper approaches. The multitaper method can be carried out by the MATLAB intrinsic function "dpss.m". This study specifies the half-bandwidth product as 2.5 and the multitaper value as 4. These specifications produce four orthogonal discrete prolate spheroidal (Slepian) sequence windows and a column weight vector (lambda). The 4 Slepian sequences are tapered to each segment, and four time-series sets ($\widehat{e_x}$ are obtained.

(3) Carry out the Fourier transform for each $\widehat{e_x}$, and calculate the Fourier coefficients E_x using the weighted mean of the four Fourier coefficient sets by multiplying them by the lambda value. And then, the instrument response is corrected in the frequency domain. Perform the same procedure for the other components (E_y, H_x, H_y, H_z). Group the 5th and 8th complex coefficients within each spectrum to perform further processing at each level.

3.2 Response Function Estimation

Several methods have been proposed to estimate the MT response function since the 1970s. MT data processing aims to discriminate signals and minimize the noise in estimating MT transfer function tensor. The first MT processing technique was developed based on the least-squares method (e.g., Sims et al. 1971). It assumes that the input channels (magnetic fields) are noise-free, the noise in the output channel has a Gaussian distribution, and the MT time series is stationary (Banks, 1998). Unfortunately, the MT data seldom satisfy the statistical assumptions in the least square method, and it can be severely disturbed by cultural noise. Through the decades, there have been two main dramatic improvements. The first is the remote reference technique by setting up a second MT remote station (Gamble et al., 1979). The second is robust algorithms (Egbert and Booker, 1986; Chave and Thomson, 1989; Smirnov, 2003; Chave and Thomson, 2003; Chave and Thomson, 2004). Combining robust and remote reference methods might generally obtain a more reliable result.

The remote reference processing can improve the estimator's performance using the cross-spectral instead of the auto-spectral when performing the regression based on the least-squares estimator. Remote reference processing can improve the MT impedance quality

when the electromagnetic noise is uncorrelated between the local and remote sites.

Robust algorithms are based on data-adaptive weighting schemes that aim at detecting and rejecting outliers from a majority of well-behaved samples. The M-estimator (Egbert and Booker, 1986) gives a small weight to reject the outlier based on the residual between the measured electric field (**E**) and the predicted electric field ($\mathbf{E_p}$), where $\mathbf{E_p}$=**ZH**, **Z** is obtained by the least-squares estimator, and **H** is the measured magnetic field. M-estimator can reduce the influence of unusual data (outliers) in the electric field but are not sensitive to exceptional input (magnetic field) data, termed leverage points. The bound influence (BI) estimator combines the robust M-estimator with leverage weighting based on the hat matrix diagonal element (Chave and Thomson, 2003; Chave and Thomson, 2004). Another single-site robust algorithm is the repeated median (RM) algorithm. The RM estimator can protect against unusual data (outlier and leverage point) maximum.

At the same time, MT researchers proposed multisite processing. Larsen et al. (1996) and Oettinger et al. (2001) proposed the signal-noise separation method (SNS). SNS used the remote magnetic field to estimate the interstation transform function as the separation tensor; they separated the local magnetic field into the signal and noise parts and then calculated the impedance. Egbert (1997) proposed a robust multivariate errors-in-variables model to separate the field data into the signal and noise using the principal component analysis. A more recent application of the method is shown in Smirnov and Egbert (2012) and Hering (2019). It is widely used in signal and noise diagnosis.

The measure of an estimator's robustness is its breakdown point. An estimator's breakdown point is the proportion of incorrect observations that an estimator can handle before giving an incorrect result. For example, for estimating the distribution center, the mean's breakdown point is 0 because even one bad observation can change the final result. The median method's breakdown point is 50 %, which can get a reliable result when the outlier is no more than half. The different method has different breakdown point in different situations.

This section focuses on reviewing the traditional MT impedance estimator.

3.2.1 Least-Square estimator

The least-squares method finds its optimum when the sum of squared residuals is minimum. The residual is obtained by the difference between the observed value (**E**), and the value predicted ($\boldsymbol{E_p}$) by the model. For the north-south component, there is the following relationship between the electric and magnetic field:

$$E_x^i = Z_{xx} * H_x^i + Z_{xy} * H_y^i, \tag{3.4}$$

where Z_{xx} and Z_{xy} denote the frequency-dependent impedance tensor, which does not depend on time. E_x^i, H_x^i, and H_y^i are the ith electric and magnetic spectra. The ith predict value E_{xp}^i is computed as follows:

$$E_{xp}^i = Z_{xx} * H_x^i + Z_{xy} * H_y^i, \tag{3.5}$$

the residual between $\mathrm{E_{xp}^i}$ and observe value E_x^i can be derived as follows:

$$\sigma = \sum_{n=1}^{N} |\, E_x^i - E_{xp}^i|^2 = \sum_{n=1}^{N} (\, E_x^i - E_{xp}^i)(E_x^i - E_{xp}^i)^{\dagger} \tag{3.6}$$

Put Eq. (3.5) into Eq. (3.6); I can get:

$$\sigma = \sum_{n=1}^{N} (\, E_x^i - (Z_{xx} * H_x^i + Z_{xy} * H_y^i))(E_x^i - (Z_{xx} * H_x^i + Z_{xy} * H_y^i))^{\dagger}, \tag{3.7}$$

where the superscript † denotes the complex conjugate, and then we can get the optimum solution of Z_{xx} and Z_{xy} when minimizing σ, and there is the following relationship :

$$\frac{\partial \sigma}{\partial Z_{xy}} = 0 \text{ and } \frac{\partial \sigma}{\partial Z_{xx}} = 0. \tag{3.8}$$

Because Z_{xx} and Z_{xy} is complex; we set the derivatives of σ for the real and imaginary parts of Z_{xx} to be zero, then we can rewrite equation (3.8) as follows:

$$\sigma = \sum_{n=1}^{N} \{ E_x^i - [(ReZ_{xx} + i * ImZ_{xx}) * H_x^i + (ReZ_{xy} + i * ImZ_{xy}] * H_y^i]\} *$$
$$\{E_x^{i\,\dagger} - [(ReZ_{xx} - i * ImZ_{xx}) * H_x^{i\,\dagger} + (ReZ_{xy} - i * ImZ_{xy}] * H_y^{i\,\dagger}\} \quad (3.9)$$

At first, the partial differential of Z_{xx} is performed.

$$\frac{\partial\sigma}{\partial ReZ_{xx}} = \sum_{n=1}^{N}\{ - H_x^i * (E_x^{i\,\dagger} - Z_{xx} * H_x^{i\,\dagger} - Z_{xy} * H_y^{i\,\dagger}) - H_x^{i\,\dagger} * (E_x^i + Z_{xx} * H_x^i + Z_{xy}H_y^i)\}, \quad (3.10)$$

$$\frac{\partial\sigma}{\partial ImZ_{xx}} = \sum_{n=1}^{N}\{ H_x^i * (E_x^{i\,\dagger} - Z_{xx} * H_x^{i\,\dagger} - Z_{xy} * H_y^{i\,\dagger}) + H_x^{i\,\dagger} * (E_x^i + Z_{xx} * H_x^i + Z_{xy} * H_y^i)\}. \quad (3.11)$$

As both Eq. 3.10 and 3.11 equal 0; finally, the following equation can be obtained:

$$\sum_{n=1}^{N}(E_x^i H_x^{i\,\dagger}) = Z_{xx} * \sum_{n=1}^{N}(H_x^i H_x^{i\,\dagger}) + Z_{xy} * \sum_{n=1}^{N}(H_y^i H_x^{i\,\dagger}) \quad (3.12)$$

Similarly, setting the derivatives of σ for the real and imaginary parts of Z_{xy} to be zero. I can get:

$$\sum_{n=1}^{N}(E_x^i H_y^{i\,\dagger}) = Z_{xx} * \sum_{n=1}^{N}(H_x^i H_y^{i\,\dagger}) + Z_{xy} * \sum_{n=1}^{N}(H_y^i H_y^{i\,\dagger}) \quad (3.13)$$

I also can calculate the residual based on the magnetic field. Similarly, I can get the following equation:

$$\sum_{n=1}^{N}(E_x^i E_x^{i\,\dagger}) = Z_{xx} * \sum_{n=1}^{N}(H_x^i E_x^{i\,\dagger}) + Z_{xy} * \sum_{n=1}^{N}(H_y^i E_x^{i\,\dagger}), \quad (3.14)$$

$$\sum_{n=1}^{N}(E_x^i E_y^{i\,\dagger}) = Z_{xx} * \sum_{n=1}^{N}(H_x^i E_y^{i\,\dagger}) + Z_{xy} * \sum_{n=1}^{N}(H_y^i E_y^{i\,\dagger}). \quad (3.15)$$

Combining Eq. 3.12 to 3.15, I can get six solutions based on the least square estimator for each impedance component.

Vozoff (1972) summarised the general solution for the impedance Z_{xx}, Z_{xy}, Z_{yx}, Z_{yy} as follows:

$$Z_{xx} = \frac{\boldsymbol{<E_xA^\dagger><H_yB^\dagger> - <E_xB^\dagger><H_yA^\dagger>}}{\boldsymbol{<H_xA^\dagger><H_yB^\dagger> - <H_xB^\dagger><H_yA^\dagger>}}, \quad (3.16)$$

$$Z_{xy} = \frac{\boldsymbol{<E_xA^\dagger><H_xB^\dagger> - <E_xB^\dagger><H_xA^\dagger>}}{\boldsymbol{<H_yA^\dagger><H_xB^\dagger> - <H_yB^\dagger><H_xA^\dagger>}}, \quad (3.17)$$

$$Z_{yx} = \frac{\boldsymbol{<E_yA^\dagger><H_yB^\dagger> - <E_yB^\dagger><H_yA^\dagger>}}{\boldsymbol{<H_xA^\dagger><H_yB^\dagger> - <H_xB^\dagger><H_yA^\dagger>}}, \quad (3.18)$$

$$Z_{yy} = \frac{\boldsymbol{<E_yA^\dagger><H_xB^\dagger> - <E_yB^\dagger><H_xA^\dagger>}}{\boldsymbol{<H_yA^\dagger><H_xB^\dagger> - <H_yB^\dagger><H_xA^\dagger>}}, \quad (3.19)$$

where **A** and **B** are $\boldsymbol{E_x, E_y, H_x, H_y}$, the superscript † denotes the complex conjugate transpose. Sims et al. (1971) discussed the effectiveness of those equations. Only 4 of those formulations are stable. The magnetic field is less noisy than the electric field in the field data. The impedance is calculated as follows:

$$Z_{xx} = \frac{\boldsymbol{<E_xH_y^\dagger><H_yH_x^\dagger> - <E_xH_x^\dagger><H_yH_y^\dagger>}}{\boldsymbol{<H_xH_y^\dagger><H_yH_x^\dagger> - <H_xH_x^\dagger><H_yH_y^\dagger>}}, \quad (3.20)$$

$$Z_{xy} = \frac{\boldsymbol{<E_xH_y^\dagger><H_xH_x^\dagger> - <E_xH_x^\dagger><H_xH_y^\dagger>}}{\boldsymbol{<H_yH_y^\dagger><H_xH_x^\dagger> - <H_yH_x^\dagger><H_xH_y^\dagger>}}, \quad (3.21)$$

$$Z_{yx} = \frac{\boldsymbol{<E_yH_y^\dagger><H_yH_x^\dagger> - <E_yH_x^\dagger><H_yH_y^\dagger>}}{\boldsymbol{<H_xH_y^\dagger><H_yH_x^\dagger> - <H_xH_x^\dagger><H_yH_y^\dagger>}}, \quad (3.22)$$

$$Z_{yy} = \frac{\boldsymbol{<E_yH_y^\dagger><H_xH_x^\dagger> - <E_yH_x^\dagger><H_xH_y^\dagger>}}{\boldsymbol{<H_yH_y^\dagger><H_xH_x^\dagger> - <H_yH_x^\dagger><H_xH_y^\dagger>}}, \quad (3.23)$$

where the brackets represent the averages of N individual auto-power and cross-power spectra. For example,

$$\boldsymbol{< H_xH_y^\dagger >} = \frac{1}{N} \sum_{j=1}^{N} H_{x\_j} H_{y\_j}^\dagger, \quad (3.24)$$

where H_{x_j} and H_{y_j} are the jth magnetic field spectrum.

The LS method can yield reliable results when the residuals are normally distributed and the magnetic field is unaffected by noise. In practice, this is rarely the case.

3.2.2 Remote reference processing

The field data consist of natural sources and local noise. The cross-power spectra are calculated as follows:

$$<\mathbf{A}\bar{\mathbf{B}}> = <(\mathbf{A}^{\mathbf{S}} + \mathbf{A}^{\mathbf{N}})(\bar{\mathbf{B}}^{\mathbf{S}} + \bar{\mathbf{B}}^{\mathbf{N}})>, \tag{3.25}$$

where $\mathbf{A}^{\mathbf{S}}$ is the signal in components $\mathbf{A}$ and $\mathbf{A}^{\mathbf{N}}$ is the noise component in $\mathbf{A}$. The same notation is used for $\mathbf{B}$. The overbar denotes the complex conjugate transpose. The cross-power spectra are calculated as follows:

$$<\mathbf{A}\bar{\mathbf{B}}>= <\mathbf{A}^{\mathbf{S}}\bar{\mathbf{B}}^{\mathbf{S}}>+<\mathbf{A}^{\mathbf{S}}\bar{\mathbf{B}}^{\mathbf{N}}>+<\mathbf{A}^{\mathbf{N}}\bar{\mathbf{B}}^{\mathbf{S}}>+<\mathbf{A}^{\mathbf{N}}\bar{\mathbf{B}}^{\mathbf{N}}>. \tag{3.26}$$

Under the assumption that the noise is uncorrelated with the signal and noise, the average value of the cross-power spectra may be neglected. Thus, the following equations can be expressed:

$$<\mathbf{A}^{\mathbf{S}}\,\bar{\mathbf{B}}^{\mathbf{N}}>=0, \quad <\mathbf{A}^{\mathbf{N}}\,\bar{\mathbf{B}}^{\mathbf{S}}>=0, \quad <\mathbf{A}^{\mathbf{N}}\,\bar{\mathbf{B}}^{\mathbf{N}}>=0. \tag{3.27}$$

The average value of the cross-power spectra yields a good approximation to the true value.

$$<\mathbf{A}\bar{\mathbf{B}}>= <\mathbf{A}^{\mathbf{S}}\bar{\mathbf{B}}^{\mathbf{S}}>. \tag{3.28}$$

I perform a similar analysis to the auto-power spectra; it is calculated as follows:

$$<\mathbf{A}\bar{\mathbf{A}}>= <\mathbf{A}^{\mathbf{S}}\bar{\mathbf{A}}^{\mathbf{S}}>+<\mathbf{A}^{\mathbf{N}}\bar{\mathbf{A}}^{\mathbf{N}}>. \tag{3.29}$$

The noise will bias the auto-power spectra a lot. The purpose of remote reference processing is to use cross-power spectra instead of auto-power spectra. The natural magnetic fields are highly correlated over distances up to several hundred kilometers, whereas the noise in the magnetic field is assumed to be uncorrelated between the local and the remote sites (Gamble et al., 1979). The electric fields can be used as a remote reference; however, electric channels are usually noisier than magnetic channels. Therefore, the magnetic field is preferred as the remote component. The impedance can be calculated as follows:

$$Z_{xx} = \frac{<E_x H_{yr}{}^{\dagger}><H_y H_{xr}{}^{\dagger}> - <E_x H_{xr}{}^{\dagger}><H_y H_{yr}{}^{\dagger}>}{<H_x H_{yr}{}^{\dagger}><H_y H_{xr}{}^{\dagger}> - <H_x H_{xr}{}^{\dagger}><H_y H_{yr}{}^{\dagger}>}, \tag{3.30}$$

$$Z_{xy} = \frac{<E_x H_{yr}{}^{\dagger}><H_x H_{xr}{}^{\dagger}> - <E_x H_{xr}{}^{\dagger}><H_x H_{yr}{}^{\dagger}>}{<H_y H_{yr}{}^{\dagger}><H_x H_{xr}{}^{\dagger}> - <H_y H_{xr}{}^{\dagger}><H_x H_{yr}{}^{\dagger}>}, \tag{3.31}$$

$$Z_{yx} = \frac{<E_y H_{yr}{}^{\dagger}><H_y H_{xr}{}^{\dagger}> - <E_y H_{xr}{}^{\dagger}><H_y H_{yr}{}^{\dagger}>}{<H_x H_{yr}{}^{\dagger}><H_y H_{xr}{}^{\dagger}> - <H_x H_{xr}{}^{\dagger}><H_y H_{yr}{}^{\dagger}>}, \tag{3.32}$$

$$Z_{yy} = \frac{<E_y H_{yr}{}^{\dagger}><H_x H_{xr}{}^{\dagger}> - <E_y H_{xr}{}^{\dagger}><H_x H_{yr}{}^{\dagger}>}{<H_y H_{yr}{}^{\dagger}><H_x H_{xr}{}^{\dagger}> - <H_y H_{xr}{}^{\dagger}><H_x H_{yr}{}^{\dagger}>}. \tag{3.33}$$

3.2.3 M-estimator

The standard robust regression is M-estimator. The scheme is an iteratively reweighted least squares regression with a Huber weight function to reduce unusual data's (outlier) influence. The robust M-estimator is first introduced to MT impedance estimation by Egbert and Booker (1986). It gives a small weight to reject the outlier based on the residual between the measured electric field ($\boldsymbol{E}$) and the predicted electric field ($\boldsymbol{E_p}$), where $\boldsymbol{E_p}=\boldsymbol{ZH}$, $\boldsymbol{Z}$ is obtained by the least-squares estimator, and $\boldsymbol{H}$ is the measured magnetic field. The solution of the M-estimator at a specific frequency is given by:

$$\boldsymbol{Z=(H^{\dagger}wH)^{-1}(H^{\dagger}\,wE)}, \tag{3.34}$$

here the weight is based on the Huber weight function; it is the minimum of $\{1, k/|r_i|\}$. Where r_i is the ith normalized residual by σ, here σ is the standard deviation of the errors. A common approach is to take $\sigma = MAR/0.44845$, where MAR is the median absolute residual. $MAR = median(\,|\, r_i\text{-}median(r_i)\,|\,)$. For the tuning constant k, the smaller value of k produces more resistance to outliers, but the efficiency is low if the errors are normally distributed. In general, k=1.5 produces 95% efficiency when the errors are normal.

Fig. 3.2 shows the distribution of Huber weight; the bigger the residual is, the smaller the weight is. In this way, I can reduce the influence of the outlier with the big residual. The

procedure of the M-estimator can be summarized as follows:

(1) For the given N estimates of the horizontal electric and magnetic fields at a frequency, perform the first regression by the least-squares estimator.

(2) Compute the $\sigma = MAR/0.44845$, the normalized residuals r, the tuning constant, and the initial residual squares sum by $\boldsymbol{rr}^{\dagger}$, where the symbol † means the complex conjugate transpose.

(3) Compute the weights $\boldsymbol{w}$ by Huber weight function and perform the regression by $\boldsymbol{Z}=(\boldsymbol{H}^{\dagger}\boldsymbol{wH})^{-1}(\boldsymbol{H}^{\dagger}\boldsymbol{wE})$. Then compute the squares residuals sum of $\boldsymbol{rwr}^{\dagger}$.

(4) Repeat steps 2 and 3 until the change of the squares residual sum becomes less than 1%.

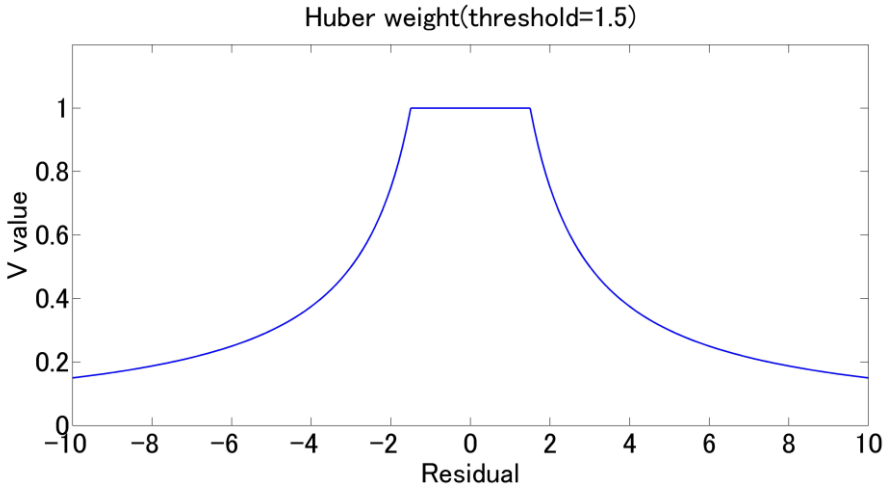

Fig. 3.2 The distribution of Huber weight, the horizontal axis denotes the normalized residual. The vertical axis denotes the wight.

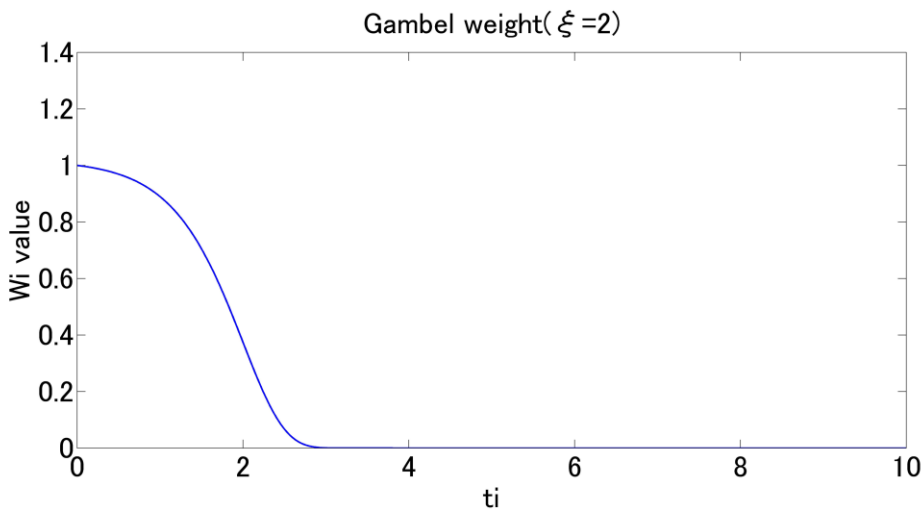

Fig. 3.3 The distribution of Gambel weight (when the tuning constant equal 2), the horizontal axis denotes the normalized residual. The vertical axis denotes the wight.

The Huber weight does not remove the outlier's influence completely. Therefore, a heavier weight proposed by Gambel is used in the last loop.

$$w_i = exp\{e^{-\xi^2}\}exp\{-e^{\xi(r^{k-1}-\xi)}\}, \tag{3.35}$$

where ξ decides the point at which down weighting begins, d is a robust estimate of the scale of the residuals, and r^{k-1} may be obtained using the residuals from the $(k-1)^{th}$ iteration. Fig. 3.3 shows the distribution of Gambel weight.

Fig. 3.4 shows a simulation to test the performance of the M-estimator. Both data sets contain an outlier P. In the upper figure, there is an outlier with a large residual. The M-estimator result almost removed the outlier's influence and recovered the true model. In contrast, there is abnormal data in the input in the lower figure, termed leverage point. The M-estimator failed to obtain the true model. The simulation shows that the M-estimator is not sensitive to the leverage point; even a single leverage point can completely dominate the estimation.

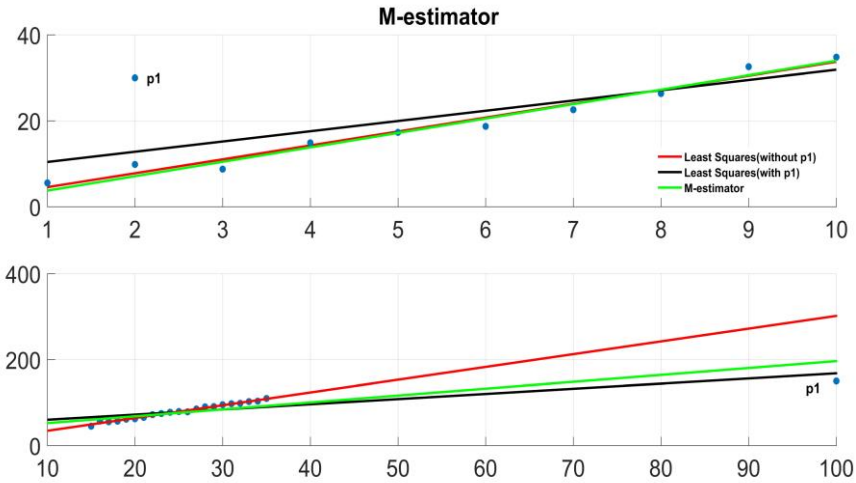

Fig. 3.4 An example of the data set includes an outlier P. Least-square result without P1 (red line). The least-square result with P1 (blue line). The M-estimator result (green line).

3.2.4 Bounded influence estimator

M-estimator can reduce the influence of unusual data (outliers) in the electric field but are not sensitive to exceptional input (magnetic field) data, termed leverage points. The hat matrix is widely used to detect abnormal large-value data. The hat matrix is *N* by *N* matrix and is defined as follows:

$$\boldsymbol{H}_{hat} = \boldsymbol{H}(\boldsymbol{H}^{\dagger}\boldsymbol{H})^{-1}\boldsymbol{H}^{\dagger}, \tag{3.36}$$

where $\boldsymbol{H}$ is the N by two matrices of the horizontal magnetic field at a specific frequency. The superscript † denotes the complex conjugate transpose. Chave and Thomson (2003) suggested that the hat matrix's diagonal element, more than several times 2/N (N denotes the sample of data), is problematic. As the noisy data is energetic, the hat matrix's diagonal element's statistical analysis is useful for detecting the noisy data.

The bounded influence (BI) estimator combines the robust M-estimator with leverage weighting ($\boldsymbol{v}$) based on the statistics of the hat matrix diagonal element (Chave et al., 2004). And the solution of BI-estimator at a specific frequency is given by:

$$\boldsymbol{Z}=(\boldsymbol{H}^{\dagger}\boldsymbol{wvH})^{-1}(\boldsymbol{H}^{\dagger}\,\boldsymbol{wvE}). \tag{3.37}$$

The detail of the BI estimator is described by Chave (2012). And the solution of remote reference based on BI-estimator at a specific frequency is given by:

$$\boldsymbol{Z}=(\boldsymbol{H}_r{}^{\dagger}\boldsymbol{wvH})^{-1}(\boldsymbol{H}_r{}^{\dagger}\boldsymbol{wvE}), \tag{3.38}$$

where $\boldsymbol{H}_r$ denotes the remote reference magnetic field data. BIRRP is an abbreviation for the Bounded Influence Remote Reference Processing method.

Fig. 3.5 compares the performance of the BI estimator and M-estimator in the presence of the leverage point. The red line denotes the result calculated by the LS method without P1; it is regarded as the true model. There is a leverage point in the dataset. The M-estimator is biased most, and the BI-estimator with Huber weight reduced the influence of the leverage point. The BI-estimator with gamble weight almost removed the influence of the leverage point and

recovered the true model. This figure showed the effectiveness of the BI estimator to the leverage point.

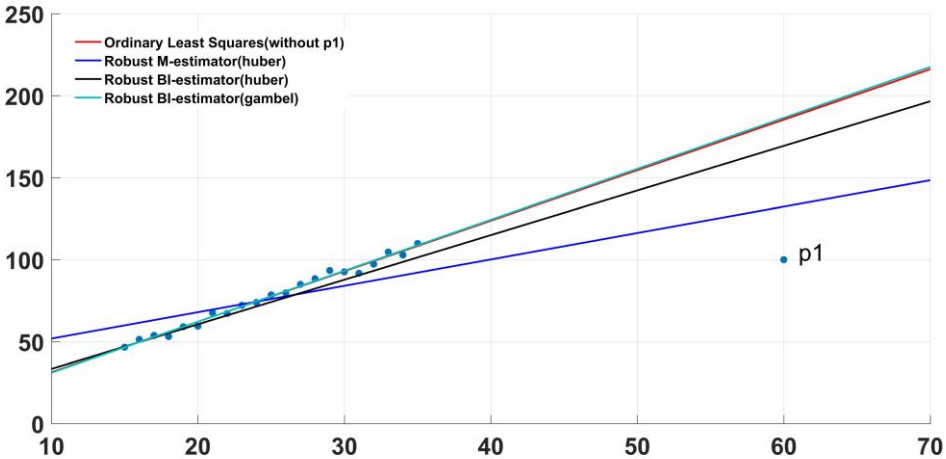

Fig. 3.5 Comparation of the BI estimator and M-estimator. The data set include a Leverage point P1. The red line denotes the result calculated by the LS method without P1. The green line denotes the result calculated by BI-estimator using the gamble weight. The black line denotes the result calculated by BI-estimator using the Huber weight. The blue line denotes the result calculated by M-estimator using Huber weight.

3.2.4 Repeated median estimator

The least-squares method can give its optimum solution if the sum of the squared residuals is minimum. The least-squares estimator (LS-estimator) is sensitive to gross errors in the data and has a breakdown point equal to zero.

M-estimator and BI-estimator are much less sensitive to outliers than the LS-estimator. Siegel's repeated median estimator has the highest possible breakdown point, equal to 50 %. The repeated median estimator will produce a reliable result when no more than half of the data is contaminated.

The repeated median estimator (RM-estimator) was proposed by Siegel (1982). Smirnov (2003) introduced RM-estimator to process the MT data. It is helpful to consider the simple linear regression model (y_i=b+ax_i+e_i), to explain the estimator in detail. The RM-estimator for the slope is then given by:

$$a= \underset{i}{med}\ \underset{k \neq i}{med} \left(\frac{y_k-y_i}{x_k-x_i}\right) \tag{3.39}$$

where *i, k=1,…, N*; when $k = i$, the inner median is skipped. N denotes the number of samples. For the solution of the slope, two points can determine a result. The inner median result for one point is the median of all the combinations with other points. The final result of the slope is the median of the first median.

In the situation of MT, if there are two datasets, I can get one result. The real and imaginary part of each impedance tensor is calculated separately. The repeat median operator is given by:

$$Re\,\{Z\} = \underset{i}{med}\ \underset{k \neq i}{med} Re\,\{(H_{ik}^{\dagger} H_{ik})^{-1}\,(H_{ik}^{\dagger} E_{ik})\}, \tag{3.40 a}$$

$$Im\,\{Z\} = \underset{i}{med}\ \underset{k \neq i}{med} Im\{(H_{ik}^{\dagger} H_{ik})^{-1}\,(H_{ik}^{\dagger} E_{ik})\}, \tag{3.40 b}$$

where the indices *i, k=1,…, N*. The inner median skips when $k = i$.

The Siegel repeat median is calculated as follows:

(1) The first median operator of Eq. (3.42) is calculated as follows:

$$S1_{Re\{Z\}_i} = \underset{k \neq i}{med} Re\{(H_{ik}^{\dagger} H_{ik})^{-1} (H_{ik}^{\dagger} E_{ik})\}; \quad (3.41\ a)$$

$$S1_{Im\{Z\}_i} = \underset{k \neq i}{med} Im\{(H_{ik}^{\dagger} H_{ik})^{-1} (H_{ik}^{\dagger} E_{ik})\}; \quad (3.41\ b)$$

The $Z_i = S1_{Re\{Z\}_i} + j*S1_{Im\{Z\}_i}$, where Z_i denotes the ith result of the inner median. The symbol j denotes the imaginary number unit.

(2) The second median operator is calculated as follows:

$$S2_{Re\{Z\}} = \underset{i}{med} S1_{Re\{Z\}_i}; \quad (3.42\ a)$$

$$S2_{Im\{Z\}} = \underset{i}{med} S1_{Im\{Z\}_i}; \quad (3.42\ b)$$

The $Z_s = S2_{Re\{Z\}} + j \cdot S2_{Im\{Z\}}$, where Z_s denotes the impedance calculated by the Siegel RM-estimator.

(3) Confidence limits are estimated from the median of absolute deviations (*MAD*):

$$Z_{MAD} = 1.483 \underset{i}{med} \{Z_i - Z_s\}. \quad (3.43)$$

This estimate of the scale parameter is straightforward to calculate and is insensitive to outliers. For Gaussian errors, a 95 % confidence limit can be defined as (error bar):

$$\Delta Z = 1.96 \cdot Z_{MAD} / \sqrt[3]{N}, \quad (3.44)$$

where *N* is the number of data.

3.2.5 Signal-noise separation

It is difficult to estimate precise MT impedance tensor with coherent noise in measured electric and magnetic time-series data. If clean remote reference data is available, it is possible to separate MT and correlated noise signals and derive unbiased MT transfer functions with the signal-noise separation method (SNS) (Larsen et al. 1996; Oettinger et al. 2001)

The field data include natural electromagnetic signals and noises. I can subdivide the noises and signals into three parts. I can rewrite the fields **E** and **H** as follow:

$$\boldsymbol{E} = \boldsymbol{E}^{MT} + \boldsymbol{E}^{CN} + \boldsymbol{E}^{IN}, \quad (3.45\ a)$$

$$\boldsymbol{H} = \boldsymbol{H}^{MT} + \boldsymbol{H}^{CN} + \boldsymbol{H}^{IN}, \quad (3.45\ b)$$

where the CN denotes the coherent noise, the IN denotes incoherent noise, and the MT denotes electromagnetic signals (Oettinger et al., 2001). Fig. 3.6 shows the flow chart of the signal-noise separation method. The first step of SNS is to separate the measured magnetic field into MT signals and noise, and I need to estimate the interstation transform function as the separation tensor T. For the north-south component: $\boldsymbol{H_x} = \boldsymbol{H_r T_x} + \boldsymbol{R_x}$; I can extend to the matrix form as follows:

$$\begin{pmatrix} H_{x\_1} \\ H_{x\_2} \\ \vdots \\ H_{x\_n} \end{pmatrix} = \begin{pmatrix} H_{xr\_1} & H_{yr\_1} \\ H_{xr\_2} & H_{yr\_2} \\ & . \\ & . \\ & . \\ H_{xr\_n} & H_{yr\_n} \end{pmatrix} \begin{pmatrix} T_{xx} \\ T_{xy} \end{pmatrix} + \begin{pmatrix} R_{x\_1} \\ R_{x\_2} \\ \vdots \\ R_{x\_n} \end{pmatrix}. \quad (3.46)$$

For homogeneous source fields and horizontally layered earth, T is unity. T is generally frequency-dependent, complex-valued and different from unity (Larsen et al.1996). The residual R_x represents the uncorrelated noise in the local magnetic variations. The separation tensor $\boldsymbol{T_x}$ can be obtained as follows:

$$\boldsymbol{T_x} = (\boldsymbol{H_r^{\dagger} H_r})^{-1} (\boldsymbol{H_r^{\dagger} H_x}), \quad (3.47)$$

where $\boldsymbol{T_x} = (T_{xx}, T_{xy})$, $\boldsymbol{H_r} = (\boldsymbol{H_{xr}}, \boldsymbol{H_{yr}})$. And then, I can estimate the $\boldsymbol{H_x^{MT}}$ by the remote

magnetic field as follows:

$$\boldsymbol{H_x^{MT}} = \boldsymbol{T_x} * \boldsymbol{Hr} \tag{3.48}$$

Thus the local magnetic field can be subdivided into two parts $\boldsymbol{H^{CU}}$ and $\boldsymbol{H^{MT}}$:

$$\boldsymbol{H^{CU}} = \boldsymbol{H} - \boldsymbol{H^{MT}} \tag{3.49}$$

Let's treat noise as the correlated noise and the uncorrelated noise as follows:

$$\boldsymbol{H^{CU}} = \boldsymbol{H^{CN}} + \boldsymbol{H^{UN}} \tag{3.50}$$

Eq. (3.47a) can be rewritten as follows:

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{E}^{MT} + \boldsymbol{E}^{CN} + \boldsymbol{E}^{UN} \\ &= \boldsymbol{Z^{MT}}\boldsymbol{H^{MT}} + \boldsymbol{Z^{CN}}\boldsymbol{H^{CN}} + \boldsymbol{E}^{UN} \\ &= \boldsymbol{Z^{MT}}\boldsymbol{H^{MT}} + \boldsymbol{Z^{CN}}\boldsymbol{H^{CN}} + \boldsymbol{Z^{CN}}\boldsymbol{H^{IN}} - Z^{CN}\boldsymbol{H^{UN}} + \boldsymbol{E}^{UN} \\ &= \boldsymbol{Z^{MT}}\boldsymbol{H^{MT}} + Z^{CN}\boldsymbol{H^{CN}} + \boldsymbol{R} \end{aligned} \tag{3.51}$$

where $\boldsymbol{R} = \boldsymbol{E}^{UN} - \boldsymbol{Z^{CN}}\boldsymbol{H^{UN}}$, finally, $\boldsymbol{Z^{MT}}$ and $\boldsymbol{Z^{CN}}$ can be obtained by the matrix as follow:

$$\begin{pmatrix} E_{x\_1} \\ E_{x\_2} \\ \vdots \\ E_{x\_n} \end{pmatrix} = \begin{pmatrix} H_{x\_1}^{MT} & H_{x\_1}^{MT} & H_{x\_1}^{CU} & H_{x\_1}^{CU} \\ H_{x\_2}^{MT} & H_{x\_2}^{MT} & H_{x\_2}^{CU} & H_{x\_2}^{CU} \\ & . & & \\ & . & & \\ & . & & \\ H_{x\_n}^{MT} & H_{x\_n}^{MT} & H_{x\_n}^{CU} & H_{x\_n}^{CU} \end{pmatrix} \begin{pmatrix} Z_{xx}^{MT} \\ Z_{xy}^{MT} \\ Z_{xx}^{CU} \\ Z_{xy}^{CU} \end{pmatrix} + \begin{pmatrix} R_{x\_1} \\ R_{x\_2} \\ \vdots \\ R_{x\_n} \end{pmatrix} \tag{3.52}$$

Analogous to the LS method, I can get:

$$\underline{\boldsymbol{Z_x}} = (\underline{\boldsymbol{H}}^{\dagger}\underline{\boldsymbol{H}})^{-1}(\underline{\boldsymbol{H}}^{\dagger}\underline{\boldsymbol{E_x}}), \tag{3.53}$$

Where the underline represents the combination. $\underline{\boldsymbol{Z_x}}$ is a tensor that combines $\boldsymbol{Z_x^{MT}}$ and $\boldsymbol{Z_x^{CU}}$, $\underline{\boldsymbol{H}}$ combines $\boldsymbol{H^{MT}}$ and $\boldsymbol{H^{CU}}$.

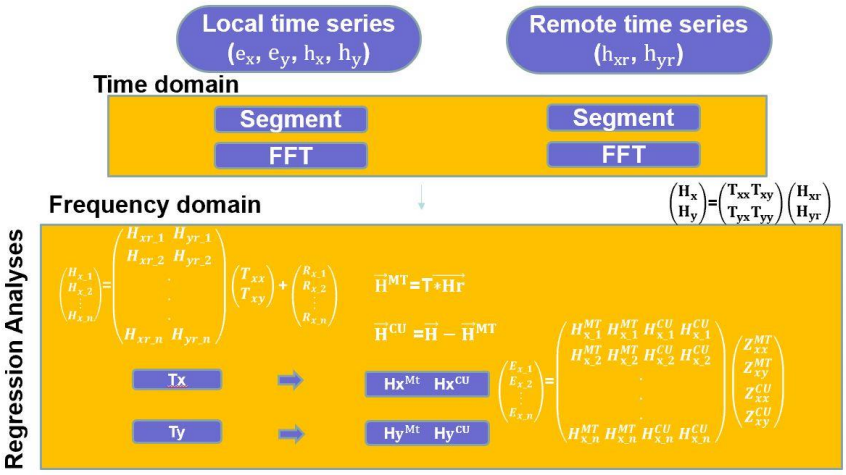

Fig 3.6 The procedure of the signal-noise separation method (SNS). The first step of SNS is to estimate the interstation transform function as the separation tensor T, separating the measured magnetic field into MT signals and noise. Finally, the impedance is calculated by the matrix of Eq. (3.52).

3.2.6 Robust multivariate errors-in-variables

Egbert and Booker (1989) proposed an errors-in-variables approach which allows for noise in the electric and magnetic fields. The processing scheme was made robust and was referred to as a robust multivariate errors-in-variables estimator (RMEV; Egbert, 1997). A more advanced method that allows the missing data are shown in Smirnov and Egbert (2012). The approach is based on the following model for an MT data matrix $\boldsymbol{X}$:

$$\boldsymbol{X_i} = \begin{pmatrix} H_{1i} \\ E_{1i} \\ \vdots \\ H_{ji} \\ E_{ji} \end{pmatrix} = \boldsymbol{U\beta_i} + \boldsymbol{V\gamma_i} + \boldsymbol{\varepsilon_i}. \tag{3.54}$$

Here E_{ji} and H_{ji} denotes Fourier coefficients of the electric and magnetic field for a specific frequency, j denotes the total number of sites, and i(=1, 2, ⋯, N) denotes the number of observations. In the case of two sites with five channels and 100 samples (Fourier coefficients) at each site, this would yield a 10 × 100 complex data matrix ($\boldsymbol{X}$). **U** refers to the matrix corresponding to the natural signal, and **V** is the matrix corresponding to the coherent noise. Following Egbert (1997), $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ define linear combination parameters of the $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. The residual matrix $\boldsymbol{\varepsilon}$ is defined by the incoherent noise.

For the linear combination of the $\boldsymbol{\beta}$ and **U**, there are relationships as follows:

$$\boldsymbol{U_{K2}\beta_{2i}} = \begin{pmatrix} \eta_{11} \\ \zeta_{11} \\ \vdots \\ \eta_{j1} \\ \zeta_{j1} \end{pmatrix} \beta_{1i} + \begin{pmatrix} \eta_{12} \\ \zeta_{12} \\ \vdots \\ \eta_{j2} \\ \zeta_{j2} \end{pmatrix} \beta_{2i}. \tag{3.55}$$

Here β_{li} defines the linearly polarized N-S (l = 1) and E-W (l =2) of the magnetic fields for the ith data segment. The **U** is K × 2 complex vectors (K=5j), which ideally represent the magnetic and electric fields observed at all sites. The MT response tensor **Z** can be derived from the individual elements of **U**(Egbert, 1997):

$$Z_j = \begin{bmatrix} \zeta_{xj1} & \zeta_{xj2} \\ \zeta_{yj1} & \zeta_{yj2} \end{bmatrix} \begin{bmatrix} \eta_{xj1} & \eta_{xj2} \\ \eta_{yj1} & \eta_{yj2} \end{bmatrix}^{-1}. \tag{3.56}$$

Here, j indicates the site number, ζ and η refer to the components of **U** related to the electric and magnetic field in x- or y- direction, respectively. Indices 1 and 2 denote the two independent source field polarizations.

In practice, natural signals and coherent noise cannot be separated easily. Hence is written as

$$X_i = \boldsymbol{W}\alpha_i + \varepsilon_i, \tag{3.57}$$

where **W** is a K × M matrix summarizing the natural signal **U** and the coherent noise **V**. This model is enough to include the true situation.

The scheme of the RMEV estimation:

1.) Estimate the initial noise level and cleaned data vectors. For this purpose, each data channel is fitted by a general data model using a robust multivariate linear regression algorithm. each data channel X_k is predicted by the remaining K-1 data channels. Regression results are then achieved by fitting the following linear model for all observations (Egbert, 1997):

$$X_{ik} = \sum_{\boldsymbol{k'} \neq k} b_{\boldsymbol{k'}} X_{i\boldsymbol{k'}} + \varepsilon_{ik}, \tag{3.58}$$

where k is the channel index, the total number of channels and the observation index.

The final noise covariance matrix (Σ_N) is calculated from the residuals of a multivariate linear regression between all cleaned data channels X_{ik}.

Finally, the residual variances $\sigma_k^2 = var(\varepsilon_{i:k})$ of all data channels can be presented by a diagonal variance matrix:

$$\boldsymbol{\Sigma_N}=\begin{pmatrix}\sigma_1^2 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_k^2\end{pmatrix}=diag(\sigma_1^2 \dots \sigma_k^2), \tag{3.59}$$

2.) The spectral density matrix **S** is obtained by calculating averaged cross- and auto-spectra from the Fourier coefficients of all (cleaned) data channels. The spectral density matrix is calculated as follows:

$$\boldsymbol{S}=\boldsymbol{X}\boldsymbol{X}^{\dagger}, \tag{3.60}$$

where the † denotes the complex conjugate transpose; for example, data measured simultaneously at two sites result in a 10×10 spectral density matrix.

3) The scaled spectral density $\mathbf{S}'$matrix is calculated by normalized **S** with $\mathbf{\Sigma_N}$:

$$\boldsymbol{S}'=\Sigma_{\boldsymbol{N}}^{-\mathbf{1/2}}\cdot\boldsymbol{S}\cdot\Sigma_{\boldsymbol{N}}^{-\mathbf{1/2}}, \tag{3.61}$$

Subsequently, the eigenvalue problem is solved as follows:

$$\boldsymbol{S}'(\boldsymbol{u})=\boldsymbol{\lambda u}, \tag{3.62}$$

the solution to the eigenvalue problem will provide information about the coherence dimension of the entire data space. In this transformed data space, incoherent noise results in an anisotropic scatter of unit variance, and only directions corresponding to coherent parts of the data will show variances significantly greater than one (Egbert, 1997). The number of eigenvalues λ significantly larger than unity will be two if the quasi-uniform plane wave source field solely represents the signal. In contrast, coherent noise sources will increase the number of dominating eigenvalues. The resulting eigenvalues are non-dimensional and can be interpreted as signal-to-noise ratios (Egbert, 1997).

4) The eigenvectors corresponding to the two largest eigenvalues are scaled back to the original data space(Hering, 2019):

$$\boldsymbol{U}=\Sigma_{\boldsymbol{N}}^{\mathbf{1/2}}\cdot\boldsymbol{u}'\cdot(\boldsymbol{u}'\cdot\Sigma_{\boldsymbol{N}}^{-\mathbf{1}}\cdot\boldsymbol{u}'), \tag{3.63}$$

The matrix $\mathbf{u}'$ consists of the two eigenvectors ($\mathbf{u_1}$ and $\mathbf{u_2}$), and the resulting matrix **U** corresponds to the natural MT source field. Suppose the coherence dimension does not exceed two. $\boldsymbol{W}$ equals the natural source field, and $\boldsymbol{U}$ can be calculated from the eigenvectors corresponding to the two largest eigenvalues. The impedance can be solved by Eq. (3.56).

3.2.7 Error estimation

The jackknife and bootstrap are resampling techniques especially useful for variance and bias estimation (https://en.wikipedia.org/wiki/Resampling_(statistics)).

The jackknife estimator systematically leaves out each observation from a dataset. Let $\hat{\theta}=f(x_1,\cdots,x_n)$ be an estimator of θ, I define the ith leave-one-out statistic as $\theta_{-i}=f(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)$. Then the jackknife estimator for the bias is calculated as follows:

$$bias_{jack}=(n-1)(\bar{\theta}_{-i}-\hat{\theta}), \tag{3. 64}$$

where $\bar{\theta}_{-i}=1/n*\sum_i\theta_{-i}$.

Bootstrapping is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample. This technique allows estimation of the sampling distribution of almost any statistic.

The bootstrap and the jackknife estimates yield similar numerical results, so each can be seen as an approximation.

3.3 Preselection Strategy

The method introduced in section 3.2 may remove the influence of noise when most data is well behaved. The solution to the problem, in which the noise content is over 50%, is to screen the high single-to-noise ratio data before the impedance estimation to the level that the robust estimator can handle.

Weckmann et al. (2005) proposed a preselection strategy. They combined several parameters

such as spectral power densities, the multiple (Gamble et al., 1979) and partial coherences, the polarization directions of the electromagnetic fields, the responses and their uncertainties in their preselection strategy. The noise can be screened before the impedance estimation in some specific situations. However, the data preselection approach proposed by Weckmann et al. (2005) is strongly dependent on the local noise condition and can not be performed automatically.

Platz and Weckmann (2019) created an automatic preselection strategy using the Mahalanobis distance. However, like all purely statistical algorithms, the criterion is limited to cases where the noise content is below 50%. If most data are contaminated, they used additional information on polarization direction (Weckmann et al., 2005) to constrain the noise. However, the polarization direction is not effective in detecting the incoherent noise, and it may fail in the presence of a large amount of incoherent noise.

The most serious situation is that the continuous noise contaminates the data; any method may fail to get a reliable result.

3.4 Comparison of the Single-Site Robust Impedance Estimators

This section tries to assess the three single-site MT impedance estimators, and I used the BIRRP code(Chave and Thomson, 2003), the M-estimator and the RM-estimator created by ourselves to test the performance of the three estimators.

3.4.1 Comparison using the synthetic data

I used the synthetic data with a known impedance Z to test the code. The way to create synthetic data is similar to the method proposed by Chen (2012). I used the 1-day magnetic data from Memambetsu (MMB) station as the synthetic magnetic time series data. MMB is one of the Magnetic Observatory stations where geomagnetic and geoelectric observations are performed in Japan. The sampling rate of synthetic data is 1 second, and its unit is nT. First, the impedances Z(ω) are calculated from our simple 1-D model and then transform the magnetic field time-series (h(t)) into the frequency domain by using the Fourier transform. Then the impedance Z(ω) is multiplied with H(ω) to determine E(ω). Subsequently, the electric field spectra are transformed into the time domain to obtain the electric time series e(t). Fig. 3.7 compares the real model's impedance curves and the result calculated by the BIRRP code, the M-estimator and the RM-estimator. All of them coincide with each other. It indicates that all of our code works well.

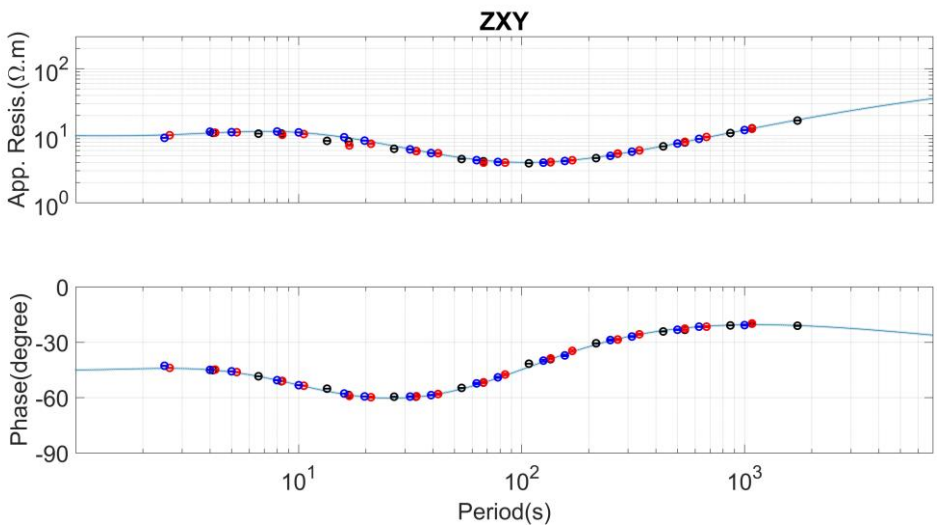

Fig. 3.7 Comparison of the impedance curves between the true model and the result calculated

from the time series. The blue line is the true curve of a simple 1-D model. Next, blue circles were calculated using the BI-estimator, and red circles were calculated using the M-estimator. Finally, black circles were calculated using the RM estimator.

3.4.2 Application to the field data

The leverage weighting is the most significant difference between the M-estimator and the BI-estimator. The leverage point corresponds to the large variation of the magnetic field. It can be an energetic signal or active noise. I used two datasets to show the two different situations. In the first dataset, the signal has a large variation when there is a geomagnetic storm, and in the second dataset, the noise has a large variation.

3.4.2.1 Comparison of the performance using the data from KAP03

The Kaapvaal 2003 (KAP03) long-period 5-component magnetotelluric time series data is used in the first dataset. Data were recorded for around a month at each site using GSC LIMS systems in 2003 as a part of the SAMTEX project. The sampling period is 5-second. The 26 long-period sites distributed in a NE-SW profile are shown in Fig. 3.8. In the middle of the survey line (KAP127-KAP145), datasets were heavily contaminated by DC noises from the power line of trains running between Kimberley and Johannesburg. In addition, there are two geomagnetic storm events in the observation periods. The first storm occurred around Oct. 29, 2003, to Oct. 31, 2003; the second storm occurred around November 20, 2003, to November 22, 2003. Both of the storms lasted about two days. I used the different period time-series datasets at noisy sites 130 and 133 to analyze the three estimators' performance; the results from site 130 are shown first.

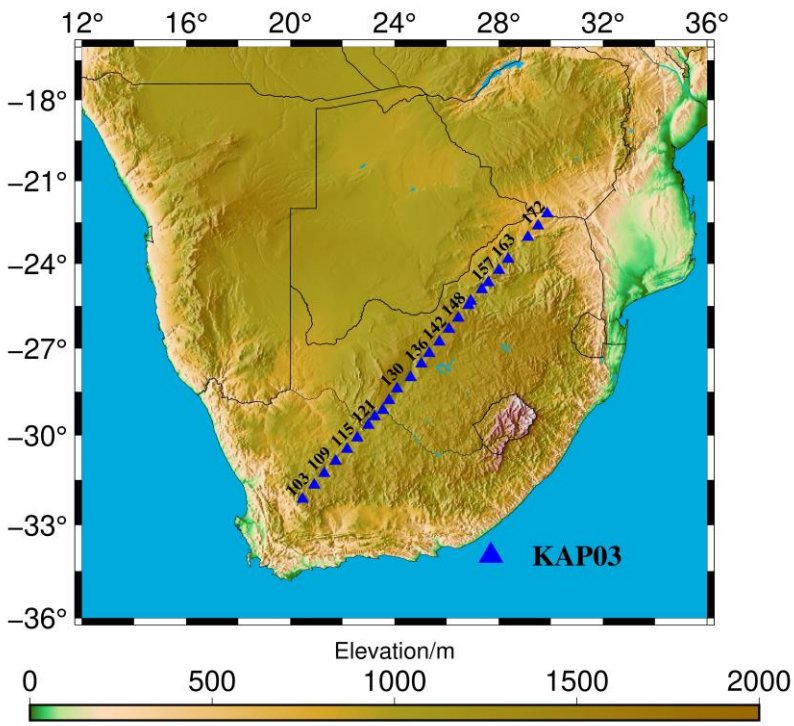

Fig. 3.8 Location map of the survey line KAP03 in South Africa.

Fig. 3.9 shows the MT time-series data observed at site 130. The red vertical lines indicate the data gaps, and the black lines indicate the 5-component MT time-series data. The sampling rate is 5-second. The electric field unit is mV/km, and the unit of the magnetic field is nT. The

blue line shows the disturbance storm time(*Dst*) index.

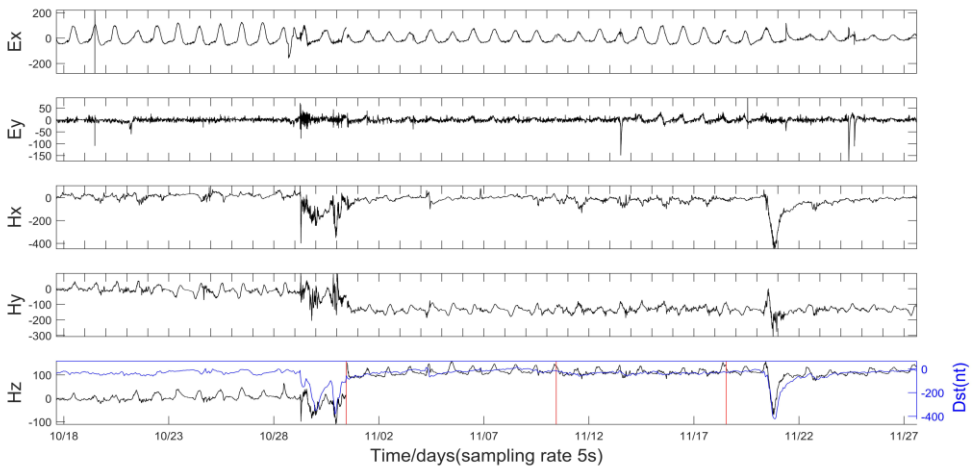

Fig. 3.9 Time series of magnetotelluric data whose sampling period is 5 seconds at site 130. The red vertical lines show the missing data section, and the black lines show the 5-component MT time series data. The blue line shows the *Dst* index of magnetic data. The electric field unit is mV/km, and the unit of the magnetic field is nT.

Fig. 3.10 shows the YX components of magnetotelluric sounding curves calculated during storm day and the non-storm day at site 130. The first column is calculated in the non-storm day data from 00:00:00 on Oct. 22 to 00:00:00 on Oct. 25. The second column is computed in the first storm from 00:00:00 on October 29 to 00:00:00 on October 31. The third column is calculated in the second storm data from 00:00:00 on Nov. 20 to 00:00:00 on Nov. 22. M-estimator calculates the red curve, and BI-estimator calculates the black curve.

When using the non-storm data, both methods failed to obtain a reliable result. The phase was close to 180º or 0º. That is the phenomenon of artificial noise (Zonge and Hughes, 1987); 180º or 0º would correspond to a dipole electric source, which could be the train line. And the long period of the XY component is very scattered. Comparing the results, I find that the YX component can be improved using storm period data. In Fig. 3.10, the first storm period data calculates the second column. Both curves are similar to each other. After comparing the result with the remote reference result, I think the result calculated by the first geomagnetic storm is the true curve. The third column is calculated by the second storm period from 00:00:00 on Nov. 20 to 00:00:00 on Nov. 22. M-estimator result was considered close to the true curve compared to the result derived from the first geomagnetic storm. However, the BI-estimator's result differs from the M-estimator's, and the phase is also close to 180º. In general, the BI-estimator performs better than the M-estimator. But in this case, it is not.

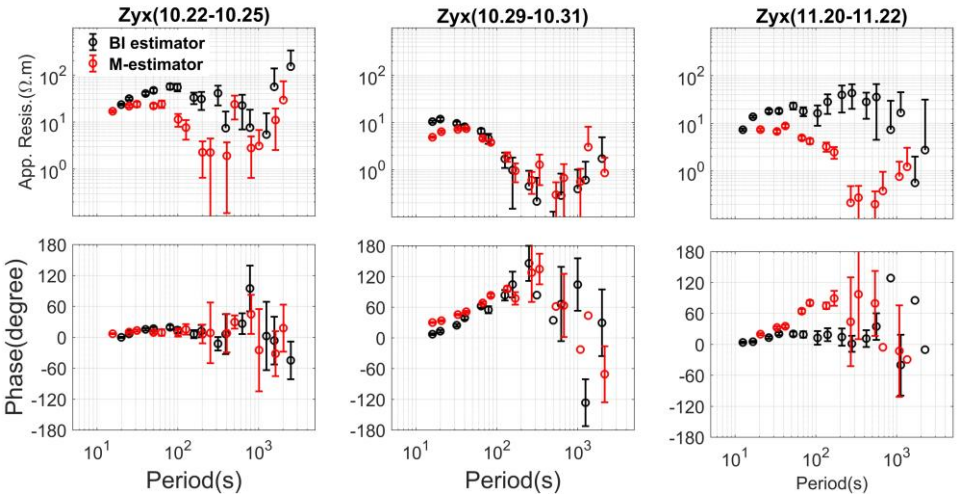

Fig. 3.10 The YX component of magnetotelluric sounding curves calculated during both storm day and the non-storm day at site 130. The M-estimator calculated the red curves, and the BI-estimator calculated the black curves. The left figures were calculated using the non-storm data, and the center figures were calculated using the second storm data. The right figures were calculated using the third storm data. The upper figures show the apparent resistivity curves, and the lower figures show the impedance phase curves.

Next, I introduced some non-storm noisy data to the first geomagnetic storm dataset to compare BI-estimator, M-estimator, and RM-estimator's performances. Fig. 3.11 shows the YX component of magnetotelluric sounding curves. The blue curves were calculated using the dataset from Oct. 27 to Oct. 31. The black curves were calculated using the dataset from Oct. 28 to Oct. 31. The red curves were calculated using the dataset from Oct. 29 to Oct. 31. The three datasets contain the geomagnetic storm data from Oct. 29 to Oct. 31. The BI results get biased with the longer non-storm data introduced. But M-estimator results almost don't change and are reliable. I think two factors lead to the results.

The most significant difference between the BI-estimator and M-estimator is the leverage weighting. I investigated the hat matrix diagonal element value situation, as shown in Fig. 3.12. The leverage points correspond to the geomagnetic storm data. The expectation value of the hat matrix diagonal element is 2/N, and N denotes the number of the data; the more the data is introduced, the smaller the expectation value is. That also means the more the leverage point will be detected. In this case, all leverage points correspond to the geomagnetic event. I need the geomagnetic storm period data to get a reliable result for the noisy site, but BI-estimator will strictly reject this data. That is the first factor. The other factor is that the more the non-storm data is introduced, the more noisy data is. The BI-estimator will deviate further from the true curve with the more non-storm data introduced by the two factors.

On the other hand, the M-estimator is not sensitive to the leverage point. As the solution of least square is based on averaging of the auto and cross power density spectra, for the north-south direction, the function can be written as follow:

$$< E_x H_x{}^{\dagger} >= Z_{xx} < H_x H_x{}^{\dagger} > + Z_{xy} < H_y H_x{}^{\dagger} >, \quad (3.65\ a)$$

$$< E_x H_y{}^{\dagger} >= Z_{xx} < H_x H_y{}^{\dagger} > + Z_{xy} < H_y H_y{}^{\dagger} >, \quad (3.65\ b)$$

where < > denotes the average of the auto-power and cross-power density spectra. In this function, the impedances are dominated by the energetic auto-power density $H_x H_x{}^{\dagger}$ and $H_y H_y{}^{\dagger}$. Therefore, even the non-storm noisy data is introduced. Still, the energic storm period data will dominate the final result, and the M-estimator is not biased by introducing the non-storm noisy data.

The RM-estimator can obtain a reliable result from the data observed from Oct. 29 to Oct. 31 and Oct. 28 to Oct. 31. In contrast, it gets biased after introducing 2-day non-storm noisy data into the geomagnetic data from Oct. 29 to Oct. 31, and I thought the data contained about one and a half-day high SNR data from Oct. 29 to Oct. 31. In this test, the M-estimator performed best, while The BI-estimator performed worst.

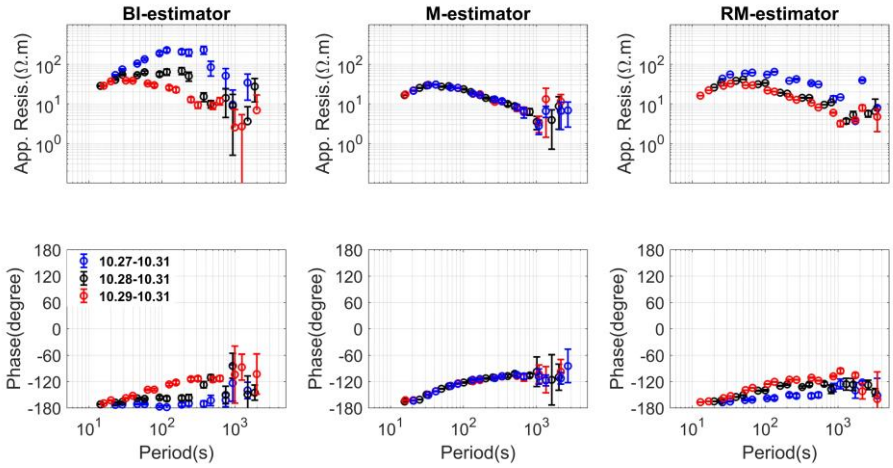

Fig. 3.11 The YX component of magnetotelluric sounding curves calculated using the different period data around the first geomagnetic storm at site 130. The left figures were calculated using the BIRRP. The second column figures were calculated using the M-estimator. The right figures were calculated using the RM-estimator. The blue curves were calculated using the dataset from Oct. 27 to Oct. 31. The black curves were calculated using the dataset from Oct. 28 to Oct. 31. Finally, the red curves were calculated using the dataset from Oct. 29 to Oct. 31.

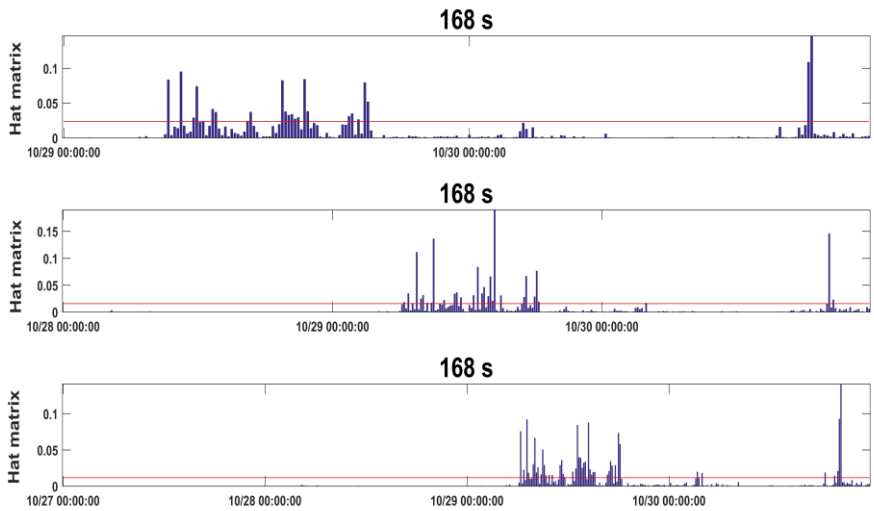

Fig. 3.12 The hat matrix diagonal element distribution at period 168 s. The upper figure shows the data from Oct. 29 to Oct. 31. The central figure shows the data from Oct. 28 to Oct. 31. The lower figure shows the data from Oct. 27 to Oct. 31. The red dotted lines denote the three times of the 2/N.

Moreover, Fig. 3.13 shows the time-series data of the second geomagnetic storm from

00:00:00 on Nov. 20 to 00:00:00 on Nov. 22. The first and third figures are the electric (e_y) and magnetic field (h_x) respectively. The second and fourth figures are the first directive of the electric and magnetic fields, respectively. The electric field has a significant transient change around 27th and 33rd hours due to the non-magnetotelluric source. The geomagnetic storm almost occurs during the 6th and 24th hours.

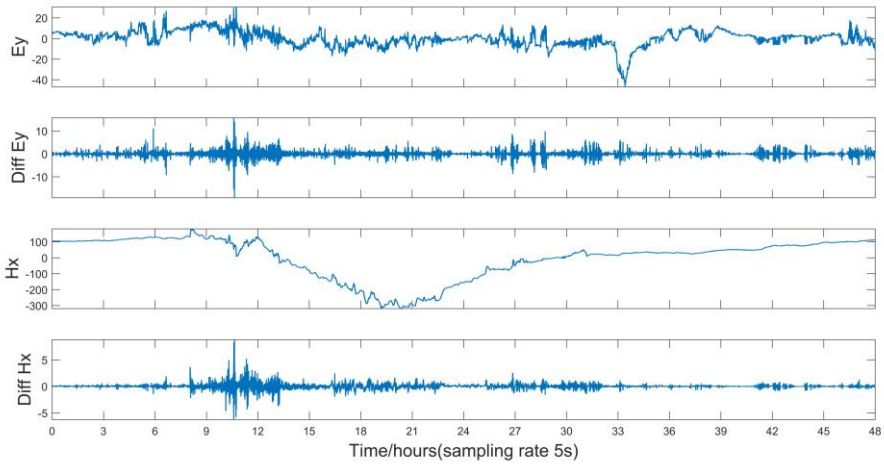

Fig. 3.13 MT time-series data at site 130 from Nov. 20 to Nov. 22. The top figure is the e_y time-series. The second figure is the first derivative of e_y. The third figure is the time series (h_x). The bottom figure is the first derivative of h_x. The electric field unit is mV/km, and the unit of the magnetic field is nT.

Fig. 3.14 shows the YX component of magnetotelluric sounding curves calculated using the different period data around the second geomagnetic storm at site 130. The black curves were calculated using the data from 06:00:00 on November 20 to 00:00:00 on November 21. The red curves were calculated using the data from 00:00:00 on November 20 to 00:00:00 on November 22. The RM-estimator could obtain a reliable result using the data from 06:00:00 on Nov. 20 to 00:00:00 on Nov. 21, but a reliable result could not be obtained using the data observed from 00:00:00 on Nov. 20 to 00:00:00 on Nov. 22. The time-series data around 27th and 33rd hours in the electric field have a great bias due to the non-magnetotelluric source, as shown in Fig. 8. Therefore, I think the noisy data focuses on the data from 00:00:00 on Nov. 21 to 00:00:00 on Nov. 22. The BI-estimator could obtain a reliable result using the geomagnetic storm data from 06:00:00 on Nov. 20 to 00:00:00 on Nov. 21. Still, it failed to obtain a reliable result after introducing the time series data from Nov. 21 to Nov. 22. The same two factors I analyzed in Fig. 3.11 lead to this result. On the other hand, the M-estimator doesn't change and get a reliable result because the energetic geomagnetic storm data dominates the final result.

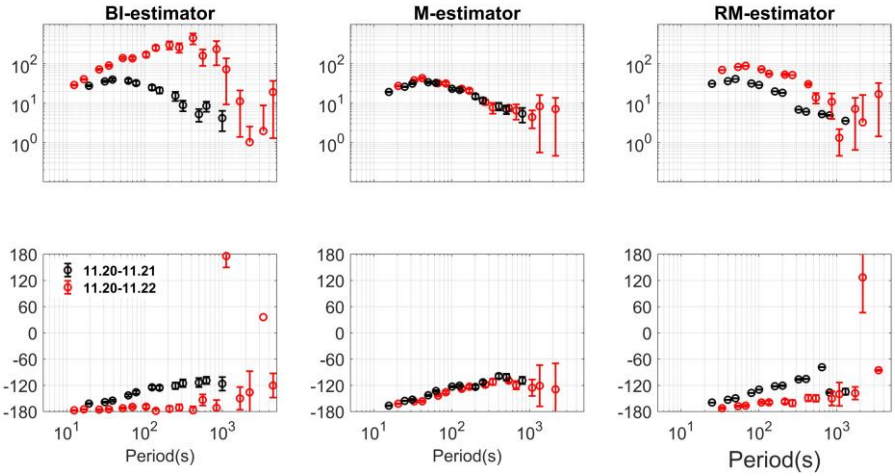

Fig. 3.14 The YX component of magnetotelluric sounding curves calculated using the different period data around the second geomagnetic storm at site 130. The left, middle and right figures were calculated using the BI-estimator, the M-estimator and the RM-estimator. The black curves were calculated using the data from 06:00:00 on November 20 to 00:00:00 on November 21. The red curves were calculated using the data from 00:00:00 on November 20 to 00:00:00 on November 22.

Fig. 3.15 shows the YX components of MT sounding curves calculated using the different periods at site 133. The red curves were calculated using the geomagnetic storm data from 00:00:00 on Oct. 29 to 00:00:00 on Oct. 31. The black curves were calculated using the data from 00:00:00 on October 28 to 00:00:00 on October 31. The blue curves were calculated using the data from 00:00:00 on October 27 to 00:00:00 on October 31. The green curves were calculated using the non-storm data from 00:00:00 on Oct. 26 to 00:00:00 on Oct. 29. The MT sounding curve calculated using the geomagnetic storm data from Oct. 29 to Oct. 31 changed reasonably. All BI-estimator, M-estimator, and RM-estimator failed to recover the result using the non-storm data(green curve). As a result, the apparent resistivity gets biased, comparing the result computed from the storm period.

I compared the results calculated using the different period data around the first geomagnetic storm in Fig. 3.15. The RM-estimator becomes deviated from the true model after introducing two-day non-storm data from Oct. 27 to Oct. 31. That indicates that the YX component has around one and a half-day high SNR data from Oct. 28 to 31. The BI-estimator further deviates from the true model by introducing more non-storm noisy data. On the other hand, the M-estimator performs robustly, and the results coincide with each other and are caused by two factors, as I analyzed in Fig. 3.11.

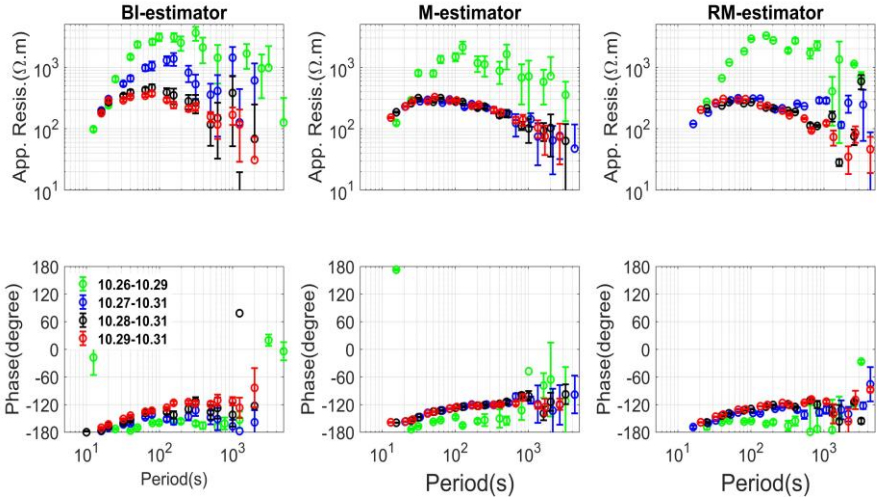

Fig. 3.15 The YX component of MT sounding curves calculated from different period data at site 133. The left figures were calculated using the BI-estimator, and the central figures were calculated using the M-estimator. The right figures were calculated using the RM-estimator. The red curves were calculated using the period from 00:00:00 on October 29 to 00:00:00 on October 31. The black curves were calculated using the period from 00:00:00 on October 28 to 00:00:00 on October 31. The blue curves were calculated using the period from 00:00:00 on October 27 to 00:00:00 on October 31. The green curves were calculated using the period from 00:00:00 on October 26 to 00:00:00 on October 31.

3.4.2.2 Comparison of the performance using the data from L6-7

The second dataset used the broadband-frequency MT time-series data observed by the Phoenix MTU-A Geophysical Instruments. This data belongs to the Institute of Geophysical and Geochemical Exploration, China Geological Survey. Fig. 3.16 shows the location map in the study area. The local survey lines are L6-1, L6-2 and L7-1, and the Y0625 is set as the remote reference site.

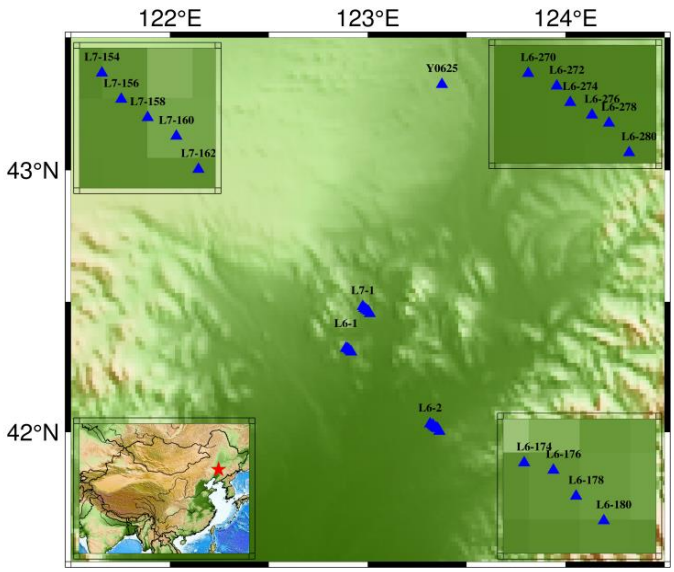

Fig. 3.16 Location map of the study area. Y0625 is the remote reference site, and the local survey lines are L6-1, L6-2, and L7-1 in the middle. The lower left map shows the survey area in China. The red star denotes the research location. The upper left map shows the detail of

survey line L7-1; The lower right map shows survey line L6-1; The upper right map shows survey line L6-2.

At first, I introduced polarization directions to estimate the background noise. The polarization directions of the electric field (α_E) and magnetic field (α_H) (Fowler et al., 1967) at a specific frequency are defined as:

$$\alpha_{E\_i} = tan^{-1}\frac{2Re(E_{x\_i}\bar{E}_{y\_i})}{|E_{x\_i}|^2-|E_{y\_i}|^2}, \tag{3.66}$$

$$\alpha_{H\_i} = tan^{-1}\frac{2Re(H_{x\_i}\bar{H}_{y\_i})}{|H_{x\_i}|^2-|H_{y\_i}|^2}. \tag{3.67}$$

I can rewrite the polarization directions as follows:

$$\alpha = tan^{-1}\frac{2Re[A_i\bar{B}_i]}{|A_i|^2-|B_i|^2} = tan^{-1}\frac{2\frac{|B_i|}{|A_i|}\cdot cos(\theta_i)}{1-\left(\frac{|B_i|}{|A_i|}\right)^2}, \tag{3.68}$$

where *i (=1,2,…, N)* is the number of spectra; A_i and B_i are H_{x_i} and H_{y_i}, respectively. The polarization direction is related to the PD and amplitude ratio (AR) between the two orthogonal fields. A variety of sources generate natural magnetic signals. These sources generate magnetic fields that vary in their incidence directions. The PD and AR between the two orthogonal fields vary with time; thus, the magnetic field has no preferred polarization direction (Weckmann et al., 2005). On the contrary, the local electromagnetic noise source usually has a constant location; the incident direction and energy have similar properties changing with time. Suppose there is a preferred polarization direction for the magnetic field; I can consider that the coherent noise contaminates the data.

This research analyzed the polarization directions at different frequencies for all sites. The intensities of natural electromagnetic fields are very weak at the dead band (1-10s); it is easy to influence the local noise. The typical polarization directions for the dead band are shown in Fig. 3.17. Fig. 3.17 shows the result of polarization directions calculated using 6-s data at site L7158. The polarization directions of the magnetic field have a preferred direction from 3:00:00 to 15:00:00. That means the data in that period is contaminated by strong noise. A similar situation occurs at all sites, including the remote site.

On the other hand, the polarization directions of the electric field changed around 16:00:00 in UTC. I thought the magnetic field was quiet from 15:00:00 to 16:00:00, but the electric field was noisy. This situation was categorized as the outlier (the unusual in the electric field). In this research, I used the data observed at site L7158 on June 26, 2016, to compare the three estimators in the second situation when the noise is energetic.

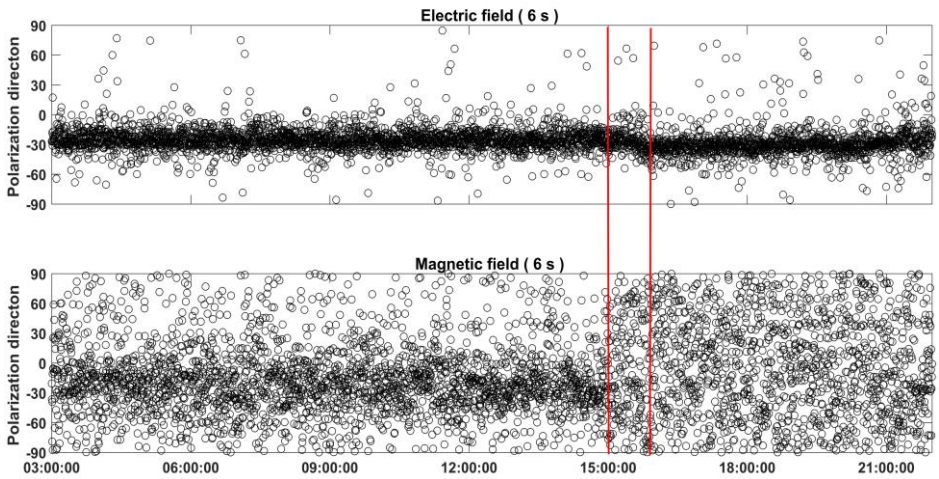

Fig. 3.17 Polarization direction for period 6 s at site L7158 from 3:00:00 to 22:00:00 in UCT. The upper figure shows the electric field's polarization directions, and the lower figure shows the magnetic field's polarization direction. The red vertical lines denote the boundary of the property of the polarization directions changed.

At first, calculate the MT sounding curve from the Phoenix software SSMT-2000 from the one-day data. I find that the result of the dead band is biased. I then calculate the time series data by BI-estimator, M-estimator, and RM-estimator using the nighttime data from 16:00:00 to 20:00:00. The black curves were calculated using the M-estimator. The red curves were calculated using the BI-estimator. The green curves were calculated using the RM-estimator. These results are shown in Fig. 3.18. The sounding curves became reliable after calculating the nighttime data.

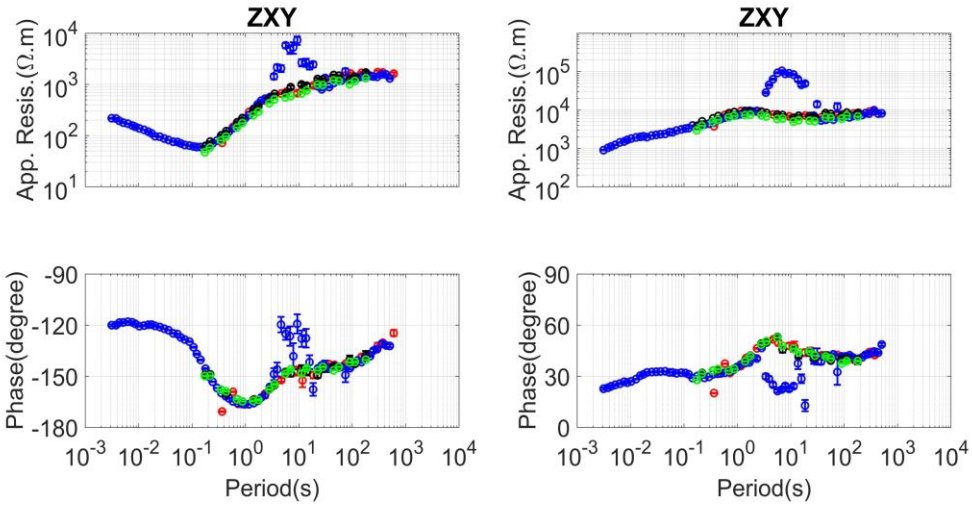

Fig. 3.18 The magnetotelluric sounding curves calculated using the data at site L7158. SSMT-2000 calculated the blue curves using one-day data. M-estimator calculated the black curves using the nighttime data from 16:00:00 to 20:00:00 in UTC. BI-estimator calculated the red curves using the nighttime data from 16:00:00 to 20:00:00 in UTC. RM-estimator calculated the green curves using the nighttime data from 16:00:00 to 20:00:00 in UTC.

In the next step, I investigated the hat matrix diagonal element. Fig. 3.19 shows the hat matrix diagonal element distribution from 14:00:00 to 20:00:00 in UTC. The magnetic noise was strong from 14:00:00 to 15:00:00, and it was relatively quiet at nighttime. I tested the BI-estimator, M-estimator and RM-estimator by introducing different lengths of noise data to the nighttime data.

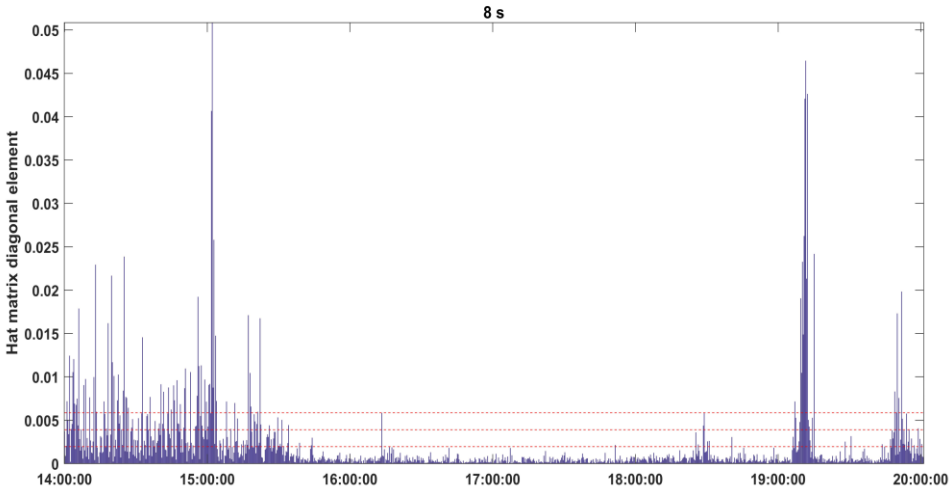

Fig. 3.19 The hat matrix diagonal element distribution from 14:00:00 to 20:00:00 in UTC at the period of 8 s. The red dotted lines denote the one, two, and three times of the 2/N, respectively.

Fig. 3.20 shows the magnetotelluric sounding curves calculated by RM-estimator using different period data at site L7158. The red curves were calculated using the nighttime data from 16:00:00 to 20:00:00 in UTC. The green curves were calculated using the data observed from 15:00:00 to 20:00:00 in UTC. The black curves were calculated using the data observed from 14:00:00 to 20:00:00 in UTC. The blue curves were calculated using the data observed from 13:00:00 to 20:00:00 in UTC. The RM-estimator result becomes biased by the data from 14:00:00 to 20:00:00. The breakdown point of the RM-estimator is 50%. That means there was about 3-hour quiet data at nighttime, and the quiet data focuses on the nighttime from 16:00:00 to 20:00:00.

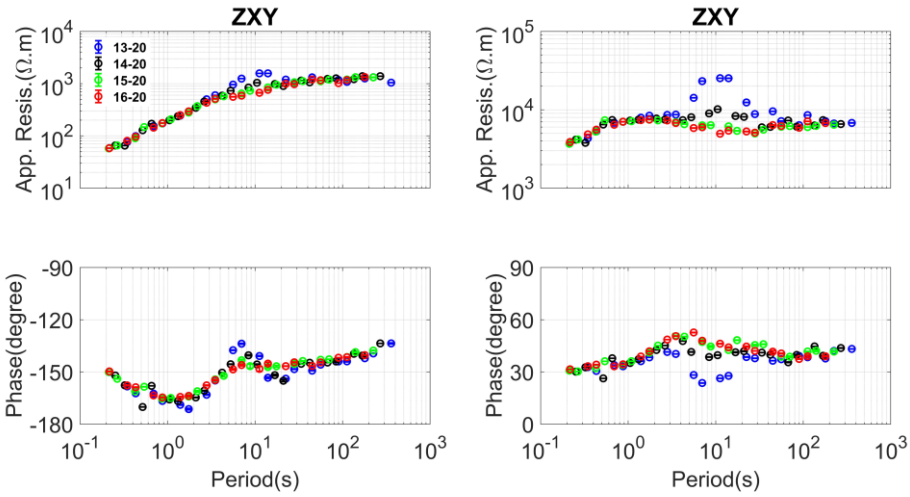

Fig. 3.20 The magnetotelluric sounding curves calculated by RM-estimator using different period data at site L7158. The red curves were calculated using the nighttime data from 16:00:00 to 20:00:00 in UTC. Next, the black curves were calculated using the data from 14:00:00 to 20:00:00 in UTC. Finally, the blue curves were calculated using the data observed from 13:00:00 to 20:00:00 in UTC.

Fig. 3.21 shows the result calculated by the M-estimator. The red curves were calculated using the nighttime data observed from 16:00:00 to 20:00:00 in UTC. Next, the black curves

were calculated using the data observed from 15:00:00 to 20:00:00 in UTC. Finally, the blue curves were calculated using the data observed from 14:00:00 to 20:00:00 in UTC. After introducing a one-hour outlier in the electric field, the M-estimator was found to work effectively. However, after introducing one more hour of leverage point data between 14:00:00 to 15:00:00, the M-estimator gave a biased result in the dead band. This study shows that the M-estimator is not sensitive to the leverage point in the magnetic field, and the leverage point will dominate the final result.

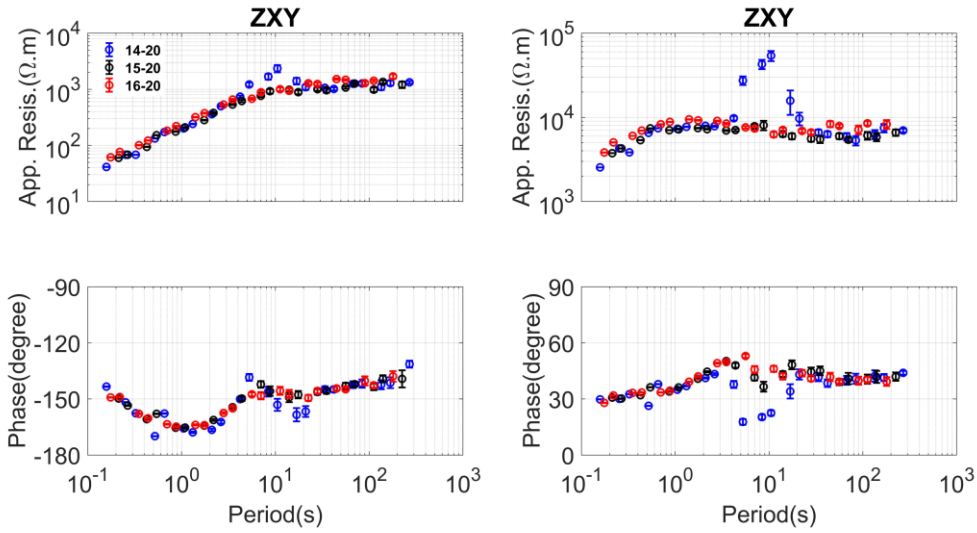

Fig. 3.21 The magnetotelluric sounding curves calculated by M-estimator using different period data at site L7158. The red curves were calculated using the nighttime data from 16:00:00 to 20:00:00 in UTC. Next, the black curves were calculated using the nighttime data from 15:00:00 to 20:00:00 in UTC. Finally, the blue curves were calculated using the data observed from 14:00:00 to 20:00:00 in UTC.

Fig. 3.22 shows the result calculated by BI-estimator. The red curves were calculated using the nighttime data from 16:00:00 to 20:00:00 in UTC. Then, the black curves were calculated using the data from 14:00:00 to 20:00:00 in UTC. Finally, the blue curves were calculated using the data observed from 13:00:00 to 20:00:00 in UTC. Compared to the result calculated by the BI-estimator and M-estimator by the data from 14:00:00 to 20:00:00 in UTC, the BI-estimator succeeded in getting a reliable result M-estimator and RM-estimator failed. The BI-estimator will work better than M when the leverage point is the active noise.

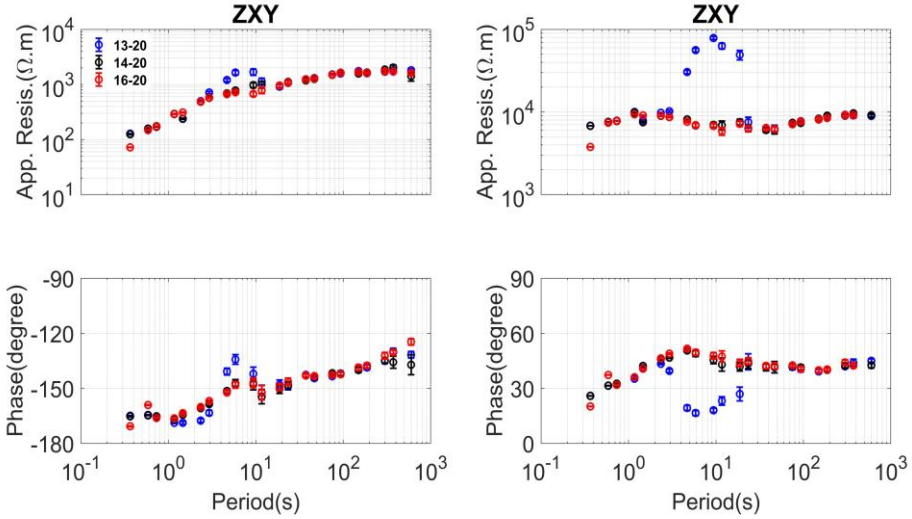

Fig. 3.22 The magnetotelluric sounding curves calculated by the BI-estimator using different period data at site L7158. The red curves were calculated using the nighttime data from 16:00:00 to 20:00:00 in UTC. The black curves were calculated using the data observed from 14:00:00 to 20:00:00 in UTC. The blue curves were calculated using the data observed from 13:00:00 to 20:00:00 in UTC.

3.5 Conclution and Discussion

I compared three single-site processing impedance estimators; the M-estimator, BI-estimator and RM-estimator. The measure of an estimator's robustness is its breakdown point. An estimator's breakdown point is the proportion of incorrect observations that an estimator can handle before giving an incorrect result. It is well known that the breakdown point of the least-squares estimator is 0; the breakdown point of the RM-estimator is close to 50%; if there is no more than half of the data is contaminated, it can produce a reliable result.

Since the BI and M-estimator's final result depends on the weighted average method(least square estimator), the M-estimator and BI-estimator's breakdown point differs in different situations. It depends on the strength of the signal and noise.

Leverage weighting is the most significant difference between the M-estimator and BI-estimator when the data site is noisy. Suppose there is no leverage point; the breakdown points of the M-estimator and BI-estimator become the same. The leverage point is the large variation of magnetic field data. The M-estimator is not sensitive to the leverage point. BI-estimator will reject the leverage point using a strict iterative weighting, but the leverage point can be an energetic signal or active noise.

In the first field data test, the leverage point corresponds to the energetic single from a geomagnetic storm event. The leverage weighting will negatively influence the reweight least square estimator. As a result, the M-estimator will perform better than the BI-estimator and RM-estimator.

The BI-estimator will work much better than the M-estimator and RM-estimator when the leverage point is the active noise at the second field data test. The breakdown point of the RM-estimator is 50%. In this test, the RM-estimator gets biased by the data from 14:00:00 to 30:00:00. I think there are three-hour high SNR time-series data from 16:00:00 to 20:00:00. The M-estimator gets biased by the data from 15:00:00 to 20:00:00. The breakdown point of the M-estimator is 40%; The RM-estimator was better than M-estimator. The BI-estimator gave biased results using the data from 14:00:00 to 20:00:00. The breakdown point of the BI-estimator is close to 50%. The breakdown point of the M-estimator and BI-estimator is different in different situations.

The breakdown point is different for the different impedance estimators. When most of the

data is well behaved, the conventional statistic methods may remove the noise. When the noise content is over 50%, the preselection strategy may produce a reliable result by screening the high SNR data before the impedance estimation. The most serious situation is that the continuous noise contaminates the data; any method may fail to get a reliable result.

However, I don't know the noise contents in the real data processing. Therefore it is necessary to know the situation of background noise. I will discuss this problem in the next chapter.

Chapter 4: A ROBUST ESTIMATOR BASED ON FFT FOR DATA QUALITY ANALYSIS AND THE MT IMPEDANCE CALCULATION

The conventional impedance estimator uses the windowed FFT to transform the time-series data into the frequency domain and uses the M-estimator to estimate the MT impedance. However, it may be biased in the presence of a large amount of artificial noise. The key to getting a reliable impedance from the noisy data is detecting and removing the noise before the impedance estimation. In theory, if the artificial noise doesn't contaminate the field data continuously. I can extract the high signal-to-noise ratio (SNR) data and get a reliable impedance.

This chapter is organized as follows. A robust M-estimator for the regression is introduced first. Then I introduce the parameter used for the data quality analysis and preselection strategy. Finally, I show the case study of data quality analysis for a specific period and the effectiveness of the data selection strategy.

4.1 The Robust M-estiamtor

This subsection introduces an M-estimator from spectrum estimation to regression.

4.1.1 Spectrum estimation

A method similar to the Bould Influence Remote Reference Processing code (BIRRP, Chave and Tompson, 2004) is adopted to estimate the spectra at different frequencies. When performing the Fourier transform, there is the following relationship:

$$f=n*fs/N, \tag{4.1}$$

where f is the frequency corresponding to the nth complex coefficient, n is the nth complex coefficient in the spectra, fs denotes the sampling rate, and N denotes the segment length. I divide the original time-series data into segments using the longest window. The segment length is variable with each level. Each segment's 5th and 8th complex coefficients are grouped to calculate the response function at each level. The segment length is then reduced by 2, and the process is repeated to calculate the impedance tensor at different frequencies.

For the real MT time series, e_x, e_y, h_x, h_y, h_z, the electric field unit is transformed into mv/km, and the magnetic field unit is transformed into nT. For each level, the procedure used to produce the accurate spectrum for each segment is as follows:

(1) Use the first difference filter to remove the trends and means; in the case of e_x, this process yields the following new series.

$$\widetilde{e_x}(k) = e_x(k+1) - e_x(k) \tag{4.2}$$

(2) Adopt a multitaper method to reduce the bias in the spectral estimation. A multitaper process is an extension of single-taper approaches. The multitaper method can be carried out by the MATLAB intrinsic function "dpss.m". This study specifies the half-bandwidth product as 2.5 and the multitaper value as 4. These specifications produce four orthogonal discrete prolate spheroidal (Slepian) sequence windows and a column weight vector (lambda). The 4 Slepian sequences are tapered to each segment, and four time-series sets ($\widehat{e_x}$) are obtained.

(3) Carry out the Fourier transform on each $\widehat{e_x}$, and calculate the Fourier coefficients E_x using the weighted mean of the four Fourier coefficient sets by multiplying them by the lambda value. Perform the same procedure for the other components (E_y, H_x, H_y, H_z). Group the 5th and 8th complex coefficients within each spectrum to perform further processing at each level.

4.1.2 Robust estimation of the response function

The regression step is similar to the methods described by Neukirch and Garcia (2014) and

Campanya et al. (2014). The impedance tensor is calculated as follows:

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{R}_x + \boldsymbol{\sigma}_x, \tag{4.3}$$

$$\boldsymbol{Z} = \left(\boldsymbol{R}_{hx}, \boldsymbol{R}_{hy}\right)^{-1}\left(\boldsymbol{R}_{ex}, \boldsymbol{R}_{ey}\right), \tag{4.4}$$

where $\boldsymbol{y}$ denotes the N-by-one matrix E_x,E_y,H_x,H_y,H_z; $\boldsymbol{x}$ denotes the N-by-two matrices, which is local (H_x, H_y) and termed $\boldsymbol{H}$ or remote-referenced $(H_{xr}$,$H_{yr})$ and termed $\boldsymbol{H}_r$; $\boldsymbol{R}_x$ is a two-by-one matrix that denotes the interstation transfer function; $\boldsymbol{R}_{ex}$, $\boldsymbol{R}_{ey}$, $\boldsymbol{R}_{hy}$, and $\boldsymbol{R}_{hx}$ denotes that when $\boldsymbol{y}$ equals E_x, E_y, H_x, H_y, respectively.

When **x** equals the local N-by-two matrices (H_x, H_y), $(\boldsymbol{R}_{hx}, \boldsymbol{R}_{hy})$ equals the identity matrix and $\boldsymbol{Z}$=$\left(\boldsymbol{R}_{ex}, \boldsymbol{R}_{ey}\right)$. The solution to $\boldsymbol{R}_x$ is a bivariate complex linear regression problem. When $\boldsymbol{y}$ equals E_x, $\boldsymbol{R}_x$ is equal to the two-by-one matrix $(Z_{xx}, Z_{xy})^T$. It is equal to E_x=$H_x Z_{xx}$+$H_y Z_{xy}$, and the function can be extended to the following complex form:

$$E_{xr_n} + jE_{xi_n} = \left(H_{xr_n} + jH_{xi_n}\right)(Z_{xxr} + jZ_{xxi}) + \left(H_{yr_n} + jH_{yi_n}\right)\left(Z_{xyr} + jZ_{xyi}\right), \tag{4.5}$$

where $n=1,2,\ldots,N$ denotes the nth number of the data, H_{xr_n} denotes the nth real part of H_x, and H_{xi_n} denotes the nth imaginary part of H_x; the same variables are used in E_x, H_y. Z_{xxr} denotes the real part of Z_{xx}, and Z_{xxi} denotes the imaginary part of Z_{xx}; these are consistent with Z_{xy}. The symbol j denotes the unit of the imaginary numbers. The total equations for all data can be transformed into a matrix form as follows:

$$\begin{pmatrix} E_{xr_1} \\ E_{xr_2} \\ \vdots \\ E_{xr_n} \\ E_{xi_1} \\ E_{xi_2} \\ \vdots \\ E_{xi_n} \end{pmatrix} = \begin{pmatrix} H_{xr_1} & -H_{xi_1} & H_{yr_1} & -H_{yi_1} \\ H_{xr_2} & -H_{xi_2} & H_{yr_2} & -H_{yi_2} \\ \vdots & \vdots & \vdots & \vdots \\ H_{xr_n} & -H_{xi_n} & H_{yr_n} & -H_{yi_n} \\ H_{xi_1} & H_{xr_1} & H_{yi_1} & H_{yr_1} \\ H_{xi_2} & H_{xr_2} & H_{yi_2} & H_{yr_2} \\ \vdots & \vdots & \vdots & \vdots \\ H_{xi_n} & H_{xr_n} & H_{yi_n} & H_{yr_n} \end{pmatrix} \begin{pmatrix} Z_{xxr} \\ Z_{xxi} \\ Z_{xyr} \\ Z_{xyi} \end{pmatrix} \tag{4.6}$$

The bivariate complex linear regression problem is converted to a real multivariate linear regression problem. The MATLAB intrinsic function "robustfit.m" can easily solve this regression problem, and the Huber weight is adopted.

When $\boldsymbol{x}$ equals the remote magnetic field of the N-by-two matrices (H_{xr}, H_{yr}), the interstation transfer function is first calculated as follows:

$$\boldsymbol{R}_e = (\boldsymbol{H}_r^{\dagger}\boldsymbol{H}_r)^{-1}(\boldsymbol{H}_r^{\dagger}\boldsymbol{E}), \tag{4.7}$$

$$\boldsymbol{R}_h = (\boldsymbol{H}_r^{\dagger}\boldsymbol{H}_r)^{-1}(\boldsymbol{H}_r^{\dagger}\boldsymbol{H}), \tag{4.8}$$

where $\boldsymbol{R}_e$ denotes when $\boldsymbol{y}$ is equal to $(E_x\ E_y)$, and $\boldsymbol{R}_h$ denotes when $\boldsymbol{y}$ is equal to (H_x, H_y). Then, the impedance tensor value is calculated using the relationship between the local transfer function and the interstation transfer function as follows:

$$\boldsymbol{Z} = \boldsymbol{R}_h^{-1}\boldsymbol{R}_e. \tag{4.9}$$

The result can be categorized into a type of two-stage processing strategy. In this way, separating the calculation into regressions between the remote reference and all channels can avoid the direct effects of the coherent noise among local channels. This method can work well if the remote magnetic field does not contain noise correlated with the local site.

Finally, the robust regression equation was bootstrapped to compute the impedance values and estimate the intrinsic data errors. Empirically, 1,024 iterations were repeated, considering a sufficient trade-off between the accuracy and computation time to evaluate the uncertainty.

4.2 The Parameters Used For Data Quality Analysis And Data Preselection

The key to getting a reliable impedance from the noisy data is detecting and removing the noise before the impedance estimation. I introduce the parameters used to examine the data quality in this subsection.

4.2.1 The phase difference and coherence between the electric and magnetic field

The MT time-series data are observed simultaneously in several channels; each channel is divided into N segments, and I can obtain *N* spectra by applying the Fourier transform to each segment. In polar coordinates, the cross-power spectra between two spectra A_i and B_i are expressed as follows:

$$A_i\bar{B}_i = |A_i|\cdot|B_i|e^{j(\varphi_{A_i}-\varphi_{B_i})}, \tag{4.10}$$

where *j* denotes the imaginary number unit, *i (=1,2,…, N)* is the number of spectra; A_i and B_i are the spectra calculated from the ith segment from the different channel; and φ_{A_i} and φ_{B_i} denote the phases of A_i and B_i, respectively. The overline denotes the complex conjugate. The auto-power spectra are expressed as follows:

$$A_i\bar{A}_i = |A_i|^2,\ B_i\bar{B}_i = |B_i|^2. \tag{4.11}$$

The phase difference between A_i and B_i (PD(A_i, B_i)) is calculated as follows:

$$\theta_i = \varphi_{A_i} - \varphi_{B_i} = arg\left(e^{j(\varphi_{A_i}-\varphi_{B_i})}\right) = arg\left(\frac{A_i\bar{B}_i}{|A_i||B_i|}\right), \tag{4.12}$$

where θ_i denotes the angle of the phase difference (PD) between the two spectra at a specific frequency, as shown in Fig. 4. 1.

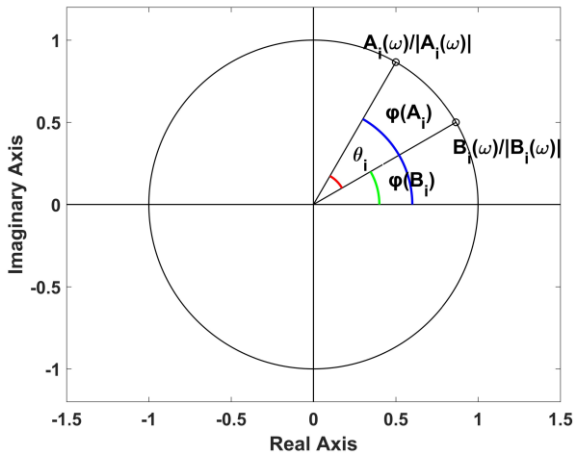

Fig. 4.1 The phase difference between two spectra at a specific frequency. θ_i denotes the angle of the PD between the two spectra at a specific frequency.

Coherency is the ratio between cross-power spectra density and the root of auto-power spectra density. For the ***A*** and ***B*** spectrum at a specific frequency, it is defined as:

$$Coh(\boldsymbol{A},\boldsymbol{B}) = \frac{|<\boldsymbol{A}\bar{\boldsymbol{B}}>|}{\sqrt{<\boldsymbol{A}\bar{\boldsymbol{A}}><\boldsymbol{B}\bar{\boldsymbol{B}}>}}, \tag{4.13}$$

where the brackets represent the averages of *N* individual auto-power spectra and cross-power spectra. For instance,

$$<\mathbf{A}\bar{\mathbf{B}}> = \frac{1}{N}\sum_{i=1}^{N} A_i\bar{B}_i. \tag{4.14}$$

When the PD is scattered, the averages of *N* individual cross-power spectra neutralize each other, and the cross-power spectra density becomes a small value; therefore, the coherence decreases and becomes close to 0. When there is a preferred direction of PD, the coherence is high and close to one. The coherence is a quantitative measure of the phase difference (PD) consistency between the two spectrums. If two spectrums are coherent, their phases must be the same or have a constant difference (Marple and Marino, 2004). Fig. 4.2 shows the example of the relationship between the PD and coherence. In this research, I introduce the variation in the PD to show the coherence. In this way, I can visualize the data quality varies with time, based on the criteria of coherence.

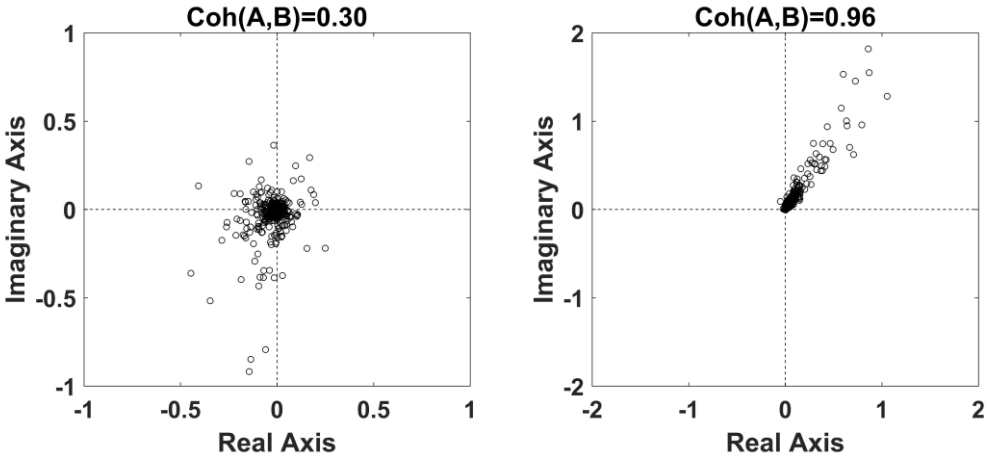

Fig. 4.2 The examples of the relationship between the phase difference and coherence value. The figures show the distribution of cross-power spectra between the $\boldsymbol{A}$ and $\boldsymbol{B}$ spectrum. Each point denotes $A_i\bar{B}_i$, where i denotes the ith point. The phase of each point is the phase difference between A_i and B_i. The title of each figure shows the coherence value.

In a certain geological situation, the impedance tensor is shown in Table 4.1. In the 1-D or 2-D (when the observation axes coincide with the geoelectrical strike) cases (1-D/2-D), there is a relationship between the orthogonal electric and magnetic field for the north-south direction as follows:

$$E_{x\_i} = Z_{xy}H_{y\_i}. \tag{4.15}$$

The orthogonal eclectic and magnetic fields have a constant PD, and the PD between the orthogonal electric and magnetic field spectra equal the impedance phase.

In specific geologic environments, the phenomenon of the phase rolling out of the quadrant (PROQ) may appear (Chouteau and Tournerie, 2000; Weckmann et al., 2003; Yu et al., 2018.). The current channelling caused by complex 3-D isotropic media can explain the PROQ phenomenon. The characteristic of PROQ is that the parallel electric and magnetic fields are coherent, while the orthogonal components are incoherent. There is a preferred direction of PD between the parallel electric and magnetic fields.

In the 2-D (when the observation axes do not coincide with the geoelectrical strike) or 3-D cases (2-D/3-D), all impedance values are nonzero, and there is a relationship between the electric and magnetic field for the north-south direction as follows:

$$E_{x\_i} = Z_{xx}H_{x\_i} + Z_{xy}H_{y\_i}. \tag{4.16}$$

There is no constant relationship between the $\boldsymbol{E_x}$ and $\boldsymbol{H_x}$, and $\boldsymbol{E_x}$ and $\boldsymbol{H_y}$.

Table 4.1 The impedance tensor characteristics in certain geological situations. In the 1-D case, the impedance of the diagonal elements is zero. In the 2-D case, when the observation axes coincide with the geoelectrical strike, the diagonal elements' impedance is zero, and the off-diagonal elements are unequal. In the 3-D case, all Z values are nonzero.

1-D case	2-D case	3-D case
$\begin{pmatrix} 0 & Z_{xy} \\ Z_{yx} & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & Z_{xy} \\ Z_{yx} & 0 \end{pmatrix}$	$\begin{pmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{pmatrix}$
$Z_{xy} = -Z_{yx}$	$Z_{xy} \neq Z_{yx}$	all $\boldsymbol{Z}$ values are nonzero

On the other hand, the time series contaminated by the coherent and incoherent noise also has a high and low coherence value. I can categorize the data into five types: 1-D/2-D, 2-D/3-D,

PROQ, the data contaminated by coherence noise, and the data contaminated by incoherent noise. However, according to the data type classification, I can't distinguish the data type by coherence. Therefore, I need to combine the other parameters to discuss the specific situation.

4.2.2 The linear coherence and amplitude ratio between the predicted and observed electric field

From the perspective of whether the data follows the linear relationship, I can subdivide the field data ($\boldsymbol{E}$ and $\boldsymbol{H}$) into three parts as follow:

$$\boldsymbol{E} = \boldsymbol{E}^{\boldsymbol{MT}}+\boldsymbol{E}^{\boldsymbol{HLN}} + \boldsymbol{E}^{\boldsymbol{LLN}}, \tag{4.17}$$

$$\boldsymbol{H} = \boldsymbol{H}^{\boldsymbol{MT}}+\boldsymbol{H}^{\boldsymbol{HLN}} + \boldsymbol{H}^{\boldsymbol{LLN}}, \tag{4.18}$$

where the *HLN* denotes the noise with a high linear relationship, the *LLN* denotes the noise with a low linear relationship, and the MT denotes natural *EM* signals with a high linear relationship. Assuming the data is in a high linear relationship, the observed electric field ($\boldsymbol{E}$) should be similar to the predicted electric field ($\boldsymbol{E_p}$), where $\boldsymbol{E_p}$=$\boldsymbol{ZH}$ and $\boldsymbol{Z}$ are obtained by the least-squares estimator. I can identify the noise with a low linear relationship by comparing the measured electric field ($\boldsymbol{E}$) and the predicted electric field ($\boldsymbol{E_p}$). A similar idea is the predicted coherence in the previous work (Gamble et al., 1979; Travassos and Beamish, 1988; Weckmann et al., 2005). The complex number has two properties: amplitude and phase. However, Weckmann et al. (2005) and Travassos and Beamish (1988) used the predicted coherence defined by the amplitude. The predicted coherence may miss the information of phase. In this study, I used the predicted linear coherence defined by phase difference to confirm the similarity in phase and used the predicted amplitude ratio to confirm the similarity in amplitude.

A new parameter, linear coherence, is proposed to express the PD in another way. The linear coherence ($Lcoh$) is defined by the cosine of the PD as follows:

$$\cos(\theta_i) = \mathrm{Re}(e^{j(\varphi_{A_i}-\varphi_{B_i})}) = \mathrm{Re}\left(\frac{A_i\bar{B}_i}{|A_i||B_i|}\right), \tag{4.19}$$

where Re denotes the real part of the complex number. The value of *Lcoh* lies in the range from -1 to 1. When the PD is close to 0°, the Lcoh value is high and close to 1. According to Euler's formula, Lcoh is also equal to the real part of $e^{j(\varphi_{A_i}-\varphi_{B_i})}$.

In this study, the predicted linear coherence between the measured electric field ($\boldsymbol{E}$) and the predicted electric Field ($\boldsymbol{E_p}$) is calculated as follows:

$$PLcoh = \mathrm{Re}\left(\frac{E_{p\_i}\bar{E}_i}{|E_{p\_i}||E_i|}\right), \tag{4.20}$$

where E_{p_i} is the predicted electric field and E_i is the measured electric field at a specific frequency calculated from the ith segment. The predicted electric field and the observed electric field should be similar, and the PD between E_p and E_{p_i} should be close to 0°; under this condition, the *PLcoh* should be high and close to one.

The high predicted linear coherence can ensure the similarity in the phase. I also used the predicted amplitude ratio to ensure the similarity in the amplitude, and it is defined as follows:

$$PAR=\begin{cases} abs(E_{p\_i})/abs(E_i), when\ abs(E_{p\_i}) < abs(E_i), \\ abs(E_i)/abs(E_{p\_i}), when\ abs(E_i) < abs(E_{p\_i}). \end{cases} \tag{4.21}$$

The *PAR* ranges from 0 to 1; the higher the *PAR* is, the higher the similarity between the two spectra in the amplitude.

There is a problem that the $\boldsymbol{Z}$ might be biased in the presence of a large amount of noise (The problem is discussed in the discussion). In this research, I divided the data into groups with specific samples, e.g. 20 samples, and assumed the property in each group was the same. In this way, I can reduce the incorrect information brought by the biased $\boldsymbol{Z}$. To differentiate the conventional predicted linear coherence and amplitude ratio; I rename them as $PLcoh_{sz}$ and

PAR_{sz}, where the *SZ* means the impedance($\boldsymbol{Z}$) calculated by the segmented field data.

4.2.3 The polarization directions between the two orthogonal magnetic fields

I also adopt the polarization direction to detect the coherent noise and the polarization direction for the magnetic field (α_{H_i}) at a specific frequency (Fowler et al., 1967; Weckmann et al., 2005) is defined as follows:

$$\alpha_{H\_i} = tan^{-1}\frac{2Re(H_{x\_i}\bar{H}_{y\_i})}{|H_{x\_i}|^2-|H_{y\_i}|^2} = tan^{-1}\frac{2\frac{|H_{x\_i}|}{|\bar{H}_{y\_i}|}\cdot cos(\theta_i)}{1-\left(\frac{|\bar{H}_{y\_i}|}{|H_{x\_i}|}\right)^2}. \quad (4.22)$$

where i (=1,2,…, N) is the number index, H_{x_i} and H_{y_i} are the spectra calculated from the ith segment; θ_i denotes the PD between H_{x_i} and H_{y_i}. The polarization direction is related to the PD and amplitude ratio (AR) between the two orthogonal fields. Various sources generate natural magnetic signals, and these sources generate magnetic fields that vary in their incident directions. The PD and AR between the two orthogonal fields vary with time; thus, the magnetic field has no preferred polarization direction (Weckmann et al., 2005). On the contrary, the local electromagnetic noise source usually has a constant location; the incident direction and energy have similar properties changing with time. Suppose there is a preferred polarization direction for the magnetic field; I can consider that the coherent noise contaminates the data. On the other hand, when the incoherent noise contaminates the field data, the magnetic field has no preferred polarization direction. Therefore, the polarization direction for the magnetic field can only detect the coherent noise.

To quantify the degree of dispersion of the polarization directions (Platz and Weckmann, 2019), the dispersion degree(DD_{pol}) is proposed as follows:

$$DD_{pol}=\frac{N_{in}}{N}. \quad (4.23)$$

N_{in} denotes the number of the samples falling in the range of (m_i+ 30°, m_i-30°), where m_i is the median for each sample x_i with its surrounding k samples, and it is calculated as follows:

$$m_i=\text{median}(x_{i-k}, x_{i-k+1}, \ldots, x_i, \ldots, x_{i+k-1}, x_{i+k}). \quad (4.24)$$

If the polarization directions randomly distribute from -90° to 90°, the DD_{pol} should be close to 1/3; and the DD_{pol} increases in the presence of a preferred direction. The DD_{pol} can be used to detect the coherent noise with strong polarization direction automatically.

Next, I explain how I set parameters properly, as the m_i is calculated by the surrounding k samples, there are two sides, and when there is no preferred direction of the polarization direction, I hope half of the data is beyond a specific range, which means the threshold is set as 0.5, and the expected value of DD_{pol} should be smaller than 0.5, and the expected value of 1/3 is chosen by our experience, which means the range of the data is 60°, and the specific range is (m_i+ 30°, m_i-30°).

4.2.4 The hat matrix

The hat matrix is the *N* by *N* matrix (*N* denotes the sample of data) and is defined as follows:

$$\boldsymbol{H}_{hat}= \boldsymbol{H}(\boldsymbol{H}^{\dagger}\boldsymbol{H})^{-1}\boldsymbol{H}^{\dagger}, \quad (4.25)$$

where $\boldsymbol{H}$ is the *N* by *2* matrices of the horizontal magnetic field ($\boldsymbol{H}_x, \boldsymbol{H}_y$) at a specific frequency. The superscript † denotes the complex conjugate transpose. The expected value of the hat matrix's diagonal element is 2/*N*. The hat matrix is widely used to detect inappropriate large-value data. As the noisy data is energetic, the hat matrix's diagonal element's statistical analysis is useful for detecting the noisy data. Chave and Thomson (2003) introduced the hat matrix into their robust processing (Bound Influence Remote Reference Processing, BIRRP). However, the inappropriate large-value data may correspond to the high natural EM signal in

the presence of a geomagnetic storm. Therefore, I should combine the parameters to discuss the specific situation. In this study, I applied the hat matrix to examine the energy variation of each segment (see the Example of Fig. 6.5).

4.2.5 The linear coherence between the local field and remote magnetic field

For the north-south direction, the linear coherence between the remote and local magnetic fields (*RLcoh*) is defined as follows:

$$RLcoh = Re\left(\frac{H_{x\_i}\bar{H}_{xr\_i}}{|H_{x\_i}||H_{xr\_i}|}\right), \tag{4.26}$$

where H_{x_i} and H_{xr_i} are the local and remote magnetic field spectra at a specific frequency calculated from the ith segment.

The field MT data include natural EM signals and noise from the local environment. I can rewrite the magnetic field $\boldsymbol{H}$ as follows:

$$\boldsymbol{H} = \boldsymbol{H^{MT}} + \boldsymbol{H^{N}}, \tag{4.27}$$

where N denotes the locale noise, and MT denotes the natural EM signals from the magnetosphere and ionosphere.The portion of the natural magnetic signals in the local (H^{MT}) and remote sites (H_r^{MT}) comes from the same source and should be similar to each other, which means the PD between the local and remote magnetic fields should be close to 0°, and the *RLcoh* value should be close to one.

RLcoh is the parameter to measure the similarity between the remote and local magnetic fields. Suppose there is a quiet remote reference site; I could use *RLcoh* to measure the similarity between the local and remote magnetic fields to evaluate the data quality at the local site changing with time. Moreover, the *RLcoh* is suitable for the selection strategy.

4.3 The Case Study For Single Site Data Processing

This subsection investigates data processing for a single site.

4.3.1 The case for the data contaminated by the intermittent incoherent noise

The first case study used the data observed at TVN48 from the USArray project (Schultz et al., 2006–2018; Kelbert et al., 2018). I downloaded the time-series data from IRIS (Incorporated Research Institutions for Seismology). Fig. 4.3 shows the footprints of MT USArray by year from 2006 to 2018. The data sampling period is 1 s. The typical recording time at each site was about two weeks. I used the data observed from June 19 to June 24, 2015. A geomagnetic storm occurred during this period (see Fig. S7 in the supplementary document).

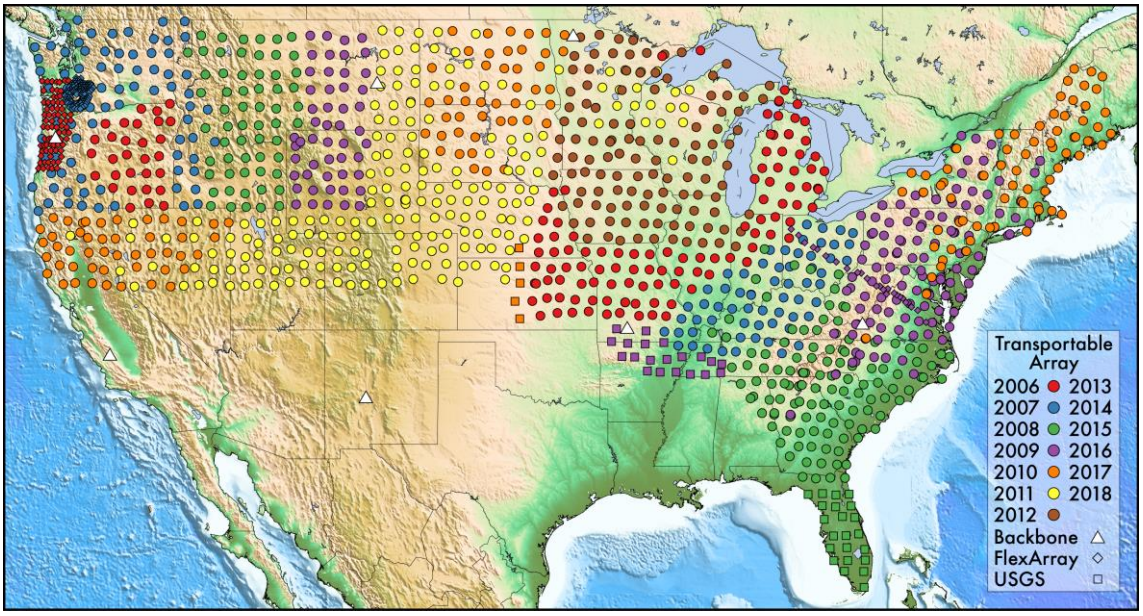

Fig. 4.3 Footprints of MT USArray by year from 2006 to 2018 (download from IRIS: http://www.usarray.org/researchers/obs/magnetotelluric/)

At first, the data quality analysis in 13.1 s is shown in Fig. 4.4; the data which both $PLcoh_{sz}$ and PAR_{sz} larger than 0.8 is shown in red, and the other data is in blue. The red color denotes the data with a high linear relationship, and the blue denotes a low linear relationship. Fig. (f) shows that the data in red has a preferred direction of PD between the orthogonal electric and magnetic fields, and the PD of the blue data is scattered. There is a constant relationship between the orthogonal electric and magnetic fields in the 1-D or 2-D (when the observation axes coincide with the geoelectrical strike). The data in red can be regarded as the high SNR data in a 1-D/2-D structure, and the blue data is dominated by incoherent noise. Fig. (d) shows that the data in red corresponds to the high hat matrix's diagonal element. The geomagnetic storm causes that (see Fig. A in the supplementary document). In conclusion, the data in blue is dominated by incoherent noise. When there is a geomagnetic storm, the nature EM signal's strength increases, and the SNR becomes high.

Fig. 4.5 shows the MT sounding curves at site TNV48. The M-estimator with the single-site processing result is down biased, and the M-estimator result using the data preselection strategy with $PLcoh_{sz}$ and PAR_{sz} coincides with the EMTF result using the remote reference technique. It shows that the $PLcoh_{sz}$ and PAR_{sz} are useful to detect incoherent noise.

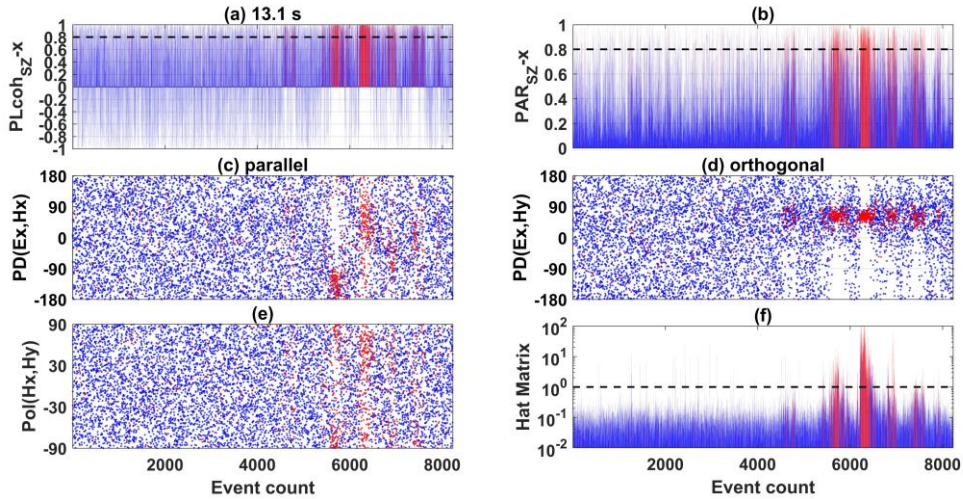

Fig. 4.4 The variation in the indices at TVN48 in the period of 13.1 s. The horizontal axis denotes the event count. Panel (a)-(b) show the variation in $PLcoh_{sz}$ and PAR_{sz} in the *x*-directions, respectively. The red color denotes the data that both $PLcoh_{sz}$–*x* and PAR_{sz}–*x* are higher than 0.8, and the other data are shown in blue. Panel (c)-(d) show the variation in PD(E_x, H_x) and PD(E_x, H_y) respectively. Panel (e) shows the variation in polarization direction for the magnetic field. Panel (f) shows the hat matrix's diagonal element variation, and the hat matrix's diagonal element is normalized by the expected value (2/*N*, *N* denotes the sample of data).

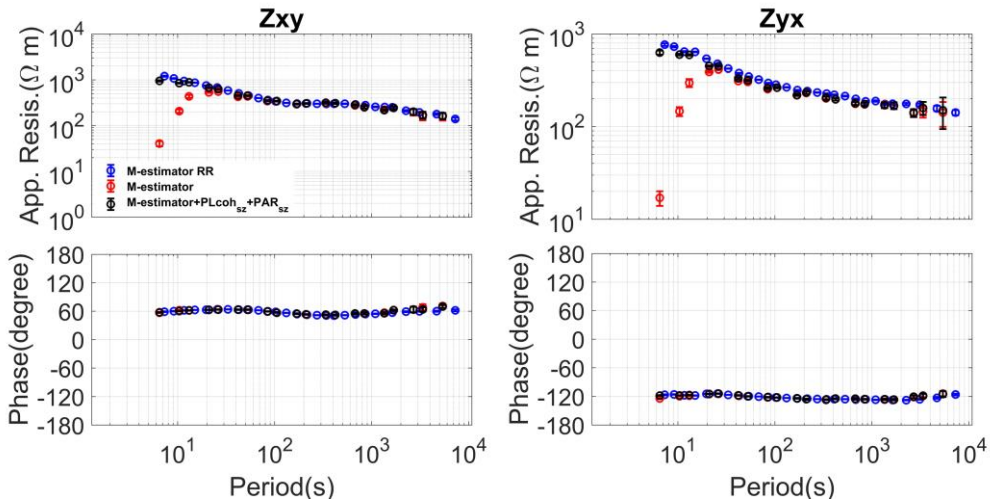

Fig. 4.5 The MT sounding curves calculated using the data observed at site TNV48. The upper figures show the apparent resistivity, and the lower figures show the impedance phase. The EMTF code calculated the blue curves using the remote reference technique provided by the USArray team. The red curve denotes the single-site processing result. The black curve denotes results using the preselection strategy with $PLcoh_{sz}$ and PAR_{sz}.

4.3.2 The case for the data contaminated by the intermittent coherent noise

The second case study used the data observed in China on June 25, 2020. Fig. 4.6 shows the local site L7-158. Phoenix Geophysics Instruments were used to observe the broadband-frequency MT time-series data. These data belong to the Institute of Geophysical and Geochemical Exploration, China Geological Survey. The low-frequency band data (15 Hz) sampled continuously was used in this case study. The observation area is in the GMT+8 time

zone, and the local midnight time is approximately 16:00. The observation period is from 3:00 to 22:00 in the local time. The field data is contaminated by a large amount of intermittent coherent noise.

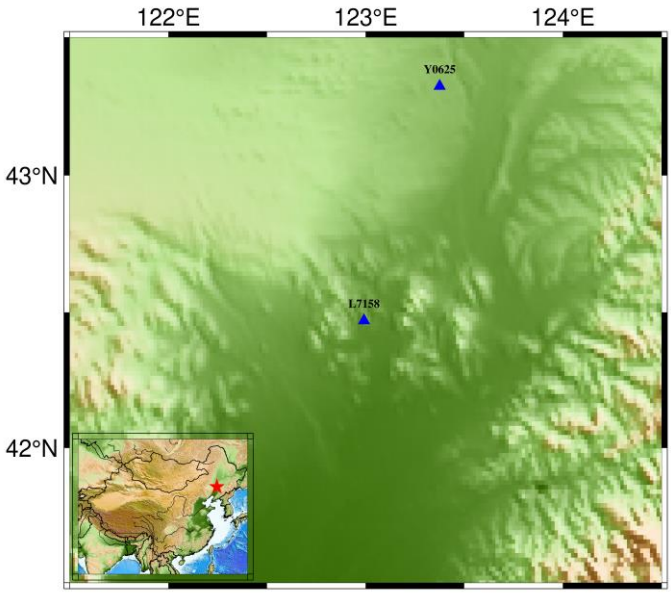

Fig. 4.6 The location map in china. Y0625 is the remote reference site, and L7158 is the local site. The map on the lower left shows the survey area in China. The red star denotes the local site.

A similar analysis with TNV48 is performed in the case study. The data quality analysis in the *x*-direction in 6.7 s is shown in Fig. 4.7; The red denotes the data with a high linear relationship, and the blue denotes a low linear relationship. The previous 2,500 data have a preferred polarization direction of around -30°, and there is a preferred direction in PD(E_x, H_x); moreover, the corresponding $Error_{sz}$ is lager. In conclusion, the data is dominated by coherent noise. And the coherent noise has a high linear relationship.

After analyzing the data quality in a different period, I find that most data is dominated by the coherent noise between 2 to 20 s. The field data is dominated by the incoherent noise between 20 to 100 s.

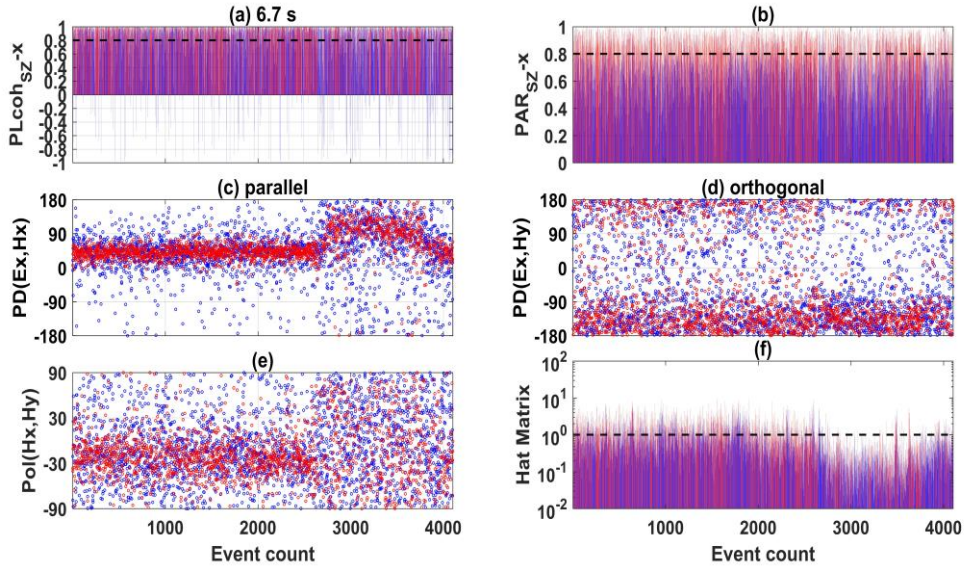

Fig. 4.7 The variation in the indices at L7-158 in the period of 6.7 s. Panel (a)-(b) show the variation in $PLcoh_{sz}$ and PAR_{sz} in the *x*-directions, respectively. The horizontal axis denotes the event count. The red color denotes the data that both $PLcoh_{sz}$–*x* and PAR_{sz}–*x* are higher than 0.8, and the other data are shown in blue. Panel (c)-(d) show the variation in PD(E_x, H_x) and PD(E_x, H_y) respectively. Panel (e) shows the variation in polarization direction for the magnetic field. Panel (f) shows the hat matrix's diagonal element variation, and the hat matrix's diagonal element is normalized by the expected value (2/*N*, *N* denotes the sample of data).

Fig. 4.8 shows the MT sounding curves calculated using the data observed at site L7-158. All the impedance result is biased because there is a rapid rise and fall between 2 to 20 s in the apparent resistivity. The M-estimator calculated the black curves using the single-site data processing is biased between the 2 to 100 s. The result between 20 to 100 s is improved after using the preselection strategy with $PLcoh_{sz}$ and PAR_{sz}. It may be that the incoherent noise dominates the data, and the preselection strategy removes the incoherent noise.

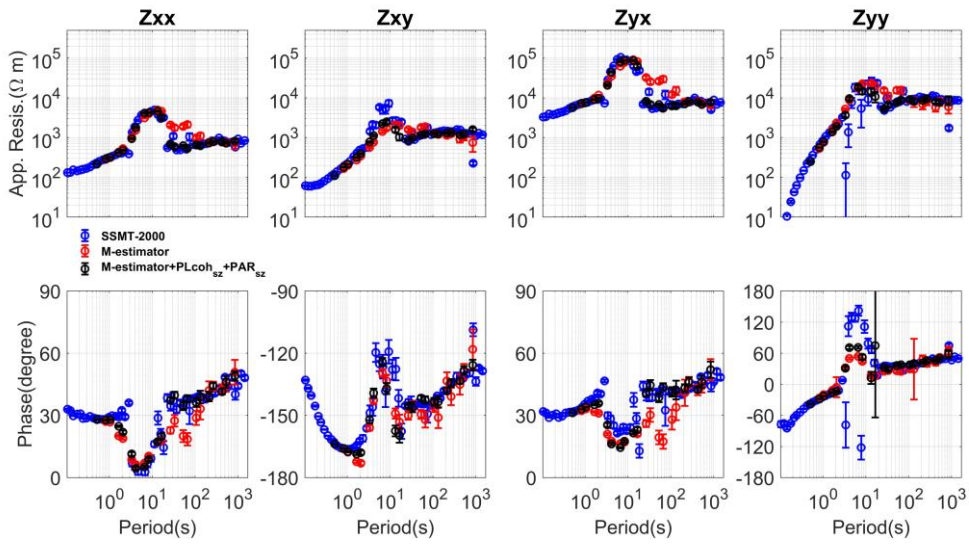

Fig. 4.8 The MT sounding curves calculated by the different methods using the data observed at site L7-158. The upper panels show the apparent resistivity, and the lower panels show the impedance phase. SSMT-2000 was used to calculate the blue curves. SSMT-2000 is one of the

standard Phoenix software sets. The M-estimator using single-site processing was used to calculate the red curves. The M-estimator using data preselection was used to calculate the black curves, $PLcoh_{sz}$ and PAR_{sz} are combined for the preselection strategy, and the threshold is set as 0.8 for both $PLcoh_{sz}$ and PAR_{sz}. The horizontal axis denotes the period in seconds.

To suppress the coherent noise, I need a different strategy. Fig. 4.9 shows the variation in polarization direction and the corresponding variation in DD_{pol} in 6.7 s and 33.4 s. the expected value of DD_{pol} is 1/3, and the data with a preferred polarization direction increases. It shows that the DD_{pol} is effective in detecting the data with a preferred direction of polarization direction.

Fig. 4.10 compares the MT sounding curves calculated by the M-estimator combining $PLcoh_{sz}$, PAR_{sz} and DD_{pol} for data preselection. The rapid rise and rapid fall are removed. It shows that the effectiveness of DD_{pol} for detecting the coherent noise with the strong polarization direction.

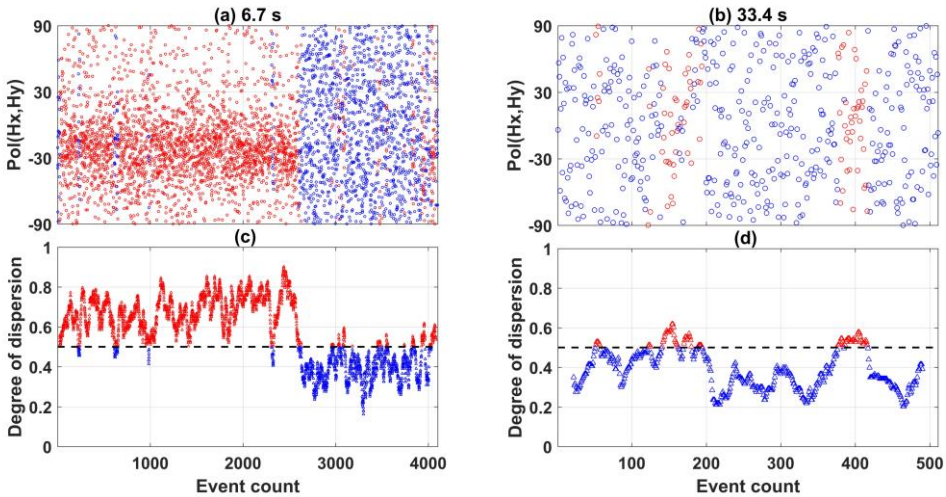

Fig. 4.9 The variation in polarization direction and the corresponding variation in DD_{pol} in the period of 6.7 s and 33.4 s. The upper panels show the variation in polarization direction, and the lower panels show the corresponding dispersion degree. The horizontal axis denotes the event count. The red color denotes the data whose dispersion degree is higher than 0.5, and the other data are shown in blue.

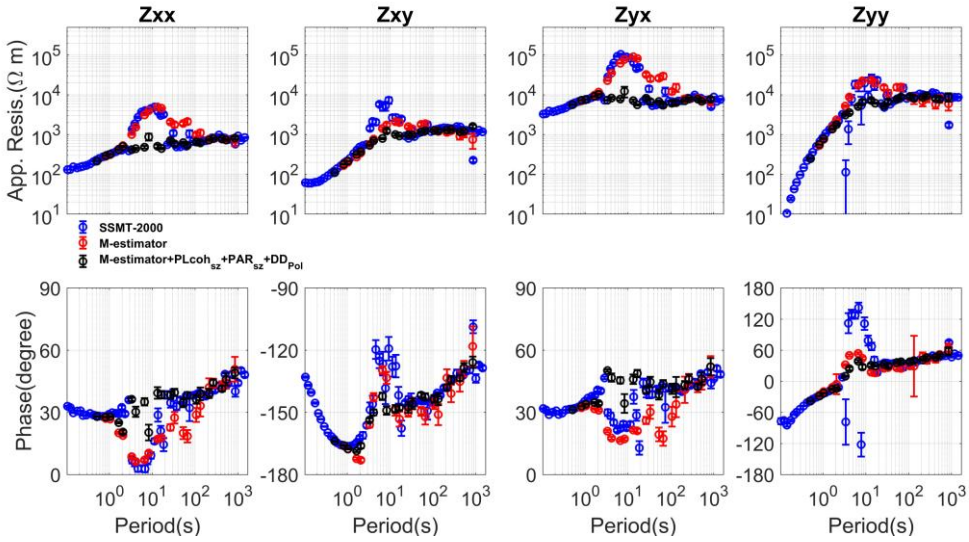

Fig. 4.10 The MT sounding curves calculated by the different methods using the data observed at site L7-158. SSMT-2000 was used to calculate the blue curves. The M-estimator using single-site processing was used to calculate the red curves. The M-estimator combining $PLcoh_{sz}$, PAR_{sz} and DD_{pol} for data preselection was used to calculate the black curves. The horizontal axis denotes the period in seconds.

4.4 The Case Study For Remote Site Data Processing

4.4.1 The case for the data in the 1-d/2-d structure and the data contaminated by the intermittent incoherent noise

The first Example used the data observed at TVN48 as the local site and ALW48 as the remote site. The observation period is from June 19 to June 24, 2015. Fig. 4.11 shows the data quality analysis in the 675 s; Fig. (a) and (b) show that the RLcoh is high and close to one in most periods. The ratio of high RLcoh data is 87%. Fig. (c) and (f) show that the parallel electric and magnetic fields are incoherent (Coh($\boldsymbol{E_x}$, $\boldsymbol{H_x}$)=0.34, Coh($\boldsymbol{E_y}$, $\boldsymbol{H_y}$)=0.22), and the PD($\boldsymbol{E_x}$, $\boldsymbol{H_x}$) and PD($\boldsymbol{E_y}$, $\boldsymbol{H_y}$) are scattered; Fig. (d) and (e) show that the orthogonal electric and magnetic fields are coherent (Coh($\boldsymbol{E_x}$, $\boldsymbol{H_y}$)=0.95, Coh($\boldsymbol{E_y}$, $\boldsymbol{H_x}$)=0.91), and there is a preferred direction of PD between the orthogonal electric and magnetic fields. Fig. (g) shows no polarization for the magnetic field. The inappropriate high values between 200 to 250 in the hat matrix diagonal elements in Fig. (h) correspond to the geomagnetic storm around June 23 (see Fig. 7 in the supplementary document). From the result of the data quality analysis, this data can be categorized into a 1-D/2-D model with good data quality in the period of 675 s.

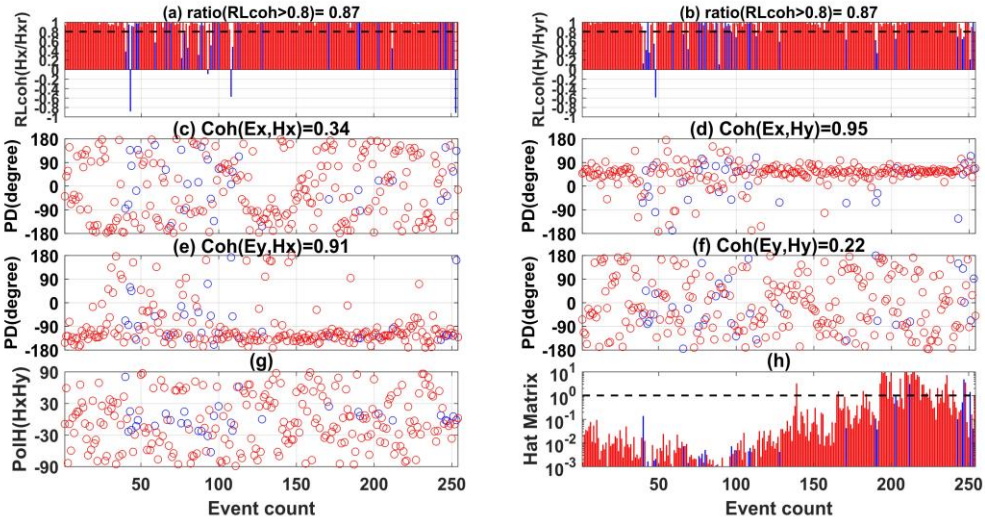

Fig. 4.11 The evaluation indices at TVN48 in the period of 675 s. Fig. (a)-(b) show the variation in RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) respectively. The red color denotes the data which both RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) are higher than 0.8, and the other data are shown in blue. The ratio of high RLcoh data is shown in the title of Fig. (a) and (b), and I can know the proportion of noisy data from the parameter. Fig. (c)-(f) show the variation in PD($\boldsymbol{E}_x$, $\boldsymbol{H}_x$), PD($\boldsymbol{E}_x$, $\boldsymbol{H}_y$), PD($\boldsymbol{E}_y$, $\boldsymbol{H}_x$) and PD($\boldsymbol{E}_y$, $\boldsymbol{H}_y$) respectively, and the corresponding coherence value calculated by the high RLcoh data is shown in the title of each figure. Fig. (g) shows the variation in polarization direction for the magnetic field. Fig. (h) shows the hat matrix diagonal element; the black dash line denotes the expected value (2/*N*).

Fig. 4.12 shows the data quality analysis in the period of 13 s. Fig. (a) and (b) show that the high RLcoh data only has 13%. Fig. (h) shows that the high RLcoh data corresponds to the high hat matrix's diagonal element. The geomagnetic storm causes that. Fig. (d) and (e) show that the PD($\boldsymbol{E}_x$, $\boldsymbol{H}_y$) and PD($\boldsymbol{E}_y$, $\boldsymbol{H}_x$) of the high RLcoh data have a preferred direction, and the coherence between the orthogonal electric and magnetic is high (Coh($\boldsymbol{E}_x$, $\boldsymbol{H}_y$)=0.97, Coh($\boldsymbol{E}_y$, $\boldsymbol{H}_x$)=0.95). Therefore, the high RLcoh data can be categorized into the 1-D/2-D case. For the low RLcoh data, the PD($\boldsymbol{E}_x$, $\boldsymbol{H}_y$) and PD($\boldsymbol{E}_y$, $\boldsymbol{H}_x$) is scattered. In conclusion, the data is dominated by incoherent noise during the noisy period. When there is a geomagnetic storm, the nature EM signal's strength increases, the SNR becomes high, and the RLcoh becomes high and close to one. On the other hand, there is no preferred polarization direction for the magnetic field in Fig. (g). It shows that the polarization direction is not suitable for detecting incoherent noise.

Fig. 4.13 shows the MT sounding curve calculated by the data observed at site TNV48. It includes the single-site processing, remote reference, and preselection strategy results. I used the RLcoh in the preselection strategy. The data in which the RLcoh is lower than 0.8 is removed before impedance estimation. The preselection strategy result coincides with the impedance result provided by the USArray team, downloaded by the IRIS(see Fig. B in the supplementary document). Here, I regard it as the true model. Fig. 4.13 shows that the result over 20 s coincides with the other. But the single-site processing result is severely down-biased lower than 20 s. The remote reference technique improved the situation but is also down biased compared with the preselection strategy result. The data quality analysis for each period showed that the data lower than 20 s is contaminated by a large amount of incoherent noise; only the data corresponding to the geomagnetic storm has a high SNR. That is why the single-site processing result is down-biased below 20 s. Only I extract the high SNR data by RLcoh can get the true model; even I used the remote reference technique also failed to get a reliable result.

This Example shows the superiority of the preselection strategy by using RLcoh.

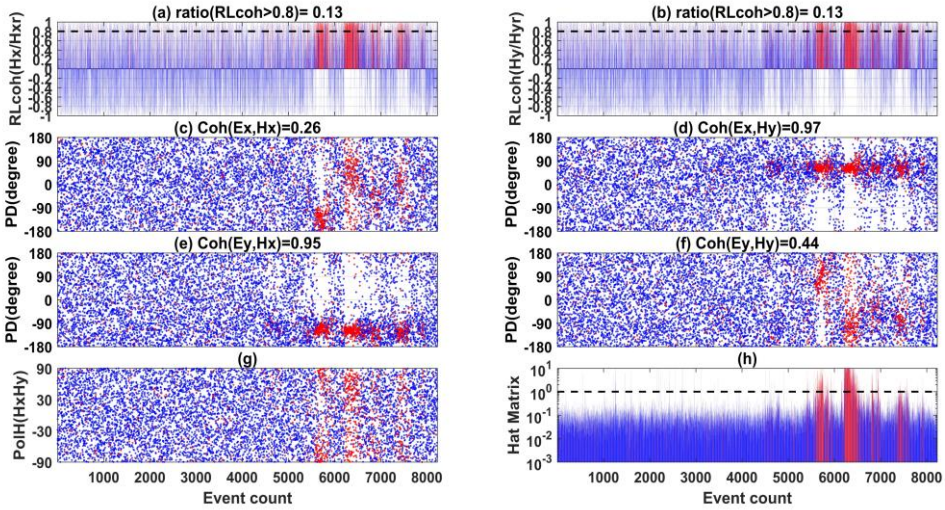

Fig. 4.12 The evaluation indices at TVN48 in the period of 13 s. Fig. (a)-(b) show the variation in RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) respectively. The red color denotes the data which both RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) are higher than 0.8, and the other data are shown in blue. The ratio of high RLcoh data is shown in the title of Fig. (a) and (b), and I can know the proportion of noisy data from the parameter. Fig. (c)-(f) show the variation in PD($\boldsymbol{E}_x$, $\boldsymbol{H}_x$), PD($\boldsymbol{E}_x$, $\boldsymbol{H}_y$), PD($\boldsymbol{E}_y$, $\boldsymbol{H}_x$) and PD($\boldsymbol{E}_y$, $\boldsymbol{H}_y$) respectively, and the corresponding coherence value calculated by the high RLcoh data is shown in the title of each figure. Fig. (g) shows the variation in polarization direction for the magnetic field. Fig. (h) shows the hat matrix diagonal element; the black dash line denotes the expected value (2/*N*).

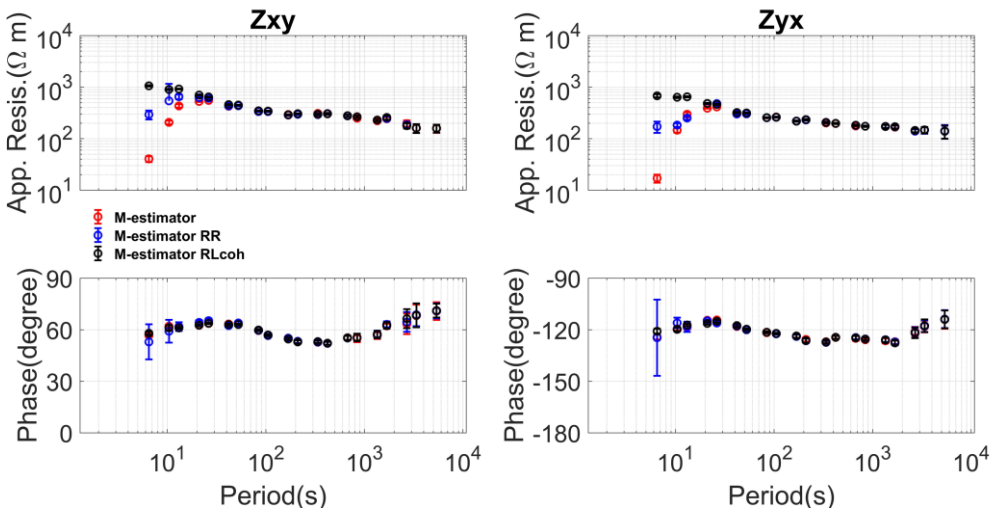

Fig. 4.13 The MT sounding curves calculated using the data observed at site TNV48. The upper figures show the apparent resistivity, and the lower figures show the impedance phase. The red curve denotes the single-site processing result, and the blue curve denotes the remote reference result. The black curve denotes the preselection strategy results using RLcoh. The threshold is set as 0.8, and the data lower than the threshold is removed before the impedance estimation.

4.4.2 The Case for the Phenomenon of Phase Rolling out of Quadrant and the Data Contaminated by the Continuous Incoherent Noise

The second Example used the data observed at WYH18 as the local site and MTG20 as the remote site. The observation period is from June 5 to June 8, 2009. Fig. 4.14 shows the data quality analysis in the 405 s. Fig. (a) and (b) show that the RLcoh is high and close to one most of the time. The high RLcoh data has 91%. Fig. (c) shows that the parallel electric and magnetic field is coherent (Coh($\boldsymbol{E_x}$, $\boldsymbol{H_x}$)=0.99); PD($\boldsymbol{E_x}$, $\boldsymbol{H_x}$) has a preferred direction. Fig. (g) shows that there is no polarization for the magnetic field. From the data quality analysis, I can categorize the data set into the phenomenon of PROQ in good data quality.

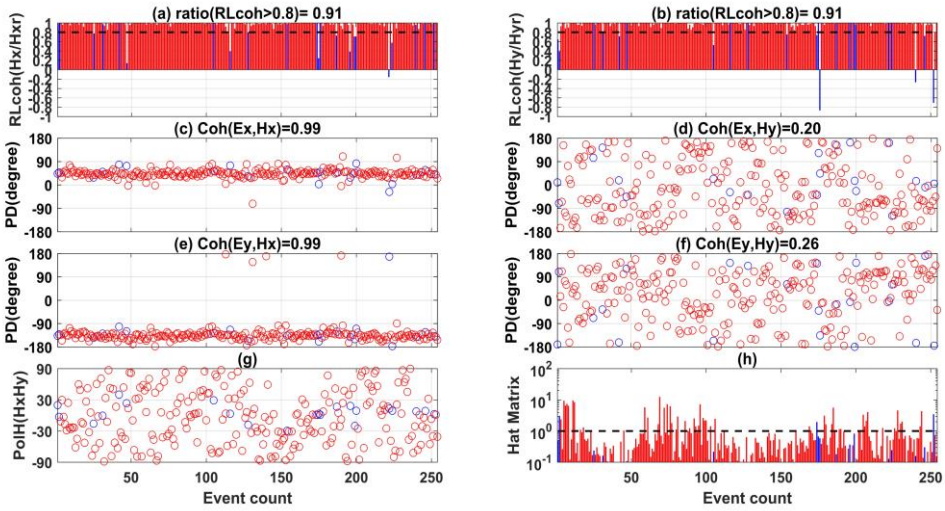

Fig. 4.14 The evaluation indices at WYH18 in the period of 405 s. Fig. (a)-(b) show the variation in RLcoh($\boldsymbol{H_x}$/$\boldsymbol{H_{xr}}$) and RLcoh($\boldsymbol{H_y}$/$\boldsymbol{H_{yr}}$) respectively. The red color denotes the data which both RLcoh($\boldsymbol{H_x}$/$\boldsymbol{H_{xr}}$) and RLcoh($\boldsymbol{H_y}$/$\boldsymbol{H_{yr}}$) are higher than 0.8, and the other data are shown in blue. The ratio of high RLcoh data is shown in the title of Fig. (a) and (b), and I can know the proportion of noisy data from the parameter. Fig. (c)-(f) show the variation in PD($\boldsymbol{E_x}$, $\boldsymbol{H_x}$), PD($\boldsymbol{E_x}$, $\boldsymbol{H_y}$), PD($\boldsymbol{E_y}$, $\boldsymbol{H_x}$) and PD($\boldsymbol{E_y}$, $\boldsymbol{H_y}$) respectively, and the corresponding coherence value calculated by the high RLcoh data is shown in the title of each figure. Fig. (g) shows the variation in polarization direction for the magnetic field. Fig. (h) shows the hat matrix diagonal element; the black dash line denotes the expected value (2/*N*).

Fig. 4.15 shows the data quality analysis in the 12 s. Fig. (a) and (b) show that the RLcoh is scattered and deviated from one. The ratio of high RLcoh data is only 4%. Fig. (b), (c), (d), and (e) show that the PD between the electric and magnetic fields is scattered. Fig. (g) shows no polarization for the magnetic field. From the data quality analysis, I think the continuous incoherent noise may contaminate the data in 12 s.

Fig. 4.16 shows the MT sounding curve calculated by the data observed at WYH18. The result below 30 s is biased even if I used the remote reference technique or data preselection strategy in this dataset. The data quality analysis for all periods shows that the continuous incoherent noise contaminates the period below the 30 s; I can not get a reliable result when the continuous noise contaminates the data. On the other hand, the MT sounding curve is smooth, and the error bar is small in a period larger than 30 s. While the phase of Z_{xy} rolls out of the normal quadrant. It is a phenomenon of PROQ. Moreover, I found that the phenomenon of PROQ appears in the sites DEM59, IDL13, RED38, REF59, REI55, REW51, SCY56, SDJ29, TTL24, VAQ57, VTE59, WAB05, WVQ53, WYH18, WYI18, and WYM24 from the USArray project; the characteristic of the phenomenon of PROQ is that there is a preferred

direction of PD between the parallel electric and magnetic components in one of the direction.

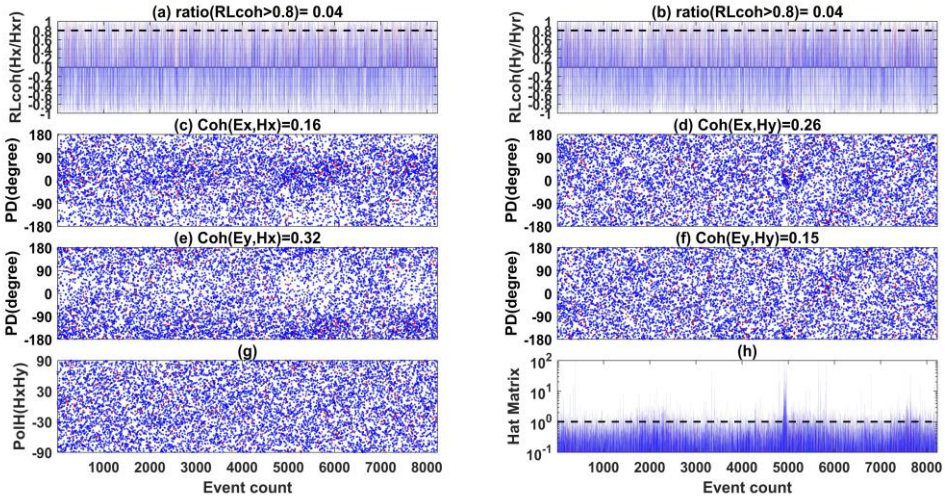

Fig. 4.15 The evaluation indices at WYH18 in the 12 s. Fig. (a)-(b) show the variation in RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) respectively. The red color denotes the data which both RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) are higher than 0.8, and the other data are shown in blue. The ratio of high RLcoh data is shown in the title of Fig. (a) and (b), and I can know the proportion of noisy data from the parameter. Fig. (c)-(f) show the variation in PD($\boldsymbol{E}_x$, $\boldsymbol{H}_x$), PD($\boldsymbol{E}_x$, $\boldsymbol{H}_y$), PD($\boldsymbol{E}_y$, $\boldsymbol{H}_x$) and PD($\boldsymbol{E}_y$, $\boldsymbol{H}_y$) respectively, and the corresponding coherence value calculated by the high RLcoh data is shown in the title of each figure. Fig. (g) shows the variation in polarization direction for the magnetic field. Fig. (h) shows the hat matrix diagonal element; the black dash line denotes the expected value (2/*N*).

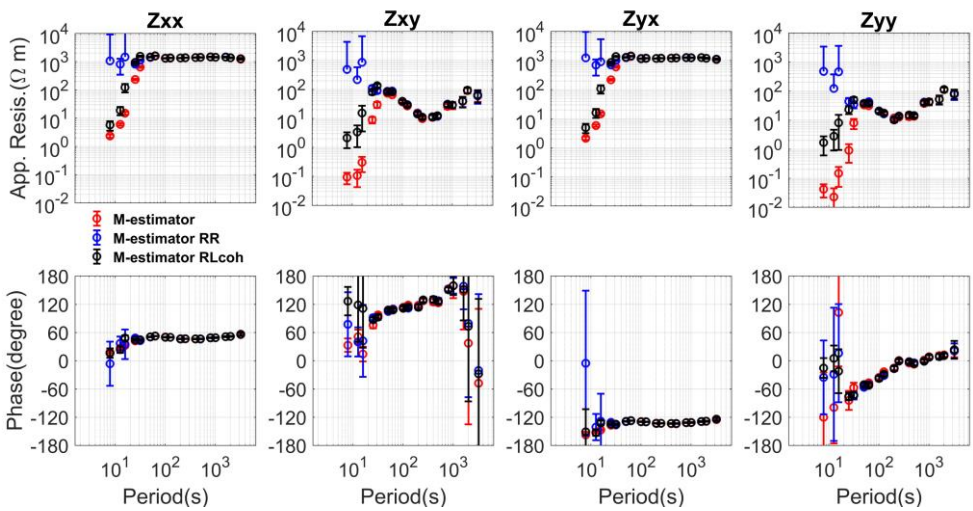

Fig. 4.16 The MT sounding curves calculated using the data observed at WYH18. The upper figures show the apparent resistivity, and the lower figures show the impedance phase.

4.4.3 The case for the data in the 2-d/3-d structure and the data contaminated by the intermittent coherent noise

The third Example used the data observed at site L7-158 in China on June 26, 2020. Fig. 4.17 shows the data quality analysis at L7-158 in the 6 s. Fig. (a) and (c) show that the high RLcoh

corresponds to the nighttime data from 16:00 to 20:00 in UTC, and its ratio is 34%. The low RLcoh data corresponds to the high hat matrix's diagonal element in Fig. (f) for the daytime data. There is a preferred direction of PD between the electric and magnetic fields close to 0° in Fig. (c) and (e). The coherent noise might cause this phenomenon. Coherent noise often appears as a spike or convex-like noise occurring simultaneously in the time domain between different channels. Moreover, there is a preferred polarization direction of around -30° for the magnetic field. In conclusion, the coherent noise contaminates the data during the daytime.

On the other hand, Fig. (c) and (e) show that the PD between the electric and magnetic fields concentrates in a narrow band in the nighttime. Moreover, the polarization direction for the magnetic field becomes scattered, and it can not be categorized into a 1-D/2-D case and PROQ. Therefore, I think the data is a phenomenon of the 2-D/3-D case for the nighttime data.

Fig. 4.18 shows the data quality analysis at L7-158 in the period of 437 s. Fig. (a) and (b) show that the ratio of high RLcoh data is 73%, and there is a preferred direction of PD between the parallel electric and magnetic fields in Fig. (c) and (e), but the coherence value is not so high (Coh($\boldsymbol{E}_x$, $\boldsymbol{H}_x$)=0.64, Coh($\boldsymbol{E}_y$, $\boldsymbol{H}_x$)=0.73). Moreover, there is no preferred polarization direction for the magnetic field. I can not categorize it into a 1-D/2-D case and PROQ. Therefore, the data can be categorized into a 2-D/3-D case.

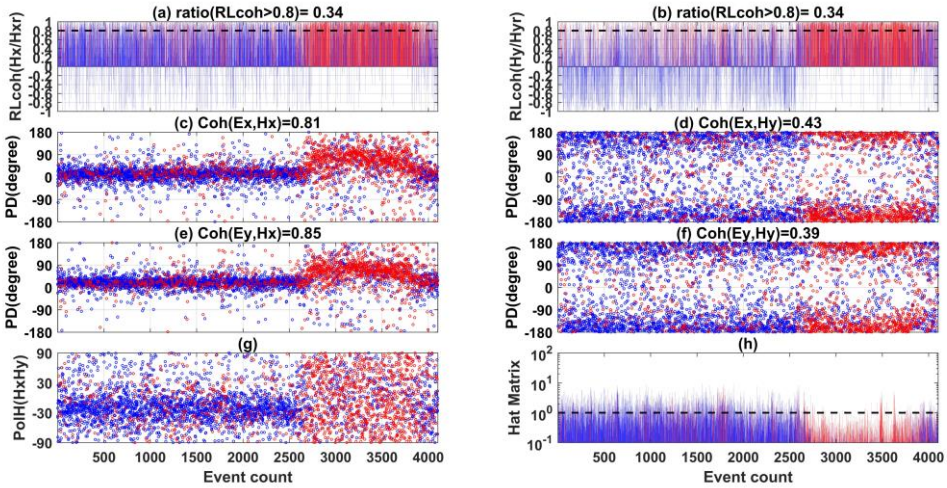

Fig. 4.17 The evaluation indices at L7-158 in the period of 6 s. Fig. (a)-(b) show the variation in RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) respectively. The red color denotes the data which both RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) are higher than 0.8, and the other data are shown in blue. The ratio of high RLcoh data is shown in the title of Fig. (a) and (b), and I can know the proportion of noisy data from the parameter. Fig. (c)-(f) show the variation in PD($\boldsymbol{E}_x$, $\boldsymbol{H}_x$), PD($\boldsymbol{E}_x$, $\boldsymbol{H}_y$), PD($\boldsymbol{E}_y$, $\boldsymbol{H}_x$) and PD($\boldsymbol{E}_y$, $\boldsymbol{H}_y$) respectively, and the corresponding coherence value calculated by the high RLcoh data is shown in the title of each figure. Fig. (g) shows the variation in polarization direction for the magnetic field. Fig. (h) shows the hat matrix diagonal element; the black dash line denotes the expected value (2/*N*).

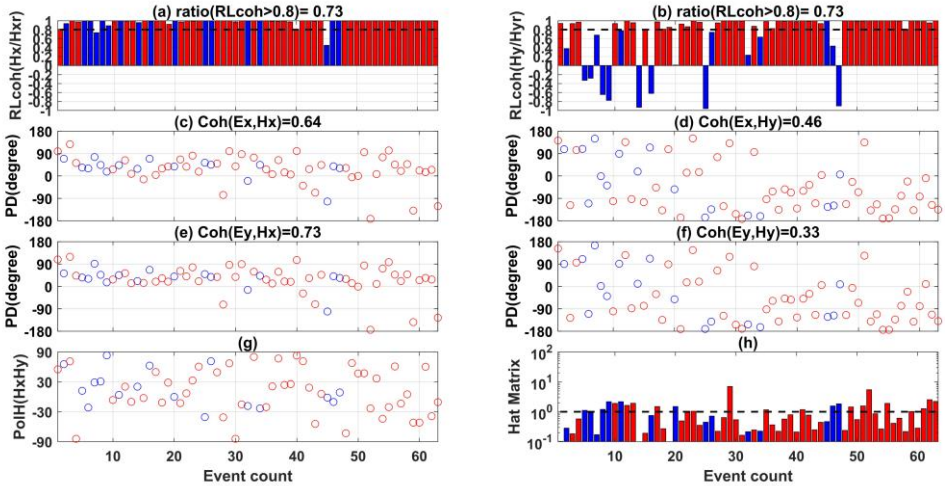

Fig. 4.18 The evaluation indices at L7-158 in the period of 427 s. Fig. (a)-(b) show the variation in RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) respectively. The red color denotes the data which both RLcoh($\boldsymbol{H}_x/\boldsymbol{H}_{xr}$) and RLcoh($\boldsymbol{H}_y/\boldsymbol{H}_{yr}$) are higher than 0.8, and the other data are shown in blue. The ratio of high RLcoh data is shown in the title of Fig. (a) and (b), and I can know the proportion of noisy data from the parameter. Fig. (c)-(f) show the variation in PD($\boldsymbol{E}_x$, $\boldsymbol{H}_x$), PD($\boldsymbol{E}_x$, $\boldsymbol{H}_y$), PD($\boldsymbol{E}_y$, $\boldsymbol{H}_x$) and PD($\boldsymbol{E}_y$, $\boldsymbol{H}_y$) respectively, and the corresponding coherence value calculated by the high RLcoh data is shown in the title of each figure. Fig. (g) shows the variation in polarization direction for the magnetic field. Fig. (h) shows the hat matrix diagonal element; the black dash line denotes the expected value (2/*N*).

Fig. 4.19 shows the MT sounding curves calculated using one-day and nighttime data at site L7-158. SSMT-2000 was used to calculate the blue curves using one-day data; SSMT-2000 is one of the standard Phoenix software sets. The M-estimator was used to calculate the red curves using the one-day data. The results calculated using the one-day data are biased. There is a rapid rise and fall between 5 to 50 s in the apparent resistivity. On the contrary, the M-estimator calculated the black curves using nighttime data recorded from 16:00:00 to 20:00:00 UTC on June 26. This MT sounding curve changes reasonably, and the rapid rise and fall are removed.

Fig. 4.20 shows the MT sounding curves calculated by the different methods using the one-day data. The single-site processing result is biased and is the same in Fig. 4.19. The remote reference curve is scattered, and the error bar is large. While the preselection result using the RLcoh becomes normal and is similar to the result calculated by the nighttime data in Fig. 4.19.

Comparing the result of Fig. 4.19, and 4.20, I should pay attention that the maximum period I can get from the preselection strategy is larger than the result obtained by the nighttime result. The 5th and 8th coefficient is used in this research, and the maximum length is set as 1/16 of the data length. The maximum period is 1/80 of the data length. The shorter the data length is, the lower the maximum period I can get. The maximum period I get from the nighttime data is around 200 s, and the maximum period obtained by the preselection strategy is around 900 s. Generally, the lower the period is, the natural EM signal strength is lower, between 1 to 1000 s. That means the local noise more easily influences the lower period. In this Example, the low-period data is noisy, while most long-period data is quiet. If I truncate the time series and use the nighttime data to calculate the impedance, I might lose the information in the long period. On the other hand, using the data preselection strategy can obtain a reliable result for a longer period in this case study.

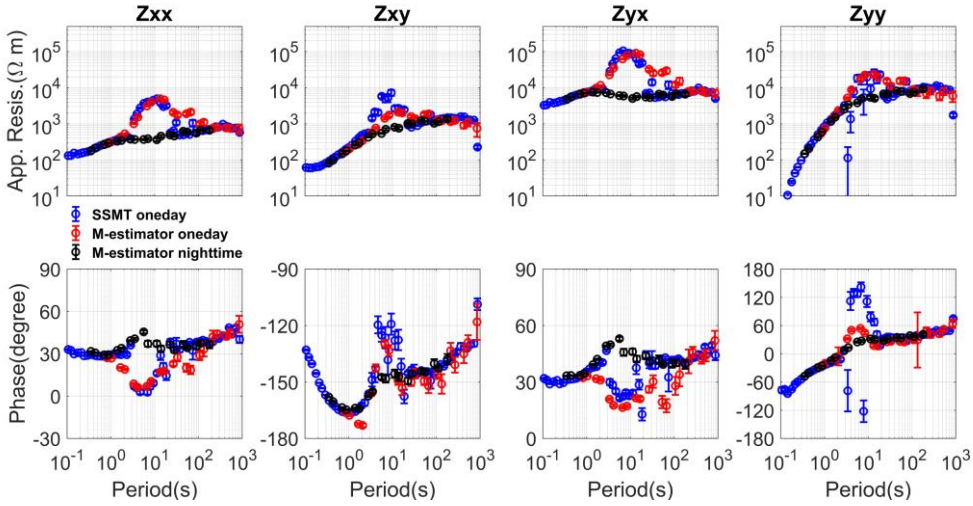

Fig. 4.19 The MT sounding curves calculated using the one-day and nighttime data at site L7-158. SSMT-2000 was used to calculate the blue curves using one-day data. The one-day and nighttime data were used to calculate the red and black curves by the M-estimator.

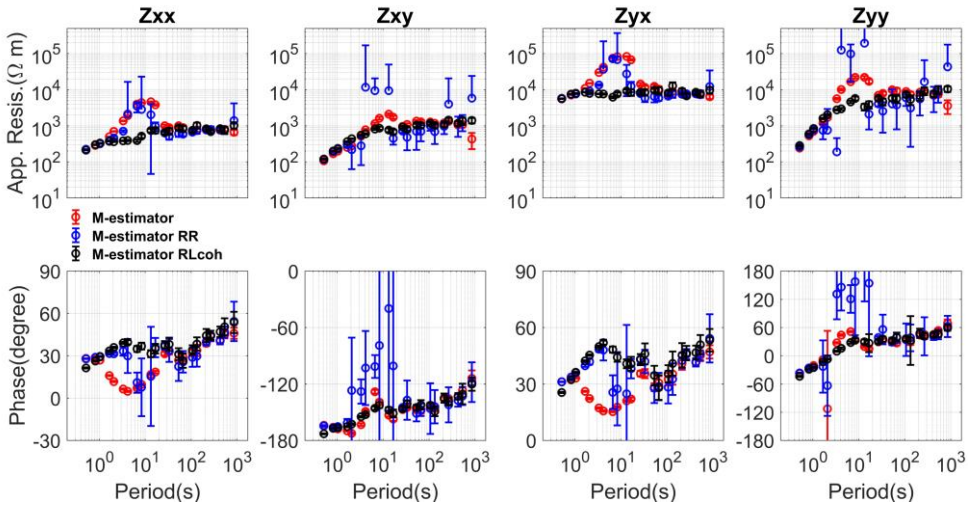

Fig. 4.20 The MT sounding curves calculated by the REFFT using the one-day data at site L7-158. The upper figures show the apparent resistivity, and the lower figures show the impedance phase. The colors are the same meaning as in Fig. 4.19.

4.5 DISCUSSION

At first, I discuss the relationship between impedance and coherence. According to the least-squares theory (Sims et al., 1971); for the north-south direction, Z_{xy} can be calculated as follows:

$$Z_{xy} = \frac{<E_x\bar{H}_y><H_x\bar{H}_x> - <E_x\bar{H}_x><H_x\bar{H}_y>}{<H_y\bar{H}_y><H_x\bar{H}_x> - <H_y\bar{H}_x><H_x\bar{H}_y>} = \frac{C-D}{E-F}, \tag{4.28}$$

For the numerator of Eq. 16, there is a relationship with the coherence as follows:

$$|C| = |< E_x\bar{H}_y >< H_x\bar{H}_x >| = coh(E_x, H_y)\sqrt{< E_x\bar{E}_x >< H_y\bar{H}_y >}\ < H_x\bar{H}_x >\ , \tag{4.29}$$

$$|D| = |< E_x\bar{H}_x >< H_x\bar{H}_y >| = coh(E_x, H_x)\sqrt{< E_x\bar{E}_x >< H_x\bar{H}_x >}\ coh(H_x, H_y)\sqrt{< H_x\bar{H}_x >< H_y\bar{H}_y >}, \tag{4.30}$$

$$\frac{|C|}{|D|} = \frac{|<E_x\bar{H}_y><H_x\bar{H}_x>|}{|<E_x\bar{H}_x><H_x\bar{H}_y>|} = \frac{coh(E_x,H_y)}{coh(E_x,H_x)coh(H_x,H_y)}, \tag{4.31}$$

For the denominator of Eq. 16, there is a relationship with the coherence as follows:

$$|E| = |< \boldsymbol{H_x}\bar{\boldsymbol{H}}_{\boldsymbol{x}} >< \boldsymbol{H_y}\bar{\boldsymbol{H}}_{\boldsymbol{y}} >| \tag{4.32}$$

$$|F| = |< \boldsymbol{H_y}\bar{\boldsymbol{H}}_{\boldsymbol{x}} >< \boldsymbol{H_x}\bar{\boldsymbol{H}}_{\boldsymbol{y}} >| = coh^2(\boldsymbol{H_x}, \boldsymbol{H_y}) \ < \boldsymbol{H_x}\bar{\boldsymbol{H}}_{\boldsymbol{x}} >< \boldsymbol{H_y}\bar{\boldsymbol{H}}_{\boldsymbol{y}} > \tag{4.33}$$

$$\frac{|E|}{|F|} = \frac{1}{coh^2(\boldsymbol{H_x}, \boldsymbol{H_y})} \tag{4.34}$$

Because the natural magnetic signals come from various sources, their incident directions vary with time, which means $\boldsymbol{H}_x$ and $\boldsymbol{H}_y$ are not coherent, and $Coh(\boldsymbol{H}_x, \boldsymbol{H}_y)$ is a small value. In the condition that the coherence value between the orthogonal component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$ is relatively high, while the coherence value between the parallel component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ is small. The C part dominates the numerator of Eq. 6.13, and the E part dominates the denominator. Therefore, in the case that $Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$ is high while $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ is small, the Z_{xy} can be rewritten as follows:

$$Z_{xy} \approx \frac{\boldsymbol{C}}{\boldsymbol{E}} = \frac{<\boldsymbol{E_x}\bar{\boldsymbol{H}}_{\boldsymbol{y}}>}{<\boldsymbol{H_y}\bar{\boldsymbol{H}}_{\boldsymbol{y}}>}. \tag{4.35}$$

In this situation, Z_{xy} is determined by the orthogonal electric and magnetic fields. The field data can be explained as the 1-D or 2-D cases.

On the contrary, if the coherence between the orthogonal component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$ is relatively low, while the coherence between the parallel component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ is high. There is no relationship of Eq. 4.35. In that case, the Z_{xy} is undetermined by the orthogonal electric and magnetic fields, and the phenomenon of PROQ may appear.

Next, I discuss how to analyze the data quality; I tried to categorize the field data into five types according to the coherence between the electric and magnetic fields, as shown in table 4.2. Type 1, 2 and 3 denote good data quality in different geological structures; types 4 and 5 denote the noisy data quality. I combined the PD between the electric and magnetic field, RLcoh, $PLcoh_{sz}$, PAR_{sz}, polarization direction for the magnetic field to identify the data type. From the data analysis, I can know the data quality varies with time and identify the dimensionality for the different periods.

The characteristic of the 1-D/2-D models is the low coherence between the parallel electric and magnetic fields and the high coherence between the orthogonal component. On the contrary, the characteristic of PROQ is the high coherence between the parallel electric and magnetic fields and the low coherence between the orthogonal component. On the other hand, the characteristic of the 2-D/3-D models is that the coherence between the orthogonal and parallel electric and magnetic fields is low.

There is a preferred direction of PD between the corresponding electric and magnetic fields when the coherence is high. In the 1-D/2-D or PROQ situations, there is always a preferred direction of PD between the corresponding electric and magnetic fields. Therefore, I can identify the data quality by finding a preferred direction of PD between electric and magnetic fields. On the other hand, the characteristic of the 2-D/3-D models is complex. There is no direct relationship between the electric and magnetic fields. When the SNR is high, the other situation of the 1-D/2-D and PROQ cases is categorized into 2-D/3-D models. Moreover, in the presence of coherent noise, the preferred direction of PD between the electric and magnetic fields tends to be 0° or 180°.

In the case of polarization, there is no preferred direction for the magnetic field in a quiet EM environment. If the magnetic field has a preferred polarization direction, I can consider that the data is contaminated by coherent noise in that period. However, when the data is contaminated by incoherent noise. The magnetic field has no preferred polarization direction. Therefore, the polarization direction is only effective in detecting the coherent noise.

From the perspective of a linear relationship, I can identify the noise with a low linear relationship by comparing the measured electric field ($\boldsymbol{E}$) and the predicted electric field ($\boldsymbol{E_p}$). I combined the $PLcoh_{sz}$ and PAR_{sz} confirm the linear relationship and suggest the situation that when both $PLcoh_{sz}$ and PAR_{sz} lagger that 0.8 can be regarded as a high linear relationship. The case study showed that the coherent noise and the good data have a high linear

relationship, and the incoherent noise corresponds to a low linear relationship.

RLcoh is a parameter to measure the similarity between the remote and local magnetic fields. Suppose there is a quiet remote reference site. I can use the RLcoh to measure the similarity between the local and remote sites to evaluate the data quality. When the SNR is high in the locale site, the RLcoh should be high and close to one. The RLcoh lies from -1 to 1; I suggest that the data larger than 0.8 can be regarded as high.

Table 4.2 The classification of the five data types and the recommending situation of the parameters in the north-south direction. Types 1, 2, and 3 denote the field data with good quality in different geological structures; types 4 and 5 denote the data contaminated by noise.

Classification	RLcoh	Liner relationship	Phase Difference (PD)	Polarization
Type 1: 1-D/2-D	≥ 0.8	High ($PLcoh_{sz}$>0.8; PAR_{sz}>0.8)	PD ($\boldsymbol{E}_x, \boldsymbol{H}_x$): scattered, PD ($\boldsymbol{E}_x, \boldsymbol{H}_y$): convergent	DD_{pol}<0.5
Type 2: PROQ	≥ 0.8	High ($PLcoh_{sz}$>0.8; PAR_{sz}>0.8)	PD ($\boldsymbol{E}_x, \boldsymbol{H}_x$): convergent, PD ($\boldsymbol{E}_x, \boldsymbol{H}_y$): scattered	DD_{pol}<0.5
Type 3: 2-D/3-D	≥ 0.8	High ($PLcoh_{sz}$>0.8; PAR_{sz}>0.8)	PD ($\boldsymbol{E}_x, \boldsymbol{H}_x$): scattered, PD ($\boldsymbol{E}_x, \boldsymbol{H}_y$): scattered	DD_{pol}<0.5
Type 4: Incoherent noise	< 0.8	Low ($PLcoh_{sz}$ <0.8; or PAR_{sz} <0.8)	PD ($\boldsymbol{E}_x, \boldsymbol{H}_x$): scattered, PD ($\boldsymbol{E}_x, \boldsymbol{H}_y$): scattered	DD_{pol}<0.5
Type 5: Coherent noise	< 0.8	High ($PLcoh_{sz}$>0.8; PAR_{sz}>0.8)	PD ($\boldsymbol{E}_x, \boldsymbol{H}_y$)/PD($\boldsymbol{E}_x, \boldsymbol{H}_y$): close to 0° or 180°	DD_{pol} >0.5

Next, I discuss the superiority of the selection strategy over remote reference processing. Remote reference processing can improve the estimator's performance by using cross-power spectra instead of auto-power spectra when performing regressions based on the least-squares estimator (Gamble et al., 1979). The Z_{xy} value can be calculated as follows:

$$Z_{xy} = \frac{\langle E_x\bar{H}_{yr}\rangle\langle H_x\bar{H}_{xr}\rangle - \langle E_x\bar{H}_{xr}\rangle\langle H_x\bar{H}_{yr}\rangle}{\langle H_y\bar{H}_{yr}\rangle\langle H_x\bar{H}_{xr}\rangle - \langle H_y\bar{H}_{xr}\rangle\langle H_x\bar{H}_{yr}\rangle}. \quad (4.36)$$

The field data consist of natural sources and local noise. The cross-power spectra are calculated as follows:

$$\langle \mathbf{A}\bar{\mathbf{B}}\rangle = \langle(\mathbf{A}^{\mathbf{S}} + \mathbf{A}^{\mathbf{N}})(\bar{\mathbf{B}}^{\mathbf{S}} + \bar{\mathbf{B}}^{\mathbf{N}})\rangle, \quad (4.37)$$

where $\mathbf{A}^{\mathbf{S}}$ is the signal in components **A** and $\mathbf{A}^{\mathbf{N}}$ is the noise component in **A**. The same notation is used for **B**. The cross-power spectra are calculated as follows:

$$\langle \mathbf{A}\bar{\mathbf{B}}\rangle = \langle \mathbf{A}^{\mathbf{S}}\bar{\mathbf{B}}^{\mathbf{S}}\rangle + \langle \mathbf{A}^{\mathbf{S}}\bar{\mathbf{B}}^{\mathbf{N}}\rangle + \langle \mathbf{A}^{\mathbf{N}}\bar{\mathbf{B}}^{\mathbf{S}}\rangle + \langle \mathbf{A}^{\mathbf{N}}\bar{\mathbf{B}}^{\mathbf{N}}\rangle. \quad (4.38)$$

Under the assumption that the noise is uncorrelated with the signal and noise, the average value of the cross-power spectra may be neglected. Thus, the following equation can be expressed:

$$\langle \mathbf{A}^{\mathbf{S}}\bar{\mathbf{B}}^{\mathbf{N}}\rangle = 0, \quad \langle \mathbf{A}^{\mathbf{N}}\bar{\mathbf{B}}^{\mathbf{S}}\rangle = 0, \quad \langle \mathbf{A}^{\mathbf{N}}\bar{\mathbf{B}}^{\mathbf{N}}\rangle = 0. \quad (4.39)$$

The average value of the cross-power spectra yields a good approximation to the true value.

$$\langle \mathbf{A}\bar{\mathbf{B}}\rangle = \langle \mathbf{A}^{\mathbf{S}}\bar{\mathbf{B}}^{\mathbf{S}}\rangle. \quad (4.40)$$

However, in the presence of a large amount of intermittent coherent noise, I cannot ensure the assumption of Eq. 4.40. Remote reference techniques rarely obtain a reliable result in such cases. Suppose there is a quiet period at the local site. It is better to extract the quiet data before the impedance estimation. When a quiet remote is available, the RLcoh is a good choice. Assuming that the remote site is unavailable, I can combine the $PLcoh_{sz}$, PAR_{sz}, and polarization direction for the magnetic field to extract the high SNR data.

At last, I want to discuss the superiority of the parameters for single-site data processing. On the one hand, I can identify the noise with a low linear relationship by comparing the measured electric field ($\boldsymbol{E}$) and the predicted electric field ($\boldsymbol{E_p}$). It is similar to considering the simple linear relationship as shown in Fig. 4.21 (a). when the linear relationship is high, the measured

output data is similar to the predicted output data. However, The complex number has two properties: amplitude and phase. Weckmann et al. (2005) and Travassos and Beamish (1988) used the predicted coherence defined by the amplitude. That parameter may miss the information on phase. In this study, I used the predicted linear coherence defined by phase difference to confirm the similarity in phase and used the predicted amplitude ratio to confirm the similarity in amplitude.

On the other hand, there is a problem that the ***Z*** might be biased in the presence of a large amount of noise. It is similar to considering the simple linear relationship as shown in Fig. 4.21. Only data set 1 has a high linear relationship in the six data sets, and the other data sets have a low linear relationship. If I calculated the model by all the data, the model would be biased, and the predicted data will be biased. To reduce the biased information from the impedance Z, I can divide the data into a short segment, and by our experience, 20 samples are chosen, and it works well.

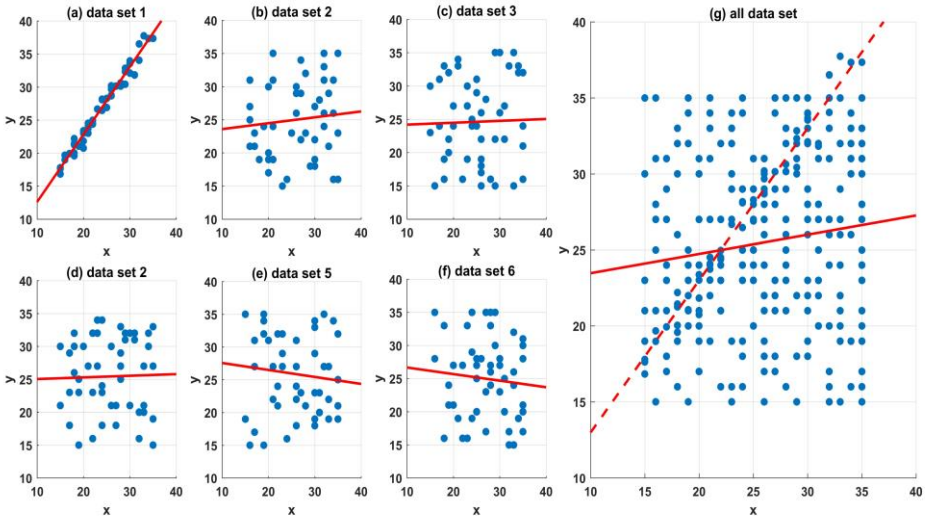

Fig. 4.21 Example of simple linear regression, Panel(a)-(f) show the examples of simple linear regression; each data set includes 50 samples, and the solid red line denotes the model calculated by the regression. Panel(g) shows the example of simple linear regression calculated by the 300 samples used in the previous data set. The solid red line denotes the model calculated by the regression, and the red dash line denotes the model calculated by the data set 1.

The effect of segmentation is shown in Fig. 4.26 using the field data at site L7158. It compares the predicted linear coherence and amplitude ratio corresponding with and without using the field data in 6.7 s at site L7158. From the data quality analysis of Fig. 4.9, I know the data is dominated by coherent noise, and the coherent noise has a high linear relationship. It shows that I can reduce misleading information from the impedance in a large amount of coherent noise by dividing the field into segments.

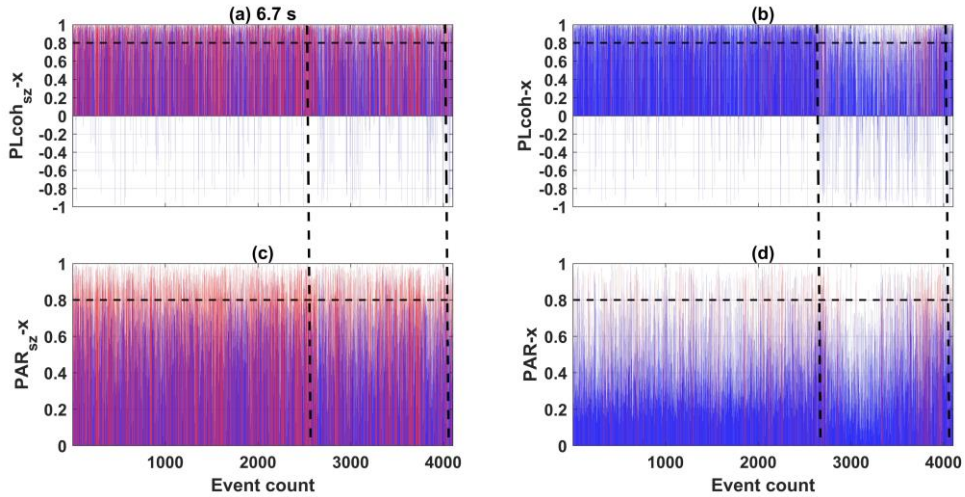

Fig. 4.22 Comparing the predicted linear coherence and amplitude ratio corresponding with and without using the segmented field data in the *x*-direction in 6.7 s. The red color denotes the data in which both the predicted linear coherence and amplitude ratio are higher than 0.8, and the other data are shown in blue. Pannel (a) and (c) shows the variation in $PLcoh_{sz}$ and PAR_{sz} with segmentation, respectively, Panel (b) and (d) show the variation in PLcoh and PAR without segmentation, respectively

4.6 CONCLUSIONS

The field data contain the signal and cultural noise. I showed the way to detect the noise from different perspectives as follows:

(1) The linear relationship

I combined the $PLcoh_{sz}$ and PAR_{sz} to confirm the linear relationship, the case study showed that the low linear relationship corresponds to the incoherent noise, and the high linear relationship can be a high signal-to-noise ratio (SNR) or strong coherent noise.

(2) The variation of PD between the electric and magnetic fields

The phase difference between the electric and magnetic fields is an important auxiliary tool to discuss the properties of the data. In the 1-D/2-D case, the orthogonal electric and magnetic fields have a preferred direction of PD, as shown in the first case study. In the 2-D/3-D case, there is no direct relationship between the electric and magnetic fields, but sometimes I can see the different properties from the noisy data, as shown in the second case study.

(3) The polarization direction for the magnetic to detect the coherent noise.

The magnetic field has no preferred polarization direction in a quiet EM environment. Suppose there is a preferred polarization direction for the magnetic field; I can consider that the coherent noise contaminates the data. I proposed a new parameter DD_{pol} to quantify the dispersion degree of the polarization directions, and it is effective to detect the coherent noise.

(4) The similarity between the local and remote site

Suppose there is a quiet remote reference site. I can use the RLcoh to measure the similarity between the local and remote sites to evaluate the data quality. When the SNR is high in the locale site, the RLcoh should be high and close to one. I recommend that RLcoh larger than 0.8 can be regarded as high.

I combined the PD between the electric and magnetic field, RLcoh, $PLcoh_{sz}$, PAR_{sz}, polarization direction for the magnetic field to discuss the data quality. From the data quality analysis, I can know the SNR varies with time and identify the dimensionality for the different periods. Usually, I can obtain sufficient samples in high-frequency data. It is suitable to apply a selection strategy to remove the contaminated data and obtain reliable results, especially for data

between the dead band (1-10 s). Local noise easily influences the data since the signal strength is very low.

This chapter introduced a robust estimator for data quality analysis and MT impedance calculation. In conclusion, I developed an enhanced robust based on the preselection strategy, as shown in Fig. 4.27. The robust estimator includes three steps. The first is the spectra analysis, a similar method to the Bounded Influence Remote Reference Processing code (BIRRP, Chave and Tompson, 2004) is adopted to extract the spectra in different frequencies. The second is the preselection strategy in the frequency domain; the third is the impedance estimation based on a robust M-estimator (Neukirch and Garcia, 2014). The robust estimator provides different parameters, for Example, $RLcoh$, $PLcoh_{sz}$ and PAR_{sz}, $Error_{sz}$, and DD_{pol} for preselection strategy. When I fail to get a reliable result from the robust estimator, I suggest confirming the background noise first and choosing a suitable parameter for the selection strategy. In theory, I can get a reliable impedance if the continuous noise doesn't contaminate the local site by the preselection strategy.

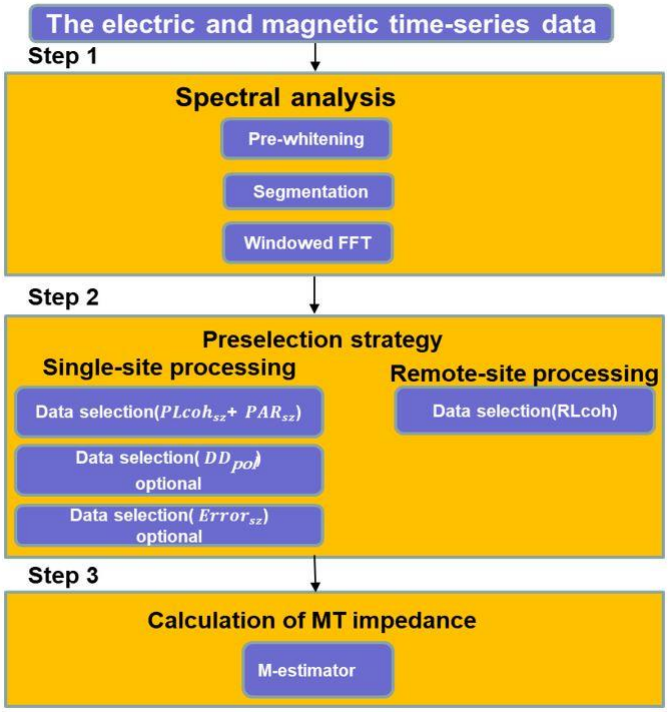

Fig. 4.27 The flow chart of the robust estimator using data preselection strategy. The robust estimator transformed the time-series data into the frequency domain by the windowed FFT and used a robust M-estimator (Neukirch and Garcia, 2014) to calculate the impedance. The robust estimator provides different parameters, the linear coherence between the remote and local magnetic fields (RLcoh), the predicted linear coherence ($PLcoh_{sz}$), the predicted amplitude ratio (PAR_{sz}), the dispersion degree of polarization direction (DD_{pol}), and the error between the predicted and observed electric field ($Error_{sz}$) to screen the data before impedance estimation.

Chapter 5: ROBUST ESTIMATOR BASED ON HILBERT HUANG TRANSFORM

In conventional MT data processing routines, the windowed FFT is used to transform the time series into the frequency domain. However, MT time-series data occasionally contain non-stationary (NS) signal and transient noise, which is not suitable for the basic requirements of conventional methods based on the Fourier transform. Impedance estimation aims to get the impedance tensor in the different frequency bands. I also can transform the time-series into the frequency domain data by the time-frequency transform technique like HHT. The estimator based on Hilbert-Huang Transform can estimate MT response functions even in NS signal and transient noise. Chen et al. (2012) used the instantaneous parameters calculated by HHT for impedance estimation. Neukirch and Garcia (2013) proved that instantaneous parameters could be used directly for impedance estimation in theory. Neukirch and Garcia (2014) also created a non-stationary processing routine named EMT. The impedance is calculated in the time-frequency domain as follows:

$$\begin{pmatrix} \sum_{n=0}^{N} E_{xh}(t_n)H_{xh}^{\dagger}(t_n) \\ \sum_{n=0}^{N} E_{yh}(t_n)H_{xh}^{\dagger}(t_n) \end{pmatrix} = \begin{pmatrix} Z_{xx}(\omega)\ Z_{xy}(\omega) \\ Z_{yx}(\omega)\ Z_{yy}(\omega) \end{pmatrix} \begin{pmatrix} \sum_{n=0}^{N} H_{xh}(t_n)H_{xh}^{\dagger}(t_n) \\ \sum_{n=0}^{N} H_{yh}(t_n)H_{xh}^{\dagger}(t_n) \end{pmatrix}, \tag{5.1}$$

and

$$\begin{pmatrix} \sum_{n=0}^{N} E_{xh}(t_n)H_{yh}^{\dagger}(t_n) \\ \sum_{n=0}^{N} E_{yh}(t_n)H_{yh}^{\dagger}(t_n) \end{pmatrix} = \begin{pmatrix} Z_{xx}(\omega)\ Z_{xy}(\omega) \\ Z_{yx}(\omega)\ Z_{yy}(\omega) \end{pmatrix} \begin{pmatrix} \sum_{n=0}^{N} H_{xh}(t_n)H_{yh}^{\dagger}(t_n) \\ \sum_{n=0}^{N} H_{yh}(t_n)H_{yh}^{\dagger}(t_n) \end{pmatrix}, \tag{5.2}$$

where $E_{xh}(t_n)$,$E_{yh}(t_n)$,$H_{xh}(t_n)$,$H_{yh}(t_n)$ denote the instantaneous values of the analytic signal corresponding to the electromagnetic field, which has the same instantaneous frequency ω at the same time point. t_n denotes the nth time point. The symbol † denotes the complex conjugate transpose. N is the total number of sampling points.

The initial technique for MT response function regression is based on the least-squares (LS) theory (Sims et al., 1971); it might be severely disturbed by cultural noise. Nevertheless, through the decades, there have been two main dramatic improvements. The first is the remote reference technique (Gamble et al., 1979). The second is the robust procedure (Egbert and Booker, 1986; Chave and Thomson, 1989; Smirnov, 2003; Chave and Thomson, 2004). The combination of them can obtain a more reliable result in general.

Table 5.1 shows the robust estimator for single-site data processing. The standard one is the maximum likelihood estimator (M-estimator). The robust M-estimator (EMTF; Egbert and Booker, 1986) gives a small weight to reject the outlier based on the residual between the measured electric field ($\boldsymbol{E}$) and the predicted electric field ($\boldsymbol{E_p}$), where $\boldsymbol{E_p}$=$\boldsymbol{ZH}$, $\boldsymbol{Z}$ is obtained by the least-squares estimator, and $\boldsymbol{H}$ is the measured magnetic field. The M-estimator can reduce the influence of unusual data in the electric field (outliers) but is not sensitive to exceptional input (magnetic field) data, termed leverage points. The bounded influence (BI) estimator combines the robust M-estimator with leverage weighting based on the hat matrix diagonal element. Chave (2004) showed that the BI-estimator would perform better than the M-estimator. The measure of an estimator's robustness is its breakdown point. An estimator's breakdown point is the proportion of incorrect observations that an estimator can handle before giving an incorrect result. These methods' breakdown point is no more than 40% - 50% (Chave and Thomson, 2004; Smirnov, 2003a). When a large amount of noise contaminates the data, those estimators might be biased.

The key point to getting a reliable impedance from the noisy data is detecting and removing the abnormal data before the impedance estimation. Herein, I introduce a robust estimator based on the Hilbert-Huang transform (REHHT), including the noise reduction in the time domain and the leverage selection in the frequency domain. It can maximumly reduce the influence of local noise.

Table 5.1 The robust estimator for single-site data processing

M-estimator	BIRRP (leverage weighting, Chave and Tompson, 2004)	**REHHT** (leverage selection, noise reduction in time-domain)
	EMTF (Egbert and Booker, 1986)	EMT (Neukirch and Garcia, 2014)
	Windowed FFT	HHT

5.1 The Hilbert-Huang Transform

Huang et al. (1998) proposed the Hilbert-Huang transform (HHT). It combines the empirical mode decomposition (EMD) and the Hilbert spectral analysis.

EMD is a technique to adaptively decompose a signal into a finite number of intrinsic mode functions (IMFs) with a trend using a process called the sifting algorithm. IMFs are time-varying mono-component functions that contain a single frequency. To obtain meaningful IPs, the IMFs have to satisfy the following two conditions (Huang et al., 1998):

(1) in the whole dataset, the number of extrema and the number of zero-crossings must either equal or differ at most by one
(2) at any point, the mean value of the envelope defined by the local minima is zero

Based on these two conditions, a signal x(t) is decomposed into IMFs through an iterative process termed sifting (Huang et al., 1998) that is described as follows:

(1) Identify the local extrema(maxima and minima) of x(t).
(2) Interpolate all local maxima and local minima by cubic splines to obtain upper and lower envelopes.
(3) Subtract the mean m(t) of the upper and lower envelopes from x(t) to obtain the signal $h(t) = x(t) - m(t)$.
(4) If h(t) fulfils the two conditions of IMF, the first IMF $c_1(t) = h(t)$. Otherwise, steps 1–3 are applied on h(t) (instead of x(t)) and are repeated i-1 times until $h_i = h_{(i-1)}(t) - m_i(t)$ fulfills the conditions of IMF. $c_1(t) = h_i(t)$ is considered as the first IMF.
(5) Subtract the first IMF $c_1(t)$ from x(t) and treat the obtained signal as a new time-series $\tilde{x}(t) = x(t) - c_1(t)$.
(6) Steps 1 to 5 are repeated to determine the subsequent IMFs from the remaining signal $\tilde{x}(t)$. The decomposition process is stopped after n iterations when the residue becomes a monotonic function, which can not be decomposed further. The final signal can then be expressed as a sum of all IMFs and the residual.

The second step of HHT is the Hilbert transform; the complex conjugate y(t) of any real-valued function *x(t)* can be determined as:

$$y(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau\,, \tag{5.3}$$

Then the time complex number can be created as:

$$z(t)=x(t)+i*y(t)=a(t)*e^{i\theta(t)}\,, \tag{5.4}$$

where $a(t)=\sqrt{x(t)^2+y(t)^2}$, $\theta(t)=arc\ tan(\frac{y(t)}{x(t)})$. Here, *a(t)* is the instantaneous amplitude, and θ is the instantaneous phase function. The instantaneous frequency is simply defined as:

$$f(t)=\frac{1}{2\pi}\frac{d\theta(t)}{dt}. \tag{5.5}$$

Finally, I apply the Hilbert transform to each IMFs and get the IPs in the time-frequency domain.

In the first test, the HHT was applied to synthetic data combining the following equations to show the property of the Hilbert-Huang Transform;

$$x_1(t)=sin(2\pi\cdot 50t)+sin(2\pi\cdot 200t),$$
$$x_2(t)=sin(2\pi\cdot 25t)+sin(2\pi\cdot 100t)+sin(2\pi\cdot 250t). \tag{5.6}$$

The sampling rate is 1,000 Hz, and each segment is 4 seconds. Both of the time segments are added with 10% Gaussian noise. The combination of the two segments is shown in Fig. 5.1. The HHT spectrum is shown in Fig. 5.2; the HHT could express the spectrum variation in the time-frequency domain.

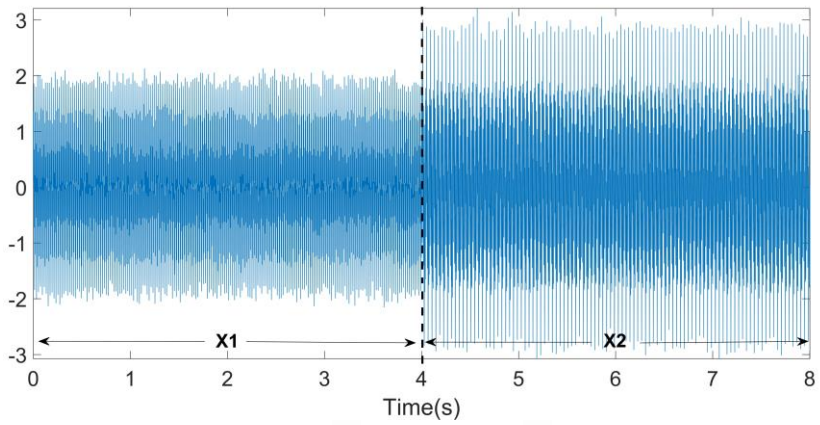

Fig. 5.1 The time-series $x_1(t) = sin(2\pi \cdot 50t) + sin(2\pi \cdot 200t)$ and $x_2(t) = sin(2\pi \cdot 25t) + sin(2\pi \cdot 100t) + sin(2\pi \cdot 250t)$. Both of the segments are added with 10% Gaussian noises.

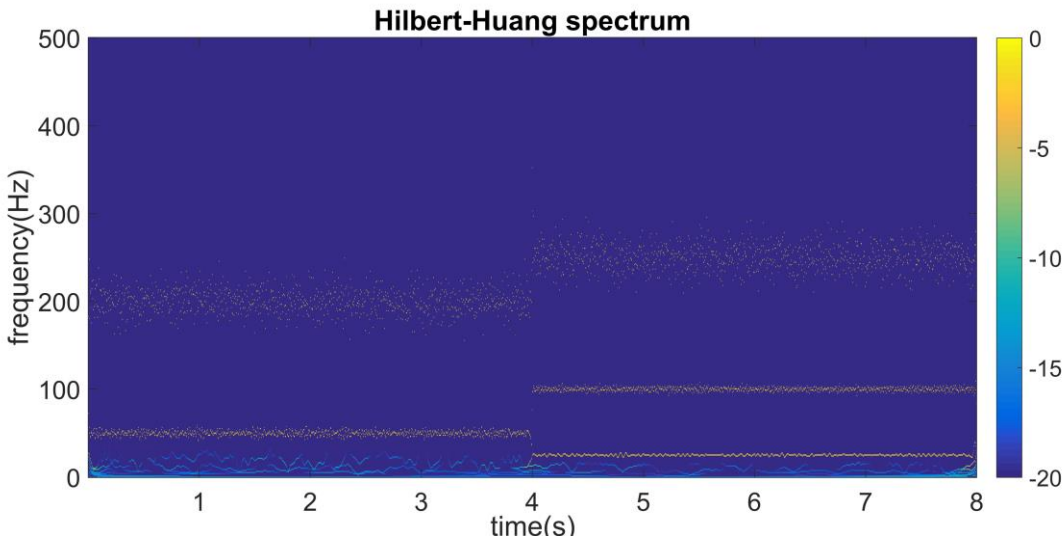

Fig. 5.2 The HHT spectrum of time-series data, shown in Fig. 1. The color denotes the value of 10·log10(amp/max), where "amp" denotes the amplitude of the IPs, and "max" denotes the maximum of the amplitude of the IPs.

5.2 Robust Estimator based on the Hilbert-Huang Transform

Fig. 5.3 shows the workflow chart to extract the electrical and magnetic field spectra into different frequency bands. The conventional impedance estimator transfer the time series into the frequency domain using the windowed FFT. I adopt the time-frequency transform technique HHT (Huang et al., 1998) to transform the time series into the frequency domain. HHT combines the empirical mode decomposition (EMD) and the Hilbert transformation. Huang et al. (1998) only presented their technique to univariate data; however, the MT data have at least four data channels, which depend on each other. Neukirch and García (2014) argued the advantage of multivariate empirical mode decomposition (MEMD). The MEMD (Rehman and Mandic, 2009) is adopted to decompose the time series into a finite number of intrinsic mode functions (IMFs),

representing its inherent oscillatory mode. For the MT time series, there are generally seven channels, e_x, e_y, h_x, h_y, h_z, r_x, r_y, local five components and two reference components, here [r_x, r_y] can be the local time series or remote time series, depending on which channel set as the reference channel; N × 7 × M matrix can be obtained by the MEMD, where N denotes the length of the time series, 7 denotes the seven-channels, M denotes the M modes of IMFs. Fig. 5.4 shows the MEMD result obtained by the 10-minutes local five components MT time series. The local magnetic field is set as the reference data. The MEDM obtains 12 modes of IMFs.

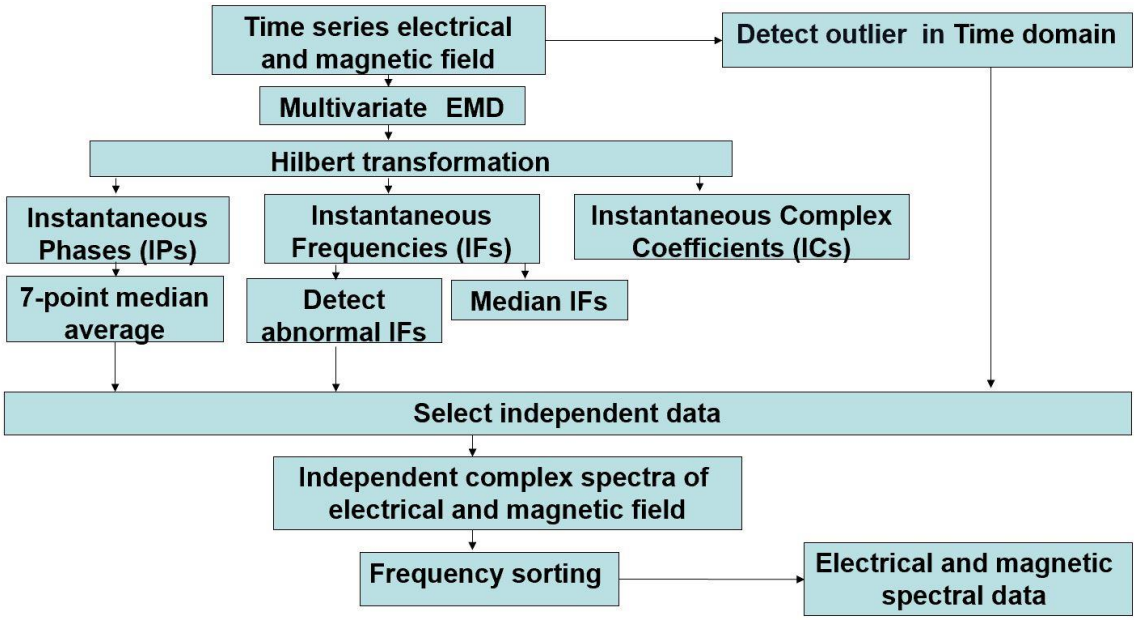

Fig. 5.3 The flow chart for extracting the electrical and magnetic field spectra into different frequency bands.

After HHT, I can get a set of instantaneous parameters, instantaneous spectra (IS), instantaneous frequencies (IFs), instantaneous phases (IPs), and instantaneous amplitudes (IAs). The instantaneous parameters are used to calculate the impedance. The key point to getting a reliable impedance is to remove the noisy data before the regression step. The procedure of the REHHT method is as follows:

(1) Transform the time-series data into the frequency-domain data by HHT
(2) Detect the outlier in the time and frequency domain, and remove the corresponding instantaneous parameters.
(3) Extract the independent instantaneous parameter into a different frequency band. I select the independent data point based on the locale maximum IPs. And then, group the independent data point into different frequency bands, typically 5 to 10 bands per decade.
(4) Remove the Leverage points by the statistical analysis of the diagonal element of the hat matrix.
(5) Estimate MT impedances by iterative least-squares regression with the Huber weight function.

This section mainly introduces the detail from step 2 to step 4.

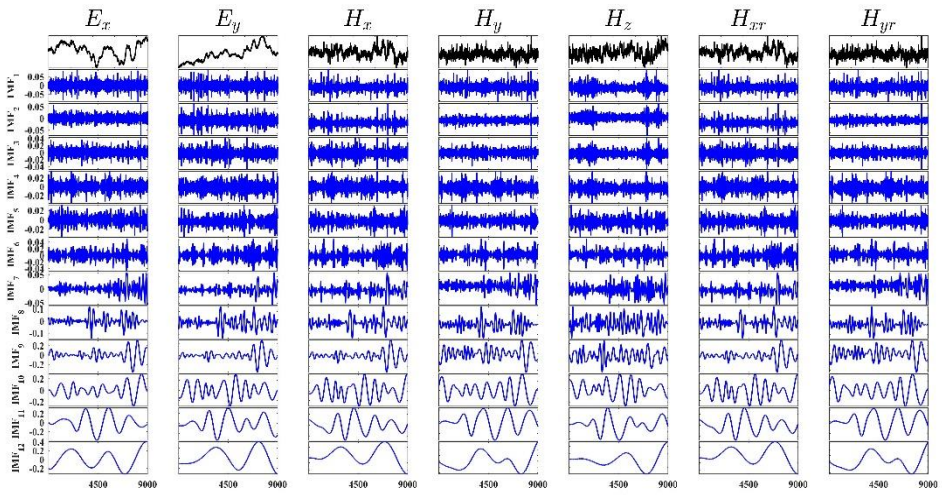

Fig. 5.4 The MEMD result obtained by the 10-minutes MT time-series data. The local magnetic field is set as the reference data, and the sampling rate is 15 Hz. The black lines denote the MT time-series data; the blue lines denote the IMFs. The horizontal axis denotes the sample number.

5.2.1 Detection of outliers in the time domain

Detecting the outlier in the time domain is an optional step. Hampel filter (Liu et al., 2004; Suomela, 2014) is useful for detecting the transient signals in the time domain. Because the transient high amplitude signal corresponding to the lighting is the signal source in the high-frequency band, the Hampel filter is adopted to detect the outlier only in the low-frequency band. The median for each sample x_i with its surrounding k samples is calculated as follows:

$$m_i = median(x_{i-k}, x_{i-k+1}, \ldots, x_i, \ldots, x_{i+k-1}, x_{i+k}). \quad (5.7)$$

In this research, the k is set to be 50. The median absolute deviation (MAD) of each sample's median (m_i) is calculated as follows:

$$\sigma_i = 1.48 median(|x_{i-k} - m_i|, \ldots, |x_{i+k} - m_i|), \quad (5.8)$$

If a sample differs from the median more than three times *MAD*, the sample is replaced with the median (m_i) and the corresponding instantaneous parameters will be removed before the step of extracting the independent data. Fig. 5.5 shows the example of the Hampel filter. In this way, I can reduce the influence of the outlier on the surrounding data when I decompose the time series by the MEMD. I will introduce the procedure by the example in Fig. 5.7 and show the effectiveness of the Hampel filter in the simulation.

Moreover, if a quiet remote reference site is available, I could further detect the abnormal data by the relationship between the local and remote magnetic fields. The local and remote magnetic field comes from the same source; they should have high similarity (Gang et al., 2018; Kappler, 2012), and the correlation value should be high in the absence of local noise. The time series is divided into a short segment, for example, a 1-minute length. The Pearson correlation between the local and remote magnetic fields below the threshold is flagged as an abnormal segment. The corresponding instantaneous parameter will be removed before extracting the independent parameter. As an option, the variance ratio between the local and remote magnetic field data can also detect the abnormal data following the method proposed by Kappler (2012).

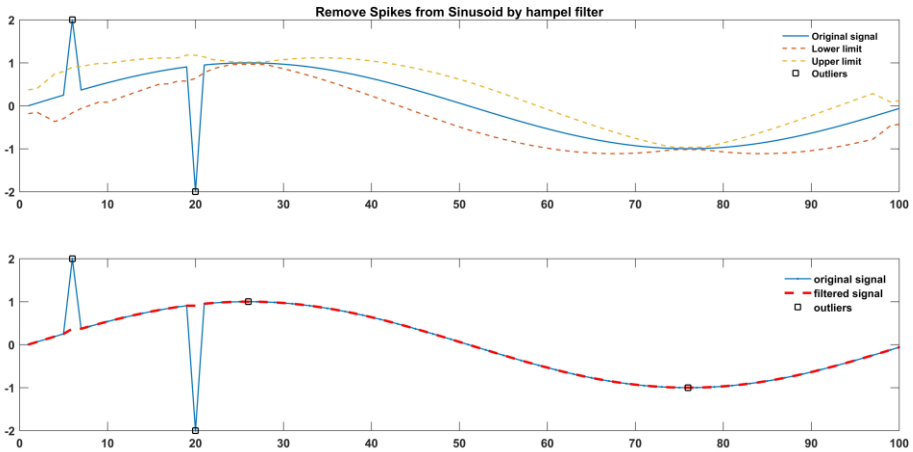

Fig. 5.5 The example to detect the outlier by the Hampel filter. The blue line denotes the time series in the upper figure, including the outlier. The yellow and brown dash line denotes the upper and lower limits. The red dash line denotes the time series after the hamper filter in the lower figure.

5.2.2 Detection of outliers based on the variance of the instantaneous frequencies

Neukirch and Garcia (2014) argued that the instantaneous frequencies (IFs) obtained from the same mode IMFs between different channels can be considered the same. But in the field data, the IFs are different between different channels sometimes; some incoherent noise or marginal effect can cause it. Herein, I will detect the outlier based on the variance of the IFs between different channels.

The variance of the IFs in the same mode between the seven channels is calculated first. Then, the instantaneous parameters, in which the variance of IFs is larger than three times the median absolute deviation (MAD), will be removed.

5.2.3 Extraction of independent instantaneous parameters and grouping them into different frequency bands

MEMD decomposes the time series into the IMFs by an iterative sifting procedure based on the local extrema. The IMF can't have more degrees of freedom than the number of extrema. All the data points between two extrema are dependent and regarded as one data. Fig. 5.6 shows the most simple IMF $\sin(2\pi\cdot 5t)$, the sampling rate is 1,000 Hz, and the segment is 2 seconds. The middle figure shows the IAs, and the lower figure shows the IPs. Here the IPs are biased at the beginning and end (the sine wave's Phase should be 0°) because of the marginal effect. There is always one maximum and minimum in the IPs between two extrema. The independent instantaneous parameters will be extracted based on the local maximum point of IPs. The IPs will be affected by the transient noise or the marginal effect and appear as the sawtooth variation, influencing the data selection and introducing some noise. I found that the seven-point median average filter was a good way to counteract this problem. After removing the abnormal instantaneous parameters introduced in the previous step, the independent instantaneous parameters will be extracted based on the local maxima point of the IPs.

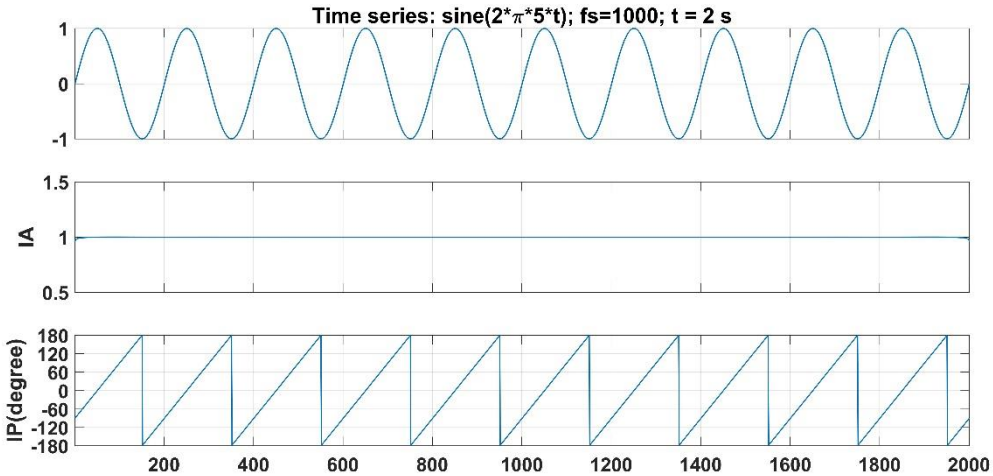

Fig. 5.6 The time-series data of sin(2π·5t) and the corresponding variations in IAs and IPs.

Fig. 5.7 shows the procedure of extracting the independent instantaneous parameters. Fig. 5.7a shows the instantaneous period of the seven components in the 9th mode of IMFs in Fig. 5.4. The black line denotes the median of the seven channels' period, and it will be regarded as the common instantaneous period of the seven channels. Fig. 5.7b shows the IPs of the Hx component. Some missing points are the removed data, which are flagged as abnormal data based on the variance of IFs. Most of the boundary data is moved, showing the effectiveness of noise reduction based on the variance of IFs. Fig. 5.7c shows the IAs of the Hx component. Fig. 5.7d shows the Hx time series.

And then, I add the convex-like noise to the Hx component between 5,000 and 6,000 points, shown as the red line in Fig. 5.7h; the corresponding IAs are shown the Fig. 5.7g in the same color. The IAs deviated from the original line (blue line). Fig. 5.7e shows the instantaneous period of the seven components in the same mode in Fig. 5.7a. Fig. 5.7f shows the IPs of the Hx component. The missing points are the removed data, flagged as abnormal based on the IFs and Hample filter (the missing data between 5,000 and 6,000 points). The time-series data modified by the Hampel filter is shown as the black line in Fig. 5.7h; the corresponding IAs are shown the Fig. 5.7g in the same color. And the curve is almost the same as the original curve, between 5,000 and 6,000. Here, I can see the effectiveness of the Hampel filter in reducing the influence of the outlier on the surrounding instantaneous parameter. But the curve is different from the blue line between 8,000 and 9,000. The Hampel filter modified the signal in other channels in that span. The channel influences each other during the MEMD; therefore, the black line deviates from the original data.

After extracting the independent data based on the local maxima point of the IPs, I will group the instantaneous parameters into different frequency bands based on the IFs. The median value of IFs obtained from the same mode IMFs in different channels is regarded as the common IF. As the impedance tensor changes smoothly with frequency, I can group the instantaneous parameters into the same frequency band, typically 5 to 10 bins per decade, to further processing.

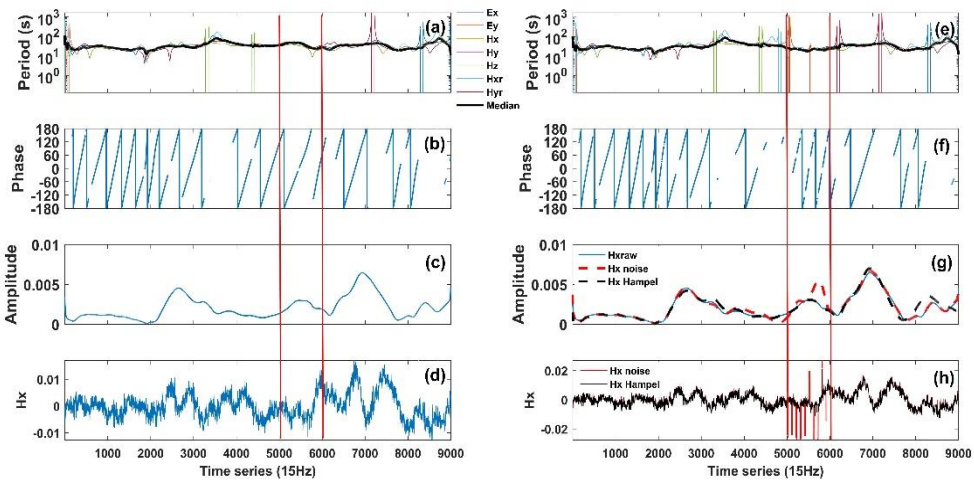

Fig. 5.7 The procedure for extracting the independent instantaneous parameters.

5.2.4 Leverage selection by hat matrix

The initial technique for MT impedance estimation is the least-squares (LS) estimator (Sims et al., 1971); it can be severely disturbed by cultural noise. The M-estimator (Egbert, 1986) gives weight to rejecting the outlier in the electric field. M-estimator can reduce the influence of unusual data (outliers) in the electric field but are not sensitive to exceptional input (magnetic field) data, termed leverage points. The hat matrix is widely used to detect unusual input data. The hat matrix is *N* by *N* matrix and is defined as follows:

$$\boldsymbol{H}_{hat} = \boldsymbol{H}(\boldsymbol{H}^{\dagger}\boldsymbol{H})^{-1}\boldsymbol{H}^{\dagger}, \tag{5.9}$$

where $\boldsymbol{H}$ is the N by two horizontal magnetic field components matrices at a specific frequency. The superscript † denotes the complex conjugate transpose. Chave and Thomson (2003) suggested that the hat matrix's diagonal element, more than several times 2/N (N denotes the sample of data), is problematic. As the noisy data is energetic, the hat matrix's diagonal element's statistical analysis is useful for detecting the noisy data. Before the estimation of impedance, the leverage point should be removed. The data where the hat matrix's diagonal element is more than 2/N will be removed.

5.3 The Strategy to Calculate the Broadband Time-series Data Based On REHHT

The strategy of processing the broadband MT time-series data by RMHHT is demonstrated in this section. The data observed by the Phenix geophysical instruments were used. The time-series data are sampled at three different sampling rates. First, the high-frequency band (2,400 Hz) was sampled for 2 seconds every 10 minutes. Second, the middle-frequency band (150 Hz) was sampled for 16 seconds every 10 minutes. Finally, the low-frequency band (15 Hz) was sampled continuously. Both the high-frequency and the middle-frequency bands were sampled intermittently. Table 5.2 shows the strategy for calculating the MT impedance for the full frequency band.

I transform time-series data into the frequency domain by the windowed FFT and group the target spectra in the frequency domain by the conventional method. In contrast, the HHT can express the spectrum variation well in the time-frequency domain. Therefore, I can combine the intermittent time-series data directly and process the discontinuous data in the same way as the continuous time-series data. The 2,400 Hz data was used to calculate the period between 4/2400 to 4/150 seconds. One continuous segment (2 seconds) contains 75 windows of the maximum period (4/150 seconds). I combine the discontinuous data directly. Even if our selection strategy has not detected the boundary data and the boundary data is regarded as noisy data. The

proportion of noise data introduced by the marginal effect is no more than 1/75. The standard M-estimator can give a reliable result when the noisy data is no more than 30 % (Smirnov, 2003a). Therefore, the boundary effect is not the main reason making the estimator collapse and can be neglected. There are 28,800 data points in 1-hour, and 100,000 points (about three and a half hours) are used to calculate the impedance. The 150 Hz data was used to calculate the period between 4/150 to 4/15 seconds. One continuous segment (16 seconds) contains 60 windows of the maximum period (4/15 seconds). The proportion of noise introduced by the marginal data is also no more than 1/60. There are 14,400 data points in 1-hour, and 100,000 points (about 7 hours) are used to calculate the impedance.

Table 5.2 Strategy for calculating the MT impedance for the full frequency band

Sampling rate (Hz)	Calculated period (s)
2400 (discontinuous)	4/2400 to 4/150
150 (discontinuous)	4/150 to 4/15
15 (continuous)	4/15 to 5
1 (continuous)	> 5

I divide the low-frequency band into two and calculate them separately to reduce the computation time. The 15-Hz data was used to calculate the period between 4/15 to 5 seconds. There are 54,000 data points in 1-hour, and 100,000 points (about 2 hours) are used to calculate, and then I downsample the continuous 15-Hz data to 1-Hz to compute the period larger than 5-second. This research used one-day data (86,400 points) to calculate the long period.

Fig. 5.8 shows the MT sounding curve calculated by SSMT-2000 and REHHT. SSMT-2000 is one of the standard Phoenix software sets to calculate the impedance. SSMT-2000 was used to calculate the blue curves, and REHHT was used to calculate the black curves. REHHT result coincides with the SSMT-2000 result. Thus, it shows that the strategy using REHHT calculates the broadband time-series data work well.

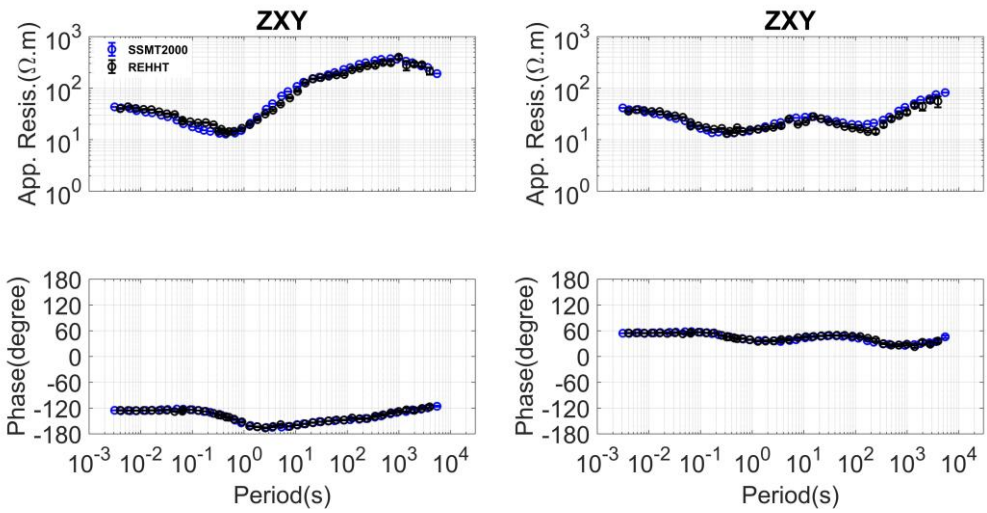

Fig. 5.8 The MT sounding curve calculated by SSMT-2000 and REHHT. SSMT-2000 was used to calculate the blue curves, and REHHT was used to calculate the black curves.

5.4 Performance Test of REHHT by The Synthetic Time-series Data

5.4.1 Test of the maximum reliable period by the synthetic time-series data

In this subsection, I try to find the maximum reliable period of REHHT using synthetic MT time-series data. The maximum reliable period is determined by the impedance calculated by a

given time series.

Creating synthetic time series is similar to the method proposed by Chen et al. (2012). The 1-day magnetic time-series data from Memambetsu (MMB) station was used as the synthetic data. MMB is one of the Magnetic Observatory stations where geomagnetic and geoelectric observations are performed in Japan. The sampling rate is 1 Hz, and its unit is nT. At first, the magnetic field h(t) is transferred into the frequency domain by the Fourier transform. Then, the impedance **Z**(ω), calculated from the simple 1-D model, is multiplied with **H**(ω) to determine **E**(ω). Subsequently, the electric field spectra are transformed back into the time domain to obtain the electric time-series data e(t).

Fig. 5.9 shows the results calculated using 1-hour, 2-hour and 3-hour synthetic MT time-series data. The light blue lines are the true model, calculated using a simple 1-D model. For the REHHT, using the 1-hour time-series data, I can get the maximum reliable period is around 400 seconds (green line). For the BIRRP, the maximum reliable result is around 100 seconds (red line) using the same 1-hour data. I set the time-bandwidth to be 1, the maximum window length to be 600 seconds, and the 3rd and 4th coefficients were used to calculate the impedance by BIRRP. Though I can get the maximum period is 200 seconds, the apparent resistivity is biased a lot; I can not see the result because it is out of the range. I also changed the parameter of time-bandwidth and the maximum window length; the maximum reliable period is around 100 s.

2-hour time-series data calculated by REHHT is good enough to get the maximum reliable period of about 600-second (purple line). Likewise, 3-hour time-series data calculated by REHHT is relatively good enough to get the maximum reliable period of about 900-second (black line). Therefore, I suggest the maximum reliable period for REHHT is around 1/10 of the time series length from the simulation. In contrast, the maximum reliable period for the method base on the windowed FFT (BIRRP) is around 1/30 of the time series length.

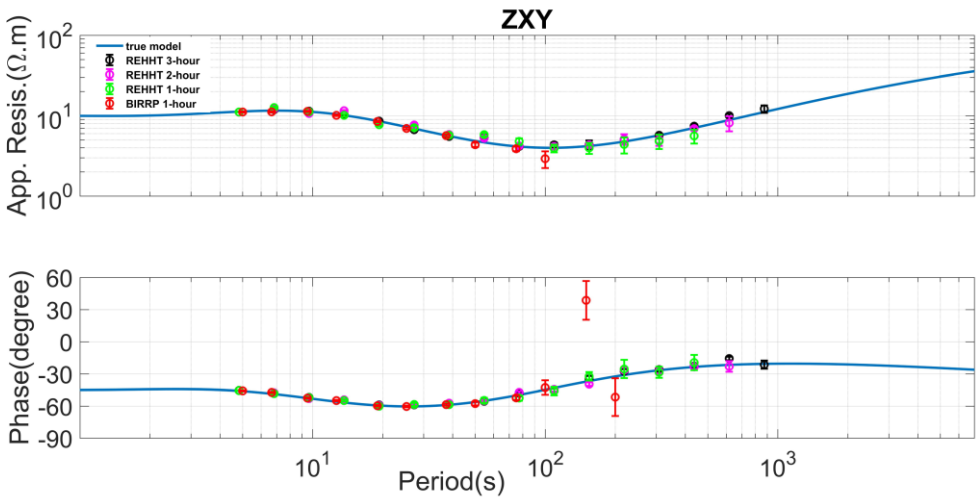

Fig. 5.9 The comparison of the impedance curves calculated by the 1-hour, 2-hour and 3-hour time-series dataset. The light blue curves are the true model curves calculated from the simple 1-D model. First, BIRRP used the 1-hour data to calculate the red circles. Next, REHHT used the 1-hour data to calculate the green circles. Then, REHHT used the 2-hour data to calculate the purple circles. Finally, REHHT used the 3-hour data to calculate the black circles. The upper figures show the apparent resistivity, and the lower figures show the impedance phase.

5.4.2 Test with contaminated synthetic time-series data

Convex-like noise often appears in MT field data. Convex-like noise is added to the magnetic field in three periods; the convex-like noise lasts 20 seconds in the 500-second interval between

each period. In Fig. 5.10, the first row shows the original Hx component. The second row shows the convex-like noise. The third row shows the synthetic, contaminated H_x component. A similar procedure is performed to the E_x, E_y, H_y and H_z component.

Fig. 5.11 shows the result calculated by the contaminated time-series data. BIRRP calculated the red circle and biased from the true model (blue line). REHHT, without using the Hample filter, calculated the green circle and biased from the true model. In contrast, REHHT with the Hample filter calculated the black circle and almost recovered the true model. From the simulation, I can confirm the effectiveness of the Hampel filter and the superiority of REHHT to the conventional method.

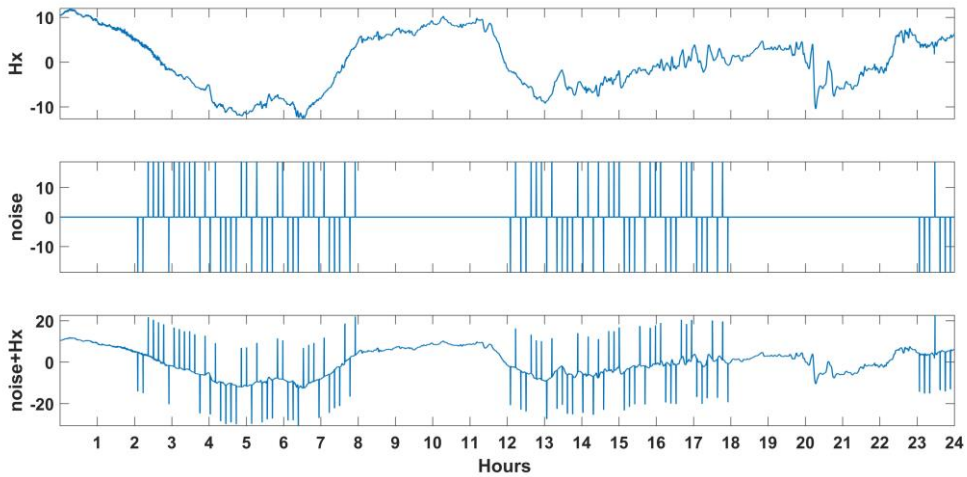

Fig. 5.10 Magnetic field data, synthetic noises and resultant magnetic field data.

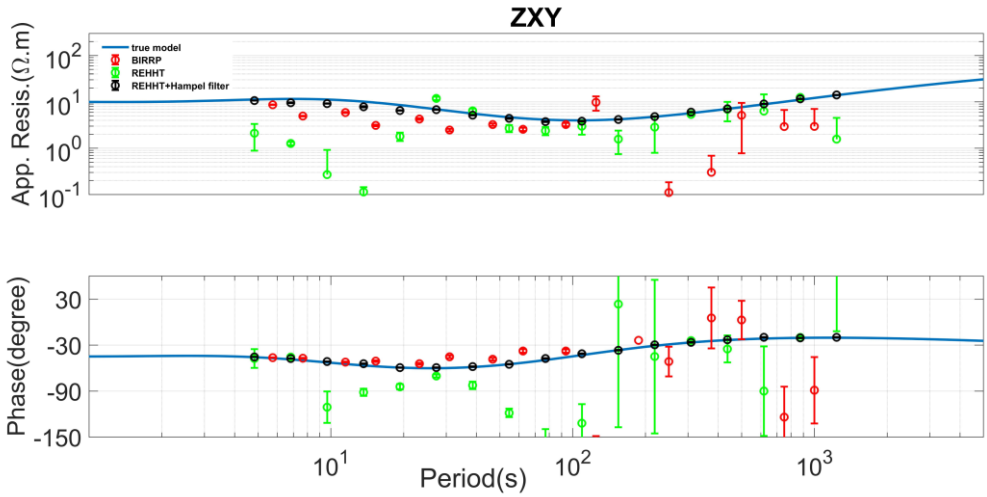

Fig. 5.11 The comparison of the impedance curves calculated by the one-day time-series dataset. The light blue curves are the true model curves calculated from the simple 1-D model. BIRRP was used to calculate the red circles. REHHT calculates the green circles without using the Hampel filter. REHHT calculates the black circles using the Hampel filter.

5.5 Applications of REHHT to the Noisy MT Field Data

The Phoenix Geophysical Instruments observed the MT time-series data. This data belongs to the Institute of Geophysical and Geochemical Exploration, China Geological Survey. The low-frequency band (15 Hz) data sampled continuously is used to show the performance of

REHHT. Fig. 5.12 shows the location map of sites in the study area. The local survey lines are L6-1, L6-2 and L7-1, and the Y0625 is set as the remote reference site. The lower left map shows the survey area in China. The red star denotes the research location. The upper left map shows the detail of survey line L7-1; The lower right map shows survey line L6-1; The upper right map shows survey line L6-2. The observation period is shown in Table 5.3. In this research, I used the data observed on June 26, 2016. The observation area is in GMT+8 time zone. The local midnight time is around 16:00.

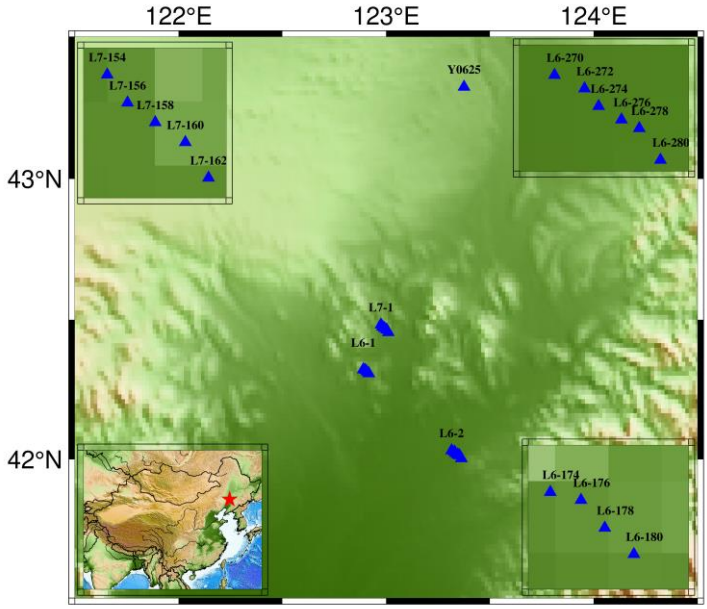

Fig. 5.12 The location map of the study area. Y0625 is the remote reference site, and L6-1, L6-2, and L7-1 in the middle are the local survey lines.

Table 5.3 The observation date of each site.

Observation Date	Site Names
2016.06.25	Y0625, L6174, L6176, L6278, L6280, L7154, L7156
2016.06.26	Y0625, L6178, L6274, L6276, L7158
2016.06.27	Y0625, L6180, L6270, L6272, L7160, L7162

The natural magnetic field in the remote and local sites comes from the same source. Therefore, the correlation between the local and remote reference magnetic field should be high and close to 1 without local noise, and it can be a good indicator of data quality.

Fig. 5.13 and Fig. 5.14 show the Hx component time-series measured simultaneously in different sites. The sampling rate is 15 Hz. Fig. 5.13 shows the time series from 3:00:00 to 22:00:00 on June 26. The time-series variation in the daytime is much stronger than that at nighttime at sites L6274 and L6276. However, the variation at other sites is almost keeping the same scale. Fig. 5.14 shows the time series in the daytime and nighttime in different sites. The left figures show the data in the daytime, and the right figures show the data at nighttime. The variation at sites L6274 and L6276 is about 100 times stronger than that at other sites in the daytime, and it is similar at nighttime. The signal-to-noise ratio (SNR) is about 0.01 in the daytime; the noise is too high and will bias the impedance tensor.

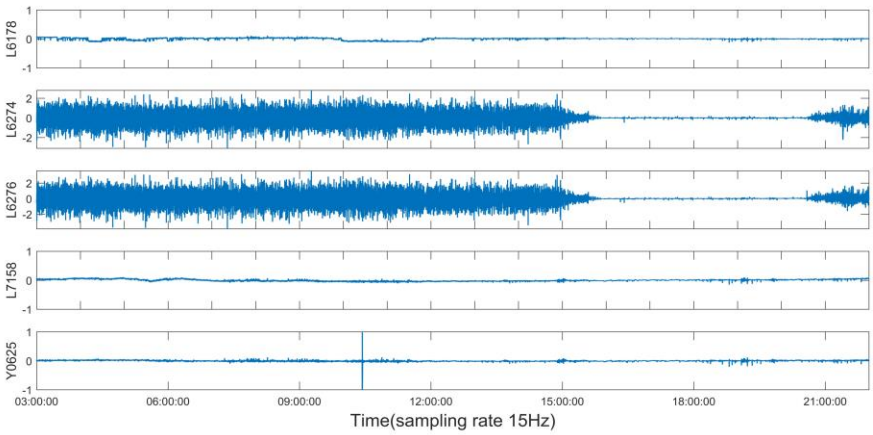

Fig. 5.13 The time variations of the Hx component whose sampling rate is 15 Hz. They were measured simultaneously from 3:00:00 to 22:00:00 in UCT time. The unit of the magnetic field is nT.

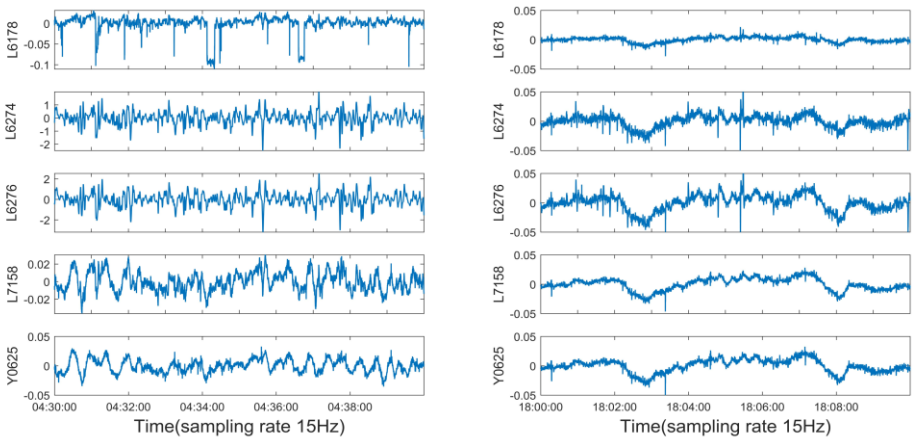

Fig. 5.14 The time variations of the Hx components whose sampling rate is 15 Hz. The left figures show the daytime data, and the right figures show the night data. Each segment is for ten minutes. The unit of the magnetic field is nT.

Fig. 5.15 shows the variation in the correlation between the local and remote magnetic field (Hx) observed at L6274 and Y0625 from 3:00:00 to 22:00:00 in UCT time on June 26. The correlation is based on the Pearson correlation using the time-series segment, and each segment's length is 10 minutes. Due to the local noise, the correlation becomes low or negative in the daytime, and the correlation becomes higher at nighttime. It means that the data in the local night is quieter than that in the daytime. However, the correlation should be high and close to 1 in the quiet situation; it is not so high (around 0.7), which means the data observed at L6274 is noisy even at nighttime.

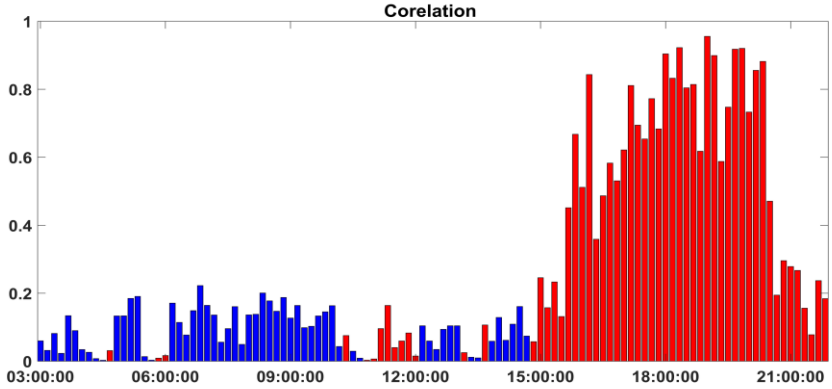

Fig. 5.15 The variation in the correlation between the local and remote magnetic field (Hx) observed at L6274 and Y0625 from 3:00:00 to 22:00:00 in UCT time on June 26. Each segment is 10 minutes. The blue indicates a negative correlation, and the red indicates a positive correlation.

The polarization direction is introduced to estimate the background noise in a different frequency band. The polarization directions for the electric field (α_E) and magnetic field (α_H) at a specific frequency (Fowler et al., 1967) are defined as:

$$\alpha_{E\_i} = tan^{-1}\frac{2Re(E_{x\_i}\bar{E}_{y\_i})}{|E_{x\_i}|^2-|E_{y\_i}|^2}, \tag{5.10}$$

$$\alpha_{H\_i} = tan^{-1}\frac{2Re(H_{x\_i}\bar{H}_{y\_i})}{|H_{x\_i}|^2-|H_{y\_i}|^2}. \tag{5.11}$$

These equations can be rewritten with the polarization directions (α) as follows:

$$\alpha = tan^{-1}\frac{2Re(A_i\bar{B}_i)}{|A_i|^2-|B_i|^2} = tan^{-1}\frac{2\frac{|B_i|}{|A_i|}\cdot cos(\theta_i)}{1-\left(\frac{|B_i|}{|A_i|}\right)^2}, \tag{5.12}$$

where *i(=1,2,…, N)* is the number of each component; A_i and B_i are the spectra calculated from the ith segment in the different channel; A_i and B_i denote H_{x_i} and H_{y_i} or E_{x_i} and E_{y_i}, respectively. The polarization direction is related to the phase difference (PD) and amplitude ratio (AR) between the two orthogonal fields. A variety of sources generate natural magnetic signals. These sources generate magnetic fields that vary in their incident directions and amplitude. The PD and AR between the two orthogonal magnetic fields vary with time; thus, the magnetic field has no preferred polarization direction. However, according to a given conductivity distribution in the subsurface, a preferred polarization direction may exist for the induced electric field (Weckmann et al., 2005). I analyzed the polarization direction at different frequencies for all sites. The signal strength at the dead band (1-10s) is very low, so the local noise can easily influence it. The typical polarization direction for the dead band is shown in Fig. 5.16. This figure shows the polarization direction calculated from the data observed at site L7158 at 6-second. The magnetic field's polarization direction has a preferred direction from 3:00:00 to 15:00:00. It indicates the data is contaminated by strong coherent noise in the daytime. A similar situation occurs at all sites, including the remote site.

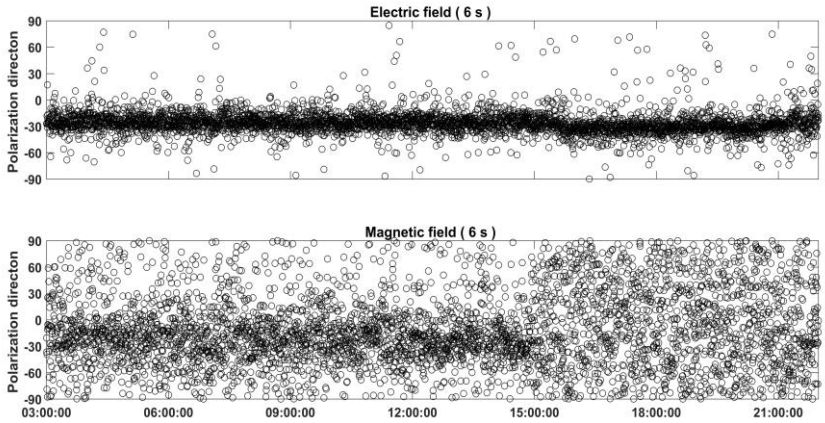

Fig. 5.16 The polarization direction for a period of 6-second at site L7158 from 3:00:00 to 22:00:00 in UCT time. The upper figure shows the electric field's polarization direction, and the lower shows the polarization direction of the magnetic field.

According to the analysis above, this data set is contaminated by typical industrial noise from a manufacturing plant. The MT time-series data at night is much quieter than in the daytime because more electric equipment is shut down during nighttime. The time-series correlation between the local and remote sites, at sites L6274 and L6276, is much higher at nighttime. The magnetic field's polarization has a preferred direction in the dead band at all sites in the daytime, which means the data is contaminated by strong noise in the dead band at all sites in the daytime.

Three different impedance estimators, BIRRP, REHHT, and SSMT-2000, are introduced to calculate the impedance. Fig. 5.17 to Fig. 5.22 show the result derived from the data observed at site L7158.

Fig. 5.17 shows the result calculated by one-day data on June 26. The apparent resistivities and Phase of all the results change unnaturally between 0.1 and 20 s. All the results have a hump in the apparent resistivities.

In Fig. 5.18, SSMT-2000 calculates the blue curves from one-day data, which is the same in Fig. 5.17. BIRRP and REHHT calculate the result by nighttime data from 16:00:00 to 20:00:00. Compared with the SSMT-2000 result, REHHT and BIRRP results calculated by the nighttime data become smooth and change naturally between 0.1 and 20 s in the apparent resistivities and Phase. Fig. 5.17 and Fig. 5.18 reflect the different properties between the daytime and nighttime between 0.1 and 20 s and are consistent with the background noise analysis in the previous.

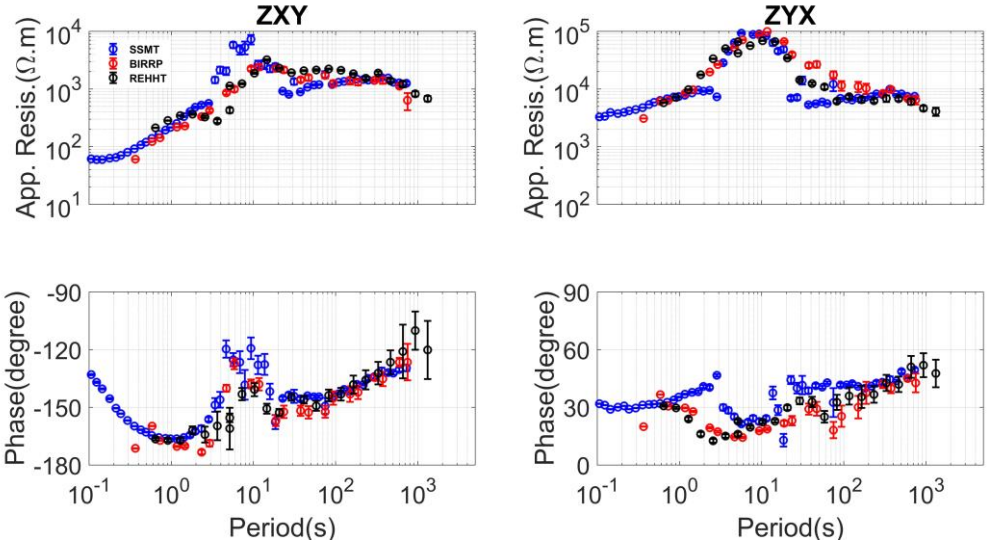

Fig. 5.17 The magnetotelluric sounding curves calculated using the one-day data observed at site L7158. SSMT-2000 was used to calculate the blue curves, BIRRP was used to calculate the red curves, and REHHT was used to calculate the black curves. The upper figures of all the results show the apparent resistivity, and the lower figures show the Phase.

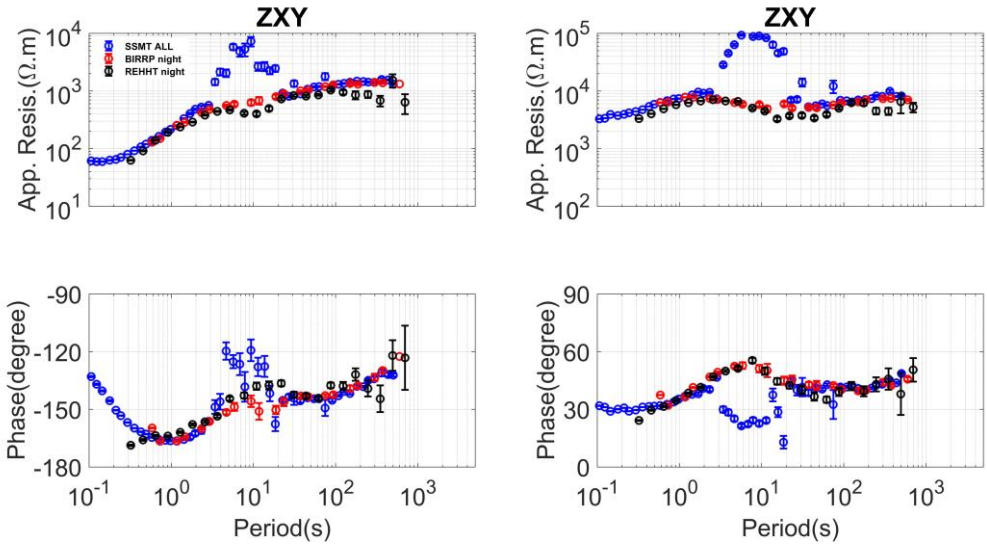

Fig. 5.18 The magnetotelluric sounding curves calculated using the data at site L7158. SSMT-2000 was used to calculate the blue curves using one-day data, and REHHT and BIRRP were used to calculate the black curves and red curves, respectively, using the nighttime data.

Fig. 5.19 shows the hat matrix diagonal element distribution at 10-second using the data from 14:00:00 to 20:00:00. Most of the value is high in the previous 800 data, corresponding to the noisy period from 14:00:00 to 16:00:00. Thus, it shows that the hat matrix's diagonal element is effective in detecting noisy data.

Fig. 5.20 compared the performance of REHHT with and without leverage selection. I used the data observed from 14:00:00 to 20:00:00, including about 2-hour noisy data. BIRRP was used to calculate the red curves. I think the BIRRP result is reliable. REHHT with and without leverage selection were used to calculate the blue and black curves. The result of REHHT with leverage selection coincides with the BIRRP result and changes naturally. This result shows the effectiveness of leverage selection. I also tried using the data from 13:00:00 to 20:00:00. Even

using BIRRP or REHHT with leverage selection failed to get a reliable result. I need an additional strategy to suppress the data when using the one-day data to calculate the impedance.

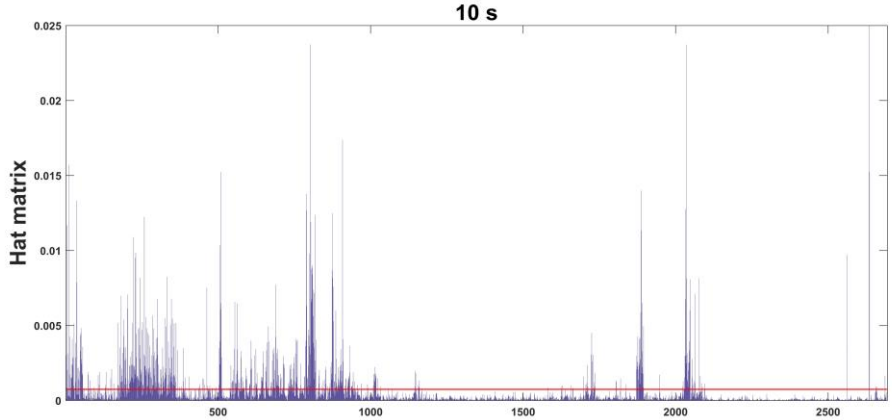

Fig. 5.19 The distribution of hat matrix diagonal element at period 10-second using the data from 14:00:00 to 20:00:00. The red line is the expected value equals 2/N, and N denotes the number of the samples.

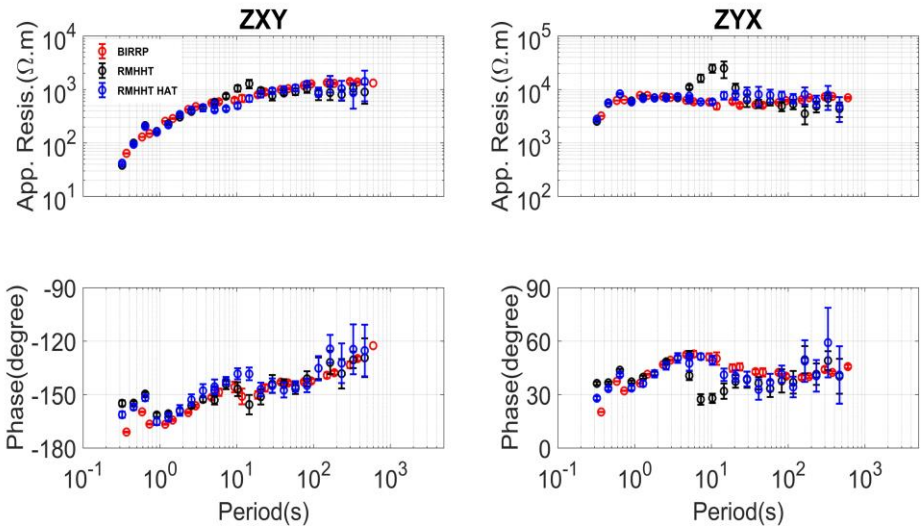

Fig. 5.20 The MT sounding curves calculated using the data from 14:00:00 to 20:00:00 observed at site L7158. BIRRP was used to calculate the red curves, and REHHT with and without leverage selection was used to calculate the blue and black curves.

Fig. 5.21 shows the variation in variance-ratio and correlation between the local and remote magnetic fields. At first, I divide the time series into a short segment with a length of 1 minute; the upper figure shows the variation in variance ratio between the local and remote magnetic fields. The data between 15:00:00 to 21:00:00 is small and stable. When I calculated the impedance by the one-day data, I should suppress the influence of the daytime data between 3:00:00 to 16:00:00 as most as possible. The variance ratio from 10:00:00 to 12:00:00 is similar to the nighttime data and makes it difficult to differentiate the noise data in the daytime from the nighttime data.

On the other hand, the correlation between the local and remote magnetic fields is better to differentiate the data between the daytime and nighttime. The lower the threshold is, the more daytime data is introduced. Therefore, a high threshold is set to suppress the data in the daytime.

The corresponding instantaneous parameters with a correlation value lower than 0.95 are removed before extracting the independent data. Fig. 5.22 compared the result calculated by the three estimators. REHHT result changes naturally and is more reliable than the BIRRP and SSMT-2000 results between 1 to 100 seconds. I also tried the threshold to be 0.9. Because the more daytime data was introduced, the result got biased.

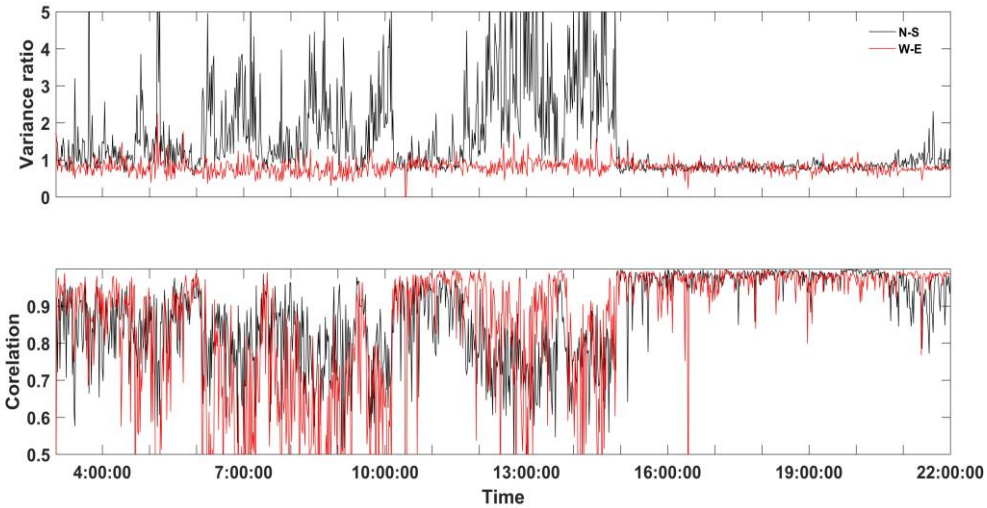

Fig. 5.21 The variation in variance-ratio and correlation between the local and remote magnetic fields. The black lines show the component in the north-south direction. The red lines show the component in the west-east direction.

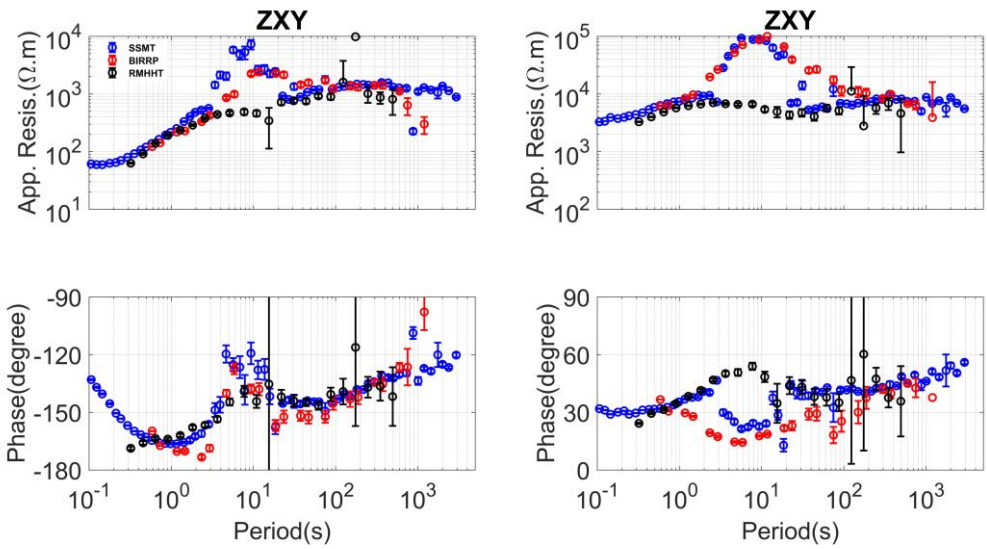

Fig. 5.22 The MT sounding curves calculated using the one-day data observed at site L7158. SSMT-2000 was used to calculate the blue curves, and BIRRP was used to calculate the red curves. REHHT with noise detection in the time domain was used to calculate the black curves.

Similar results were obtained at the L6174, L6178, L6180 and Y0626. Fig. 5.23 and Fig. 5.26 show the result calculated using the data at sites L6174, L6178, L6180 and Y0626, respectively. The SSMT-2000 calculates the blue curve, and BIRRP calculates the red curve. And the REHHT calculate the black curve using the data selection based on the correlation between local and remote magnetic field. The threshold is 0.9 when calculating the impedance using the

data observed at L6174, L6178, and L6180 using the Y0625 as the remote site. The threshold is 0.95 at Y0625 and uses the L7158 as the remote site.

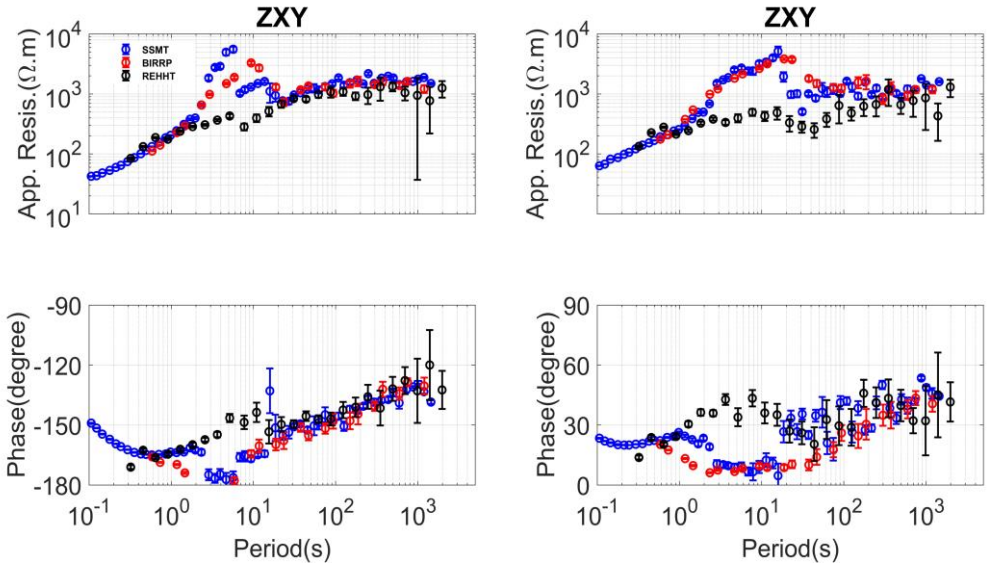

Fig. 5.23 The magnetotelluric sounding curves calculated using the data at site L6174. The color denotes the same meaning as Fig. 5.22.

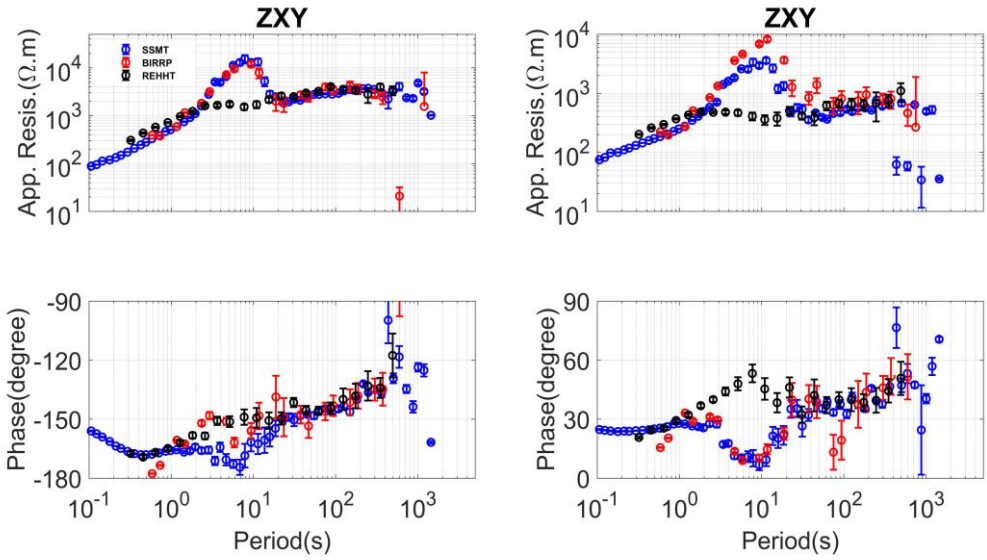

Fig. 5.24 The magnetotelluric sounding curves calculated using the data at site L6178. The color denotes the same meaning as Fig. 5.22.

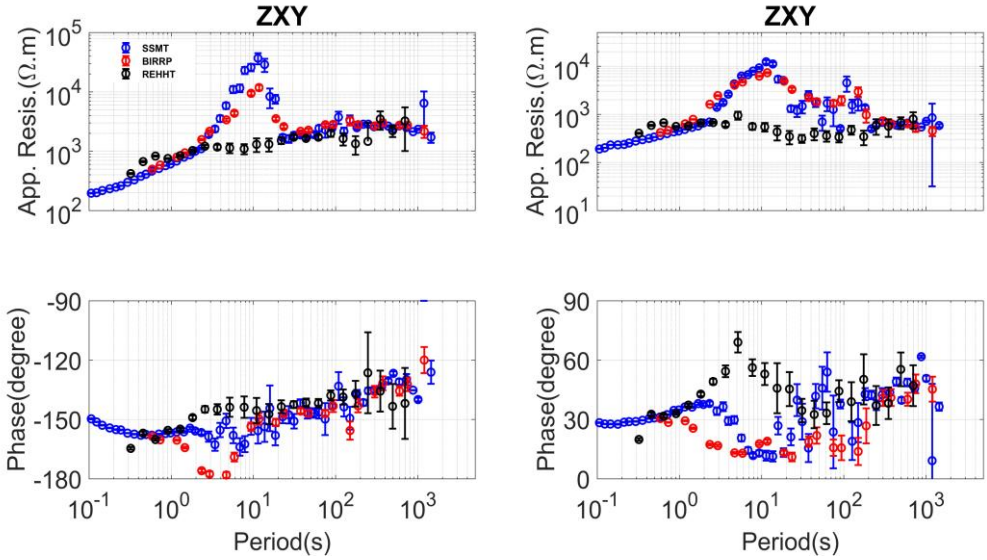

Fig. 5.25 The magnetotelluric sounding curves calculated using the data at site L6180. The color denotes the same meaning as Fig. 5.22.

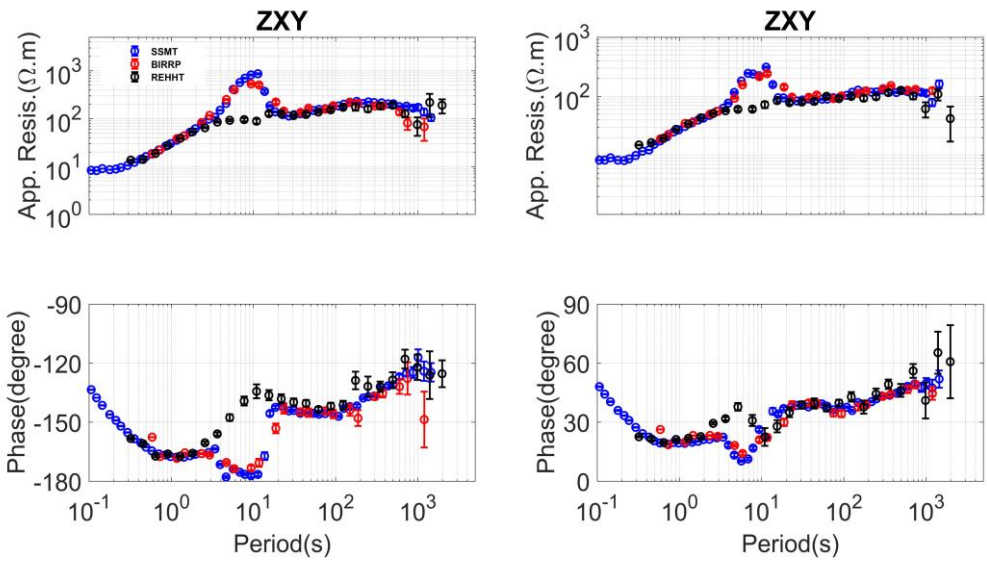

Fig. 5.26 The magnetotelluric sounding curves calculated using the data at site Y0626. The color denotes the same meaning as Fig. 5.22.

5.6 Conclusions And Discussion

I introduced a robust impedance estimator based on the HHT (REHHT). I demonstrated a strategy to process the broadband MT time-series data properly by the REHHT estimator. The time-series data observed by the Phoenix system contain discontinuous and continuous data. I combined the discontinuous time-series data in the time domain and processed the spliced, discontinuous data in the same way as the continuous data. I showed that the strategy worked well using the field data and discussed that the noise introduced by the marginal effect could be neglected in the case study. The problem of REHHT is time-consuming, and by the strategy of downsampling, the computation time can be reduced.

I tested the maximum reliable period calculated from a given time series using the quiet synthetic data. I conclude that the maximum reliable period is around 1/10 of the time series length for the RMHHT. In contrast, the maximum reliable period is around 1/30 of the time

series length for the method based on the windowed FFT (BIRRP). That means calculating the same maximum period, the shorter the quiet data I need. For example, to get a reliable result until 1,000 seconds, the data I input to the estimator should contain around 30,000-second quiet data for BIRRP and 10,000-second quiet data for REHHT.

Let us assume the breakdown point of BIRRP is 40% (Chave and Thomson, 2004; Smirnov, 2003a), which means getting a reliable result, the low signal-to-noise ratio (SNR) data should be no more than 40% in the data. Therefore, it was worthwhile to do a careful background noise analysis and mask noisy data before MT data processing for the conventional method. For example, I introduced polarization direction and correlation to analyze the background noise. I found a relatively quiet time window during the local nighttime. I used the data from this quiet period to estimate impedances and get a reliable result. The breakdown point of the robust estimator itself does not exceed 50%. In contrast, the breakdown point of REHHT may be over 50% when there is a quiet remote site.

I showed the superiority of REHHT by the contaminated synthetic time series data and the noisy field data. The effectiveness of the Hampel filter in reducing the influence of transient noise is shown using the synthetic time series. Fig. 5.19 and Fig. 5.20 also showed the hat matrix's effectiveness in detecting the noise in the frequency domain. But both the Hampel filter and leverage selection failed to get a reliable result from the one-day noisy data. I also need an additional method to reduce the influence of noise. The step of noise detection based on the correlation between the local and remote magnetic fields is useful. By this step, no matter how much noise the data contains, I may get a reliable result if the data contain about ten quiet oscillations, and the ten oscillations don't need to be continuous. I used one-day data observed at sites L7158, L6174, L6178, L6180 and Y0626, including a large amount of noise, to show the performance of REHHT with the noise detection based on the parameter of correlation between the local and remote magnetic field.

All in all, any impedance estimators need regular data at the regression step. The key point to getting a reliable impedance from the noisy data is to remove or reduce the noise before or during the impedance estimation. REHHT can maximally reduce the influence of noise by selecting the data in the time and frequency domain.

Chapter 6: THE INFLUENCE OF GEOMAGNETIC STORMS ON THE QUALITY OF MAGNETOTELLURIC IMPEDANCE

Not all MT time-series data include usable information on the electrical conductivity distribution at depth. A low signal-to-noise ratio (SNR) can occur when the natural signal level is comparable to or below the instrument noise level or in the presence of some types of cultural noise (Chave and Jones, 2012). The first step in MT data processing is estimating the frequency-domain impedance tensor from the measured time-series data. All MT data interpretations are based on the impedance. Therefore, it is very important to obtain a reliable impedance. When the noise continuously contaminates the site, it is not easy to obtain a reliable impedance from the current MT data processing technology(Chave et al., 1987; Chave and Thomson, 2003, 2004; Egbert, 1997; Egbert and Booker, 1986; Smirnov, 2003; Weckmann et al., 2005).

It is well known in MT communities that the natural EM signal increases significantly during a strong geomagnetic storm. However, examples of evaluating the effects of geomagnetic storms on the MT impedance quality are rare. On the other hand, the plane wave assumption is generally acceptable at midlatitudes (Lezaeta et al., 2007; Viljanen et al., 1993). However, it is violated at high magnetic latitudes because the source field is nonuniform during geomagnetic storms (Mareschal, 1981; Viljanen et al., 1993; Garcia et al., 1997; Lezaeta et al., 2007). There is another issue that the geomagnetic pulsations (Pc's) in the Pc3-4 band (~10-100 s) at mid-latitudes associated with field-line resonances can violate the fundamental assumption of the MT method over resistive regions, where the skin depths are large (Egbert et al., 2000; Murphy and Egbert, 2018). However, this issue is not the focus of this paper. Possible biases in the MT transfer function due to the source effect are considered only at long periods (> 1000 s) and near the auroral or equatorial electrojets (Murphy and Egbert, 2018).

Our hypothesis of the geomagnetic storm is beneficial to MT impedance quality at midlatitudes. In a quiet EM environment, no matter whether there is a geomagnetic storm, I can get a reliable MT impedance tensor. On the other hand, in a noisy EM environment, I may fail to get a reliable impedance tensor during a non-storm day, while the SNR becomes high during a storm day, and a reliable impedance result may be obtained by using the storm data. This paper investigates this hypothesis by comparing the data quality and MT impedance quality during the storm and non-storm days using field datasets observed at midlatitudes.

6.1 Method to Evaluate MT Impedance Quality

Coherence is an important parameter often used for analyzing data quality (e.g., Egbert and Livelybrooks, 1996; Fontes et al., 1988; Jones and Jödicke, 1984). In general, I can combine the performance of impedance and coherence to estimate the MT impedance quality across all periods; I also combine the phase difference between the electric and magnetic fields, the polarization direction, and the hat matrix to discuss the data quality for a specific period simultaneously.

6.1.1 Impedance Tensor

In the MT method, the magnetic field ($\boldsymbol{H}$) and the electric field ($\boldsymbol{E}$) have a linear relationship in the frequency domain. The impedance tensor $\boldsymbol{Z}$ relates the electric and magnetic fields as follows:

$$\boldsymbol{E}=\begin{pmatrix}\boldsymbol{E_x}(\omega)\\ \boldsymbol{E_y}(\omega)\end{pmatrix}=\begin{pmatrix}Z_{xx}(\omega)\ Z_{xy}(\omega)\\ Z_{yx}(\omega)\ Z_{yy}(\omega)\end{pmatrix}\begin{pmatrix}\boldsymbol{H_x}(\omega)\\ \boldsymbol{H_y}(\omega)\end{pmatrix}=\boldsymbol{ZH}, \tag{6.1}$$

where $\boldsymbol{E}$ and $\boldsymbol{H}$ are the horizontal electric and magnetic fields at a specific frequency, respectively, ω denotes the angular frequency. The suffix x denotes the north-south direction,

and *y* denotes the east-west direction.

The bounded influence remote reference processing code (BIRRP; Chave et al., 1987; Chave and Thomson, 2003, 2004) is a typical conventional robust estimator to calculate the impedance tensor based on windowed FFT. I used the BIRRP code to calculate the impedance in this research and adopted the standard M-estimator (Chave et al., 1987).

The MT sounding curves are the most important result for discussing MT impedance quality. I expect MT sounding curves to be smooth as a function of frequency from the underlying physics because EM propagation in the conductive Earth is a diffusive process.

6.1.2 Phase Difference and Coherence between the Electric Field and Magnetic Field

The MT time-series data are observed simultaneously in several channels; each channel is divided into *N* segments, and I can obtain *N* spectra by applying the Fourier transform to each segment. In polar coordinates, the cross-power spectra between two spectra A_i and B_i are expressed as follows:

$$A_i\bar{B}_i = |A_i|\cdot|B_i|e^{j(\varphi_{A_i}-\varphi_{B_i})}, \tag{6.2}$$

where *j* denotes the imaginary number unit, *i* (=1,2,…, N) is the number index; A_i and B_i are the spectra calculated from the ith segment from the different channel; and φ_{A_i} and φ_{B_i} denote the phases of A_i and B_i, respectively. The overline denotes the complex conjugate. On the other hand, the auto-power spectra are expressed as follows:

$$A_i\bar{A}_i=|A_i|^2,\ B_i\bar{B}_i = |B_i|^2. \tag{6.3}$$

The phase difference between A_i and B_i ($PD(A_i,\ B_i)$) is calculated as follows:

$$\theta_i = \varphi_{A_i} - \varphi_{B_i} = arg\left(e^{j(\varphi_{A_i}-\varphi_{B_i})}\right) = arg\left(\frac{A_i\bar{B}_i}{|A_i||B_i|}\right), \tag{6.4}$$

where θ_i denotes the phase difference (PD) between the two spectra at a specific frequency.

Coherence is the ratio between cross-power spectra density and the root of auto-power spectra density. For the $\boldsymbol{A}$ and $\boldsymbol{B}$ spectrum at a specific frequency, it is defined as:

$$Coh(\boldsymbol{A},\boldsymbol{B}) = \frac{|<\boldsymbol{A}\bar{\boldsymbol{B}}>|}{\sqrt{<\boldsymbol{A}\bar{\boldsymbol{A}}><\boldsymbol{B}\bar{\boldsymbol{B}}>}}, \tag{6.5}$$

where the brackets represent the averages of *N* individual auto-power spectra and cross-power spectra. For instance,

$$<\boldsymbol{A}\bar{\boldsymbol{B}}> = \frac{1}{N}\sum_{i=1}^{N} A_i\bar{B}_i\ . \tag{6.6}$$

The coherence is a quantitative measure of the phase difference (PD) consistency between the two channels. If two channels are coherent, their phases must be the same or have a constant difference (Marple and Marino, 2004). That means there is a preferred direction of the phase difference when two channels are coherent.

In a certain geological situation, the impedance tensor is shown in Table 2. The diagonal element of impedance is zero in the 1-D or 2-D models (when the observation axes coincide with the geoelectrical strike). There is a relationship between the orthogonal electric and magnetic field as follows:

$$E_{x\_i} = Z_{xy}H_{y\_i}. \tag{6.7}$$

In the case of the 1-D, or 2-D structure in a quiet environment (1-D/2-D case), the orthogonal eclectic and magnetic fields are coherent and have a constant PD, and the PD equals the impedance phase.

In the 2-D (when the observation axes do not coincide with the geoelectrical strike) or 3-D cases, all impedance values are nonzero, and for the north-south direction, there is a relationship between the electric and magnetic fields as follows:

$$E_{x\_i} = Z_{xx}H_{x\_i} + Z_{xy}H_{y\_i}. \tag{6.8}$$

There is no direct relationship between $\boldsymbol{E}_x$ and $\boldsymbol{H}_x$ and between $\boldsymbol{E}_x$ and $\boldsymbol{H}_y$, which is very complex in real field data.

On the other hand, the case of phase rolling out of the quadrant (PROQ) (Chouteau and Tournerie, 2000; Weckmann et al., 2003; Yu et al., 2019) is increasingly recognized in MT field

surveys. The channeling current caused by complex 3-D isotropic media can explain the PROQ phenomenon. The characteristic of PROQ is that the parallel electric and magnetic fields are coherent, while the orthogonal components are incoherent.

Although PD and coherence are different in different situations, they are useful parameters for discussing data quality.

Table 6.1 Impedance tensor characteristics in certain geological situations. In the 1-D case, the impedance of the diagonal elements is zero. In the 2-D case, when the observation axes coincide with the geoelectrical strike, the impedance is zero in the diagonal elements, and the off-diagonal elements are unequal. In the 3-D case, all $\boldsymbol{Z}$ values are nonzero.

1-D structure	2-D structure	3-D structure
$\begin{pmatrix} 0 & Z_{xy} \\ Z_{yx} & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & Z_{xy} \\ Z_{yx} & 0 \end{pmatrix}$	$\begin{pmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{pmatrix}$
$Z_{xy} = -Z_{yx}$	$Z_{xy} \neq Z_{yx}$	all $\boldsymbol{Z}$ values are nonzero

6.1.3 Polarization Directions

Weckmann et al. (2005) showed the effectiveness of using the polarization directions to estimate the background noise. The polarization directions for the magnetic field (α_H) at a specific frequency (Fowler et al., 1967) are defined as follows:

$$\alpha_{H\_i} = tan^{-1}\frac{2Re(H_{x\_i}\bar{H}_{y\_i})}{|H_{x\_i}|^2-|H_{y\_i}|^2} = tan^{-1}\frac{2\frac{|H_{x\_i}|}{|\bar{H}_{y\_i}|}\cdot cos(\theta_i)}{1-\left(\frac{|\bar{H}_{y\_i}|}{|H_{x\_i}|}\right)^2}, \tag{6.9}$$

where H_{x_i} and H_{y_i} are the spectra calculated from the ith segment in the magnetic field and θ_i denotes the PD between H_{x_i} and H_{y_i}. The polarization direction is related to the phase difference and amplitude ratio between the two orthogonal fields.

A variety of sources generate natural magnetic signals. These sources generate magnetic fields that vary in their incident directions and amplitude. The phase difference and amplitude ratio between the two orthogonal magnetic fields vary with time; thus, the magnetic field has no preferred polarization direction (Weckmann et al., 2005).

In contrast, the local EM noise source has a constant location; the incident direction and energy have similar properties changing with time. Suppose there is a preferred polarization direction for the magnetic field; I can consider that the coherent noise contaminates the data. On the other hand, when the incoherent noise contaminates the field data, the magnetic field has no preferred polarization direction. Therefore, the polarization direction for the magnetic field can only detect coherent noise.

6.1.4 Hat matrix

The hat matrix is an *N* by *N* matrix (*N* denotes the sample of data) defined as follows (Chave and Thomson, 2003, 2004):

$$\boldsymbol{H}_{hat} = \boldsymbol{H}(\boldsymbol{H}^{\dagger}\boldsymbol{H})^{-1}\boldsymbol{H}^{\dagger}, \tag{6.10}$$

where $\boldsymbol{H}$ represents *N* by two matrices of the horizontal magnetic field ($\boldsymbol{H}_x, \boldsymbol{H}_y$) at a specific frequency. The superscript † denotes the complex conjugate transpose. The expected value of the hat matrix's diagonal element is 2/*N*. The hat matrix is widely used to detect abnormal large-value data. I adopted the hat matrix to examine the energy variation of each segment (see the example in Fig. 6.3), and the data whose hat matrix's diagonal element was over the expected value (2/*N*) were defined as the leverage point in this research.

6.2 Case Studies in Three MT Fields

Three field datasets are used to evaluate the influence of geomagnetic storms on the MT impedance. Fig. 6.1 shows the location map of the three field datasets (KAP03, USArray, Sawauchi). The first field dataset used the long-period 5-component MT time-series data observed at Kaapvaal 2003 (KAP03). The sampling period of this survey is a 5-second. The data were recorded for almost a month at each site using GSC LIMS systems in 2003 as a part of the SAMTEX project. The 26 long-period sites distributed in a NE-SW profile are shown in the right corner of Fig. 6.1. Data for the sites in the middle of the profile (site 127 - site 145 of KAP03) were heavily contaminated by DC signals from the DC train line running between Kimberley and Johannesburg (MTNET, see the website in references). The second field dataset was observed at site TNV48 from the USArray project. The third field dataset was observed at Sawauchi station, Japan. All of the filed data include geomagnetic storms.

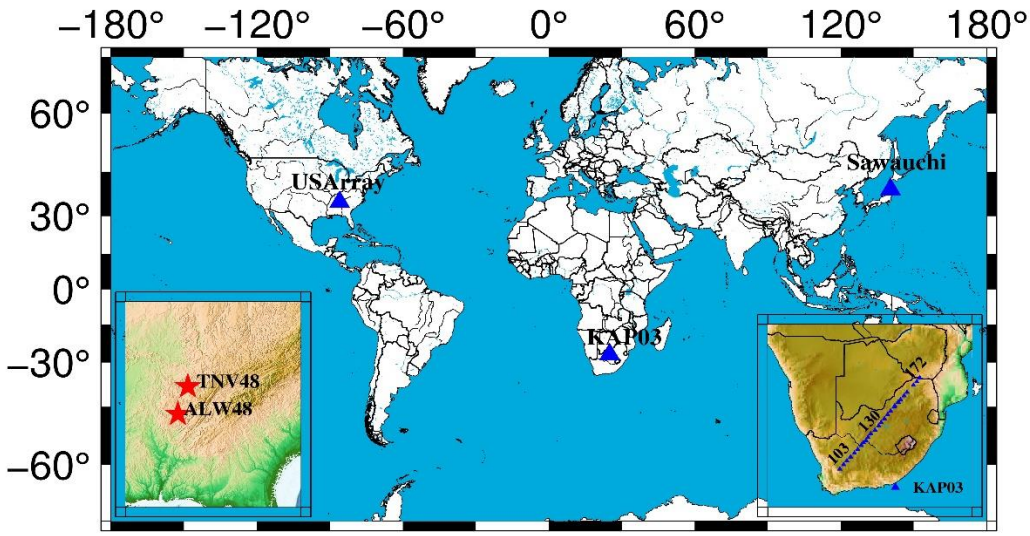

Fig. 6.1 The location map of the three field data (KAP03, USArray, Sawauchi). The left map shows the detailed site location used in USArray, and the right map shows the survey line of KAP03.

6.2.1 Case study 1: KAP03, South Africa

At first, I investigate the relationship of the variation between the time series data and the Dst index. The following first difference filter is used to remove the trends and means of the signal $\boldsymbol{x}$:

$$Diff(x_i)=x_i-x_{i-1}, \tag{6.11}$$

where i denotes the number of the time-series data. The absolute value of $Diff(\boldsymbol{x})$ can represent the strength of the fluctuation. In addition, the threshold k ($k = 1.5\sigma$) is used to detect abnormally high values, where σ is the standard deviation. A common approach is to take $\sigma = MAD/0.44845$, where MAD is the median of the absolute deviations from the data's median and $MAD = median(|x_i-median(\boldsymbol{x})|)$(Chave and Thomson, 2003). Fig. 6.2 shows the time variations in the Hx component at site 163 and the Dst index. The red dashed line in Panel (d) denotes the k used to detect the abnormally high value. The high absolute value of $Diff(Dst)$ corresponds to the high variation of $Diff(Hx)$ and the storm. Therefore, I could use $Diff(Dst)$'s absolute value to detect the high variation data in the presence of a storm. This geomagnetic storm started at approximately 06:00 on October 29 and ended at approximately noon on October 31. I extracted two-day data from 06:00 on October 29 to 06:00 on October 31 as the storm data.

Moreover, I investigate the relationship between the magnetic field amplitude and the hat matrix's diagonal elements. Fig. 6.3 compares the variations in the magnetic field amplitude and the hat matrix's diagonal elements in the period of 101 s using the data observed at site 163 from 00:00 on October 29 to 00:00 on November 4. It shows that the energy of the magnetic field is much stronger than that during the non-storm days. On the other hand, the variation of the hat matrix's diagonal elements has the same trend as the magnetic field amplitude. Therefore, I use the hat matrix to show the energy variation in the magnetic field and combine the phase difference between the electric and magnetic fields and polarization direction to discuss how the data quality varies with the energy in a specific period. In this research, the leverage point whose hat matrix's diagonal element exceeds the expected value (2/*N*) corresponds to a geomagnetic storm, and the corresponding data are shown in red in the following data quality analysis.

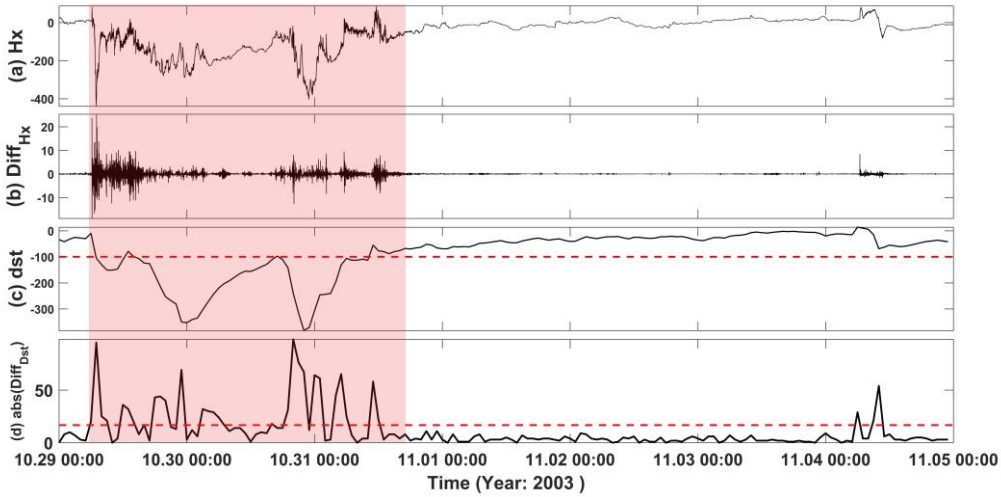

Fig. 6.2 Variations in the *Hx* component of site 163 and the *Dst* index. Panel (a) shows the variation in the *Hx* component. Panel (b) shows the first derivative of the *Hx* component. Panel (c) shows the variation in the *Dst* index. Panel (d) shows the first derivative of the *Dst* index's absolute value; the red dashed line denotes 1.5σ (σ is the standard deviation). The horizontal axis denotes the time in UTC. The shadow in red represents the data in the storm period.

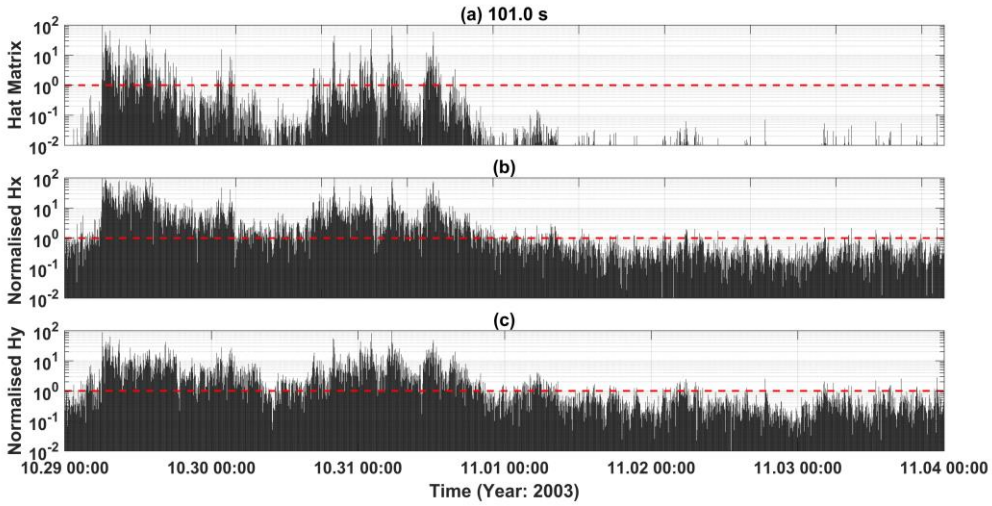

Fig. 6.3 Variations in the magnetic field amplitude and the hat matrix's diagonal elements in the period of 101 s using the data observed at site 163 from 00:00 on October 29 to 00:00 on

November 4. Panel (a) shows the hat matrix's diagonal elements normalized by the expected value of 2/*N*. Panels (b) and (c) show the magnetic fields normalized by the standard deviation in the *x*- and *y*-directions, respectively. The variation in the hat matrix's diagonal elements has the same trend as the magnetic field amplitude.

6.2.1.1 The influence of geomagnetic storms on the data in a 1-D/2-D structure in a quiet environment (site 163)

Next, the data quality analysis at site 163 in 101 s is shown in Fig. 6.4. The data observed from 00:00 on October 29 to 00:00 on November 4 is used. Panels (b) and (c) show that the PD between the orthogonal electric and magnetic fields ($PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$) has a preferred direction, while the PD between the parallel electric and magnetic fields ($PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_y)$) is scattered all the time. The PD is more concentrated during the storm day than the non-storm day, and it coincides with the increase in the coherence value during the storm in Fig. 6.5 (j) and (k).

Next, I combine the impedance result and coherence distribution to discuss the MT impedance quality at site 163. Fig. 6.5 shows the MT sounding curves and coherence distributions during the storm and non-storm days. The data observed from 00:00 on November 1 to 00:00 on November 4 are used as the non-storm data. The coherence values between the orthogonal electric and magnetic fields ($Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$ and $Coh(\boldsymbol{E}_y, \boldsymbol{H}_x)$) are high, while those between the parallel components ($Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ and $Coh(\boldsymbol{E}_y, \boldsymbol{H}_y)$) are low across all periods. The data can be categorized into a 1-D/2-D case in a quiet environment during the storm and non-storm days. Because the environment is quiet, the proportion of high SNR data is high enough to obtain a reliable impedance during the storm and non-storm days; therefore, the MT impedance results calculated by storm and non-storm data coincide. I also compared the result with the remote reference results and found that both are similar. This example shows that the storm does not influence the impedance quality in a quiet EM environment.

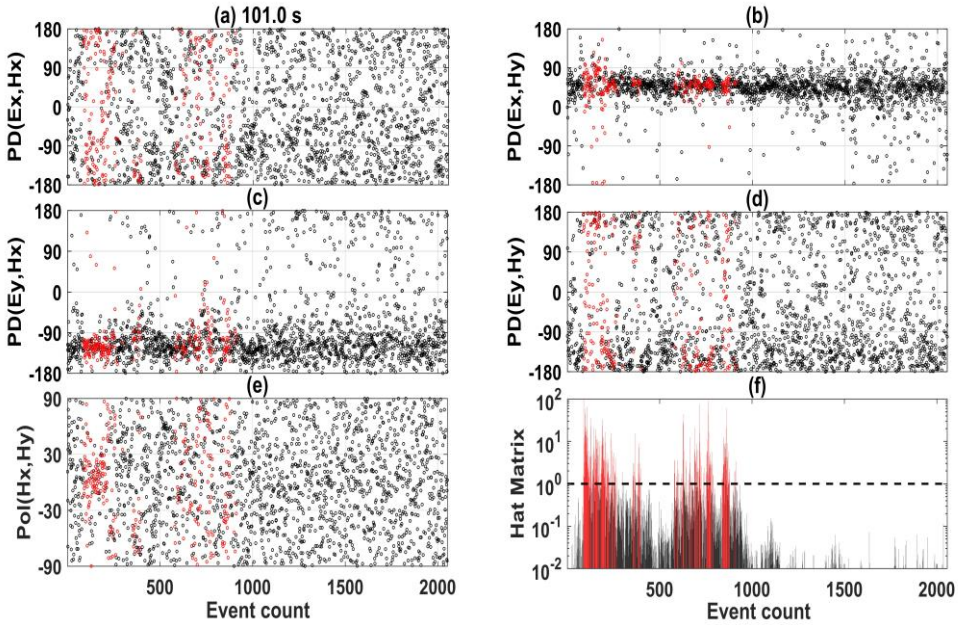

Fig. 6.4 Evaluation indices in 101 s using the data observed at site 163 from 00:00 on October 29 to 00:00 on November 4. Panels (a)-(d) show the variation in $PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$, and $PD(\boldsymbol{E}_y, \boldsymbol{H}_y)$; Panel (e) shows the variation in polarization direction for the magnetic field, and Panel (f) shows the variation in the normalized hat matrix's diagonal

element. The data in red correspond to the leverage point, whose hat matrix diagonal element exceeds the expected value (2/*N*). One data in the frequency domain is regarded as one event. The horizontal axis denotes the event count.

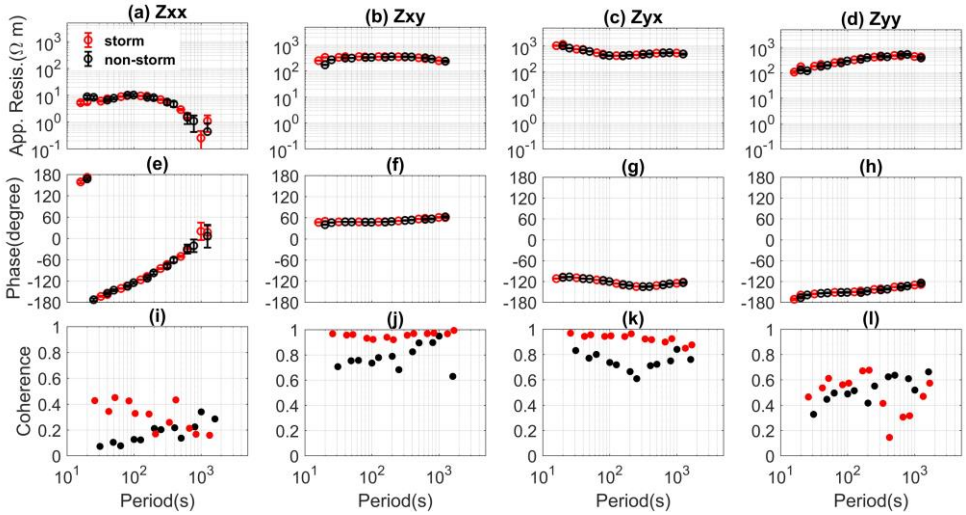

Fig. 6.5 MT sounding curves and coherence distributions during the storm and non-storm days using the data observed at site 163. The black color indicates the results calculated by the non-storm data; the red color indicates the results calculated by storm data. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. Panels (i)-(l) show the distribution of coherence. The horizontal axis denotes the period in seconds.

6.2.1.2 The influence of geomagnetic storms on the data in a 1-D/2-D structure contaminated by the incoherent noise (site 142)

And then, a similar analysis is performed at site 142. The data quality analysis in 84 s is shown in Fig. 6.6. The data observed from 00:00 on October 26 to 00:00 on October 31 is used. The natural EM signal also varies significantly during the storm, as shown in Panel (f). The SNR depends on the strength of the natural EM signal and noise. Therefore, the data during the storm may also contain data dominated by noise. Panels (b) and (c) show that the PD between the orthogonal electric and magnetic fields ($PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$) has a preferred direction corresponding to the leverage point during the storm day. The data in red can be categorized into a 1-D/2-D case with a high SNR. The SNR of the black data is low, and all the PD is scattered. Incoherent noise may contaminate the data during the non-storm day.

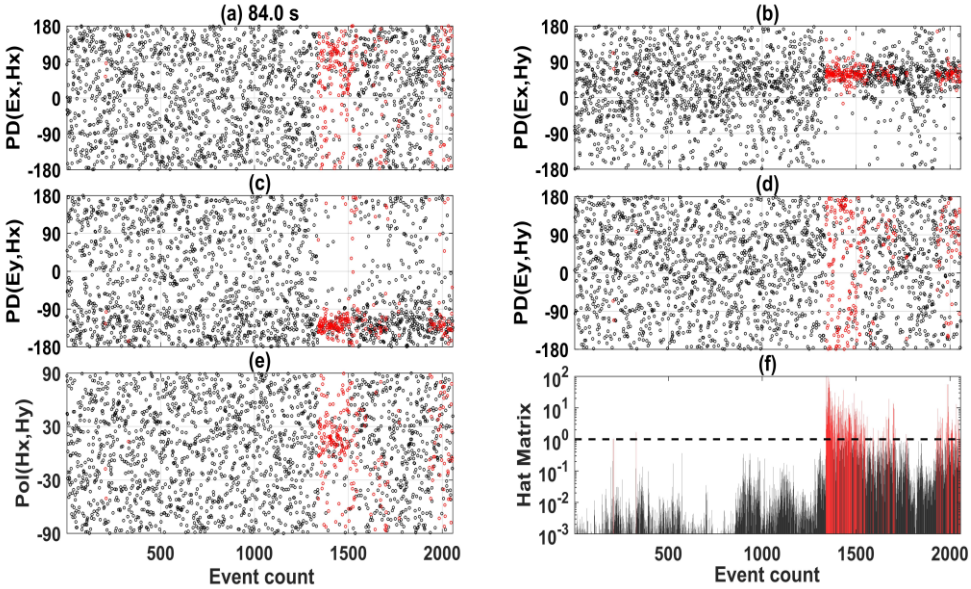

Fig. 6.6 Evaluation indices in 84 s using the data observed from 00:00 on October 26 to 00:00 on October 31 at site 142. Panels (a)-(d) show the variation in $PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$, and $PD(\boldsymbol{E}_y, \boldsymbol{H}_y)$, respectively; Panel (e) shows the variation in polarization direction for the magnetic field, and Panel (f) shows the variation in the normalized hat matrix's diagonal element. One data in the frequency domain is regarded as one event. The horizontal axis denotes the event count. There are preferred directions of the phase difference between the orthogonal electric and magnetic fields corresponding to the leverage data, as shown in Panels (b) and (c).

Next, I discuss the MT impedance quality during the storm and non-storm days at site 142. Fig. 6.7 shows the MT sounding curves and coherence distributions. The data observed from 00:00 on October 26 to 00:00 on October 29 are used as the non-storm data. The coherence is low during non-storm days, and it coincides with the PD being scattered during non-storm days, as shown in Fig. 6.6 (a)-(d). It increased considerably and was close to one across all periods during the storm in the XY and YX components. The low coherence during the non-storm days could be attributed to incoherent noise. The storm MT sounding curve is smoother than the non-storm result in the XY and YX components, and the error bar is small. Both the XX and YY component is unstable. This storm result can be interpreted as a 1-D/2-D case with a high SNR.

Moreover, I compared the MT sounding curves calculated by single-site and remote reference data processing (see Fig. S1 in the supplementary document). It shows that the storm results are reliable. The impedance is biased using the non-storm data even though I used the remote reference data due to the continuous incoherent noise. The data quality is improved in a storm, and I can get reliable results using the storm data.

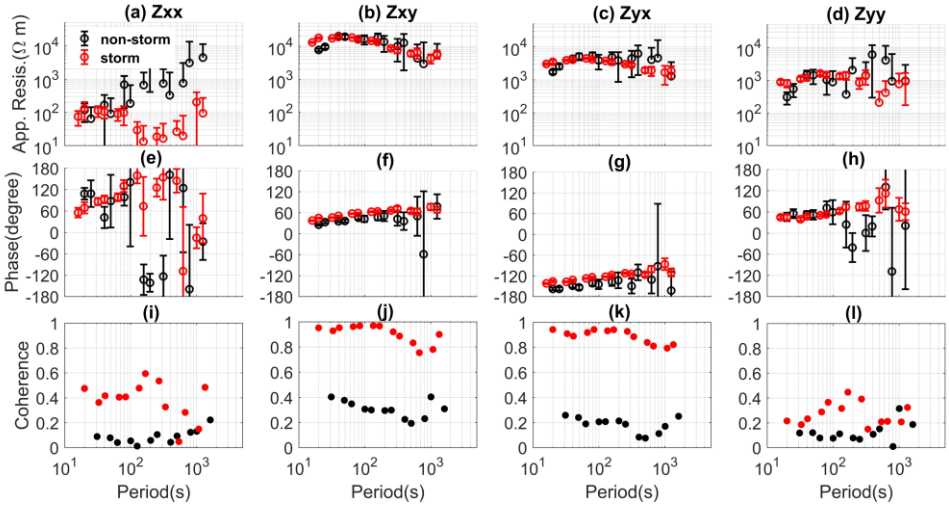

Fig. 6.7 MT sounding curves and coherence distributions during the storm and non-storm days using the data observed at site 142. The black color indicates the results calculated by the non-storm data; the red color indicates the results calculated by storm data. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. Panels (i)-(l) show the distribution of coherence. The horizontal axis denotes the period in seconds.

6.2.1.3 The influence of geomagnetic storms on the data in a PROQ case contaminated by the coherent noise (site 130)

Next, I do a similar analysis at site 130. The data quality analysis in 84 s is shown in Fig. 6.8. The data observed from 00:00 on October 26 to 00:00 on October 31 is used. Panels (a), (b), (c) and (d) show a preferred direction of PD close to 0° or 180°, and there is a preferred direction of polarization at approximately -30° during the non-storm day. The polarisation direction variation indicates that the non-storm data is dominated by coherent noise during the non-storm day. On the other hand, the PD between the electric and magnetic fields ($PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$ and $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$) has a preferred direction corresponding to the leverage point during the storm day. The enhanced natural EM signal may cause it.

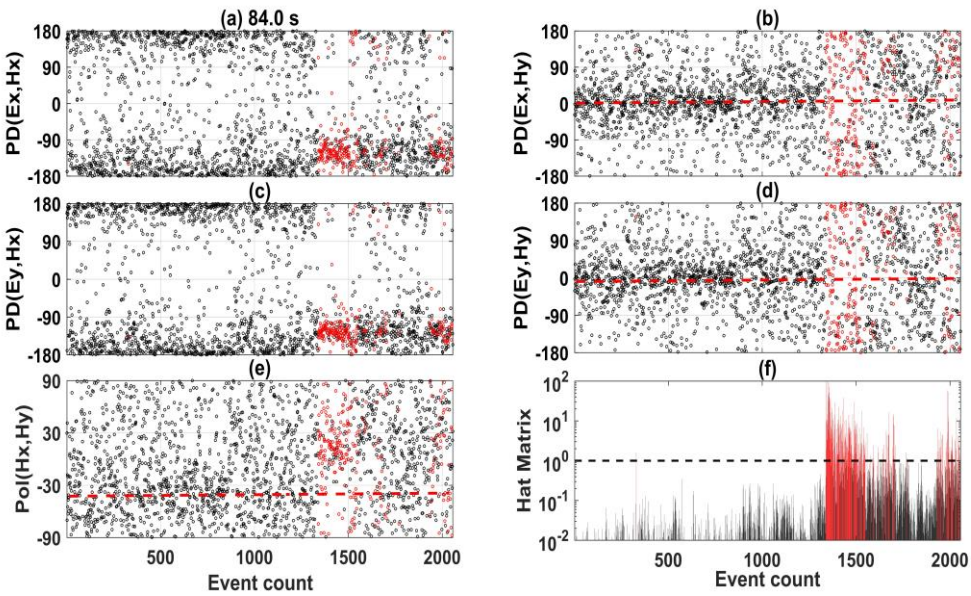

Fig. 6.8 Evaluation indices in 84 s using the data observed at site 130 from 00:00 on October 26

to 00:00 on October 31. Panels (a)-(d) show the variation in $PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$, and $PD(\boldsymbol{E}_y, \boldsymbol{H}_y)$; Panel (e) shows the variation in polarization direction for the magnetic field, and Panel (f) shows the variation in the normalized hat matrix's diagonal element. The data in red correspond to the leverage point, whose hat matrix diagonal element exceeds the expected value (2/*N*). One data in the frequency domain is regarded as one event. The horizontal axis denotes the event count. There are preferred directions of the phase difference between the electric and magnetic fields corresponding to the leverage data, as shown in Panels (a) and (c).

Next, I discuss the MT impedance quality during the storm and non-storm days at site 130. Fig. 6.9 shows the MT sounding curve and coherence distributions. The data observed from 00:00 on October 26 to 00:00 on October 29 are used as the non-storm data. The coherence values in $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ and $Coh(\boldsymbol{E}_y, \boldsymbol{H}_x)$ are high between 10 and 200 seconds during non-storm days; however, the *XX* and *YX* phases calculated by non-storm data are close to -180°, and the apparent resistivity increases as a line on the log scale between 10 and 200 seconds, which represents a phenomenon of local noise (Zonge and Hughes, 1987). A -180° would correspond to the grounded dipole noise sources (e.g., DC trains). On the other hand, the coherence value between the parallel electric and magnetic field $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ is high, while the coherence value between the orthogonal electric and magnetic field $Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$ is relatively low during storm days. Moreover, the result of Z_{xx} is stable, while that of Z_{xy} is unstable, and the phase of Z_{xy} rolls out of the normal quadrant. The storm result can be interpreted as the phenomenon of the PROQ case with a high SNR.

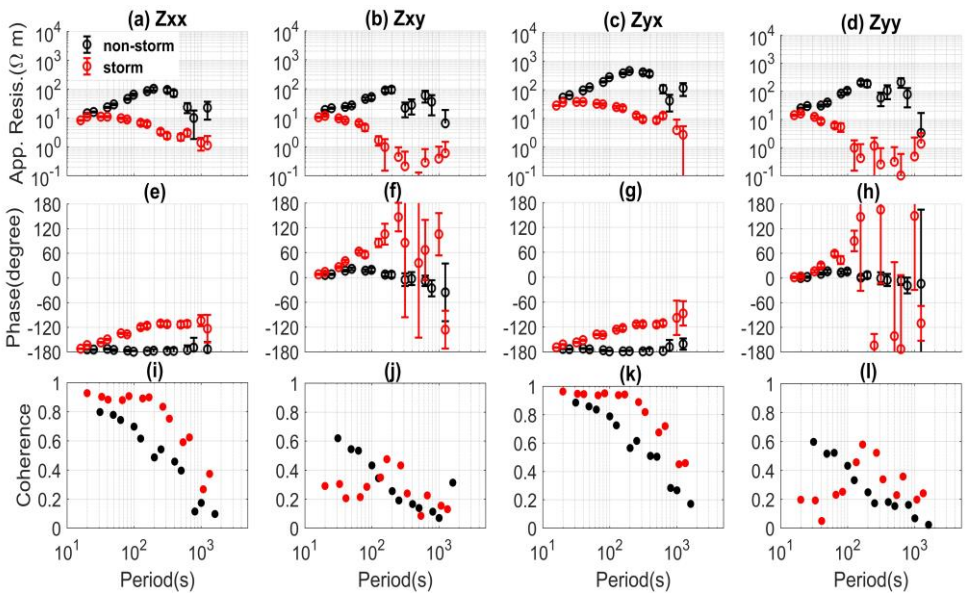

Fig. 6.9 MT sounding curves and coherence distributions during the storm and non-storm days using the data observed at site 130. The black color indicates the results calculated by the non-storm data; the red color indicates the results calculated by storm data. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. Panels (i)-(l) show the distribution of coherence. The horizontal axis denotes the period in seconds.

Moreover, I compare the MT sounding curves calculated by single-site and remote reference data processing at site 130 (see Fig. S2 in the supplementary document). All the results coincide with each other. However, the non-storm-RR result is more scattered, and the error bar is larger than the storm result. Comparing the single-site and remote reference processing results shows

that I can obtain a reliable result despite using single-site processing calculated by storm data.

6.2.1.4 The influence of geomagnetic storms on the data in a PROQ case contaminated by the incoherent noise (site 136)

Fig. 6.10 shows the data quality analysis in 84 s using the data observed at site 136 from 00:00 on October 26 to 00:00 on October 31. Fig. 6.10 (a) and (c) show that the PD between the electric and magnetic field ($PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$) have a preferred direction during the storm period. And all component of the PD is scattered during the non-storm day.

Fig. 6.11 shows the MT sounding curve and coherence distributions using the data observed during the storm and non-storm days at site 136 of KAP03. The non-storm data observed from October 26 to October 29 is used. The impedance curve is scattered, and the coherence is relatively low during the non-storm days. On the contrary, the coherence between the parallel component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ is high, while the coherence between the orthogonal component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$ is low during the storm day. Moreover, comparing the single-site and remote reference data processing results (see Fig. S3 in the supplementary document) show that the storm results are reliable. The impedance is biased using the non-storm data even though I used the remote reference data due to the continuous incoherent noise. For the north-south component, the result of Z_{xx} is stable while Z_{xy} is unstable, and the phase of Z_{xy} rolls out of the normal quadrant. The storm result can be explained as the phenomenon of PROQ in a quiet environment.

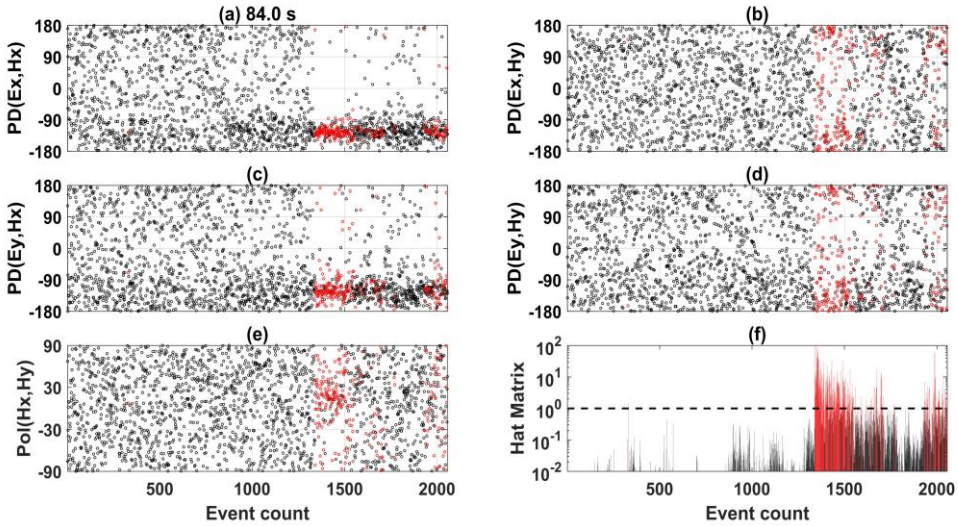

Fig. 6.10 The evaluation indices at site 136 in 84 s using the data observed from 00:00 on October 26 to 00:00 on October 31. Panels (a)-(d) show the variation in $PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$, and $PD(\boldsymbol{E}_y, \boldsymbol{H}_y)$; Panel (e) shows the variation in polarization direction for the magnetic field, and Panel (f) shows the variation in the normalized hat matrix's diagonal element. The data in red correspond to the leverage point, whose hat matrix diagonal element exceeds the expected value (2/*N*). One data in the frequency domain is regarded as one event. The horizontal axis denotes the event count.

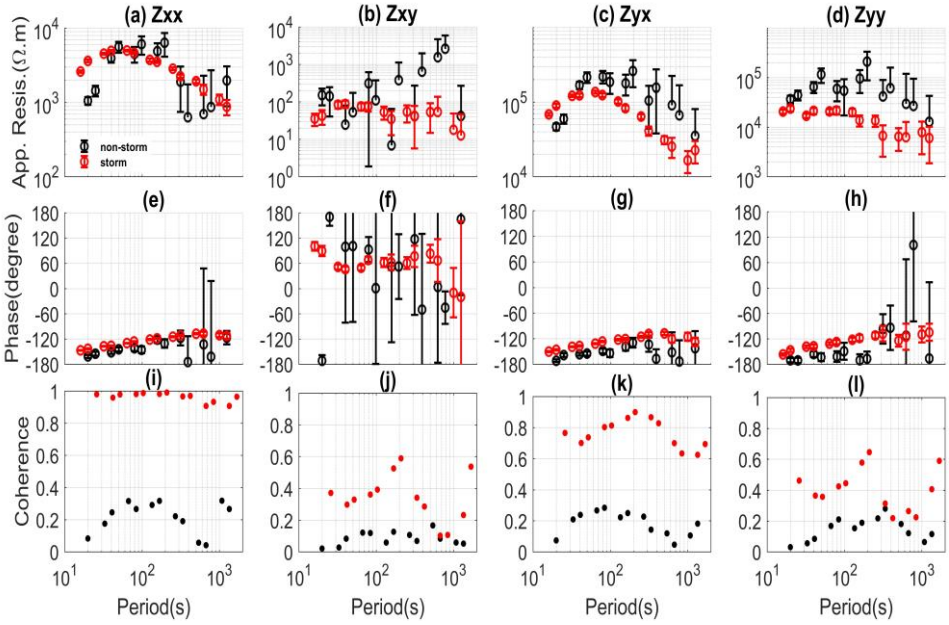

Fig. 6.11 The MT sounding curves and coherence distributions using the data observed during the storm and non-storm days at site 136 of KAP03. The black color indicates the results calculated by the non-storm data; the red color indicates the results calculated by storm data. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. Panels (i)-(l) show the distribution of coherence. The horizontal axis denotes the period in seconds.

6.2.1.5 The influence of geomagnetic storms on the data in a 2-D/3-D case contaminated by the incoherent noise (site 139)

Fig. 6.12 shows the data quality analysis in 84 s using the data observed at site 139 from 00:00 on October 26 to 00:00 on October 31. Fig. 6.12 (a) and (c) show that the PD between the electric and magnetic field ($PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$) have a preferred direction during the storm period. It is similar to the result at site 136. But the other component of PD also has a preferred direction, but not so concentrated. Therefore, I categorize it into a 2-D/3-D case during the storm day.

Fig. 6.13 shows the MT sounding curve and coherence distributions using the data observed during the storm and non-storm days at site 139 of KAP03. The non-storm data observed from October 26 to October 29 is used. The impedance curve is scattered, and the error bar is large; on the other hand, the coherence is relatively low during the non-storm days. On the contrary, the coherence between the parallel component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ is high, while the coherence between the orthogonal component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$ is low during the storm day. Moreover, comparing the single-site and remote reference data processing results (see Fig. S4 in the supplementary document) show that the storm results are reliable. The impedance is biased using the non-storm data even though I used the remote reference data due to the continuous incoherent noise. The storm result can be explained as a 2-D/3-D case in a quiet environment.

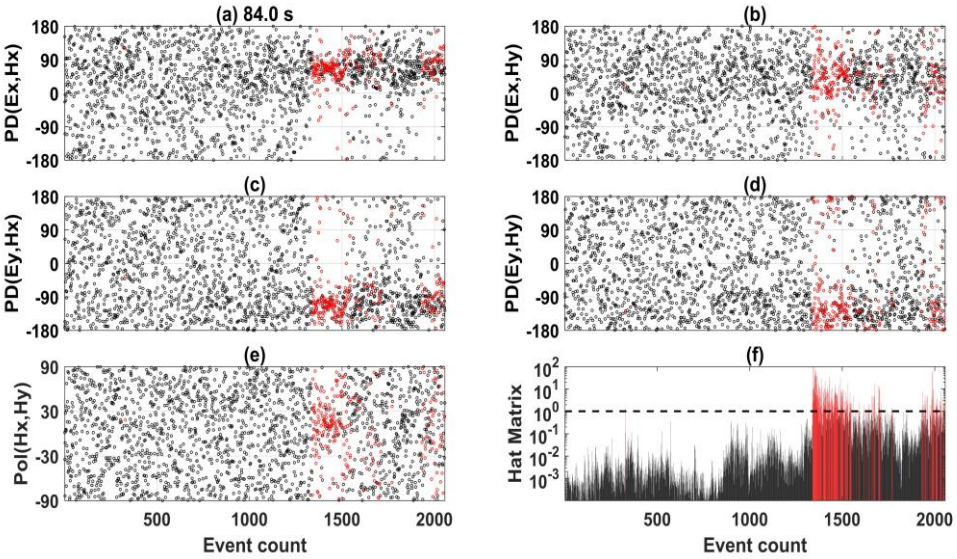

Fig. 6.12 The evaluation indices at site 139 in 84 s using the data observed from 00:00 on October 26 to 00:00 on October 31. Panels (a)-(d) show the variation in $PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$, and $PD(\boldsymbol{E}_y, \boldsymbol{H}_y)$; Panel (e) shows the variation in polarization direction for the magnetic field, and Panel (f) shows the variation in the normalized hat matrix's diagonal element. The data in red correspond to the leverage point, whose hat matrix diagonal element exceeds the expected value (2/*N*). One data in the frequency domain is regarded as one event. The horizontal axis denotes the event count.

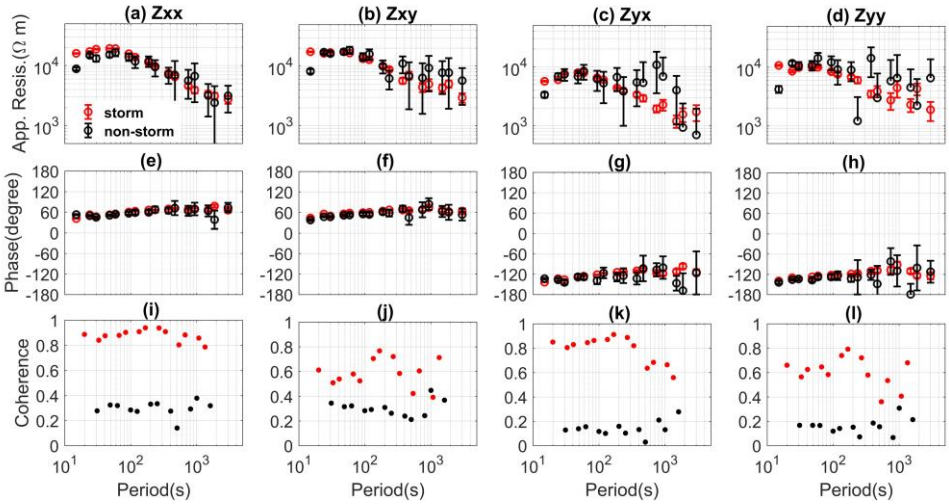

Fig. 6.13 The MT sounding curves and coherence distributions using the data observed during the storm and non-storm days at site 139. The black color indicates the results calculated by the non-storm data; the red color indicates the results calculated by storm data. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. Panels (i)-(l) show the distribution of coherence. The horizontal axis denotes the period in seconds.

6.2.1.6 The influence of geomagnetic storms on the data in a 2-D/3-D case contaminated by the coherent noise (site 133)

Fig. 6.14 shows the data quality analysis in 84 s using the data observed at site 133 from 00:00 on October 26 to 00:00 on October 31. Fig. 6.14 (a), (b), (c), and (d) show that the PD between

the electric and magnetic fields have a preferred direction, but it is different during the storm and non-storm period. And there is a preferred direction of polarization around -30° during the non-storm day. The coherent noise dominates the non-storm data.

Fig. 6.15 shows the MT sounding curve and coherence distributions using the data observed during the storm and non-storm days at site 133. The non-storm data observed from October 26 to October 29 is used. The coherence value in $Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$ are high between 10 and 400 seconds during the non-storm days; however, the *XY* and *YX* phases calculated by non-storm data are close to 0°, and the apparent resistivity increases as a line on the log scale. That is the phenomenon of local noise. Moreover, comparing the single-site and remote reference data processing results (see Fig. S5 in the supplementary document) show that the storm results are reliable. I can't find any unusual in the impedance result. The impedance is biased using the non-storm data even though I used the remote reference data due to the coherent noise. On the other hand, the coherence distribution during the storm can't be explained as the 1-D/2-D case or PROQ. Therefore I categorized the storm result as a 2-D/3-D case in a quiet environment.

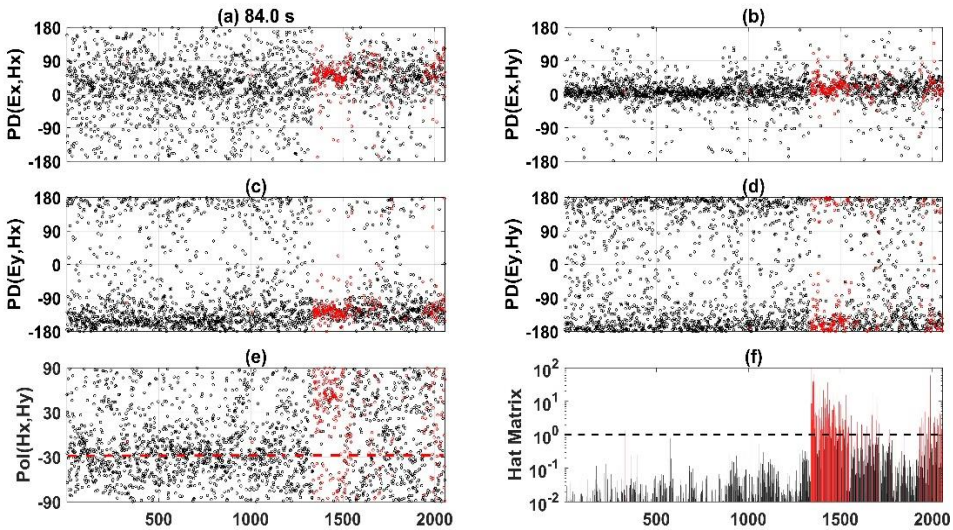

Fig. 6.14 The evaluation indices at site 133 in 84 s using the data observed from 00:00 on October 26 to 00:00 on October 31. Panels (a)-(d) show the variation in $PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$, and $PD(\boldsymbol{E}_y, \boldsymbol{H}_y)$; Panel (e) shows the variation in polarization direction for the magnetic field, and Panel (f) shows the variation in the normalized hat matrix's diagonal element. The data in red correspond to the leverage point, whose hat matrix diagonal element exceeds the expected value (2/*N*). One data in the frequency domain is regarded as one event. The horizontal axis denotes the event count.

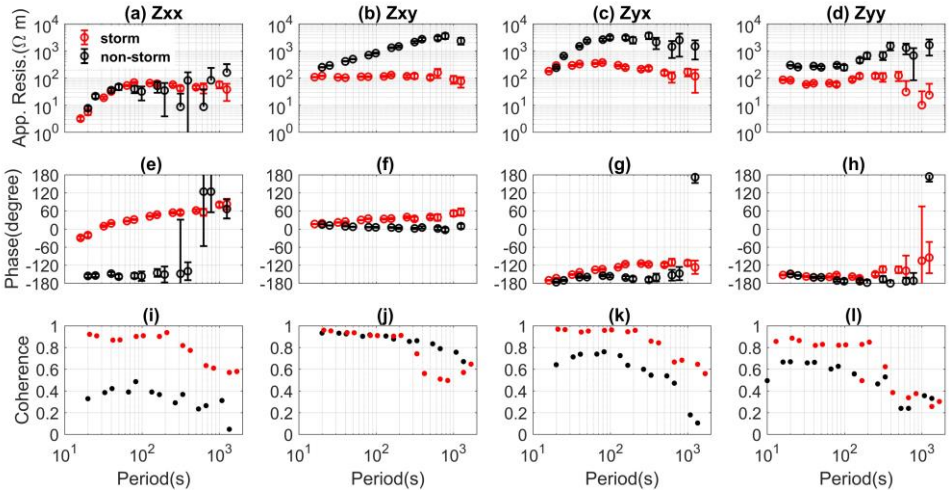

Fig. 6.15 The MT sounding curves and coherence distributions using the data observed during the storm and non-storm days at site 133. The black color indicates the results calculated by the non-storm data; the red color indicates the results calculated by storm data. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. Panels (i)-(l) show the distribution of coherence. The horizontal axis denotes the period in seconds.

6.2.2 Case study 2: USArray, USA

The second field data used the long-period MT time-series data observed at TNV48 from the USArray project. The data sets were recorded with a 1-second sampling period for around two weeks in 2015. The geomagnetic storm started at around 18:00 on June 22 and ended at around 00:00 on June 24, as shown in Fig. 6.16.

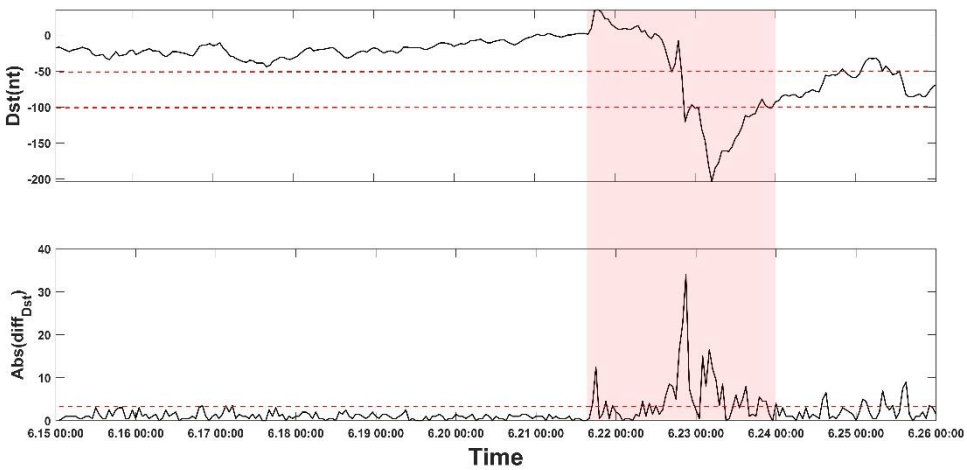

Fig. 6.16 The variation in the *Dst* index. The upper figure shows the variation in the *Dst* index. The lower figure shows the absolute value of *Diff(Dst)*; the red dash line denotes 1.5σ (σ is the standard deviation). The horizontal axis denotes the time in UTC. The geomagnetic storm started at around 18:00 on June 21 and ended at around 00:00 on June 24.

Fig. 6.17 shows the data quality analysis in 13.1 s using the data observed from June 19 to June 24. Panel (b) and (c) show that $PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$ and $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$ have a preferred direction

corresponding to the leverage point. Because the nature signal around the dead band (0.1 -10 s) is relatively low, the local noise can easily influence it. When there is a geomagnetic storm, the nature EM signal's strength increases and the signal-to-noise ratio (SNR) becomes high. A preferred direction of PD between the orthogonal electric and magnetic fields appears during the storm. The data is dominated by incoherent noise during the non-storm period.

Fig. 6.18 shows MT sounding curves and coherence distributions during the storm and non-storm days. The non-storm data observed from June 19 to June 21 and storm data from June 22 to June 24 are used. The coherence value between the orthogonal electric and magnetic fields increased obviously from 4 to 40-second and 400 to 2,000-second during the storm. The low coherence during the non-storm day can be attributed to the incoherent local noise. The non-storm result's apparent resistivity in the period between 8 s and 30 s is severely down-biased, and the phase's error bar is bigger than that of the storm result. I think the incoherent noise makes the impedance severely down-biased between 8 to 30-second during the non-storm days.

Moreover, comparing the single-site and remote reference data processing (see Fig. S7 in the supplementary document) shows that the storm results using the remote reference processing are reliable. Although a more reliable impedance can be obtained using the storm data with single site data processing. But the data also contain much intermittent incoherent noise. And the remote reference can remove such kinds of noise. And a more reliable result can be obtained by remote reference processing.

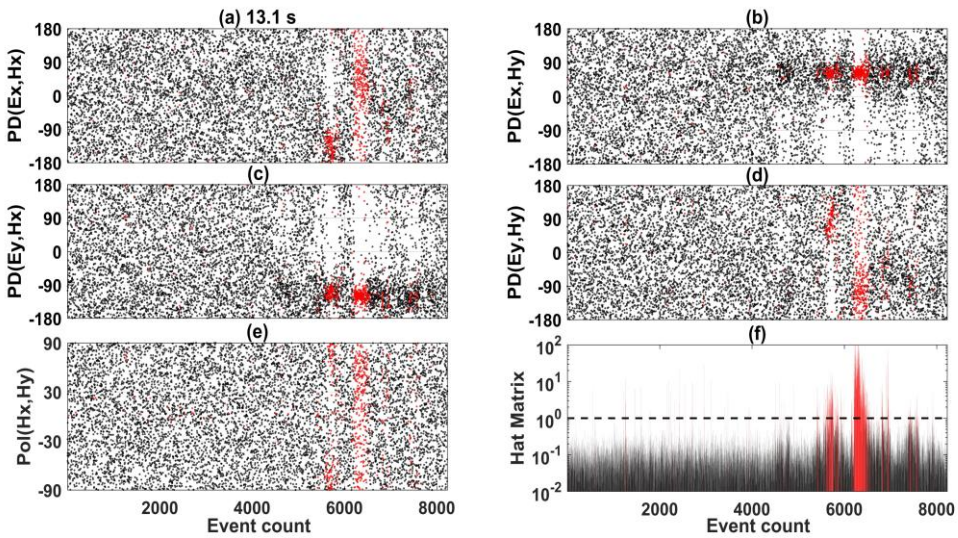

Fig. 6.17 The evaluation indices at site TNV48 in the period of 13.1 s using the data observed from 00:00 on June 19 to 00:00 on June 24. Panels (a)-(d) show the variation in $PD(\boldsymbol{E}_x, \boldsymbol{H}_x)$, $PD(\boldsymbol{E}_x, \boldsymbol{H}_y)$, $PD(\boldsymbol{E}_y, \boldsymbol{H}_x)$, and $PD(\boldsymbol{E}_y, \boldsymbol{H}_y)$; Panel (e) shows the variation in polarization direction for the magnetic field, and Panel (f) shows the variation in the normalized hat matrix's diagonal element. The data in red correspond to the leverage point, whose hat matrix diagonal element exceeds the expected value ($2/N$). One data in the frequency domain is regarded as one event. The horizontal axis denotes the event count.

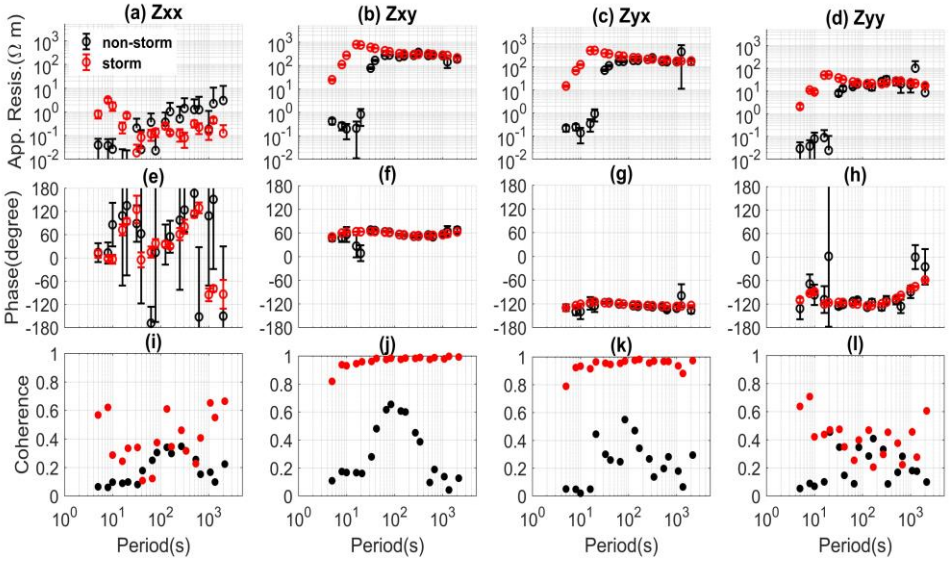

Fig. 6.18 The MT sounding curves and coherence distributions during the storm and non-storm days using the data observed at TNV48 from USArray. The black color indicates the results calculated by the non-storm data; the red color indicates the results calculated by storm data. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. Panels (i)-(l) show the distribution of coherence. The horizontal axis denotes the period in seconds.

6.2.3 Case study 3: Sawauchi, Japan

The Phoenix Geophysics system's broadband frequency 5-component MT time-series data were used in this case study. The data were observed from August 20 to August 28, 2018, at Sawauchi station, Japan. The geomagnetic storm occurred around August 26. The low-frequency data (15 Hz) sampled continuously was used in the research.

I calculated the MT impedance using each day's data at first. Fig. 6.19 shows typical MT sounding curves and the coherence distributions using the data observed during the storm day (August 26) and non-storm day (August 23). In Fig. 6.12, the coherence value between parallel component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ is much higher than the coherence value between orthogonal component $Coh(\boldsymbol{E}_x, \boldsymbol{H}_y)$. Moreover, the impedance phase in the period over than 200-second rolls out of the quadrant. That is the phenomenon of PROQ. Because the environment is quiet, the high SNR data's proportion is high enough during the storm and non-storm days; therefore, the impedance results calculated by the storm and non-storm data are similar.

Comparing the two impedance results in detail, the coherence value in $Coh(\boldsymbol{E}_x, \boldsymbol{H}_x)$ increased during the storm between 4 and 30 seconds. That may be caused by the increased intensity of the natural EM signal. Because the natural EM signal strength is very low in the dead band (0.1-10 seconds), local noise can easily influence it. The enhancement of the natural EM signal may produce a more reliable impedance result in that period. Next, I investigate the change in impedance result during the storm and non-storm days at 10-second. Fig. 6.20 shows the time variations in the *XY* component calculated by each hour's data in the period of 10-second. The impedance curve becomes more stable and correlated with the geomagnetic storm event.

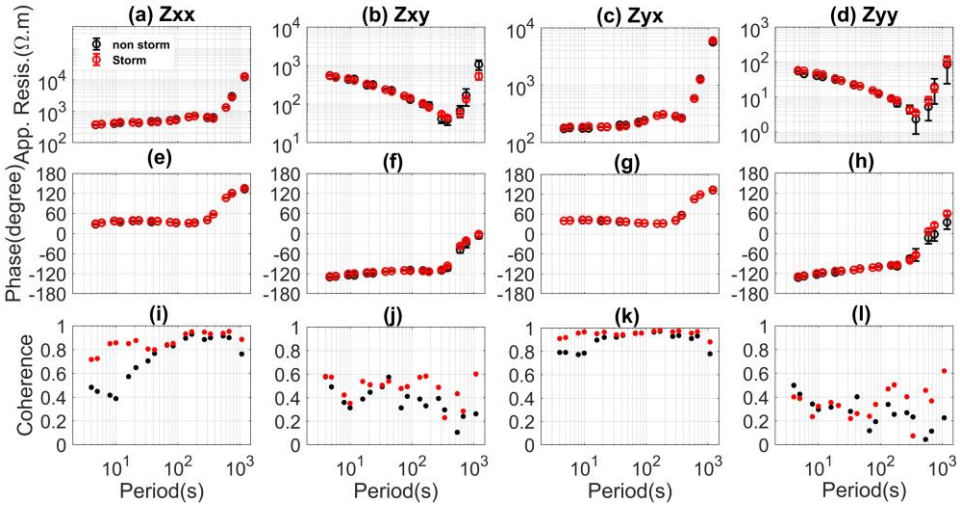

Fig. 6.19 The MT sounding curves and coherence distributions during storm days (August 26) and non-storm days (August 23). The black color indicates the results calculated by the non-storm data; the red color indicates the results calculated by storm data. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. Panels (i)-(l) show the distribution of coherence. The horizontal axis denotes the period in seconds.

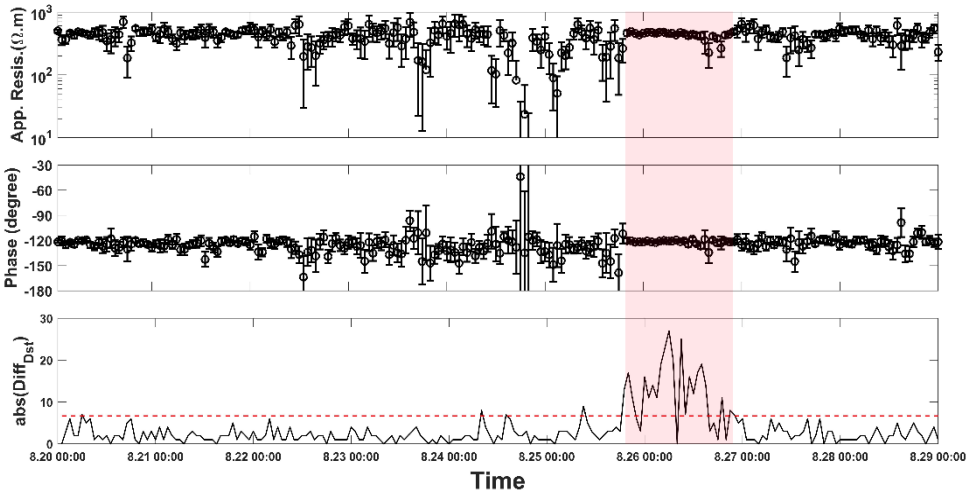

Fig. 6.20 The time variation of the apparent resistivity and phase calculated using each hour's time-series data in the period of 10-second. One result was calculated using one-hour data. The upper figures show the apparent resistivity. The second-row figures show the impedance phase. The lower figure shows the absolute value of *Diff(Dst)*; The horizontal axis denotes the time in UTC.

6.3 Discussion

First, I discuss how to analyze the data quality; the characteristics of the 1-D/2-D case present low coherence between the parallel electric and magnetic fields and high coherence between the orthogonal components. In contrast, the characteristics of PROQ present high coherence between the parallel electric and magnetic fields and low coherence between the orthogonal components. When the coherence is high, there is a preferred direction of PD. For 1-D/2-D or PROQ situations with good quality, there is always a preferred direction of PD between the corresponding electric and magnetic fields. On the other hand, the time-series data contaminated

by coherent noise also have high coherence values and a preferred direction in PD. Therefore, I combine the polarization direction to confirm whether the coherent noise contaminates the data. There is no preferred polarization direction for the magnetic field in a quiet EM environment. If the magnetic field has a preferred polarization direction, I can consider that the data are contaminated by coherent noise in that period.

Moreover, the hat matrix is a parameter used to visualize the energy variation (see the example in Fig. 6.3). Therefore, I can combine the phase difference between the electric and magnetic fields, the polarization direction and the hat matrix to discuss how the data quality varies with energy changes for a specific period.

Finally, I discuss the different influences of strong and weaker storms on the impedance calculation. According to the least-squares theory (Sims et al., 1971), for the north-south direction, Z_{xy} can be calculated as follows:

$$Z_{xy} = \frac{<E_x\bar{H}_y><H_x\bar{H}_x> - <E_x\bar{H}_x><H_x\bar{H}_y>}{<H_y\bar{H}_y><H_x\bar{H}_x> - <H_y\bar{H}_x><H_x\bar{H}_y>} , \tag{6.12}$$

There are three cross-power spectra densities ($< \boldsymbol{E_x\bar{H}_x} >, < \boldsymbol{E_x\bar{H}_y} >, < \boldsymbol{H_x\bar{H}_y} >$) and two auto-power spectra densities ($< \boldsymbol{H_x\bar{H}_x} >, < \boldsymbol{H_y\bar{H}_y} >$) . The energy of the signal and noise vary with time; therefore, the SNR varies with time, and I can rewrite the cross-power and auto-power spectra densities as follows:

$$<\mathbf{A}\bar{\mathbf{B}}> = <\mathbf{A}^S\bar{\mathbf{B}}^S> + <\mathbf{A}^N\bar{\mathbf{B}}^N>, \tag{6.13}$$

$$<\mathbf{A}\bar{\mathbf{A}}> = <\mathbf{A}^S\bar{\mathbf{A}}^S> + <\mathbf{A}^N\bar{\mathbf{A}}^N>, \tag{6.14}$$

where $<\mathbf{A}^S\bar{\mathbf{A}}^S>$ and $<\mathbf{A}^S\bar{\mathbf{B}}^S>$ are the auto-power and cross-power spectra densities calculated by the high SNR data and $<\mathbf{A}^N\bar{\mathbf{B}}^N>$ and $<\mathbf{A}^N\bar{\mathbf{A}}^N>$ are the auto-power and cross-power spectra densities calculated by the low SNR data. The noise influences the cross-power and auto-power spectra density between the electric and magnetic fields, influencing the impedance calculation of Eq. 6.12.

In a very strong storm, such as the first case study, the energy also varies intensely during the storm. On the one hand, the amplitude of the leverage point can sometimes be 100 times that with low-energy data, as shown in Fig. 4. The influence of the high SNR data on the spectra is 10,000 times that of the low SNR data. On the other hand, the storm lasts a long time in a very strong storm, which means that the proportion of data with high SNR is high. At that time, $<\mathbf{A}^N\bar{\mathbf{B}}^N>$ and $<\mathbf{A}^N\bar{\mathbf{A}}^N>$ can be neglected compared with $<\mathbf{A}^S\bar{\mathbf{A}}^S>$ and $<\mathbf{A}^S\bar{\mathbf{B}}^S>$. Examples of sites 130 and 142 of KAP03 show that the single-site result coincides with the remote reference result calculated by storm data. Although using single-site processing, I can obtain a reliable result in a very strong storm.

In the case of a less intense storm, such as the second case study, the intensity is smaller than that in the first case study and lasts a shorter time, which means that the proportion of high SNR data is low. In this situation, $<\mathbf{A}^N\bar{\mathbf{B}}^N>$ and $<\mathbf{A}^N\bar{\mathbf{A}}^N>$ cannot be neglected compared with $<\mathbf{A}^S\bar{\mathbf{A}}^S>$ and $<\mathbf{A}^S\bar{\mathbf{B}}^S>$. Even if high SNR data are observed, it is not sufficient to obtain a reliable result using single-site processing; in this case, remote reference technology(Gamble et al., 1979) may reduce the influence of noise. I can see the example of site TNV48 of USArray. Assuming that the remote site is unavailable, I can also use the preselection strategy, e.g., the method proposed by Weckmann et al. (2005), to extract the high SNR data and obtain a reliable impedance.

6.4 Conclusions

I used the three field datasets observed at midlatitude to investigate the influence of a geomagnetic storm on the MT impedance from the perspective of SNR. The following is a summary of the finding from our case studies.

In a quiet EM environment, the high SNR data's proportion is high enough to get a reliable impedance during the storm and non-storm days; therefore, the impedance results calculated by

the storm and non-storm data are similar. This situation is shown by the data observed at site Sawauchi and site 163 of KAP03. Moreover, the positive effect of the geomagnetic storm on the MT impedance quality is shown at 10-second around the dead band. Because the natural EM signal is relatively low, the local noise can easily influence it. When there is a geomagnetic storm, the nature EM signal's strength increase and, the SNR become higher, the impedance curve becomes more stable and reliable at 10-second.

In a noisy EM environment, the artificial noise dominates the data quality during non-storm days. In the presence of a geomagnetic storm, the high SNR data increases; the noisy data (data types 4 and 5) convert into good data (data types 1, 2, and 3). Therefore, I can get a reliable impedance using the storm data. This situation is shown by the data observed at site TNV48 of USArray and sites 130, 133, 136, 139, and 142 of KAP03.

MT field data include natural EM signal and artificial EM noise. Especially in urban areas, artificial disturbances to EM observations are serious. The observation occasionally contains continuous noise, which is difficult to remove by the current technique. I recommend performing the MT campaign during the geomagnetic storm if I need to re-survey at such sites. The geomagnetic storm has a seasonal and 11-year solar cycle and doesn't frequently occur (see the statistical analysis of geomagnetic storms in chapter 2). The space weather forecast information makes it possible to get storm data by predicting the geomagnetic storm. The Space Weather Prediction Center (SWPC; see the website in references) provides information about space weather in the coming three days. Utilizing the data observed during a strong geomagnetic storm may overcome the local noise's influence and obtain a more reliable and interpretable impedance, although the site is contaminated by continuous noise.

Chapter 7 : SUMMARY

This thesis mainly studied the algorithm of the MT impedance estimation, and it is changing to get a reliable impedance result in the presence of noise. MT field data include natural signal sources and noise (e.g., instrumental, humanmade, and so on). Not all time segments contain information on the depth resistivity distribution. Before MT data processing, noisy data should be considered in the impedance estimation. Both the amplitude of the signal and noise vary with time.

The initial MT impedance estimator is based on the least-squares theory (Sims et al., 1971), and the MT impedance can be severely disturbed by cultural noise. Removing these disturbances is mainly based on robust statistics (Egbert and Booker, 1986; Chave and Tompson, 1989; Smirnov, 2003; Chave and Tompson, 2003; Chave and Thomson, 2004), remote reference processing (Gamble et al., 1979) or their combination.

Remote reference processing uses the cross-power spectra instead of auto-power spectra when performing regressions based on the least-squares estimator (Gamble et al., 1979). Remote reference processing can improve the MT impedance quality when the electromagnetic noise is uncorrelated between the local and remote sites. On the other hand, suppose the local site is contaminated by continuous noise; the remote reference method rarely produces a reliable result.

Robust algorithms aim to detect and reject outliers by a data-adaptive weighting scheme. These methods require reasonable proportions of normal data to yield reliable results, e.g., data with no more than 50% contamination (Smirnov, 2003b).

At the same time, MT researchers proposed multisite processing. Larsen et al. (1996) and Oettinger et al. (2001) proposed the signal-noise separation method (SNS). SNS used the remote magnetic field to estimate the interstation transform function as the separation tensor; they separated the local magnetic field into the signal and noise parts and then calculated the impedance. Egbert (1997) proposed a robust multivariate errors-in-variables estimator to separate the field data into the signal and noise using the principal component analysis. A more recent application of the method is shown in Smirnov and Egbert (2012). Both RMEV and SNS methods used the robust approaches mentioned above in their data processing. Those methods may bias when the local noise content is more than half.

All statistical algorithms mentioned above can obtain a reliable result when most of the recorded data is well-behaved. When the noise content is more than 50%, it is practical to use a preselection strategy to reduce the EM noise to a level that the robust estimator can deal with (Weckmann et al., 2005).

In real data processing, I don't know the content of the background noise and which method is effective for a specific case study. Therefore, it is necessary to know the situation of the background noise. We proposed a new method to estimate the data quality of the MT time-series data. By this method, we can extract the high signal-to-noise ratio (SNR) data automatically. We can get a reliable result assuming the field data is not contaminated by the continuous noise. We also used the data quality analysis method to investigate the influence of geomagnetic storms on the MT impedance calculation. We found that the geomagnetic storm is beneficial to data quality; the high SNR data may appear during a geomagnetic storm. And we can get a more reliable result using the data observed during the geomagnetic storm in a noisy site. Utilizing geomagnetic storm data may give us a reliable result despite the site being contaminated by continuous noise.

Another issue of noise reduction to MT field data is noise identification and modification (Kappler, 2012; Wang et al., 2017). They replace the abnormal time series by the relationship between the locale and remote site in the time domain, and the corresponding abnormal spectra are changed to the normal data. But when a large amount of noise contaminates the data, those methods may fail to get reliable results.

In conclusion, comparing all the methods in the MT time-series processing, I think the preselection strategy is the most effective and attractive method. Assuming the continuous noise doesn't contaminate the field data, it is possible to get a reliable result. The calculation of the impedance tensor is very dependent on the data quality. In the future, I hope the data quality analysis method proposed by this thesis can be applied in the fieldwork. While doing the field observation, I can know the situation of background noise, making sure there are high-quality data during the observation period.

THE SUPPLEMENTARY INFORMATION

1 The supplementary information for site 142 of KAP03.

Fig. S1 compares the MT sounding curves with and without the remote reference processing using the storm and non-storm data observed at site 142. The non-storm data from October 26 to October 29 is used. Site 151 is used as the remote reference site. The results calculated by storm data with and without remote reference processing coincide, while the non-storm-RR result is scattered, and the error bar is also large. The impedance is biased using the non-storm data despite remote reference data being used due to the continuous incoherent noise, according to the data quality analysis in Fig. 6.4. The data quality is improved during the storm, and I can obtain a reliable impedance even using single-site data processing.

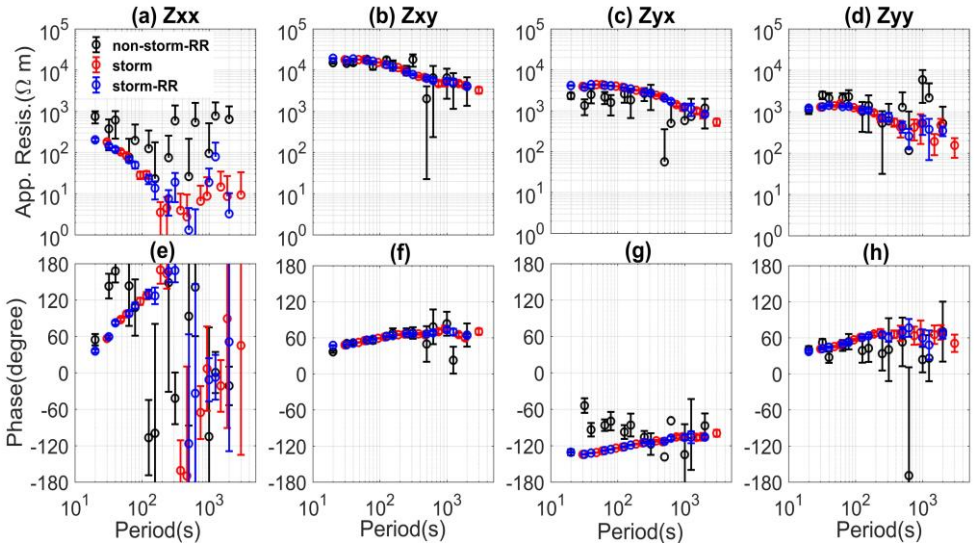

Fig. S1 Comparison of the MT sounding curves calculated by single-site and remote reference data processing using the storm and non-storm data observed at site 142. The black curves are calculated by the non-storm data with remote reference processing (non-storm-RR); the blue curves are calculated by storm data with remote reference processing (storm-RR), and the red curves are calculated by storm data with single-site data processing (storm). Panel (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. The horizontal axis denotes the period in seconds.

2 The supplementary information for site 130 of KAP03.

Fig. S2 compares the MT sounding curves with and without the remote reference processing using the storm and non-storm data observed at site 130. The non-storm data from October 26 to October 29 is used. Site 151 is used as the remote reference site. All the results coincide with each other. However, the non-storm-RR result is more scattered, and the error bar is larger than the storm result. Comparing the single-site and remote reference processing results shows that I can obtain a reliable result despite using single-site processing calculated by storm data.

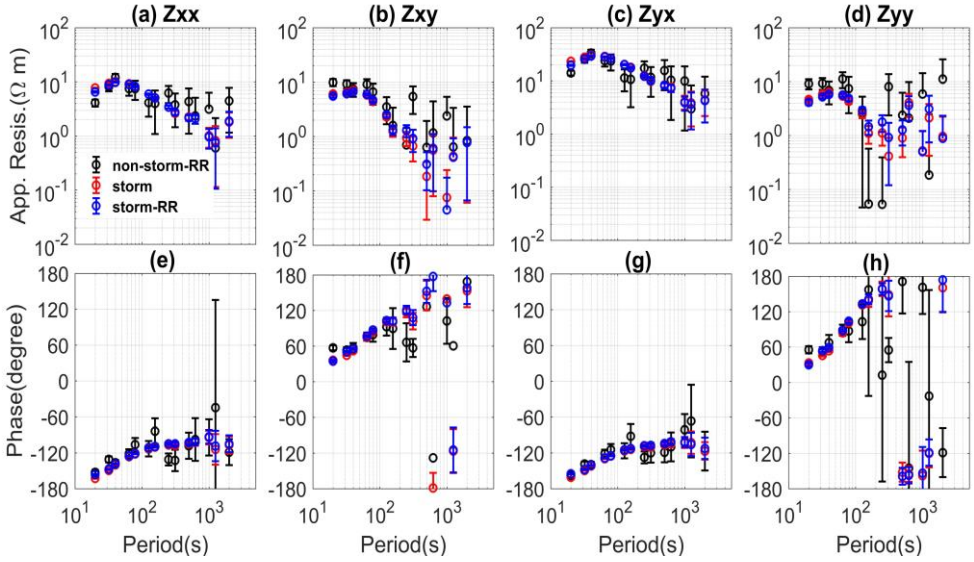

Fig. S2 Comparison of the MT sounding curves calculated by single-site and remote reference data processing using the storm and non-storm data observed at site 130. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. The horizontal axis denotes the period in seconds. The storm result and storm-RR result coincide and cover each other.

3 The supplementary information for site 136 of KAP03.

Fig. S3 compares the MT sounding curves with and without the remote reference processing using the storm and non-storm data observed at site 136. The non-storm data from October 26 to October 29 is used. Site 163 is used as the remote reference site. The storm result with and without the remote reference processing coincide. While the remote result using the non-storm day is scattered, and the error bar is large. The phase rolls out normal quadrant appear in *XY* component. It shows the storm result is reliable.

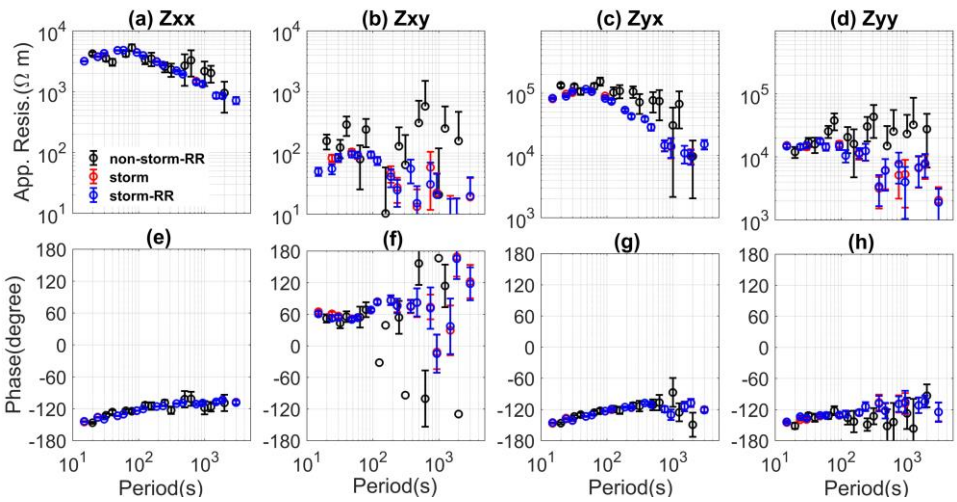

Fig. S3 Comparison of the MT sounding curves with and without the remote reference processing using the storm and non-storm data using the data observed at site 136. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase.

4 The supplementary information for site 139 of KAP03.

Fig. S4 compares the MT sounding curves with and without the remote reference processing using the storm and non-storm data observed at site 139. The non-storm data from October 26 to

October 29 is used. Site 163 is used as the remote reference site. The storm result with and without the remote reference processing coincide. While the remote result using the non-storm day is scattered, and the error bar is large. It shows that the storm result is reliable.

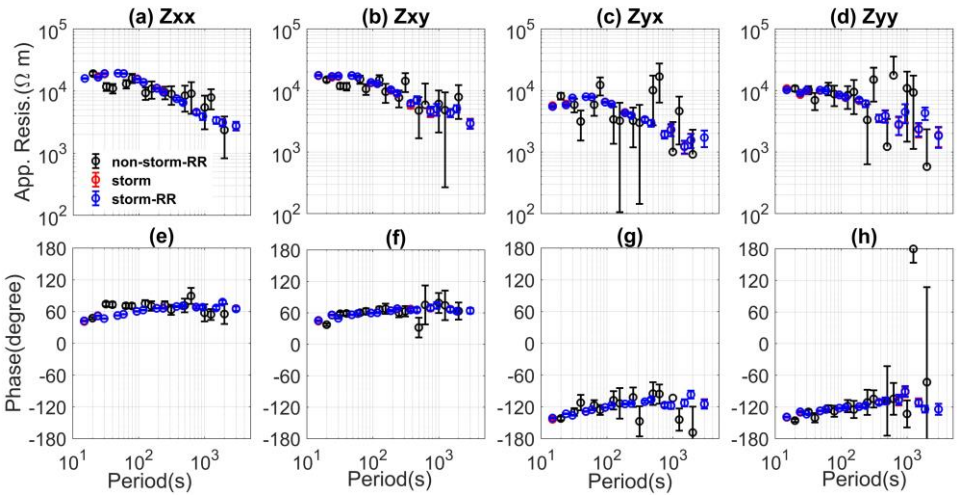

Fig. S4 Comparison of the MT sounding curves with and without the remote reference processing using the storm and non-storm data using the data observed at site 139. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase.

5 The supplementary information for site 133 of KAP03.

Fig. S5 compares the MT sounding curves with and without the remote reference processing using the storm and non-storm data observed at site 133. The non-storm data from October 26 to October 29 is used. Site 163 is used as the remote reference site. The storm result with and without the remote reference processing coincide. While the remote result using the non-storm day is scattered, and the error bar is large. It shows that the storm result is reliable.

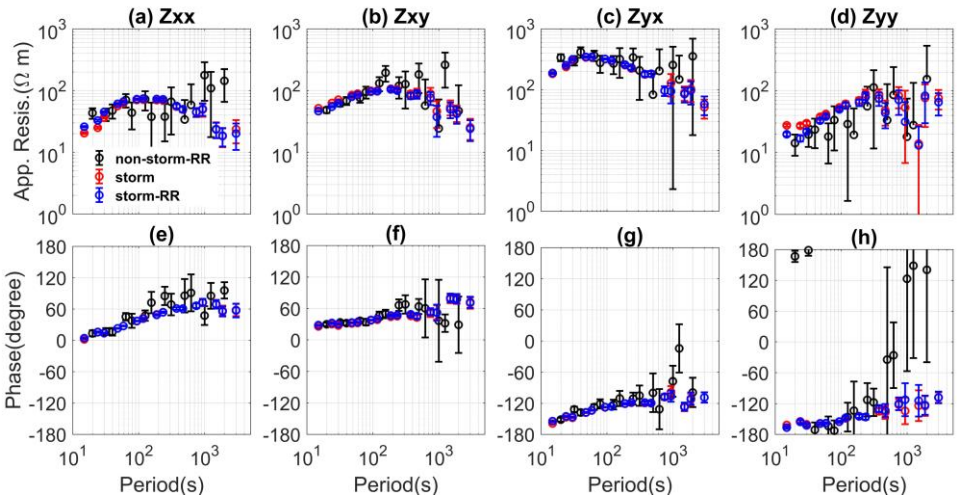

Fig. S5 Comparison of the MT sounding curves with and without the remote reference processing using the storm and non-storm data using the data observed at site 133. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase.

6 The supplementary information for TNV48.

The disturbance storm time (*Dst*) index is an index of geomagnetic activity used to estimate the averaged change of the horizontal component of the Earth's magnetic field. When the *Dst* index

is less than -50 nT, it is categorized as a geomagnetic storm. When the *Dst* index is less than -100 nT, it is categorized as a strong geomagnetic storm. Fig. S6 shows the variation in the *Dst* index in July 2015. There was a geomagnetic storm around June 23.

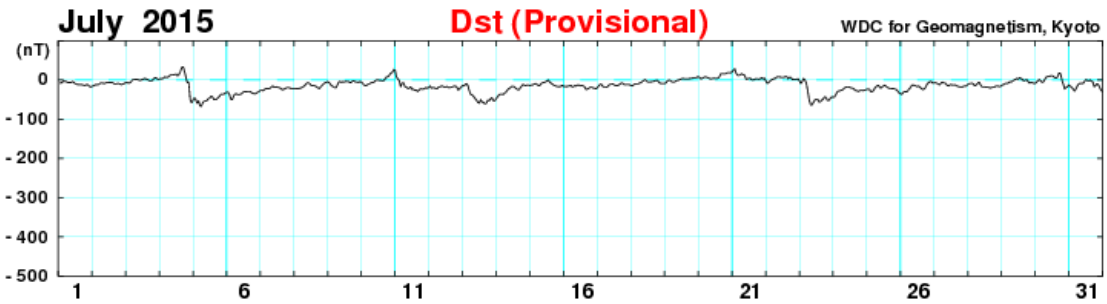

Fig. S6 The variation in the *Dst* index in July 2015 (Download from the homepage of WDC Kyoto Observatory: http://wdc.kugi.kyoto-u.ac.jp/dstdir/index.html).

Fig. S7 compares the MT sounding curves with and without the remote reference processing using the storm and non-storm data. The data observed at site ALW48 is used as the remote site. And the non-storm data from June 19 to June 21 and the storm data from June 22 to June 24 are used. All the result coincides in a period longer than 40 s. the incoherent noise below 40 s contaminates the data. Comparing all the results and the data quality analysis in 13.1 s, I think the storm result with the remote reference processing is reliable. The incoherent noise leads the apparent resistivity down biased, and the remote reference data processing can remove such noise.

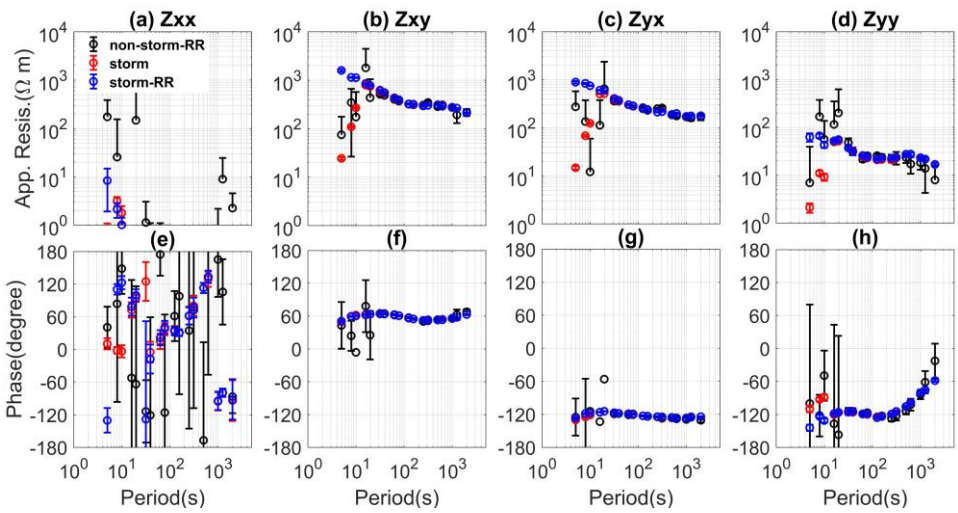

Fig. S7 Comparison of the MT sounding curves calculated by single-site and remote reference data processing using the storm and non-storm data observed at site TNV48. Panels (a)-(d) show the apparent resistivity. Panels (e)-(h) show the impedance phase. The horizontal axis denotes the period in seconds. The MT sounding curves longer than 40 s are similar and cover the XY and YX components.

ACKNOWLEDGEMENT

First of all, I appreciate my supervisor, Pro. Mizunaga. He allowed me great freedom to develop my skills. He taught me how to write the paper and helped me modify the manuscript carefully. I hope to use this thesis to reward his trust and help.

I also thank my tutor, Tanaka sensei. Whenever I came to him with any query, he welcomed me warmly and helped me solve the problem.

I thank my friend Yutian Zhang; he taught me useful tools, like downloading paper and books. We also will discuss some problems in our studies and daily life.

I thank the senior Gang Wang, Lei Zhou, and Lingqiang Zhao give me some valuable field data to do the research.

I also thank my senior, Peng Han, for giving me useful comments on my studies. I also thank him for giving me some part-time jobs and lending me money.

I also thank Lily, who gave me some part-time jobs and provided me with the chance to practice myself.

I offer my special thanks to Maik Neukirch for his kindness. I sent him my initial manuscript containing many mistakes, and he patiently gave me many useful comments on the manuscript and my research. Moreover, Maik provided the EMT code to investigate, making me deeply understand the MT data processing.

I also want to thank the Institute of Geophysical and Geochemical Exploration, China Geological Survey, for providing the data observed in China. I also thank Nittetsu Mining Consultants Co., Ltd. for providing the broadband frequency MT time-series data observed at Sawauchi, Japan. I thank all SAMTEX and USArray team members for providing the time-series data used in this study. I thank the INTERMAGNET (International Real-time Magnetic Observatory Network) for providing the magnetic time-series data observed at the KAK station. I thank the WDC for Geomagnetism, Kyoto, for providing the *Dst* index data to analyze the geomagnetic storm.

Finally, I also thank Ute Weckmann, Louise Alexander, Benjamin Murphy, Alan Chave and several anonymous reviewers who gave us constructive comments to improve our studies.

REFERENCES

Cagniard, L., 1953. The basic theory of the magneto-telluric method of geophysical prospecting. *Geophysics*, **18**, 605–635.

Constable, C. G. and Constable, S.C., 2004. Satellite magnetic field measurements: applications in studying the deep Earth. State Planet Front. Chall. *Geophysics*, **19**, 147.

Constable, C., 2016. Earth's electromagnetic environment. *Surv, Geophysics*, **37**, 27–45.

Campanya, J., Ledo, J., Queralt, P., Marcuello, A., & Jones, A. G., 2014, A new methodology to estimate magnetotelluric (MT) tensor relationships: Estimation of Local transfer-functIons by Combining Interstation Transfer-functions (ELICIT). *Geophysical Journal International*, **198(1)**, 484-494.

Chave, AD and Thomson, D.J., 1989, Some comments on magnetotelluric response function estimation, *J.Geophys.Res.*, **94**, 14215–14225.

Chave, A. D. and Thomson, D. J., 2003, A bounded influence regression estimator based on the statistics of the hat matrix, *J. R. Stat. Soc., Ser. C*, **52**, 307–322.

Chave, AD, and Thomson, D.J., 2004, Bounded influence estimation of magnetotelluric response functions, *Geophys. J. Int*, **157**, 988–1006.

Chave, A.D., Jones, A.G., 2012. The magnetotelluric method: Theory and practice. Cambridge University Press.

Chouteau, MTournerie, B., 2000. Analysis of magnetotelluric data showing phase rolling out of quadrant (PROQ), SEG Technical Program Expanded Abstracts 2000. Society of Exploration Geophysicists, pp. 344–346.

Chen, J., Heincke, B., Jegen, M. and Moorkamp, M. (2012): Using empirical mode decomposition to process marine magnetotelluric data, *Geophys. J. Int.*, **190**, 293–309.

Chandorkar, M., 2019. Machine Learning in Space Weather. Université of Eindhoven.

Egbert, G. D. and Booker, J. R., 1986, Robust estimation of geomagnetic transfer functions, *Geophys. J. R. Astr. Soc.*, **87**, 173-194.

Egbert, G.D., Eisel, M., Boyd, O.S., Morrison, H.F., 2000. DC trains and Pc3s: Source effects in mid-latitude geomagnetic transfer functions. *Geophys. Res. Lett.* **27**, 25–28.

Egbert, G.D., Livelybrooks, D.W., 1996, Single station magnetotelluric impedance estimation: Coherence weighting and the regression M-estimate. *Geophysics*, **61**, 964–970.

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

Egbert, G.D., Eisel, M., Boyd, O.S., Morrison, H.F., 2000. DC trains and Pc3s: Source effects in mid-latitude geomagnetic transfer functions. *Geophys. Res. Lett.* 27, 25–28.

Fowler, R.A., Kotick, B.J., Elliott, R.D., 1967. Polarization analysis of natural and artificially induced geomagnetic micropulsations. *J. Geophys. Res*. **72**, 2871–2883.

Fontes, S.L., Harinarayana, T., Dawes, G.J.K., Hutton, V.R.S., 1988. Processing of noisy magnetotelluric data using digital filters and additional data selection criteria. *Phys. Earth Planet. Inter*. **52**, 30–40.

Gamble, T. D., Goubau, W. M. and Clarke, J., 1979, Magnetotellurics with a remote reference, *Geophysics*, **44**, 53–68.

Garcia, X., Chave, A.D., Jones, A.G., 1997. Robust processing of magnetotelluric data from the auroral zone. *J. Geomagn. Geoelectr*. **49**, 1451–1468.

Garcia, X. and Jones, A. G., 2002. Atmospheric sources for audio-magnetotelluric (AMT) sounding, *Geophysics*, **67**, 448-458.

Gang, Z., Xianguo, T., Xuben, W., Song, G., Huailiang, L., Nian, Y., Yong, L., Tong, S., 2018, Remote reference magnetotelluric processing algorithm based on magnetic field correlation. *Acta Geod. Geophys*. **53**, 45–60.

Huang, N. E., Wu, Z., Long, S. R., Arnold, K. C., and Chen, X., 2009. On instantaneous frequency, *Adv. Adapt. Data Anal*., **1**, 77.

Hennessy, L., Macnae, J., 2018. Source-dependent bias of sferics in magnetotelluric responses.

Geophysics, **83**, E161–E171.
Hering, P., 2019. Advances in Magnetotelluric Data Processing, Interpretation and Inversion, Illustrated by a Three-dimensional Resistivity Model of the Ceboruco Volcano. Johann Wolfgang Goethe-Universität Frankfurt am Main.
Junge, A., 1996. Characterization of and correction for cultural noise. *Surv. Geophys.* **17**, 361–391.
Jones, A.G., Jödicke, H., 1984. Magnetotelluric transfer function estimation improvement by a coherence-based rejection technique, in: SEG Technical Program Expanded Abstracts 1984. Society of Exploration Geophysicists, pp. 51–55.
Kappler, K.N., 2012. A data variance technique for automated despiking of magnetotelluric data with a remote reference, *Geophys. Prospect.*, **60,** 179–191.
Liu, H., Shah, S., Jiang, W., 2004. Online outlier detection and data cleaning. Comput. Chem. Eng. 28, 1635–1647.
Lezaeta, P., Chave, A., Jones, A.G., Evans, R., 2007. Source field effects in the auroral zone: Evidence from the Slave craton (NW Canada). *Phys. Earth Planet. Inter*. **164**, 21–35.
Larsen, J.C., Mackie, R.L., Manzella, A., Fiordelisi, A., Rieven, S., 1996. Robust smooth magnetotelluric transfer functions. *Geophys. J. Int*. **124**, 801–819.
Mareschal, M., 1981. Source effects and the interpretation of geomagnetic sounding data at sub-auroral latitudes. *Geophys. J. Int*. **67**, 125–136.
Mareschal, M., 1986. Modelling of natural sources of magnetospheric origin in the interpretation of regional induction studies: a review. *Surv. Geophys*. **8**, 261–300.
Marple, S.L., Marino, C., 2004. Coherence in signal processing: a fundamental redefinition, Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004. IEEE, pp. 1035–1038.
McPherron, R.L., 2005. Magnetic pulsations: their sources and relation to solar wind and geomagnetic activity, *Surv. Geophysics*, **26**, 545–592.
Murphy, B.S., Egbert, G.D., 2018. Source biases in midlatitude magnetotelluric transfer functions due to Pc3-4 geomagnetic pulsations. *Earth Planets Space*, **70**, 1–9.
MTNET, https://www.MTnet.info/data/kap03/kap03.html.
Neukirch, M. and Garcia, X, 2013. Nonstationary Time-series Convolution: On the Relation Between the Hilbert–Huang and Fourier Transform. *Advances in Adaptive Data Analysis*, **5**, 1350004.
Neukirch, M. and Garcia, X. 2014. Nonstationary magnetotelluric data processing with the instantaneous parameter, *J. Geophys. Res. Solid Earth*, **119**, 1634–1654.
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**, 659–669.
Platz, A., Weckmann, U., 2019. An automated new preselection tool for noisy Magnetotelluric data using the Mahalanobis distance and magnetic field constraints. *Geophys. J. Int.* **218**, 1853–1872.
Rikitake, T., 1948. 1. Notes on electromagnetic induction within the Earth. *Bull. Earthquake Res. Inst.*, **24**, 1-9.
Rehman, N., and Mandic D. P., 2009. Multivariate empirical mode decomposition, Proc. Phys. Soc. London, Sect. A, **466**, 1291–1302.
Sims, W. E., Bostick, F. X., and Smith, H. W., 1971, The estimation of magnetotelluric impedance tensor elements from measured data, *Geophysics*, **36**, 938–942.
Siegel, A.F., 1982. Robust regression using repeated medians. *Biometrika*, **69**, 242–244.
Szarka, L., 1988. Geophysical aspects of man-made electromagnetic noise in the Earth—A review. *Surv. Geophys*. **9**, 287–318.
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**, 1423–1438.
Suomela, J., 2014. Median filtering is equivalent to sorting. ArXiv Preprint. ArXiv14061717.
SWPC (The Space Weather Prediction Center), https://www.swpc.noaa.gov/content/wmo/geomagnetic-activity.
Tikhonov, A. N., 1950. On determining electrical characteristics of the deep layers of the Earth's crust, in Doklady, Citeseer, 295–297.
Tsurutani, B.T., Gonzalez, W.D., Gonzalez, A.L., Guarnieri, F.L., Gopalswamy, N., Grande, M., Kamide, Y., Kasahara, Y., Lu, G., Mann, I., 2006. Corotating solar wind streams and recurrent geomagnetic activity: A review. *J. Geophys. Res. Space Phys*. **111**.
Travassos, J.M., Beamish, D., 1988, Magnetotelluric data processing—A case study. *Geophys. J. Int*. **93**, 377–391.
Vozoff, K., 1972. The magnetotelluric method in the exploration of sedimentary basins. *Geophysics*, **37**, 98–141.
Viljanen, A., Pirjola, R., & Hákkinen, L., 1993. An attempt to reduce induction source effects at high latitudes, *Journal of geomagnetism and geoelectricity*, **45**(9), 817-831.
Wight, D., and Bostick, F., 1980, Cascade decimation-A technique for real-time estimation of power spectra. In ICASSP'80. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. **5**, pp. 626-629). IEEE.
Weckmann, U., Ritter, O., Haak, V., 2003. A magnetotelluric study of the Damara Belt in Namibia: 2. MT phases over 90 reveal the internal structure of the Waterberg Fault/Omaruru Lineament. *Phys. Earth Planet. Inter*. **138**, 91–112.
Weckmann, U., Magunia, A. and Ritter, O., 2005. Effective noise separation for magnetotelluric single site data processing using a frequency domain selection scheme, *Geophys. J. Int.*, **161**, 635–652.
Wang, H., Campanyà, J., Cheng, J., Zhu, G., Wei, W., Jin, S., and Ye, G., 2017, Synthesis of natural electric and magnetic Timeseries using Interstation transfer functions and time-series from a Neighboring site (STIN): Applications for processing MT data, *J. Geophys. Res. Solid Earth*, **122**, 5835–5851.
WDC Kyoto Observatory, http://wdc.kugi.kyoto-u.ac.jp/dstdir/index.html.
Yu, G., Xiao, Q., Li, M., 2018. Anisotropic Model Study for the Phase Roll Out of Quadrant Data in Magnetotellurics.
Zonge, K. L. and Hughes, L. J., 1987. Controlled source audio-frequency magnetotellurics, in Electromagnetic Methods in Applied Geophysics. Applications (ed. Nabighian, M. N.), SEG, 713–809.