Data Quality Analysis on China Permanent Seismic Network by Repeating Earthquakes

Xuchao CHAI[1], Pei ZHANG[1], Chuang WANG[1], and Qingliang WANG[1]

[1]The Second Monitoring and Application Center, CEA

November 21, 2022

Abstract

Highly similar waveforms recorded from repeating earthquakes can be utilized to evaluate the data quality of a seismic station. We used a hypothesis testing method to establish a data quality detection model based on repeating earthquakes. The model effectiveness was verified using repeating earthquake data from 109 stations in the Global Seismic Network. A total of 842 permanent broadband stations in mainland China were evaluated using this model. Eighteen anomalies were found mainly attributed to calibration, instrument noise, mass recentering, and regional long-period interference. We found that most of the stations function well. Moreover, utilizing repeating earthquakes to analyze the waveform quality can circumvent the need for extensive forward calculations, as well as greatly reduce the influence of source parameter uncertainties and structural complexity on the seismogram. Additionally, the need for detection in other datasets in different regional networks has broadened the scope of these applications.

GEOPHYSICAL RESEARCH LETTERS

Supporting Information for

Data Quality Analysis on China Permanent Seismic Network by Repeating Earthquakes

Xuchao Chai[1], Pei Zhang[1,2], Chuang Wang[1], Qingliang Wang[1]

[1] The Second Monitoring and Application Center, CEA, Xi'an, 710054, China.

[2] Institute of Geophysics, China Earthquake Administration, Beijing 100081, China

Contents of this file

Figures S1 to S6

Introduction

This supplementary material contains 6 figures (Figure S1 to S6). Anomaly recording station detection process is demonstrated in Figure S1. Calculation results of correlation coefficient of abnormal recording stations for GSN and mainland China seismic network for the repeating earthquakes are illurstrated in Figure

S2, Figure S3, respectively. Figure S4 to S6 shows the original waveform of potential anomalous stations, which are utilized to confirm anomalous stations.

Anomaly recording station detection process

According to Figure S1, the detection process of abnormal recording stations is as follows:

Step1, acquisition of repeating earthquake event data. We prepare a list of repeating earthquakes and retrieve the waveform data of the station to be detected from the NEDBC database (Chai et al., 2020).

Step 2, calculation of the correlation coefficient. We retrieved relevant event waveforms, and performed waveform pre-processing such as demean, detrend, band-pass filtering (0.01-0.05 Hz), and normalization. Then, we calculated the CC and σ_{SF} (variance of scale factor) with a sliding window method. The statistical parameter hypothesis testing method is used to obtain the confidence interval thresholds η_{CC} and η_{SF} of CC and σ_{SF}, respectively.

Step3, filtering of potentially anomalous stations. After determining the detection threshold of each channel, if the CC of the channel falls outside the confidence interval, the station is recorded as a potential abnormal station. While for cases where the CC values fall in the confidence area, the variance σ_{SF} of the 3-channel scale factor SF is calculated to determine whether the σ_{SF}falls within the confidence interval. If the σ_{SF} falls within the confidence interval, the station is recorded as a normal record station, otherwise it will be considered as a candidate abnormal station.

Step 4, confirmation of abnormal recording station. Combining the original waveforms and the PSD curve characteristics, we verified the anomalous stations and classified them into four categories.

Calculation results of correlation coefficient of abnormal recording stations

Figure S2 represents calculation results of three channels' CC and σ_{SF} for mainland China permanent stations with the repeating earthquakes near Japan island. The values of CC and σ_{SF} of the repeating earthquake waveforms recorded by 842 stations in mainland China are listed, of which 18 stations exceed the thresholds.

Figure S3 demonstrates calculation results of three channels CC and σ_{SF} for seismic networks in the southeast of China with repeating earthquakes near Taiwan island. The calculation results of three channels CC and σ_{SF} for repeating earthquakes near Taiwan island are listed; and stations with abnormal records were detected, of which one station had three channels outside of the confidence interval (HI.LSH) and five had one or two channels outside of the confidence area (HI.SLL, SC.TQ, FJ.PTLC, SC.BZH, SC.DFU).

Analysis of potentially abnormal recording stations

Figure S4 to S6 shows the original waveform of potential anomalous stations. In addition, we caculated the correlation coefficients of the waveform records of the same seismic event at other stations within 100 km to review record quality of a abnormal station.

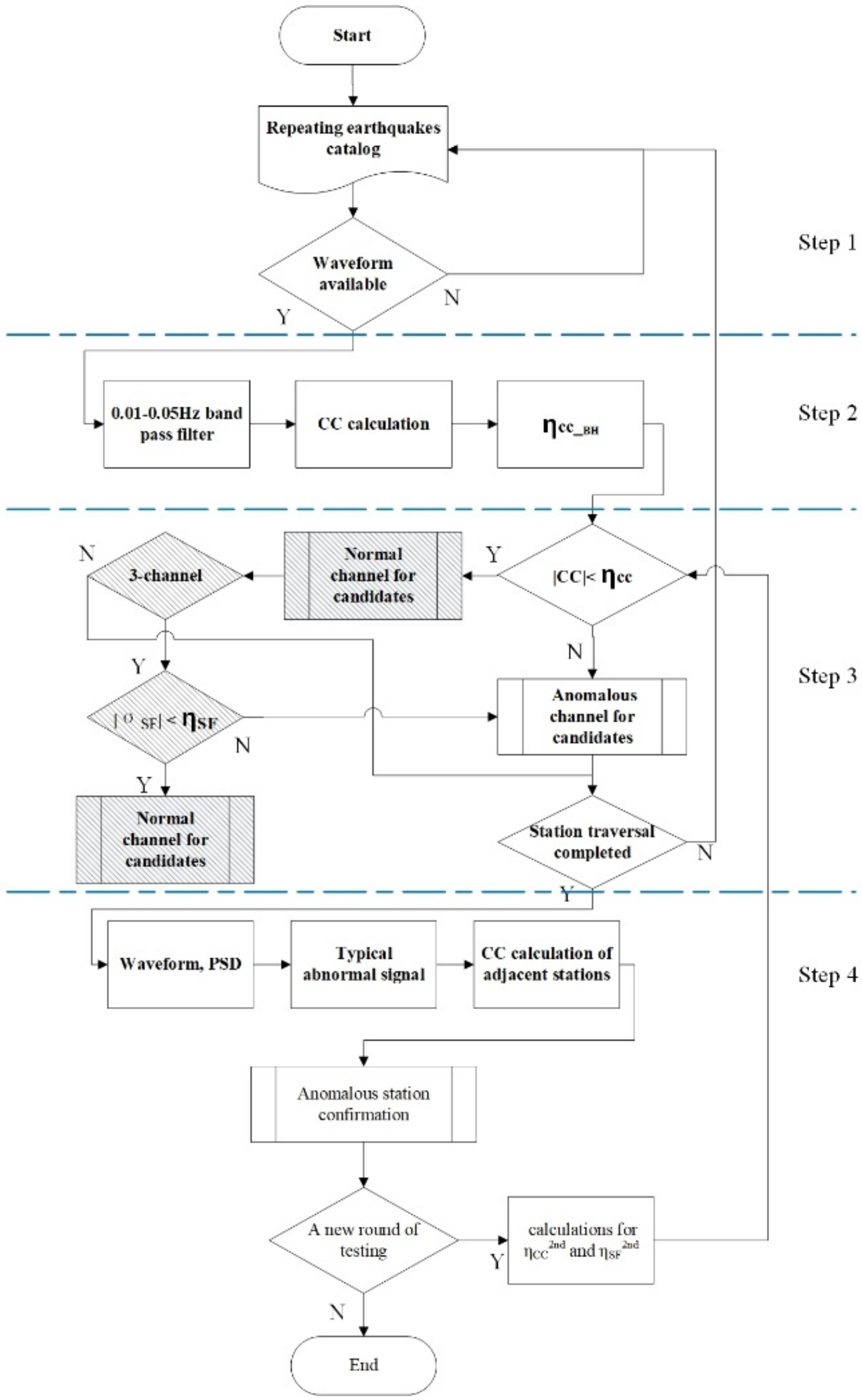


Figure S1. Flow of anomaly recording station detection process.

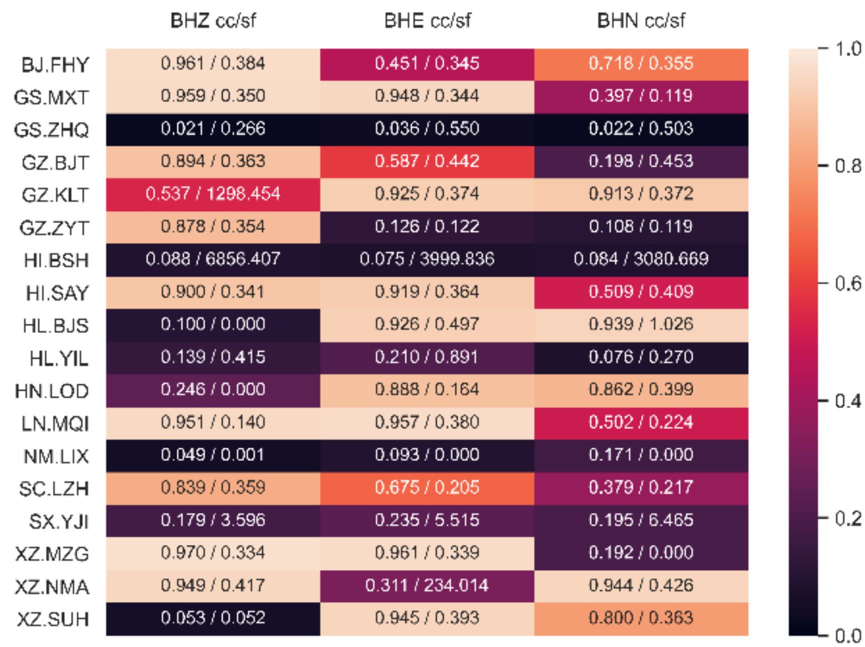
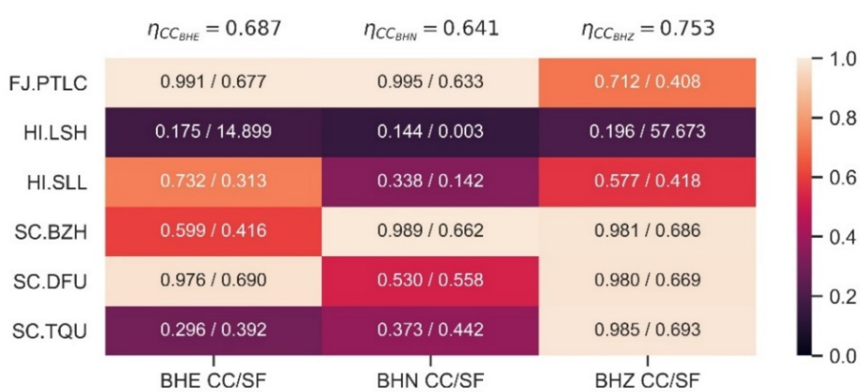
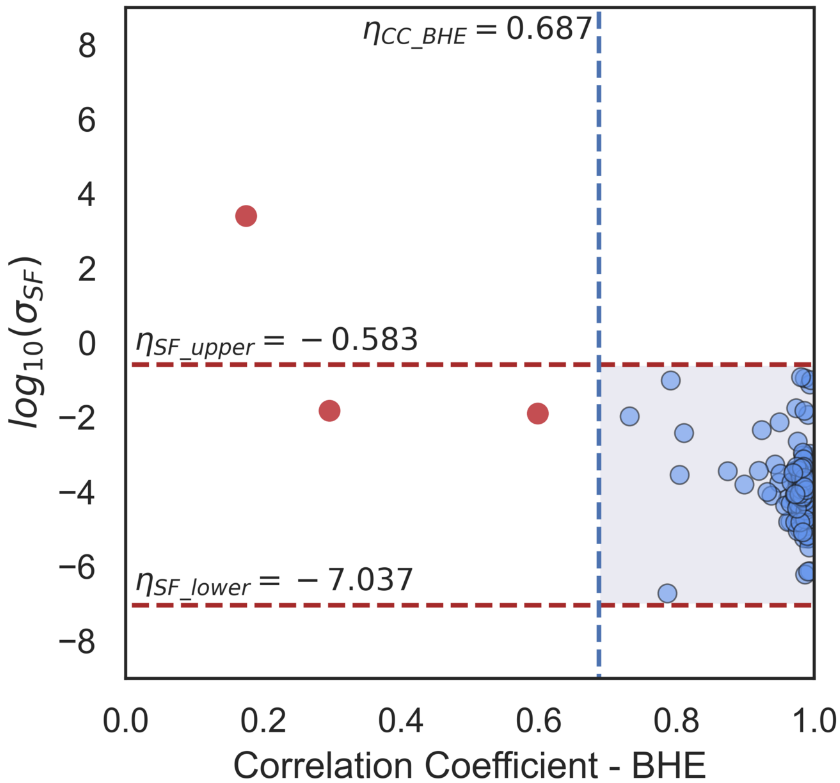


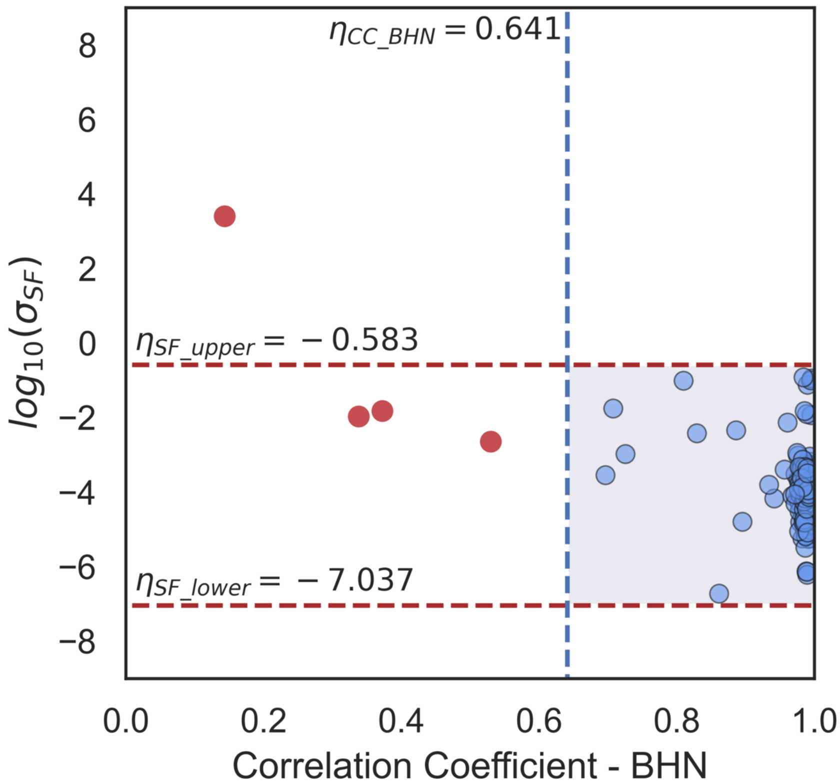
Figure S2. Abnormal results of three channels CC and σ_{SF} for mainland China permanent stations with repeating earthquakes near Japan island. The value of CC and $\sigma_{SF}$of the repeating earthquake waveforms recorded by anomaly stations in mainland China are listed, and CC values are color coded, which dark color represents lower cross-correlation value.

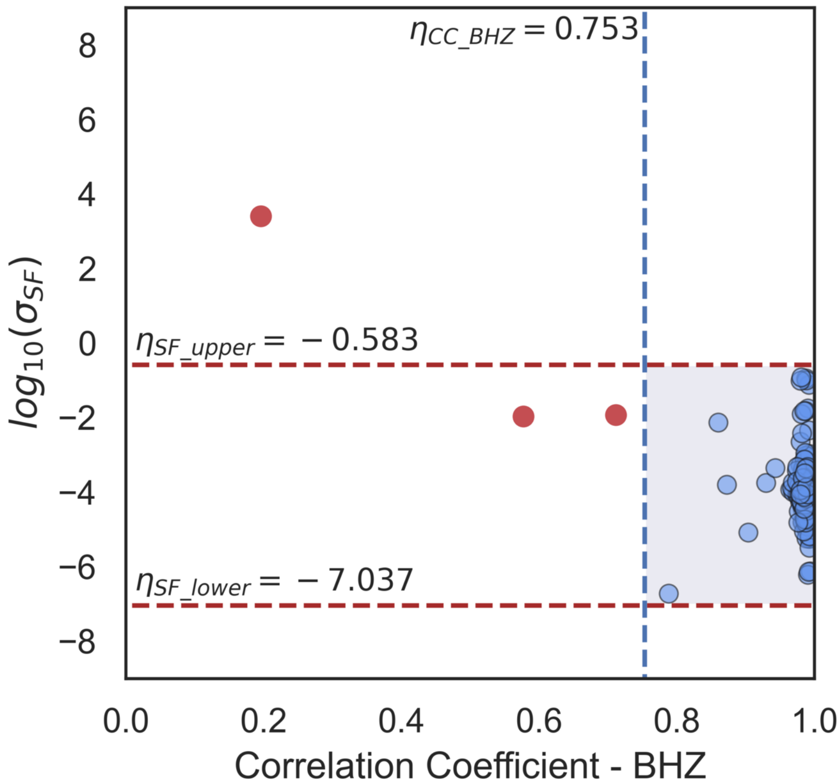


(a)



$\eta_{CC\_BHE} = 0.687$
$\eta_{SF\_upper} = -0.583$
$\eta_{SF\_lower} = -7.037$
$log_{10}(\sigma_{SF})$
8
6
4
2
0
−2
−4
−6
−8
0.0
0.2
0.4
0.6
0.8
1.0
Correlation Coefficient - BHE


$\eta_{CC\_BHN} = 0.641$
$\eta_{SF\_upper} = -0.583$
$\eta_{SF\_lower} = -7.037$
$log_{10}(\sigma_{SF})$
Correlation Coefficient - BHN
8
6
4
2
0
−2
−4
−6
−8
0.0
0.2
0.4
0.6
0.8
1.0



(b) (c)(d)

Figure S3. Regional scale network dectection results using repeating earthquakes near Taiwan island. (a) Calculation results of three channels CC and σ_{SF} for repeating earthquakes near Taiwan island with a bandpass filter of 0.05-0.1 Hz. The value of CC and σ_{SF} of the repeating earthquake events recorded by 173 permanent stations of the GD, FJ, HI, GX, GZ, SC network stations in Mainland China are listed. (b)-(d) The CC and σ_{SF}distributions of the detection results of repeating earthquakes recorded by regeonal-scale permanent seismic stations in mainland China. The red outliers represent the 3, 4, and 3 anomalous stations in BHN, BHE, and BHZ, respectively.

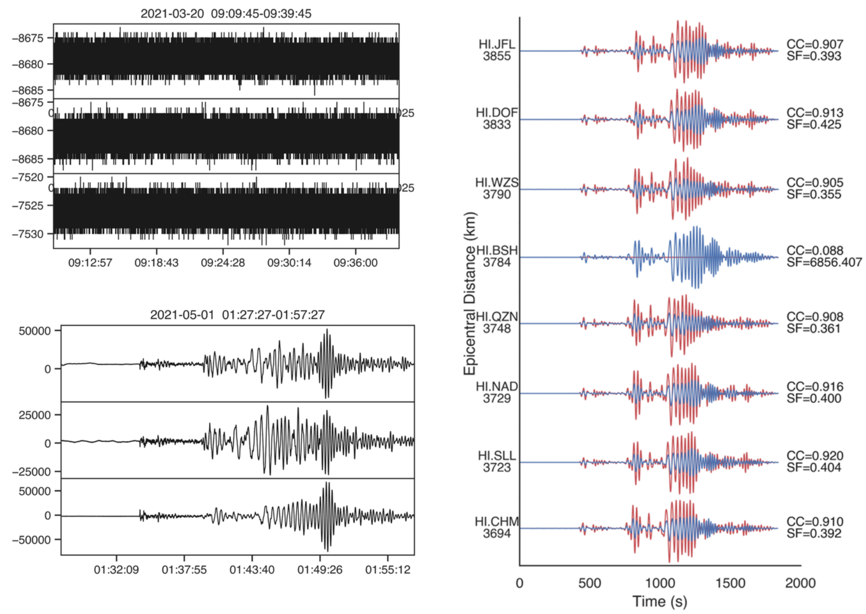


Figure S4. Comprehensive analysis of anomalous stations with instrument self-noise (no record in a certain channel of the station). (a) Repeating eathquake records of abnormal station HI.BSH. (b) Waveform cross-correlation images of the same repeating seismic pair at other stations within 100 km of the station HI.BSN at BHZ channel. The red curves represent the records of March 20th 2021 earthquake (Mw=7.0) at different stations, while the May 1st, 2021 earthquake (Mw=6.9) waveforms are shown with blue lines. At the far right of each station are the cross-correlation and scale factor for the repeating earthquakes, respectively..

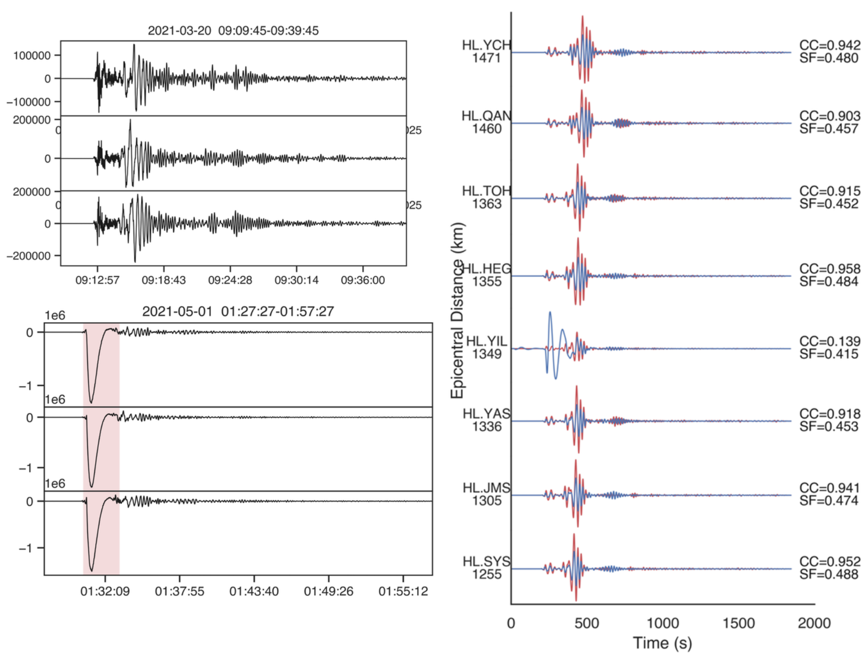


Figure S5. Comprehensive analysis of anomalous stations with mass recentering. (a) Repeating eathquake records of abnormal station HL.YIL. (b) Waveform cross-correlation images of the same repeating seismic pair at other stations within 100 km of the station HL.YIL at BHZ channel. The same as Figure S4(b) but for different stations.

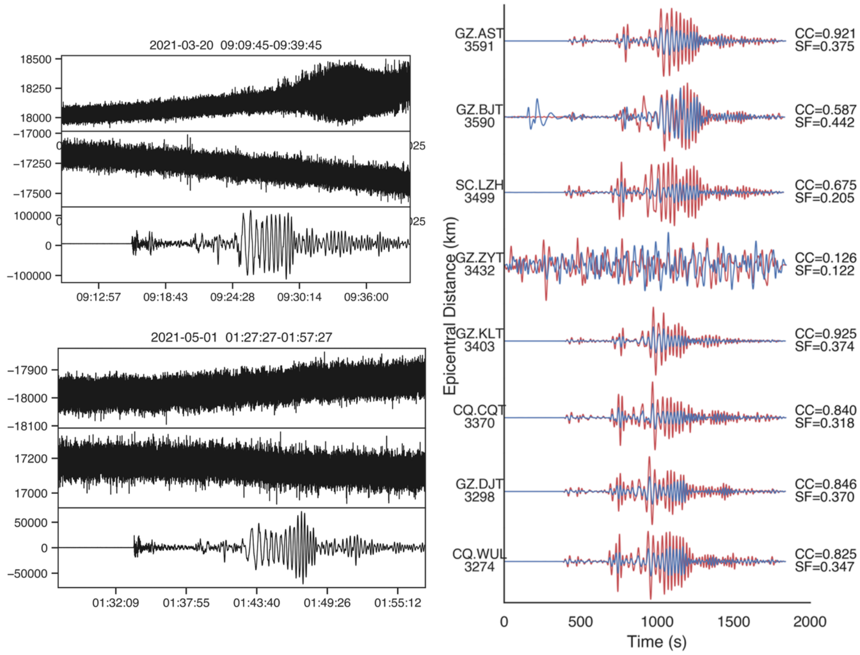


Figure S6. Comprehensive analysis of anomalous stations with regional long-period interference. (a) Repeating eathquake records of abnormal station GZ.ZYT. (b) Waveform cross-correlation images of the same repeating seismic pair at other stations within 100 km of the station GZ.ZYT at BHN channel. The same as Figure S4(b) but for different stations.

Hosted file

`si tables_0928.docx` available at https://authorea.com/users/531593/articles/597682-data-quality-analysis-on-china-permanent-seismic-network-by-repeating-earthquakes

Data Quality Analysis on China Permanent Seismic Network by Repeating Earthquakes

Xuchao Chai[1], Pei Zhang[1,2], Chuang Wang[1], Qingliang Wang[1]

[1] The Second Monitoring and Application Center, CEA, Xi'an, 710054, China.

[2] Institute of Geophysics, China Earthquake Administration, Beijing 100081, China

Corresponding author: Xuchao Chai (chai_xc@126.com), Pei Zhang (zhp1017@mail.ustc.edu.cn)

Address: The Second Monitoring and Application Center, CEA, 316 Xiying Road, Xi'an, China, 710054

Key Points:

- The proposed waveform data quality detection method based on repeating earthquakes can be utilized for waveform quality control.
- Hypothetical testing parameters were utilized to quantitatively filter gross errors and improve the accuracy of the proposed method.
- Data quality control for the permanent broadband seismic stations of mainland China is realized based on repeating earthquake records.

Abstract

Highly similar waveforms recorded from repeating earthquakes can be utilized to evaluate the data quality of a seismic station. We used a hypothesis testing method to establish a data quality detection model based on repeating earthquakes. The model effectiveness was verified using repeating earthquake data from 109 stations in the Global Seismic Network. A total of 842 permanent broadband stations in mainland China were evaluated using this model. Eighteen anomalies were found mainly attributed to calibration, instrument noise, mass recentering, and regional long-period interference. We found that most of the stations function well. Moreover, utilizing repeating earthquakes to analyze the waveform quality can circumvent the need for extensive forward calculations, as well as greatly reduce the influence of source parameter uncertainties and structural complexity on the seismogram. Additionally, the need for detection in other datasets in different regional networks has broadened the scope of these applications.

Plain Language Summary

An application model for the quality control of waveform data from repeating earthquakes was proposed. The model was validated with waveform data from the GSN network, and the data quality of the permanent seismic stations in mainland China was quantitatively described from the perspective of the ability to record earthquakes. Previous studies have rarely directly elucidated the waveform quality of natural earthquakes recorded by stations based on a certain aspect or metric. Thus, we utiliezed the fact that repeating earthquakes exhibit extremely similar waveform records at an individual station to evaluate the data quality for recorded seismic events. This allowed rapid evaluation of the quality of seismic stations. Most of the stations in the broadband seismic network in mainland China exhibit a good recording performance. Station anomalies are mainly caused by calibration issues, instrument noise, mass recentering, and regional long-period

interferences. This method can be utilized for the quality control of seismic datasets in different locations according to the magnitudes and epicenters.

1. Introduction

Seismology is a discipline based on observational data; high-quality observation data is a vital prerequisite for seismological research. After the recent implementation of seismic network engineering (under the Tenth Five-Year Plan) and background field projects (under the Eleventh Five-Year Plan) of China Earthquake Administration (CEA), a seismic observation network has been established in China that includes more than 1,200 permanent stations (Liu et al., 2008). Additionally, the Himalayan Observation Project includes more than 1,400 mobile observation stations (Song et al., 2012). With the continuous accumulation of seismic data, rapidly and accurately assessing the quality of these records and easily understanding the status of the observation system have become key issues in the construction of the current seismic network.

In recent decades, global geoscience institutions and researchers have focused on seismic data quality and developed some effective waveform data quality control systems, such as the Albuquerque Seismological Laboratory's Data Quality Analyzer (Ringler et al., 2015) and the Incorporated Research Institutions for Seismology's (IRIS) Modular Utility for STAtistical kNowledge Gathering (Magana-Zook et al., 2016; Casey et al., 2018). These tools can be used to conduct a comprehensive and detailed analysis of data quality based on sensor issues; station equipment, timing, and metadata problems; calibration errors, and station security challenges (Katherine et al., 2021). However, few methods have examined the waveform quality of natural earthquakes using a specific metric. We used the characteristics of repeating earthquakes to assess the quality of large-scale station data from the perspective of recording seismic events, thereby enabling rapid analysis of all data from the original records to the final data examination. The application of repeating earthquakes has developed rapidly since its first proposal by Isacks et al. (1967). It has been widely used in the estimating deep slip rates for faults (Schmittbuhl et al., 2016; Uchida et al., 2007; Yoshida et al., 2015; Li et al., 2011; Ma et al., 2014), monitoring temporal changes in the structure and properties of the Earth (Li et al., 2006; Schaff & Beroza, 2004), studying the Earth's inner core (Zhang et al., 2008; Wen, 2006; Yu, 2016a, 2016b; Yang & Song, 2020) and predicting earthquake (Matsuzawa et al., 2002; Khoshmanesh et al., 2015). We can use the highly similar waveforms and focal mechanisms of these earthquakes (Nadeau et al., 1995) to evaluate the waveform quality in earthquake records at seismic stations.

The use of repeating earthquakes for waveform similarity detection not only avoids the variations caused by different seismographs or digital equipment used at diverse stations, but also greatly reduces the impact of seismic sources and subsurface structures on seismic waveforms. Such a study has yet to be conducted . To evaluate seismic datasets in different regions efficiently and accurately, we proposed a data quality detection method based on repeating earthquakes. First, we verified the effectiveness of the model with records from 109 stations in the GSN network. Second, we assessed the data quality of 842 permanent broadband seismic stations in mainland China (Figure 1) and distinguished 18 stations with anomalous records of their earthquake recording capabilities. Additionally, by implementing anomaly detection in the GD, FJ, GX, SC, and XZ networks (See Table S2 for China Earthquake Network Code) using a pair of repeating earthquakes of smaller magnitudes in Taiwan, we confirmed the applicability of this approach on regional and global scales. Consequently, this method can be utilized for waveform quality control of datasets in different locations according to the magnitudes and epicenters.

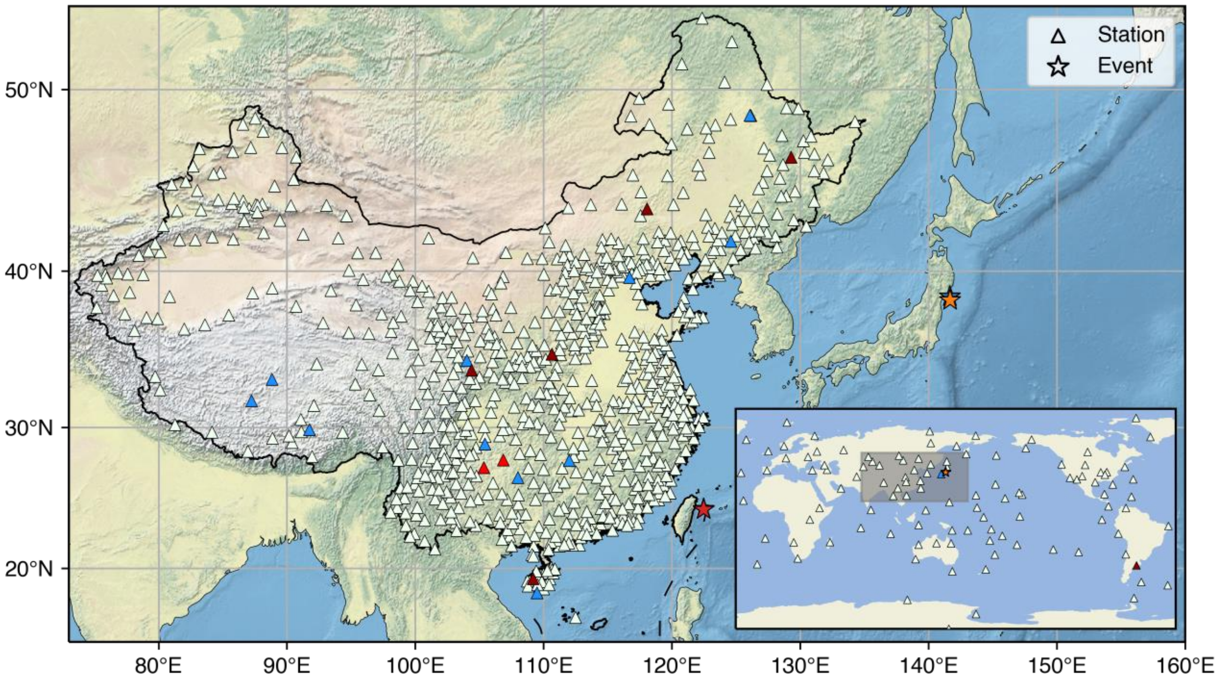


Figure 1. Distribution of permanent seismic stations in mainland China (China Earthquake Administration, CEA). Pentagrams represent the two sets of repeating earthquake events that occurred in northeastern Japan and northeastern Taiwan. Triangles represent seismic stations. Light blue, red, and brown triangles represent stations with one, two, and three abnormal channels, respectively.The inset in the lower right corner shows the distribution of repeating earthquakes and the stations of the GSN network selected in this study; the blue and brown triangle represent the IU.MAJO and IU.TRQA station,respectively, which recorded two and three channel anomalies.

2. Data and Methods

2.1 Waveform data quality control with repeating seismic records

Repeating earthquakes with highly similar waveforms are the cornerstone of this study. Hypocenter location and waveform similarity are two main methods for identifying repeating earthquakes (Poupinet et al., 1984; Frémont & Malone, 1987; Dodge et al., 1995; Shearer et al., 1997; Lees, 1998; Philips, 2000; Moriya et al., 2003; Uchida, 2019); here, we mainly utilize the latter approach. We focus on earthquakes with magnitudes greater than *Mw* 5.0 in the Preliminary

Determination of Epicenters (PDE) and consider a correlation coefficient of greater than 0.8 as a requirement for repeating earthquakes (Schaff & Richards, 2004) (see Table S1 in the supporting information for more details). To identify repeating earthquakes effectively and accurately, the correlation coefficient (CC) was calculated for every two events with epicenters that were less than 0.5 ° apart (Yang & Song,2020). After preprocessing and calculating CC, we identified a series of repeating earthquakes, in which the pair with the largest magnitude occurred in northeastern Japan on 20 March (*Mw=6.7*) and 1 May 2021(*Mw=7.0*).

In addition to the correlation coefficient of the repeating earthquakes at the same station, the difference in the amplitude scale factor (SF) among the three channels at each station was also used to determine whether there was abnormal channel energy at that station (Ekström et al., 2006; Ringler et al., 2012). Therefore, we employed statistical parameter hypothesis testing through correlation coefficients and waveform scale factors to filter out gross errors from large-scale station records. If the waveform records from repeating earthquakes at a station were highly correlated and the scale factor was within the range of the energy difference, then the performance of the station for recording earthquakes during that period was considered good. In contrast, a lower waveform correlation coefficient or a single-channel scale factor that exceeded or fell short of that of other channels by an order of magnitude indicated that the station was working abnormally, and the cause of the abnormality was further investigated.

Although the confidence interval threshold for the correlation coefficient obtained by this method may be lower than the common judgment threshold for repeating earthquakes, it avoids many false triggers resulting from excessive detection sensitivity.

2.2 Method verification

2.2.1 Filtering of potential anomalous stations

When the seismic signal recorded by a station was abnormal and the cause was unknown, we used parameter hypothesis testing to determine gross errors of waveform similarity from the system observation variables and screen potentially abnormal stations (Akaike, 1974; Lehmann & Joseph, 2008). The process of detecting abnormal stations can be divided into four steps, acquisition of repeating earthquakes data, calculation of the correlation coefficient and scale factor, filtering out potentially anomalous stations, and confirmation of potentially anomalous stations (the description of the specific steps are illurstrated in Figure S1 in the Supporting Information). We regarded the seismic observation network as a single network system composed of multiple sensors, where each station was a sensor. If the waveform signals of a pair of repeating earthquake events output by the station are considered a sample output of the system, a correlation coefficient can be obtained for the data.

The data sequence composed of CC after angular transformation approximately obeys the Gaussian distribution. We used the simple PauTa criterion (3σ criterion), which is widely used in statistics, automatic control, and industrial quality control theories, to obtain gross errors (Hui et al., 2002; Xiong & Wu, 2010; Hua et al., 2013; Ding & Cai, 2019). If the CC of repeating earthquake waveforms recorded by one channel of a station is recorded as X_i, the absolute error ΔCC is calculated as follows:

$$|\Delta CC_i|=|CC_i-\mu_j|>3\sigma_j \quad (1)$$

where μ is the mean value of the CC of a channel, σ is the standard deviation, i and j are the station code and channel code, respectively. If ΔCC is greater than 3σ, the station record is considered a gross error. The gross error threshold of the channel (η_j) can be expressed as

$$\eta_j=\mu_j-3\sigma_j \tag{2}$$

As shown in Figure 2(a), the red outliers of the correlation coefficient IU.TRQA ($CC_{\_BHZ}$, $CC_{\_BH1}$, $CC_{\_BH2}$ channels are 0.165, 0.207, 0.165, respectively) and IU.MAJO ($CC_{\_BHZ}$ is 0.152) for the repeating earthquakes are lower than the gross error thresholds of each channels (η_{BHZ}, η_{BH1}, η_{BH2} are 0.683, 0.699, 0.567, respectively); therefore, these samples are statistically gross errors (the related values are listed in Table S3). In addition, if a station has a three-channel record, we can calculate the scale factor of the three-channel amplitude as a necessary condition to further determine whether there is an abnormality in each channel of the station. To avoid differences in instrument response between stations due to different seismographs, we only analyzed the scale factor standard deviation σ_{SF} of all the channels in the same station using the following equation:

$$\sigma_{SF\_i} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\sigma_{SF\_j}-\mu_i)^2} \tag{3}$$

This analysis can show the stations with obvious energy differences among channels. Here, σ_{SF_j} is the variance of the single-channel SF, and μ_i is the arithmetic mean of the three-channel SF of a station.

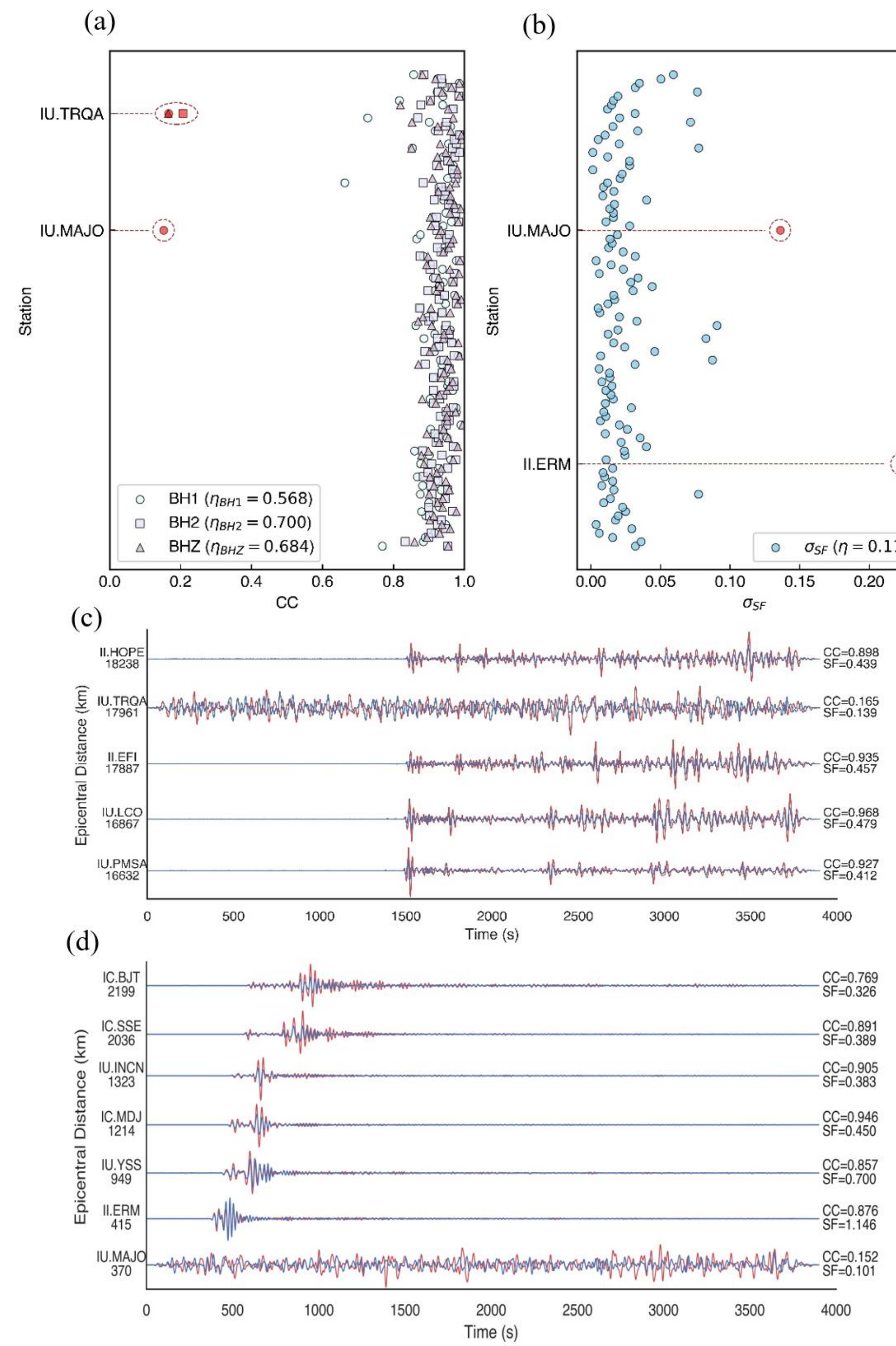


Figure 2 Distribution of CC and σ_{SF} of the GSN stations. (a) The scatter diagram of the correlation coefficients of GSN stations, where the red outliers represent relevant channels of IU.TRQA (BHZ, BH1, and BH2) and IU.MAJO (BHZ) stations. (b)The outlier of σ_{SF} results shown are consistent with (b). (c)-(d) The correlation coefficient between the IU.TRQA, IU.MAJO stations and other GSN stations within 3000 km, respectively. The red curves represent the records of March 20 2021 earthquake (Mw=7.0) at different stations, while the May 1, 2021 earthquake (Mw=6.9)

waveforms are shown with blue lines. At the far right of each station are the cross-correlation and scale factor for the repeating earthquakes, respectively.

2.2.2 Confirmation of the abnormal stations

After filtering out the potentially anomalous GSN stations from a statistical perspective, we confirmed these anomalies by analyzing the original waveform, considering instrumental and environmental noise, and applying other traditional seismological methods. The three channels of IU.TRQA and BH1 of IU.MAJO stations are outliers and thus potentially abnormal (Figure 2a). BHZ channel of these two stations recorded the repeating earthquake waveforms as shown in Figure 2(c) and (d), but neither of them contained obvious seismic signals (such as random noise with small amplitude changes). The Power Spectral Density (PSD) curves showed that the amplitudes were only a few counts and vary from sample to sample. It was most likey caused by instruments failures, such as a seismometer lockout, or due to an excessive distance from the epicenter. By comparing the correlation coefficients with those of the stations within 3000 km of IU.TRQA and IU.MAJO, and the analysis of the PSD curve (see Section 4 for more discussion on this), the abnormal recordings at IU.TRQA and IU.MAJO might have been likely caused by instruments failure.

In Figure 2(a) and (b), II.ERM with normal CC and abnormal σ_{SF} can be regarded as a potentially anomalous station (Ringler et al., 2012), and the σ_{SF_ERM} was significantly higher than other stations in comparison. By analyzing the original waveform, we found that the amplitude of the BH1 channel on March 20 was much lower than that of other channels, which contributed to the variance diffuse. We will further discuss σ_{SF} in the following inspections of permanent seismic stations in mainland China.

3. Waveform Quality of Permanent Seismic Stations in Mainland China

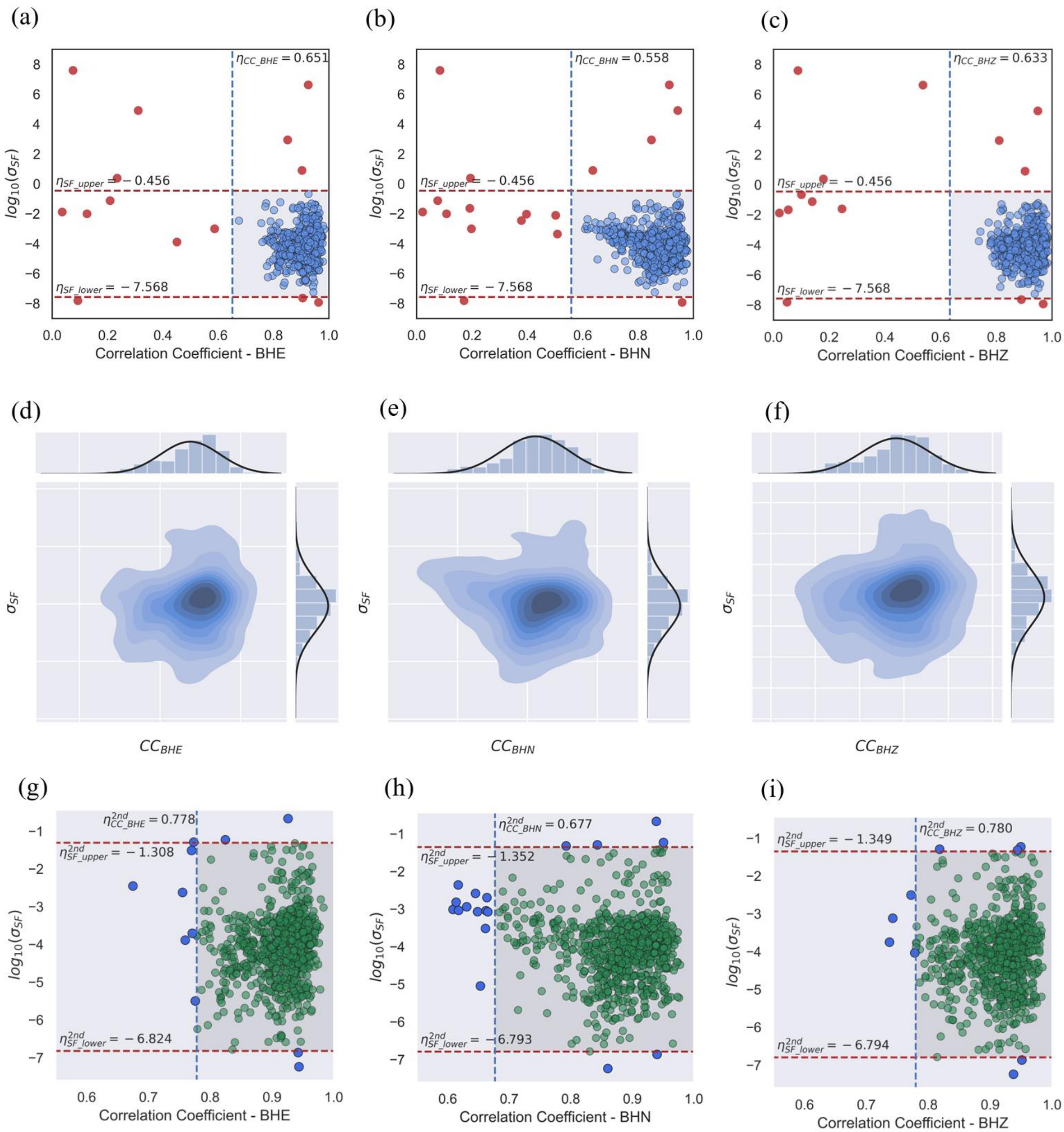


Figure 3 The CC and σ_{SF} distributions of the detection results of the repeating seismic records from the permanent seismic observation stations in mainland China.The outliers among the 842 seismic stations in different channels: (a) BHN, (b) BHE, and (c) BHZ. The red outliers outside the thresholds represent the 14, 17, and 14 potential anomalous stations in BHN, BHE, and BHZ, respectively. The single-sided risk threshold η_{cc} on the left is μ_{cc}-3σ_{cc}, and η_{SF} is the gross error threshold of the three-channel scale factor sample sequence at the same station. (d)-(f) The two-dimensional

joint probability distribution of the CC and σ_{SF} of each channel for the remaining stations after excluding the gross errors (blue dots in (a)-(c)). The upper and right sides are the normal distribution curves fitted by the two marginal probability densities CC and SF, respectively; and the standard deviation of the scale factor of the three channels at the same station is σ_{SF}. (g)-(i) The results of the second round of parameter testing for the normal stations (blue dots in (a)-(c)). The detection results reveal that BHN, BHE, and BHZ consist of 11, 17, and 9 potentially abnormal stations, respectively (represented in blue).

A 0.01-0.05 Hz band-pass filter was used to process the original waveforms, with the duration window lasting 30 minutes from the origin time. The two-dimensional joint probability distribution of the waveform CC and SF recorded by 842 permanent seismic observation stations in mainland China were obtained, as shown in Figure 3. Similar to the method mentioned in Section 2.1, the statistical characteristics of the approximate normal distribution were obtained by first reversibly transforming CC and σ_{SF}, and then using the parameter hypothesis testing to rapidly filter out the potentially abnormal stations. As shown in Figure 3(a)–(c), each channel has a two-dimensional joint distribution constructed by the CC and the σ_{SF} for the station. The marginal distributions of $CC \sim \boldsymbol{N}(\mu_{CC}, \sigma_{CC}^2)$ and $SF \sim \boldsymbol{N}(\mu_{SF}, \sigma_{SF}^2)$ are independent.

Using equation 2, the risk thresholds of the correlation coefficient and the scale factor were obtained for the marginal distribution. As shown in Figure 3(a)-(c), η_{CC} (a one-sided detection shown as a blue dashed line) and σ_{SF} (a two-sided detection shown as red dashed lines) reveal outliers outside the confidence interval. The distribution of 842 seismic stations that recorded the repeating earthquakes in mainland China is shown in Figure 1. For these earthquakes, 14, 17, and 14 channels in BHN, BHE, and BHZ fell outside the confidence interval. Moreover, five stations (i.e., GS.ZHQ, HI.BSH, HL.YIL, NM.LIX, and SX.YJI) had three abnormal channels, five stations had two abnormal channels, and eight stations had one abnormal channel (see Figure S2 for more details).

As shown in Figure 3(a)–(c), a total of 18 stations falling outside the confidence interval were filtered through the hypothesis test. However, there were some stations in each of these channels that were judged to be "normal" for recording near the threshold. The channels within the confidence zone but close to the risk threshold may show various anomalies; therefore, they warrant further review. Accordingly,the remaining stations in the confidence area (after removing the gross errors) could be filtered and analyzed repeatedly using the same method. The two-dimensional joint probability distribution of CC and σ_{SF} in the gray confidence region in Figure 3(a)–(c) corresponds to (d)–(f), respectively. The results of abnormal stations obtained through the second round of assessment are shown in Figures (g)–(i). This process will not be repeated further.

For a smaller-scale regional network, we used a pair of repeating earthquakes that occurred in eastern Taiwan on 13 June ($Mw = 5.4$) and 26 July 2020 ($Mw = 5.2$) to detect the data quality at 173 stations in south China near the epicenters. Consequently, six stations with abnormal records were detected, of which one station had three channels outside of the confidence interval (HI.LSH) and five had one or two channels outside of the confidence area (Figure S3). Therefore, using repeating earthquakes successfully implements data quality assessment of regional seismic networks at different scales.

4. Anomalous Station Categorization

The potentially abnormal stations were identified through the above-mentioned parameter verification and analysis. However , the factors that contributed to the anomalies, such as human activities, environmental factors, and instrument failure, need to be further verified. Combining the original waveforms and the PSD curve characteristics, we divided the anomalous stations into four categories. The original waveforms and correlation calculation results for all abnormal stations in this section can be seen in Figure S4 - S6.

4.1 Calibration signal interference

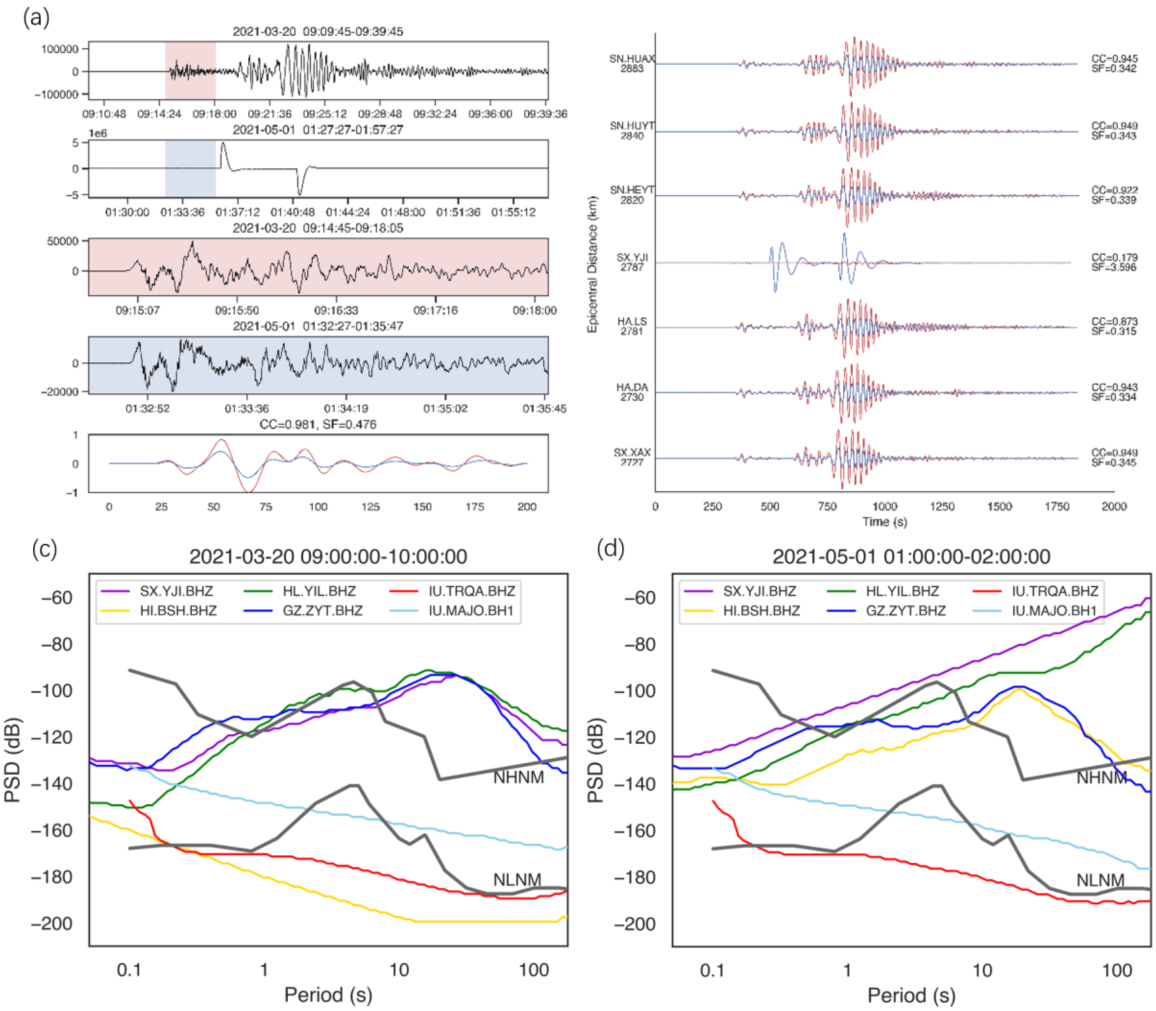


Figure 4 Analysis of potentially abnormal recording stations. (a) Repeating earthquake records of abnormal station SX.YJI. (b) The result of the correlation coefficient of repeating earthquake waveforms of each station within 100 km of SX.YJI at BHZ channel. (c) (d) The Power Spectral Density (PSD) curves (McNamara & Buland, 2004) of abnormal stations recorded for 1 hour during the occurrence of repeating earthquakes on (c) 20 March 2021 and (d) 1 May 2021. The black curves represent NHNM (New High Noise Model) and NLNM (New Low Noise Model).

None of the three channels at station SX.YJI effectively recorded the earthquake that occurred in northeastern Japan on May 1. We calculated the correlation coefficients of the waveform records of the same seismic event at other stations within 100km (Figure 4b). SX.YJI and other stations recorded a relatively high correlation coefficient for the earthquake that occurred on 20 March, while the correlation coefficient for the earthquake on 1 May at SX.YJI station was much lower than those at other stations (Figure 4b). The original waveforms show notable waveform signals before the calibration signal of the station (shown in the red box in Figure 4a). We extracted this signal and found that it is highly similar to the P wave signal of the March 20 earthquake. The correlation coefficient of these two truncated waveforms is 0.981, which falls within the confidence zone. The PSD curve (purple diagonal line in Figure 4d) also showed step calibration signal characteristics. Therefore, there was no problem with the seismograph at this station; the absent earthquake records resulted from the large amplitude of the seismograph calibration signal, which suppressed the seismic signal.

The three channels of GS.ZHQ were also affected by the step calibration signal. This station was calibrated every day for two weeks before the earthquake. Station XZ.SUH shows a single-channel anomaly, which was mainly caused by the square-wave calibration signal of the vertical component.

4.2 Instrument noise

The waveform of the HL.BSH station for 20 March indicated no record of the seismic signal and only recorded noise with little amplitude change. However, the May 1 earthquake was recorded normally. Additionally, the PSD curve in Figure 4(c) (the yellow line) for the HI.BSH station is significantly lower than that of NLNM. Templeton (2014) and Wang et al. (2019) found that when the seismometer noise level was below NLNM, only the self-noise of the instrument was recorded. Therefore, the seismometer at the station failed to work normally during the March 20 earthquake. Both the NM.LIX and HN.LOD stations also showed this type of abnormality.

4.3 Mass recenter

The original waveform for the HL.YIL station on 1 May showed that although each channel recorded the earthquake event, a long-period interference with a large amplitude was observed before the seismic signal. The correlation coefficient of the waveform showed that this station had interference signals before the earthquake. The PSD curve (the dark green curve in Figure 4d) exceeds that of NHNM. Comprehensive station waveform records show that this long-period interference is consistent with the signal characteristics of an instrument with a mass recenter command (McNamara & Buland, 2004), which can also be identified in the NM.DSM, GS.MXT, BJ.FHY, GZ.BJT, and XZ.SNA stations.

4.4 Regional long-period interference

The correlation coefficients for the horizontal components of the two events recorded by GZ.ZYT were very low. The original seismic waveform showed that the anomaly was mainly caused by long-period interference in the horizontal direction during the entire seismic recording process (Zürn & Widmer, 1995). However, other stations within 100 km normally recorded both the repeating earthquake events. As shown in the blue color curves in Figure 4(c) and (d), there is a larger PSD peak that is higher than the NHNM baseline from 10 s to 100 s. This is consistent with the signal characteristics of long-period interference sources (Wang et al., 2019). Therefore, the abnormal recording at the GZ.ZYT station was affected by the long-period interference in the region. The recording anomalies at the HI.SAY, GZ.KLT, HL.BJS, HN.LOD, XZ.MZG, and XZ.NMA stations were also due to similar regional long-period interference signals.

4.5 False detection probability and confidence threshold

The confidence interval threshold of the correlation coefficient obtained by the statistical parameter test in this study might be lower than that derived from studying the repeating earthquakes. Although this threshold avoids many false triggers caused by over-sensitivity in the detection model, there may also be a probability of missed detections. Therefore, the threshold can be adjusted to 2σ or even 1σ to increase the sensitivity of the system when filtering abnormal records. However, this will inevitably increase the workload during the confirmation of abnormal stations. For this study, 35, 48, and 35 potential abnormal stations would be obtained with a 2σ threshold for each of the three channels of broadband stations in mainland China, respectively; while 25, 29, and 23 potential abnormal stations would be filtered out after two round of the tests with a 3σ threshold. Although with different thresholds， we finally confirmed that the abnormal stations are almost the same, it can be considered that the workload of the latter can be reduced by nearly 30%. Therefore, it is valuable to continue filtering and analyzing the two-dimensional joint probability distribution of the correlation coefficient of each channel and the variance of the scale factor of the same station, as shown in Figure 3(d)–(i). In this work, obtaining potentially anomalous stations through statistical methods is a "falsification" process, while searching for repeating earthquakes and other previous studies is mostly a "verification" work. Therefore, the conditions for selecting the key parameters for thresholds, such as correlation coefficients and scale factors, are different.

4.6 Model application promotion

In this study, we utilized a group of relatively large magnitude (approximately Mw=6.8 and 7.0) repeating earthquakes in northeastern Japan, to identify possible instrumentation issues across the global networks and broadband permanent seismic stations in mainland China. In addition, a pair of repeating earthquakes with smaller magnitudes (Mw=5.2 and 5.4) occurred in northeastern Taiwan were used to verify the quality of the small-scale networks in the southeast of China, which are close to the epicenter, thereby expanding the scope of application of this method. Consequently, this method can be used on other datasets in different regions according to magnitudes and epicenters of repeating earthquakes: when the magnitude is as small as 5.0, it can be exploited to the small-scale regional network nearby, such as local or provincial networks. Certainly, relatively large magnitudes can be further applied to any global network for seismic data quality control.

5. Summary and Conclusions

A data quality assessment model based on repeating earthquakes was established by calculating the waveform correlation coefficient. The proposed model was applied to various situations and found to be effective for networks with different apertures. Statistical hypothesis testing of parameters was then utilized to determine the gross errors of the station records and quantitatively judge the stations with abnormal records in the seismic networks.

The earthquake record data quality for 842 permanent broadband seismic stations in mainland China was examined, of which 18 exhibited anomalous records. The results show that the data quality of most permanent seismic observation stations in mainland China is good, and that the data anomalies were mainly caused by calibration signals, instrument self-noise, mass recentering, and regional long-period interferences.

Using our concise filtering method, the quality of large-scale seismic station records can be quickly assessed using repeating earthquakes with highly similar seismic waveforms. This method not only reduces the amount of calculation as compared with that required for forward simulation, but also minimizes the impact of source parameter uncertainty and subsurface inhomogeneity on seismic waveforms. It can also be used on datasets in different regions according to magnitudes and epicenters of repeating earthquakes, especially suitable for regional-scale quality control work by repeating earthquakes with high frequency and small magnitude. In addition, repeating earthquakes are useful for many other geophysical analysis methods, and this study could provide additional insight in these applications as well.

Although data quality can be determined quickly and effectively using this method, it can only describe the quality during the time interval of two or more repeating earthquakes due to data limitations. Therefore, the versatility of the proposed method can be further improvemed. A more universal and flexible quality control model might be achievable by combining this method with forward simulation strategies and fine Earth structure model in future works. The improvement of seismic data quality requires the long-term joint efforts of seismic instrument managers, data centers, and researchers, along with international geoscience organizations.

Acknowledgments

The Monitoring and Forecasting Department of the CEA has played a significant role in promoting the project of this work. At the same time, this work has been helped and supported by many seismologists and geophysicists. Prof. Liu Ruifeng, and Mu Leiyu of Institute of Geophysics, CEA (IGCEA) gave careful guidance and great help in the research process for a long time. Prof. Ai Yinshuang of Institute of Geology and Geophysics, Chinese Academy of Sciences (IGGCAS) provided valuable advices on the work of waveform quality control and seismograph status monitoring. Prof. Su Jinrong of the Sichuan Earthquake Agency and Wang Honglei of Hebei Earthquake Agency have devoted a lot of effort to improve network performance and record quality mentioned in this article. The authors are grateful here.

Data Availability Statement

Data are obtained from http://service.iris.edu/fdsnws/dataselect/1/ for GSN Waveform data; http://service.iris.edu/fdsnws/station/1/ for StationXML; https://www.globalcmt.org/CMTsearch.

html for focal mechanism; https://www.sciencebase.gov/catalog/item/588b90dae4b0ad6732402989 for the PDE catalogs. Waveform data of permanent stations of mainland China can be accessed at https://dataverse.harvard.edu/privateurl.xhtml?token=fd62d4e1-3036-40ef-a2bc-139f9363ec26.

References

Casey R., Templeton M. E., Sharer G., et al. (2018). Assuring the quality of IRIS data with MUS−TANG. *Seismological Research Letters*, 89(2A): 630-639. https://doi.org/10.1785/0220170191

Ding J, Cai J. (2019) Two-side coalitional matching approach for joint MIMO-NOMA clustering and BS selection in multi-cell MIMO-NOMA systems. *IEEE Transactions on Wireless Communications*, 19(3): 2006-2021. https://dio.org/10.1109/twc.2019.2961654

Dodge, D., Beroza, G., Ellsworth, W. (1995). Foreshock sequence of the 1992 Landers, California, earthquake and its implications for earthquake nucleation. *Journal of Geophysical Research: Solid Earth*, 100(B6), 9865-9880. https://dio.org/10.1029/95JB00871

Ekström G., Dalton C., Nettles M. (2006). Observations of time-dependent errors in long-periodinstrument gain at global seismicstations. *Seismological Research Letters*, 77(1): 12-22. https://dio.org/10.1785/gssrl.77.1.12

Frémont, M., Malone, S. (1987). High precision relative locations of earthquakes at Mount St. Helens, Washington. *Journal of Geophysical Research: Solid Earth*, 92(B10), 10223-10236. https://dio.org/10.1029/JB092iB10p10223

H. Akaike. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 6(19): 716-723. https://dio.org/10.1109/TAC.1974.1100705

Hua C., Zhang Q, Xu G. , Zhang Y., Tao X. (2013). Performance reliability estimation method based on adaptive failure threshold. *Mechanical Systems and Signal Processing*, 36(2): 505-519. https://dio.org/10.1016/j.ymssp.2012.10.019

Hui N., Heydt G., Mili L. (2002). Power system stability agents using robust wide area control, *IEEE Transactions on Power Systems*, 17(4): 1123-1131. https://dio.org/10.1109/TPWRS.2002.805016

Isacks B. L., Sykes L. R., Oliver J. (1967). Spatial and temporal clustering of deep and shallow earthquakes in the Fiji-Tonga-Kermadec region. *Bulletin of Seismological Society of America*, 57(5): 935-958. https://doi.org/10.1785/BSSA0570050935

Katherine A., Jessica B., Anthony A., Phillip K. (2021). Pycheron: A Python-Based Seismic Waveform Data Quality Control Software Package. *Seismological Research Letters*, https://doi.org/10.1785/0220200418

Khoshmanesh M., Shirzaei M., Nadeau R. M. (2015). Time-dependent model of aseismic slip on the central San Andreas Fault from InSAR time series and repeating earthquakes. *Journal of Geophysical Research: Solid Earth*, 120(9): 6658-6679. https://dio.org/10.1002/2015JB012039

Lees, J. M. (1998). Multiplet analysis at Coso geothermal. *Bulletin of the Seismological Society of America*, 88(5), 1127-1143. https://dio.org/10.1029/98JB00594

Lehmann E., Joseph P. (2008). Testing statistical hypotheses. *Springer Science & Business Media*.

Li L, Chen Q., Niu F., et al. (2011). Deep slip rates along the Longmen Shan fault zone estimated from repeating microearthquakes. *Journal of Geophysical Research: Solid Earth*, 116(B9). https://dio.org/10.1029/2011JB008406

Li Y., Chen P, Cochran E., et al. (2006). Seismic evidence for rock damage and healing on the San

Andreas fault associated with the 2004 M 6.0 Parkfield earthquake. *Bulletin of the Seismological Society of America*, 96(4B): S349-S363. https://dio.org/10.1785/0120050803

Liu R., Gao J., Chen Y., Wu Z., Huang Z., Xu Z. and Sun L. (2008). Construction and development of digital seismograph networks in China. *Acta Seismologica Sinica*, 30(5): 533--539. https://dio.org/CNKI:SUN:DZXB.0.2008-05-013

Ma X., Wu Z., Jiang C. (2014). ‘Repeating earthquakes’ associated with the WFSD1 drilling site. *Tectonophysics*, 619: 44-50. https://dio.org/10.1016/j.tecto.2013.07.017

Magana-Zook S., Gaylord J., Knapp D., Dodge D., Ruppert S. (2016). Large-scale seismic waveform quality metric calculation using Hadoop, *Computers & Geosciences*, 94: 18-30. https://doi.org/10.1016/j.cageo.2016.05.012

Matsuzawa T., Igarashi T., Hasegawa A. (2002). Characteristic small-earthquake sequence off Sanriku, northeastern Honshu, Japan. *Geophysical Research Letters*, 29(11): 38-1-38-4. https://dio.org/10.1029/2001GL014632

McNamara, D., and Buland, R. (2004). Ambient noise levels in the continental United States. *Bulletin of the seismological society of America*, 94(4), 1517-1527. https://doi.org/10.1785/012003001

Moriya, H., Niitsuma, H., & Baria, R. (2003). Multiplet-clustering analysis reveals structural details within the seismic cloud at the Soultz geothermal field, France. *Bulletin of the Seismological Society of America*, 93(4), 1606-1620. https://dio.org/10.1785/0120020072

Nadeau R., Foxall W., McEvilly T. (1995). Clustering and periodic recurrence of microearthquakes on the San Andreas fault at Parkfield, California. *Science*, 267(5197): 503-507. https://dio.org/10.1126/science.267.5197.503

Poupinet G., Ellsworth W., Frechet J. (1984). Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras Fault, California. Journal of Geophysical Research: Solid Earth, 89(B7): 5719-5731. https://dio.org/10.1029/JB089iB07p05719

Ringler, A., Gee, L., Marshall, B., Hutt, C. R., & Storm, T. (2012). Data quality of seismic records from the Tohoku, Japan, earthquake as recorded across the Albuquerque seismological laboratory networks. *Seismological Research Letters*, *83*(3), 575-584. https://doi.org/10.1785/gssrl.83.3.575

Ringler A., Hagerty M., Holland J. (2015). The data quality analyzer: A quality control program for seismic data. *Computers & Geosciences*, 76: 96-111. https://doi.org/10.1016/j.cageo.2014.12.006

Schaff D., Beroza G. (2004). Coseism ic and postseismic velocity changes measured by repeating earthquakes. *Journal of Geophysical Research: Solid Earth*, 109(B10). https://dio.org/10.1029/2004JB003011

Schaff, David P., and Paul G. Richards. (2004). Repeating seismic events in China. *Science* 303(5661): 1176-1178. https://dio.org/10.1126/science.1093422

Schmittbuhl J., Karabulut H., Lenglin é O., et al. (2016). Long lasting seismic repeaters in the Central Basin of the Main Marmara fault. *Geophysical Research Letters*, 43(18): 9527-9534. https://dio.org/10.1002/2016GL070505

Shearer, P. M. (1997). Improving local earthquake locations using the L1 norm and waveform cross correlation: Application to the Whittier Narrows, California, aftershock sequence. *Journal of Geophysical Research: Solid Earth*, 102(B4), 8269-8283. https://dio.org/10.1029/96JB03228

Song L.,Yang W., Ge H.,Yuan S., Ouyang B. (2012). The current status and progress of portable seismic array observation in China. *Recent Development in World Seismology*, 3: 16-21. https://dio.org/10.3969 /j.issn.0235-4975.2012.03.004

Templeton M. (2014). Waveforms and their power spectral density expressions. Seattle, DC: IRIS DMC. Retrieved from https://ds.iris.edu/ds/nodes/dmc/tutorials/waveforms-and-their-power-spectral-density-expressions

Uchida, N. (2019). Detection of repeating earthquakes and their application in characterizing slow fault slip. *Progress in Earth and Planetary Science*, 6: 40. https://doi.org/10.1186/s40645-019-0284-z

Uchida N., Matsuzawa T., Ellsworth W., Imanishi K., Okada T., & Hasegawa, A. (2007). Source parameters of a M4. 8 and its accompanying repeating earthquakes off Kamaishi, NE Japan: Implications for the hierarchical structure of asperities and earthquake cycle. *Geophysical Research Letters*, 34(20):):153-175. https://dio.org/10.1029/2007GL031263

Wang F., Wang W., Long J., Mu L., Fu L. (2019). Seismic noise characteristics of broad-band seismic networks in Chinese mainland. *Acta Seismologica Sinica*, 41(5): 569−584. https://dio.org/10.11939/jass.20190031

Wen L. (2006). Localized temporal change of the Earth's inner core boundary. *Science*, 314(5801): 967-970. https://dio.org/10.1126/science.1131692

Xiong Y.,Wu X. (2010). The Generalizing Application of Four Judging Criterions for Gross Errors. *Physical Experiment of college*, 23(01): 66-68. https://dio.org/10.14139/j.cnki.cn22-1228.2010.01.008

Yang Y., Song X. (2020). Temporal changes of the inner core from globally distributed repeating earthquakes. *Journal of Geophysical Research: Solid Earth*, 125(3): e2019JB018652. https://dio.org/10.1029/2019JB018652

Yoshida S., Kato N., Fukuda J. (2015). Numerical simulation of the Kamaishi repeating earthquake sequence: change in magnitude due to the 2011 Tohoku-oki earthquake. *Tectonophysics*, 651: 44-57. https://dio.org/10.1016/j.tecto.2015.03.012

Yu W. (2016a). Time-dependent inner core structures examined using repeating earthquakes in subduction zones of the southwest Pacific. *Geophysical Journal International*, 2016a, 204(2): 1204-1215. https://dio.org/10.1093/gji/ggv508

Yu W. (2016b). Detectability of temporal changes in fine structures near the inner core boundary beneath the eastern hemisphere. *Geophysical Research Letters*, 2016b, 43(13): 6924-6931. https://dio.org/10.1002/2016GL069664

Zhang J., Richards P., Schaff D. (2008). Wide-scale detection of earthquake waveform doublets and further evidence for inner core super-rotation. *Geophysical Journal International*, 174(3): 993-1006. https://dio.org/10.1111/j.1365-246X.2008.03856.x

Zürn W., Widmer R. (1995). On noise reduction in vertical seismic records below 2 mHz using local barometric pressure. *Geophysical Research Letters*, 22(24): 3537-3540. https://dio.org/10.1029/95GL03369

References From the Supporting Information

Chai X. C., Wang Q. L., Chen W. S., Wang W. Q., Li Y. (2020). Research on a Distributed Processing Model Based on Kafka for Large-Scale Seismic Waveform Data. *IEEE Access*, PP(99):1-1. https://dio.org/10.1109/ACCESS.2020.2976660