Supporting Information for "Earthquake magnitude with DAS: a transferable data-based scaling relation"

Jiuxun Yin[1], Weiqiang Zhu[1], Jiaxuan Li[1], Ettore Biondi[1], Yaolin Miao[2], Zack J. Spica[2], Loïc Viens[3], Masanao Shinohara[4], Satoshi Ide[5], Kimihiro Mochizuki[4], Allen L. Husker[1], Zhongwen Zhan[1]

[1]Seismological Laboratory, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA

[2]Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, USA

[3]Los Alamos National Laboratory, Los Alamos, New Mexico, USA

[4]Earthquake Research Institute, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo, 113-0032, Japan

[5]Department of Earth and Planetary Science, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033, Japan

Contents of this file

Text S1. Conversion of raw DAS phase shift data to strain

A DAS system measures the phase/phase shift of Rayleigh back-scattered laser signal. When the DAS amplitude information is the focus, conversion from phase to strain is required:

$$d\phi = \frac{4\pi n G \xi}{\lambda}\epsilon, \tag{1}$$

where $d\phi$ and ϵ are the phase and strain, respectively. $n \approx 1.468$ and $\lambda = 1550$ nm are the refractive index of sensing fiber and optical wavelength, respectively. $\xi = 0.78$ is the photo-elastic scaling factor and G is the gauge length. Among all the parameters, only the gauge length G can be configured. All other parameters are related to cable properties and regarded as constants. We list all the instrumental parameters in Table S1.

Text S2. Event detection and phase arrival-time picking using PhaseNet-DAS on the California arrays

Fast and accurate detection and picking of seismic phase arrivals are critical to an effective earthquake early warning (EEW) system. We used a deep learning model, PhaseNet-DAS (Zhu et al., 2023), to detect and pick the arrival times of both P and S phases from earthquakes. Deep-learning-based phase-picking models, such as PhaseNet (Zhu & Beroza, 2019), have dramatically improved earthquake detection and phase picking on conventional seismic stations. The DAS-tailored PhaseNet-DAS (Zhu et al., 2023) model is based on semi-supervised learning to transfer deep learning models trained on large seismic datasets to DAS data (Zhu & Beroza, 2019). We use the two California DAS arrays (i.e., the Ridgecrest and Long-Valley arrays) to train PhaseNet-DAS so it can di-

rectly process 2-D spatio-temporal DAS data. The trained model achieves a high-picking accuracy and good earthquake detection performances on DAS data.

Text S3. Waveform Similarity Search on the Sanriku array

PhaseNet-DAS cannot be directly applied to the submarine Sanriku DAS array because it is trained based on terrestrial data. Therefore, we apply a Waveform Similarity Search (WSS), which utilizes the spatial coherency of earthquake waveforms across DAS channels for detection from the Sanriku dataset. We collect 10,379 high-SNR S-wave waveforms from 34 nearby Hi-net seismometers (Aoi et al., 2020), and cross-correlate them with continuous DAS data to find similar events. Before cross-correlating waveforms, the entire dataset is downsampled from 500 to 25 Hz and bandpass filtered between 1-8 Hz, which is the average dominant frequency band of earthquakes recorded along the array. Cross-correlations are finally computed independently for each individual DAS channel. A detection is triggered when the cross-correlation value exceeds nine times the median absolute deviation of the cross-correlation function at a single channel (Shelly et al., 2007). Then, a new event is kept if it matches at more than 40 channels. This relatively high threshold guarantees a large spatial consistency (i.e., an earthquake is detected over at least a 208-m section of the cable) and excludes non-coherent detections. In total, we detect 10,321 events over the 12-day period.

We then associate these events with the Japan Meteorological Agency (JMA) catalog to find their epicenter locations and magnitude information. We first compute the theoretical arrival time based on the 1-D preliminary reference Earth model (Dziewonski et al., 1981). We also apply an amplitude attenuation threshold to filter out cataloged earthquakes that

are likely too weak to be recorded. A body wave geometrical spreading model is applied: $A(r) = A_0 e^{-Br}/r$, where r is the hypocentral distance, A_0 is the amplitude at the source and B is a constant when assuming all earthquakes coming from different azimuth with a constant frequency (i.e., 2 Hz) as well as a homogeneous medium. This allows us to constrain further and refine the association process and only keep high-probability events in our analysis. Finally, a total of 464 earthquakes were selected as detected earthquakes for further analysis.

Text S4. Peak strain rate from DAS

With the event picking, we further extract the peak amplitude. We apply a series of quality control steps to ensure reliable peak amplitude extraction. Because of the different picking methods on the land (California) and submarine (Sanriku) DAS data, their processings are slightly different.

The California DAS arrays use the OptaSense ODH Plexus interrogator unit (IU), which gives the phase-converted raw measurement of strain. We down-sample the data to 100 Hz and convert strain data to strain rate to remove the low-frequency noise and instrumental drifts. No further filtering is applied to the land DAS data. The Sanriku DAS array is probed with an AP Sensing N5200A IU, which is different from that used in California but also measures the strain. Some detailed instrumental parameters are shown in Table S1. submarine DAS data are contaminated by oceanic noise at low frequencies (<0.5 Hz), especially for the channels near the coast (Spica et al., 2020). Therefore, we apply a 0.5 Hz high-pass filter to remove most of the ocean noise. Notice that this filtering process can

potentially generate magnitude saturation for large magnitude earthquakes, and needs to be considered in the future applications, especially for large earthquakes.

Because of the nature of the earthquake signals recorded by a DAS array, coherent signals should appear on most DAS channels as seismic waves propagate through the cable within a short period (less than the cable length divided by the apparent wave speed). We inspect the event picking and exclude events that are only detected by a few channels (≤ 100) in the DAS array. If the waveforms of an earthquake are only detected by a few channels, the detection is likely a false detection, and the recorded waveforms are mostly from local noise signals. Including those false-detected waveforms can lead to a magnitude overestimation of many small earthquakes ($M2-3$). We also tune this threshold of detection channel number to make sure the channel number we use can give the optimal results, regarding the qualified event number and final results of magnitude estimation.

We further calculate the signal-to-noise ratio (SNR) for P and/or S waves of each channel with the detected events. In this study, SNR is defined as $10\log_{10}(||S||^2/||N||^2)$, which is the average power ratio of the signal window (S) to the noise window (N) in decibel (dB). For the California data with clear P and/or S arrivals, the noise window is chosen as a 2-second time window ending 1 second before the detected P-wave arrival. The signal windows are the 2-second time window after the P and/or S direct arrivals, respectively. For the Sanriku DAS array, the situation is different. The Sanriku events are mainly detected by template-matching of S-waves, and it is difficult to get clear P phase arrivals. Therefore, we approximate the noise window as a 10-second-long window

ending 10 seconds before the detected event time. The signal window is chosen as 10-second long centered around the detected event time after we carefully check the event waveforms to ensure the SNR is robustly estimated. For the California data, we only keep the channels from M2+ earthquakes with SNR $>$ 10dB to ensure a good signal quality. For the Sanriku data, which is mainly used for validation, we only keep M2+ events with SNR values higher than 5dB.

After quality control, we measure the peak DAS strain rate for all available channels of the qualifying events. For the California DAS arrays with clear P and/or S pickings, we measure the peak amplitude of strain rate 2 seconds after the corresponding phase arrivals. We also test other window lengths up to 10 seconds, and the magnitude and distance coefficients using different window length are listed in Table S2. Although the values of scaling coefficients vary slightly, the corresponding site terms and the final results of magnitude estimation do not vary much with window lengths. Using shorter time windows can significantly help to suppress incorrect measurements due to noise from vehicle traffic. We show the results from the California DAS arrays using a 2-second window length in the main manuscript. For the Sanriku DAS array, we directly measure the peak S-wave amplitude from the 20-s long signal window centered at the event detection time.

Text S5. Iterative regression analysis

Based on the strong correlations between the peak amplitude and earthquake catalog magnitude and hypocentral distance (Figure 2), we fit for the empirical relations between earthquake magnitude, hypocentral distance, and peak amplitude (strain rate) for both P and S waves. Previous results on strainmeters (Barbour & Crowell, 2017; Barbour et al.,

2021) have validated the use of a generalized functional model to describe the observed peak values of dynamic strain:

$$\log_{10} E_i = aM + b \log_{10} D_i + K_i, \tag{2}$$

where E is the observed peak amplitude of dynamic strain/strain rate, D is the hypocentral distance in kilometers to each station/channel and M is the earthquake magnitude. The subscript i corresponds to each channel, and K_i is the corresponding site calibration term that compensates for the combined local effects such as instrumental coupling, fiber material properties, geological features, and noise. The goal is to fit the corresponding magnitude coefficient a, distance coefficient b, and K_i. We apply an iterative regression method to obtain the coefficients. Firstly, we assume that all channels in a DAS array share a constant site calibration term K_0. With the peak amplitude measurements and the targeting scaling relation, we apply regression to the data to fit for the coefficients a, b and the constant site calibration term K_0. Secondly, we fix the coefficients a and b, and fit for the specific site calibration term K_i for each channel to minimize the data misfit. Thirdly, we fix the site calibration terms K_i and further update the coefficients a and b. The second and third steps are repeated until the data misfit does not improved. We found that our dataset only need 3-5 iterations for the misfit values to converge within 1%. The regression can be done flexibly for either individual DAS arrays or multiple arrays at the same time. We test all cases and show our final coefficients a, b, and site calibration terms in Figure S1 and Figure S2, respectively.

The dynamic strain signal may also include earthquake-specific source terms (Barbour & Crowell, 2017; Barbour et al., 2021). For real time EEW applications, however, such prior information on the source process is difficult to obtain. Therefore, we do not explicitly fit for the source terms.

Text S6. Site calibration terms

Through our regression, we can also obtain the site calibration terms. Unlike conventional seismic sensors, which have standardized sensor designs and well-quantified instrumental responses, DAS instrument response is not as well constrained. The DAS cables used in this study are all dark fibers of the telecommunication optical fibers, and the cable constructions and installations vary significantly with regions. Both local conditions and cable installation properties greatly affect the recorded DAS data. Potential coupling issues are commonly noticed in the data(Ajo-Franklin et al., 2019; Lindsey et al., 2020; Trainor-Guitton et al., 2019; Paitz et al., 2020), but challenging to characterize from the instruments.

Our fully data-driven methodology, however, can directly quantify the local differences of DAS channels by introducing the site calibration terms K_i measured from earthquakes. The site-calibration terms K_i aim at quantifying all local effects that can change the measured amplitude, and are functions of channel locations. The obtained K_i are shown in Figure S2. We find that the values of K_i vary significantly along the cables in different regions. There are a few spikes of K_i values along the cables, which are caused by poorer data quality at local channel, likely due to fiber loops or the fiber not being coupled to the ground. Moreover, we find that the patterns of site calibration terms from P- and S-

waves are similar. Understanding the local variations of K_i is essential to characterize the local cable properties. Neverthess, we emphasize that the site calibration terms are just calibration terms that integrate many different local factors, such as the cable properties, instrumental coupling, and local geology. It is non-trivial to interpret K_i as a proxy of some specific factor, although we do see strong correlations between K_i and local shallow velocity structure(Spica et al., 2020; Viens, Bonilla, et al., 2022; Viens, Perton, et al., 2022) or wave amplification(Yang et al., 2022).

We also notice that the land (Figure S1) and submarine DAS arrays (Figure S4) are quite different in terms of the local site effects. The site calibration term values from the California arrays are all above 1 except for a few channels located at fiber loops. However, site calibration terms of the Sanriku array present larger variations. The site calibration terms in Sanriku are mostly less than 1 and indicate a local attenuation in the DAS-recorded amplitude. Further investigations of the differences between the land and submarine DAS and the transition from amplification to attenuation along DAS arrays would be an important future direction to explore.

Text S7. Validation of strain rate measurements and magnitude estimation

We first validate the scaling relation by comparing the measured peak strain rate with that calculated by the scaling relation Eq. (2) with the catalog magnitude M and hypocentral distance D (Figure S3). Most of the calculated values of peak strain rate are consistent with the measured values. The difference between predicted and measured values is less than one in logarithmic scale for all arrays. This validation guarantees that the regression is done properly, and the fitted scaling relation can robustly explain features in the data.

We can then use the determined scaling relation to estimate earthquake magnitude by reorganizing the scaling relation:

$$M_i = (\log_{10} E_i - b \log_{10} D_i - K_i)/a. \tag{3}$$

Given the distance D_i and measured peak amplitude E_i, the magnitude can be calculated at each individual DAS channel to get an estimation M_i, and the final magnitude M can be obtained by calculating the mean and median values of all M_i.

Text S8. Transferring scaling relation from California to Sanriku

We find that different regions have similar values of the scaling coefficients a and b (Figure S1). The regional differences mainly lie in the regional site calibration terms K_i (Figure S2). This implies that the DAS-recorded strain rate data follow the same magnitude scaling relation that can be transferred/extrapolated to other DAS arrays in different regions.

To test this hypothesis, we transfer the scaling relation obtained solely from California data to the Sanriku region, where the tectonic setting is different. We fix the magnitude and distance coefficients to the same as the values from California. Then, we randomly choose n events from the 47 qualified earthquakes in the Sanriku dataset as the fitting Set 1. Peak measurements of events in Set 1 are used to constrain the local site calibration term $K^S_{i(Sanriku)}$. The remaining events are used as validation Set 2 for magnitude estimation. This allocation of data sets allows us to test both the validity and transferability of the obtained scaling relation Eqs.(1)-(3) at the same time. Finally, we measure the percentage of good estimation for Set 2 events, which is defined as the percentage of events

whose magnitude is estimated within 0.5 unit of its catalog magnitude, as the metric to quantify how well the transferred scaling relation performs.

We systematically explore the event allocation: we increase the number of events n in Set 1 from 2, 3, ... to 30. For each n, we repeat the test for 50 times to measure the average percentage of good estimation. The variation of percentage is shown in Figure S5.

Our results show that only a few events are needed to calibrate the regional site terms (Figure S5), then the updated scaling relation can be used to estimate the earthquake magnitude (Figure 3). On average, two events give about 80% of good estimation percentage; and 5 events give relatively stable percentage from most random tests. Theoretically, we only need one well-cataloged earthquake measurement for each channel to measure the corresponding site calibration. Considering the uncertain data quality in a real situation, a few events with clear waveforms are sufficient to robustly constrain the site terms.

Text S9. Real-time magnitude estimation

We provide an idealized experiment to illustrate the application of our scaling relation for EEW. We assume that we can immediately detect and locate earthquakes. When the P wave arrives and the earthquake is detected, the system begins to measure the peak P-wave amplitude from the incoming DAS waveforms, and calculates the corresponding magnitude with the P-wave scaling relation Eq.(4) for the each available channel until 2 seconds after the P wave arrival. If the S-wave is also detected, the system also measures the peak S-wave amplitude and uses the S-wave scaling relation to estimate the magnitude until 2 seconds after the S wave arrival. If one single channel has both P-wave and S-wave

estimated magnitude, the mean value is taken. Our scaling relations are obtained with the peak amplitude in the 2-second window after P- or S- arrivals, but this time window can be easily adjusted based on how the scaling relations are built.

In this way, the incoming DAS data at each channel can be efficiently converted to real-time magnitude estimation. Finally, the magnitude estimations at all available channels are averaged to give the final magnitude estimation for the earthquake, and the standard deviation of magnitude estimation is taken as the uncertainty estimation. We tested on many events, including one event outside of our regression data sets, and find that all of them can give an accurate estimation of the magnitude.

References

Ajo-Franklin, J. B., Dou, S., Lindsey, N. J., Monga, I., Tracy, C., Robertson, M., ... Li, X. (2019). Distributed acoustic sensing using dark fiber for near-surface characterization and broadband seismic event detection. *Scientific Reports*, *9*(1), 1328. doi: 10.1038/s41598-018-36675-8

Aoi, S., Asano, Y., Kunugi, T., Kimura, T., Uehira, K., Takahashi, N., ... Fujiwara, H. (2020). Mowlas: Nied observation network for earthquake, tsunami and volcano. *Earth, Planets and Space*, *72*(1), 126. doi: 10.1186/s40623-020-01250-x

Barbour, A. J., & Crowell, B. W. (2017). Dynamic strains for earthquake source characterization. *Seismological Research Letters*, *88*(2), 354–370. doi: 10.1785/0220160155

Barbour, A. J., Langbein, J. O., & Farghal, N. S. (2021). Earthquake magnitudes from dynamic strain. *Bulletin of the Seismological Society of America*, *111*(3), 1325–1346. doi: 10.1785/0120200360

Dziewonski, A. M., Chou, T.-A., & Woodhouse, J. H. (1981). Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *Journal of Geophysical Research: Solid Earth*, *86*, 2825–2852. doi: 10.1029/JB086iB04p02825

Lindsey, N. J., Rademacher, H., & Ajo-Franklin, J. B. (2020). On the broadband instrument response of fiber-optic das arrays. *Journal of Geophysical Research: Solid Earth*, *125*(2), e2019JB018145. doi: 10.1029/2019JB018145

Paitz, P., Edme, P., Gräff, D., Walter, F., Doetsch, J., Chalari, A., ... Fichtner, A. (2020). Empirical investigations of the instrument response for distributed acoustic sensing (das) across 17 octaves. *Bulletin of the Seismological Society of America*, *111*(1), 1–10. doi: 10.1785/0120200185

Shelly, D. R., Beroza, G. C., & Ide, S. (2007). Non-volcanic tremor and low-frequency earthquake swarms. *Nature*, *446*(7133), 305–307. doi: 10.1038/nature05666

Spica, Z. J., Nishida, K., Akuhara, T., Pétrélis, F., Shinohara, M., & Yamada, T. (2020). Marine sediment characterized by ocean-bottom fiber-optic seismology. *Geophysical Research Letters*, *47*(16), e2020GL088360. doi: 10.1029/2020GL088360

Trainor-Guitton, W., Guitton, A., Jreij, S., Powers, H., & Sullivan, B. (2019). 3d imaging of geothermal faults from a vertical das fiber at brady hot spring, nv usa. *Energies*, *12*(7), 1401. doi: 10.3390/en12071401

Viens, L., Bonilla, L. F., Spica, Z. J., Nishida, K., Yamada, T., & Shinohara, M. (2022). Nonlinear earthquake response of marine sediments with distributed acoustic sensing. *Geophysical Research Letters*, *49*(21), e2022GL100122. doi: 10.1029/2022GL100122

Viens, L., Perton, M., Spica, Z. J., Nishida, K., Yamada, T., & Shinohara, M. (2022). Understanding surface wave modal content for high-resolution imaging of submarine sediments with distributed acoustic sensing. *Geophysical Journal International*, *232*(3), 1668–1683. doi: 10.1093/gji/ggac420

Yang, Y., Atterholt, J. W., Shen, Z., Muir, J. B., Williams, E. F., & Zhan, Z. (2022). Sub-kilometer correlation between near-surface structure and ground motion measured with distributed acoustic sensing. *Geophysical Research Letters*, *49*(1), e2021GL096503. doi: 10.1029/2021GL096503

Zhu, W., & Beroza, G. C. (2019). Phasenet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273.

Zhu, W., Biondi, E., Li, J., Yin, J., Ross, Z. E., & Zhan, Z. (2023). Seismic arrival-time picking on distributed acoustic sensing data using semi-supervised learning. *arXiv preprint*.

	Cable length	Channel number	Gauge length (m)	Spatial sampling (m)	Pulse repetition rate (Hz)
Ridgecrest	10km	1250	16.34	8.00	5000
Long Valley North	50km	4300	20.42	11.63	2000
Long Valley South	50km	4600	20.42	10.87	1000
Sanriku	50km	9600	40.78	5.21	500

Table S1. Instrumental configuration parameters of each study area.

	P, 2 s	P, 3 s	P, 4 s	S, 2 s	S, 4 s	S, 6 s	S, 10 s
magnitude coef. (a)	0.437	0.471	0.487	0.690	0.680	0.681	0.678
distance coef. (b)	-1.269	-1.220	-1.200	-1.588	-1.520	-1.508	-1.494

Table S2. Scaling coefficients using different time windows for the California DAS arrays.

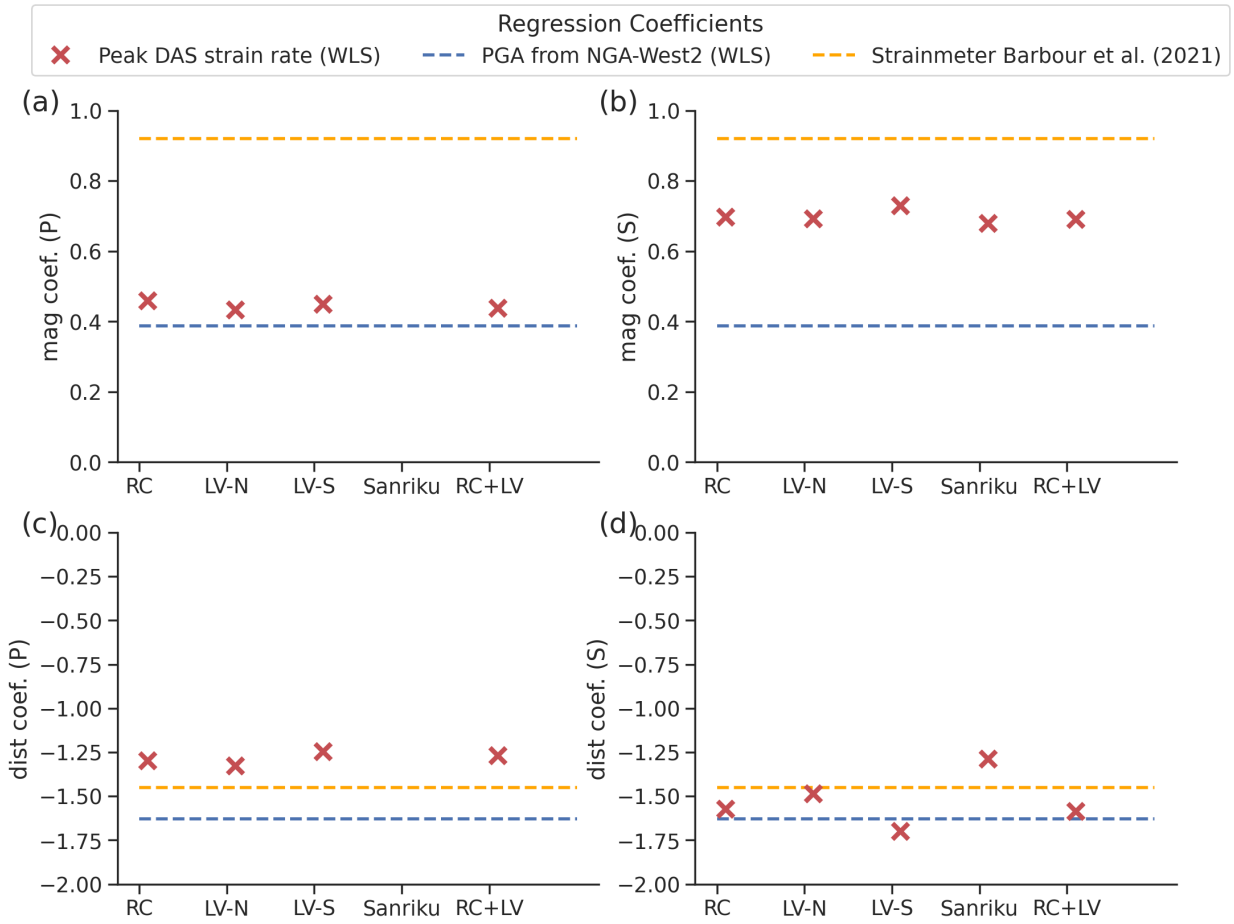

Figure S1. Regression coefficients from different data sets: (a) the P wave magnitude coefficients; (b) the S wave magnitude coefficients; (c) the P wave hypocentral distance coefficients; (d) the S wave hypocentral distance coefficients. RC is for Ridgecrest data only; LV-N is for Long-Valley northern array data only; LV-S is for Long-Valley southern array data only; Sanriku is for Sanriku data only; RC+LV are the results from combining RC, LV-N and LV-S arrays' data. The dashed lines also indicate the coefficients (distance: -1.45 and magnitude: 0.92) from strainmeter data (Barbour et al., 2021) and fit the same model Eq.(1) with the NGA-West 2 PGA dataset (distance: -1.63 and magnitude: 0.39), respectively.

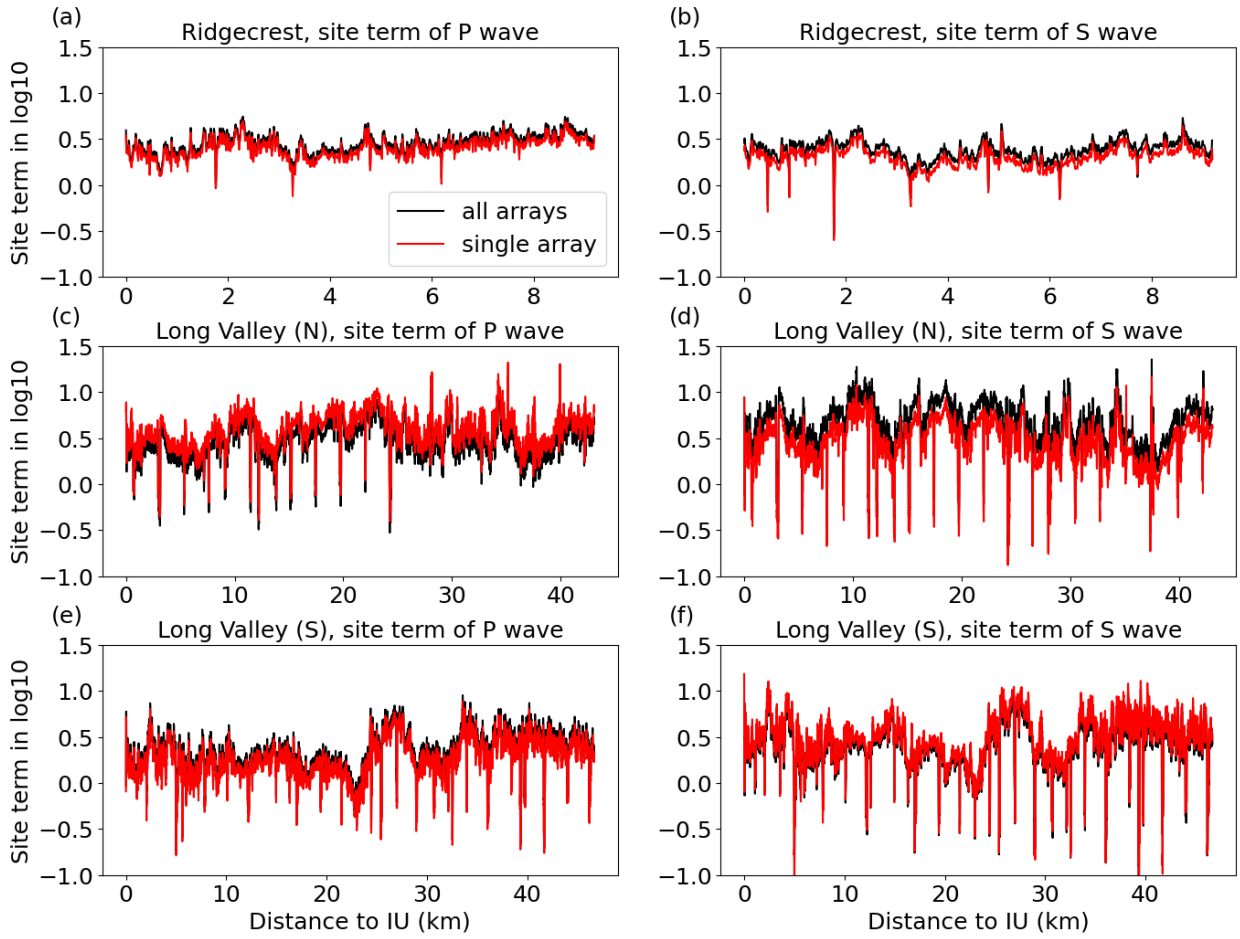

Figure S2. Site calibration terms of arrays: (a) Ridgecrest array, P wave; (b) Ridgecrest array, S wave; (c) Long-Valley Northern array, P wave; (d) Long-Valley Northern array, S wave; (e) Long-Valley Southern array, P wave; (f) Long-Valley Southern array, S wave. Black lines are results from fitting all arrays and red lines are results from fitting individual array data.

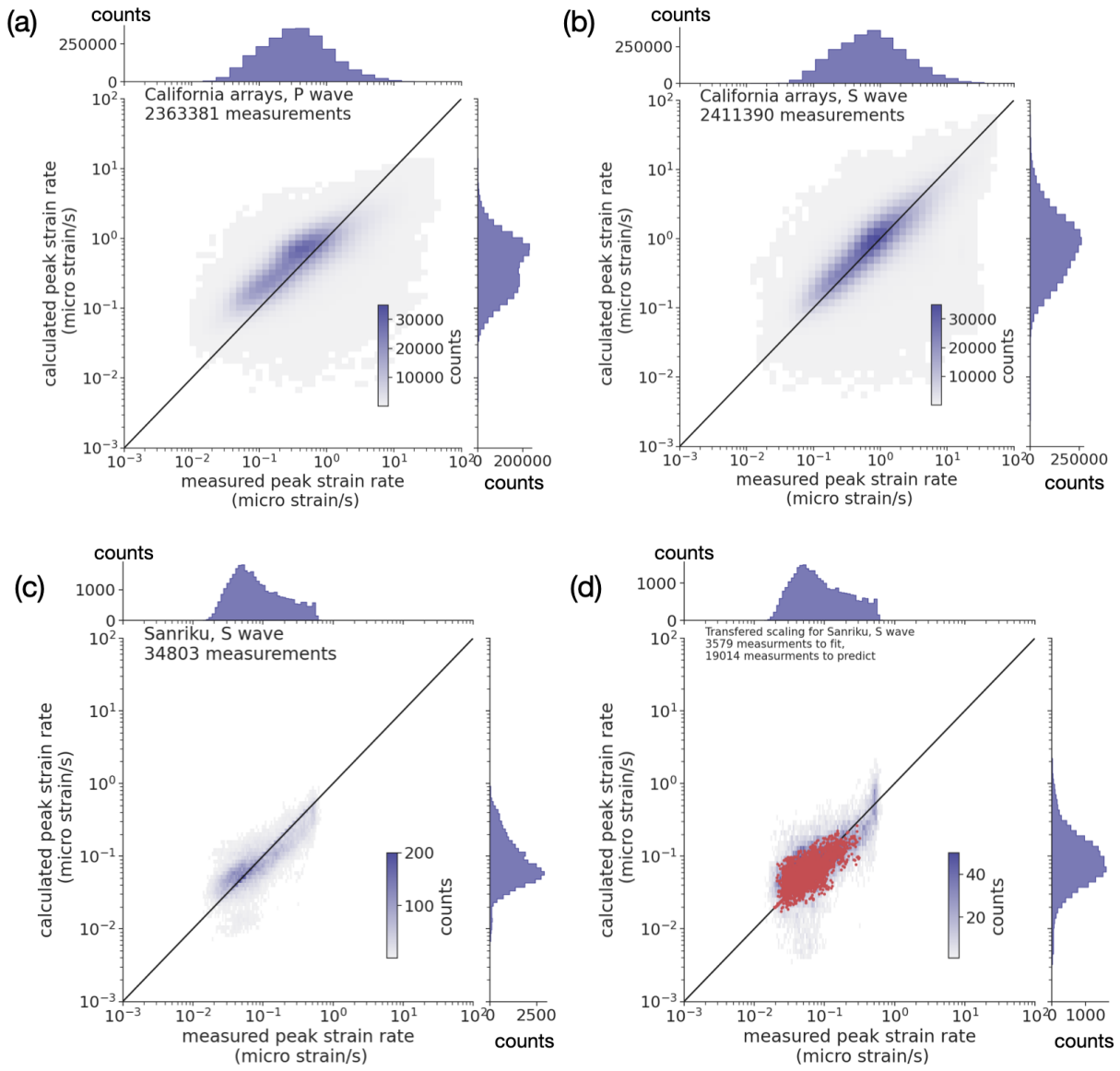

Figure S3. Validation on the peak DAS strain rate by comparing the measured strain rate and calculated peak strain rate based on the scaling relations. (a) Validation on the P-wave scaling relation applied to the California data. The scaling relation is from all three California DAS arrays. (b) Validation on the S-wave scaling relation applied to the California data. The scaling relation is from all three California DAS arrays. (Caption continued on next page)

Figure S3. (Caption continued from previous page) (c) Validation on the S-wave scaling relation applied to the Sanriku data. The scaling relation is from the Sanriku array. (d) Validation on the S-wave scaling relation applied to the Sanriku data. The scaling relation is transferred from California DAS arrays. Red dots highlight measurements that are used to calibrate the local site terms. Black solid lines indicate the accurate estimation.

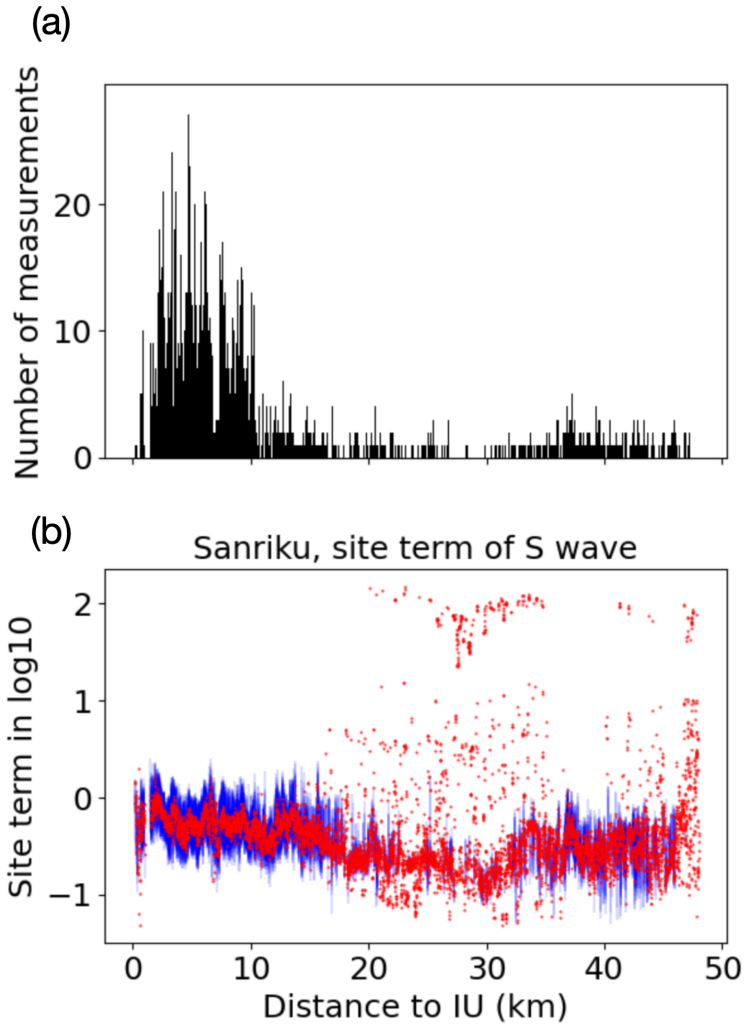

Figure S4. Site calibration terms of Sanriku array. (a) Number of peak DAS strain rate measurements at each channel. (b) Best fit site calibration term at each channel is shown by the red dots. The standard deviation is indicated by the blue error bars.

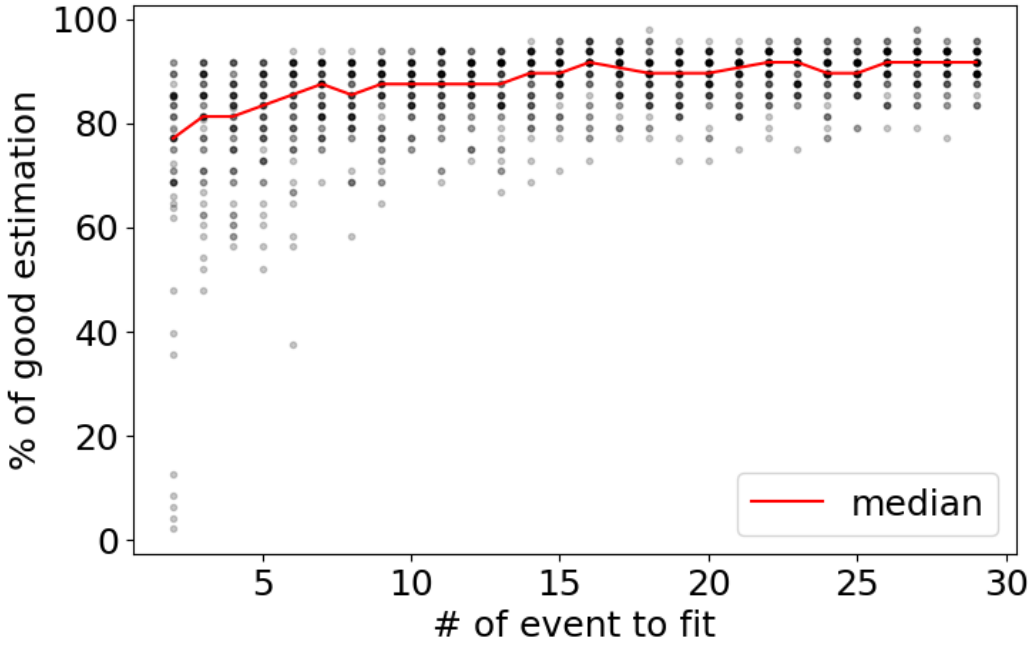

Figure S5. Number of events for transferring scaling relation. Each black dot corresponds to results of one random test. The red line is the average percentage of good magnitude estimation with uncertainty less than 0.5 units of magnitude.

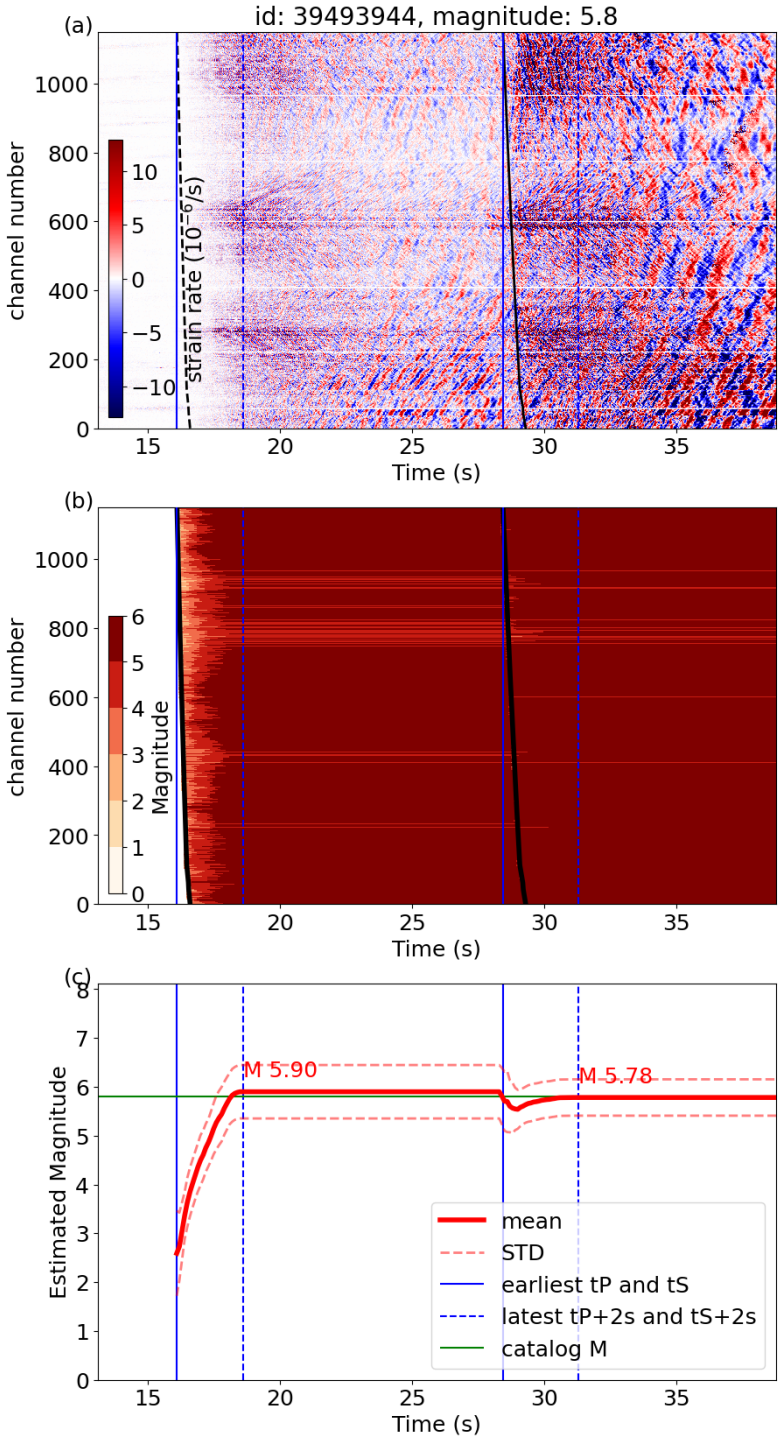

Figure S6. Same magnitude estimation as Figure 4, but for the M5.8 earthquake that occurred in the Ridgecrest region.