Geophysical Journal International

Geophys. J. Int. (2022) **230,** 1725–1750
Advance Access publication 2022 April 13
GJI Seismology

https://doi.org/10.1093/gji/ggac137

Full-waveform centroid moment tensor inversion of passive seismic data acquired at the reservoir scale

Qiancheng Liu,[1,2] Umair bin Waheed,[3] Dmitry Borisov,[4] Frederik J. Simons,[1] Fuchun Gao[5] and Paul Williamson[5]

[1] *Department of Geosciences, Princeton University, Princeton, NJ* 08544*, USA. E-mail: qiancheng.liu@kaust.edu.sa*
[2] *Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing* 100029*, China*
[3] *Department of Geosciences, King Fahd University of Petroleum and Minerals, Dhahran* 31261*, Saudi Arabia*
[4] *Kansas Geological Survey, Lawrence, KS* 66047*, USA*
[5] *TotalEnergies SE Research and Technology, Houston, Texas TX* 77002*, USA*

Accepted 2022 April 8. Received 2022 February 2; in original form 2021 June 24

SUMMARY

Passive seismic inversion at the reservoir scale offers the advantages of low cost, negligible environmental impact and the ability to probe a target area with low-frequency energy not afforded by even the most modern active-source seismic technology. In order to build starting models suitable for full-waveform wave speed tomography, characterization of earthquake sources is an indispensable first step. We present a workflow for the centroid moment tensor (CMT) inversion of seismic events identified in a passive seismic data set acquired by a large and dense array of three-component broad-band seismic sensors in a mountainous setting in the Himalayan foothills. The data set comprised 256 instruments operating for 2×4 months over an area of 8000 km^2. An initial 3-D wave speed model was determined for the region via the analysis of first-arriving traveltime picks. Of the 2607 identified seismic events that were well recorded at frequencies between 0.2–50 Hz, 86 with magnitudes $1.3 \leq M \leq 3.0$ initially had their CMT focal mechanisms determined by a waveform fitting procedure built on a Green's function approach in a 1-D layered average wave speed model, for stations within an offset of 10 km, in the frequency range 0.2–1.4 Hz. Here, we obtain updated CMT mechanisms for the 86 events in that catalogue via multicomponent full-waveform inversion in the 3-D wave speed model. Our workflow includes automated data- and model-driven data selection using a combination of different metrics derived from signal-to-noise considerations and waveform-fitting criteria, and relies upon spectral-element simulations of elastic wave propagation in the 3-D wave speed model, honouring topography. Starting from the initial CMT solutions, we seek improvement to the data fit within the frequency band 0.5–2.5 Hz by minimizing the waveform difference between observed and synthetic data, while accommodating wave speed-model errors by allowing for small time-shifts. We balance uneven data coverage and tune their contributions via data-space weighting functions. We quantify the improvements to the data fit in terms of different metrics. We summarize the changes to the CMT solutions, and present and analyse the resulting catalogue for the region, including their breakdown into double-couple and non-double couple components, and their relation to mapped faults.

Key words: computational seismology; waveform inversion; earthquake source parameters.

1 INTRODUCTION

In areas of the world where active-source acquisition is challenging, naturally occurring (micro-)earthquakes may provide useful information for hydrocarbon exploration (Saltzer *et al.* 2011; Tselentis *et al.* 2011). Using spontaneous seismicity is cost-effective and environmentally friendly, and offers access to low-frequency and large-offset data that are vital (Virieux & Operto 2009) to building good starting models for full-waveform inversion (FWI). Equally important for seismic imaging as currently practiced across scales (e.g. Liu & Gu 2012; Tromp 2020) is the characterization of the source. To use earthquakes as sources for tomographic inversion (e.g. Tape *et al.* 2009; Bozdağ *et al.* 2016; Lei *et al.* 2020), knowledge of the moment tensor (Jost & Herrmann 1989; Tape & Tape 2013) is crucial. The focal mechanism furthermore plays a central role in understanding regional tectonics (e.g. Vicente *et al.* 2008;

Hingee *et al.* 2011), the evolution of the stress field in scenarios of induced seismicity (e.g. Li *et al.* 2011), and for studies of probabilistic seismic hazard (e.g. Convertito & Herrero 2004).

The centroid moment tensor (CMT) represents a seismic point source by means of the six components of the symmetric moment tensor, the three space coordinates of the centroid, and a source-time function (Ekström *et al.* 2012). Waveform-based inversions for CMT mechanisms may include deterministic least-squares methods (e.g. Liu *et al.* 2004; Kim *et al.* 2011), statistical sampling approaches (e.g. Vackář *et al.* 2017; Fichtner & Simutė 2018), neural networks and pattern recognition (e.g. Käufl *et al.* 2014, 2015), and may involve any of a number of observables (both seismic and geodetic) and numerical methods (e.g. normal-mode summation, WKBJ synthetics, finite-difference, finite-element, spectral-element approaches) to carry out the forward modelling of the seismogram in simplified or realistically heterogeneous media. Specific issues arising from the shallowness of earthquake seismic sources and their inversion based on high-frequency seismic waveforms were discussed by Hejrani & Tkalčić (2020). As regards the generally smaller events (microseisms) and often exotic mechanisms of induced seismicity or hydraulic fracturing, questions of resolvability from a variety of observational vantage points (e.g. surface monitoring or borehole acquisition) have been addressed by Eyre & van der Baan (2017), and Willacy *et al.* (2019) developed a full-waveform workflow for event location and moment tensor inversion in those settings. Particularities related to the automatization of regional moment-tensor calculations using high-performance computing infrastructure have recently been discussed by Triantafyllis *et al.* (2021).

In our particular problem setting, we have access to an initial 3-D wave speed model and initial CMT solutions. Both of these were derived under restrictive assumptions from limited data and modelling efforts (i.e. automated traveltime picks, polarity analysis, near-offset records, 1-D average wave speed models). They require updating and extending to realize the benefits of passive-source data acquisition for resource exploration and reservoir imaging, and for tectonic interpretation. Our well-curated and quality-controlled data set is derived from a large and dense array of three-component broad-band seismic sensors. We follow the framework of Liu *et al.* (2004) and Kim *et al.* (2011) to improve on the CMT solutions by minimizing the least-squares waveform differences between the observations and synthetics calculated via spectral-element modelling (Komatitsch & Tromp 1999) in the 3-D elastic initial model. The unstructured mesh, generated internally by the open-source code `SPECFEM3D` (Peter *et al.* 2011) honours the topography of the mountainous Khatlon Region in the Himalayan foothills.

Beyond the complexity of the wave-propagation modelling, the low magnitudes of the seismic events, $1.3 \leq M \leq 3.0$, present challenges for full-waveform CMT inversion due to the inherent low signal-to-noise ratios (SNRs) of the data and their contamination by ambient noise, which requires careful filtering and proper data selection. We select time-windowed waveforms around the distinct body-wave arrivals determined from automated high-frequency picking of *P* arrivals on the vertical (Saragiotis *et al.* 2002) and of *S* arrivals on the horizontal components (Lois *et al.* 2013), carrying them over across components where data quality permits. Data segments are retained for analysis when the initial fit between the observed and synthetic waveforms is high, as determined by various similarity metrics. Data weights take into account epicentral distance and back azimuth (see, e.g. Ruan *et al.* 2019), in a way that favours contributions from the near offsets and balances the record counts across azimuthal quadrants.

In this paper, we first briefly discuss the background of the data collection and the construction of the initial models both for wave speed structure and for the preliminary CMT mechanisms of the subset of events. Next, we review the theory of least-squares full-waveform CMT inversion in the context of spectral-element simulations. We discuss the rationale for and implementation of a suite of data- and model-driven metrics to aid in data selection, and discuss the choice of weighting functions. Finally, we illustrate our methodology by deriving updated CMT mechanisms for 86 moment tensor earthquakes in the Bokhtar region, and quantify the degree to which they represent an improvement over the previous state of knowledge using a variety of metrics.

2 DATA AND INITIAL MODELS

We report on an experiment conducted for TotalEnergies SE by SeismoTech Ltd. in 2015. The campaign consisted of two phases of about four months (121 d) each, with 252 and 247 continuously recording receivers, respectively, installed as shown in Fig. 1(a). The area of study, some 8000 km^2 in extent, is marked by significant rugged topography, with elevations between 0.5–3.5 km. The average station spacing was about 5 km, resulting in a station density in each phase of about 1 station per 25 km ^{2}. For context, the average station spacing in the North American USArray was about 70 km.

The Geobit C-100/S-100 borehole sensors were placed in shallow (<5 m) drill holes, oriented vertically. Data quality control consisted of two procedures for each station. The first was a calibration, orientation and polarity weight-drop test, performed in the field immediately after installation of the sensors. In the second, data acquired during the first few days were analysed in order to estimate the station noise levels according to the procedures laid out by McNamara & Buland (2004). Stations or components that did not pass quality control (some 7 per cent of the stations in Phase I and some 10 per cent in Phase II) were not retained for analysis. An additional quality control during Phase I consisted in ascertaining that a certain teleseismic event (a moment-magnitude 4.9 event that ruptured at 100 km depth below the Afghanistan-–Tajikistan border on 2015 March 21, and was assigned the Global CMT code 201503211744A), was indeed recorded by the totality of the stations.

The bandwidth of the equipment ranges from 0.2 to 96 Hz and their velocity response is approximately flat between 1 and 96 Hz. All data were recorded in a frequency band from 0.2 to 50 Hz, with a sampling rate of 10 ms. Some 2607, about 10 per recording day, relatively small-scale passive seismic events were identified by SeismoTech using a three-component energy-based STA/LTA detection analysis (Withers *et al.* 1998; Trnkoczy 2012). Accurate *P*- and *S*-wave arrivals within the segments were determined using kurtosis-based detection (Saragiotis *et al.* 2002) on the vertical component, and time-domain polarization attributes via eigenvalue analysis of the three-component seismic record (Lois *et al.* 2013), respectively.

The arrivals picked automatically were reviewed and cross-checked by an analyst at SeismoTech. Subsequently they were used to determine hypocentre source locations and origin times for all these events, the majority of which occurred within the array bounds, and to obtain the 3-D wave speed model shown in Fig. 1(b), with its laterally averaged velocities against the 1-D velocity profile shown in Fig. 1(c), the 3-D model mesh honouring the topography shown in Fig. 1(d), and the moment-tensor solutions for a subset of 86 events with magnitudes $M > 1.3$ shown in Fig. 1(e).

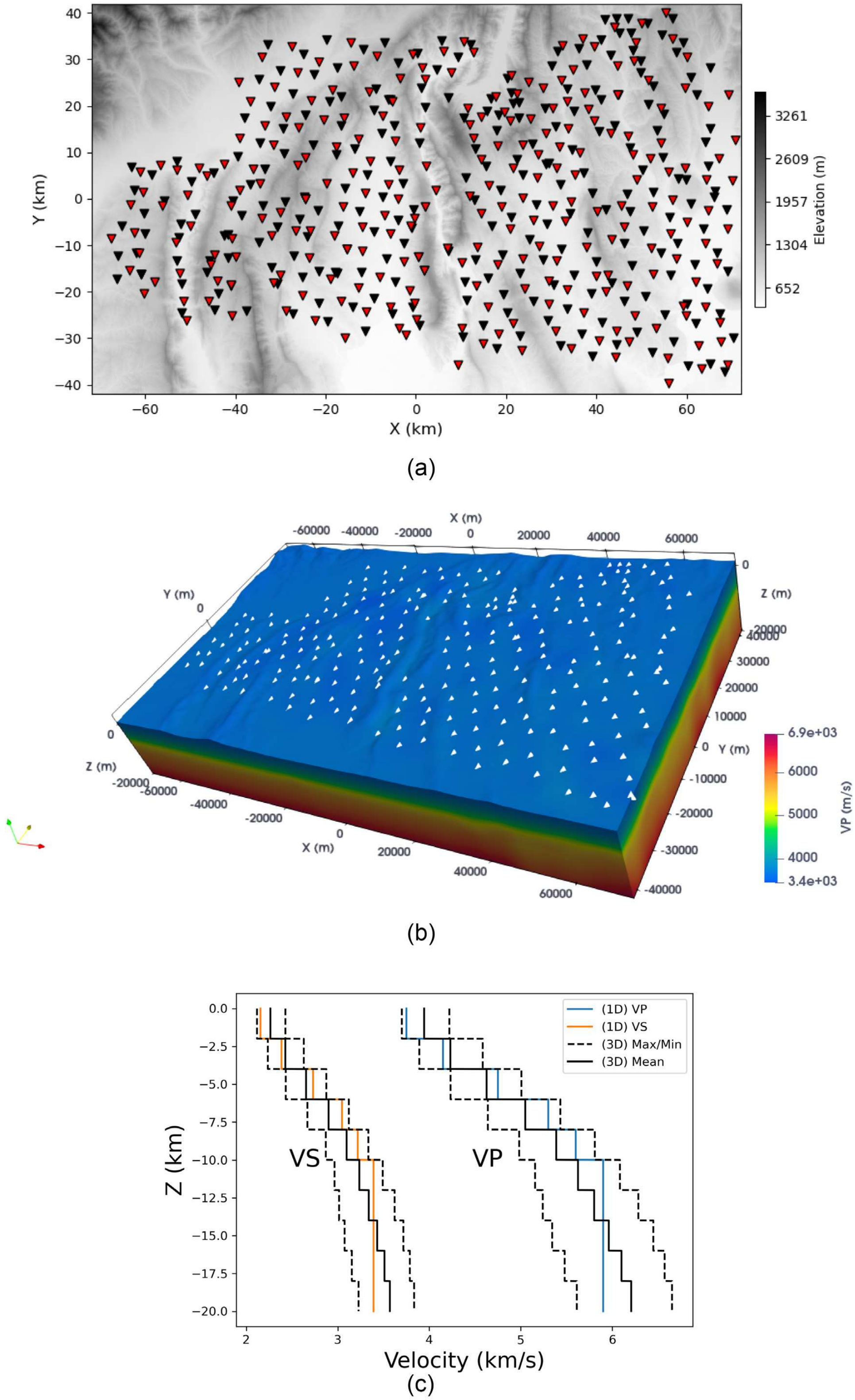
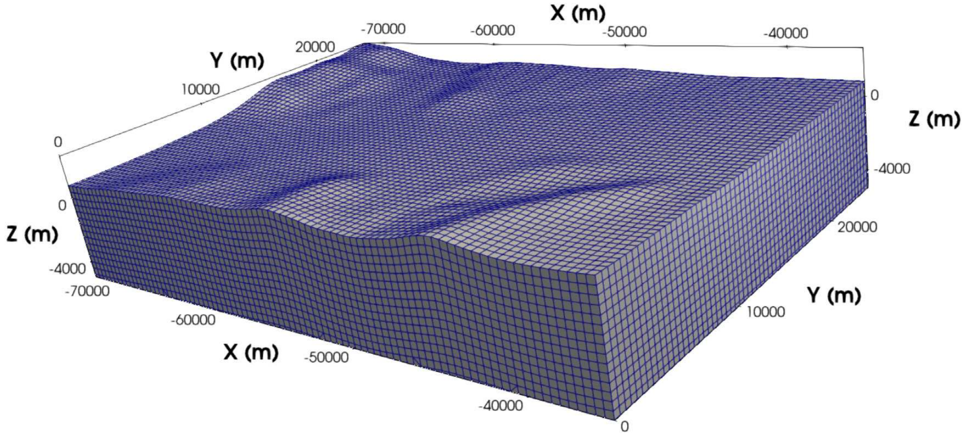
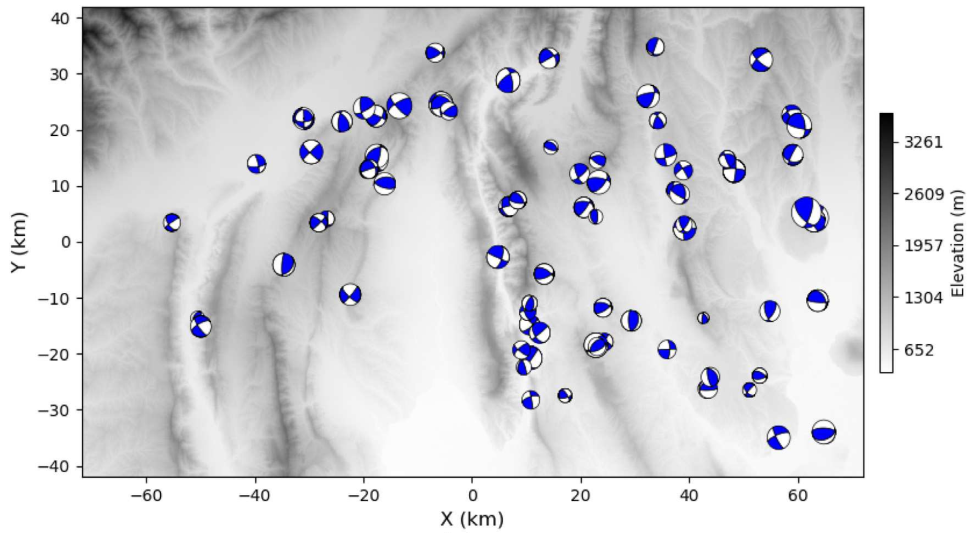


Figure 1. (a) Array geometry and topography of the study region. The stations of the Phase I deployment are in black, and those of Phase II in red triangles. (b) Side view of the 3-D initial *P* wave speed model, with white triangles indicating the location of the Phase I stations. (c) The 1-D *P* and *S* wave speed profiles and the depth averages of the 3-D velocity model shown in (b). (d) Side view of a part of the 3-D mesh honouring the topography. (e) Initial focal mechanisms for the 86 earthquakes in our catalogue, derived from the preliminary analysis by SeismoTech Ltd.



(d)



(e)

Figure 1. Continued.

Source locations and origin times were preliminarily determined by SeismoTech from an iterative traveltime inversion (Lee & Lahr 1972) within the 1-D starting wave speed model shown in Fig. 1(c), and subsequently refined by a joint hypocentre and 3-D wave speed inversion (e.g. Aki & Lee 1976; Crosson 1976) conducted by SeismoTech via ray-tracing (Um & Thurber 1987) and damped least-squares singular-value decomposition inversion (Thurber 1983). We refer to the resulting model of compressional, V_{P}, and shear wave speeds, V_{S}, as the 'initial 3-D wave speed model'—see Fig. 1(b). The majority of these events are shallower than 10 km in depth.

Focal mechanisms were preliminarily estimated by SeismoTech starting from first motions using the fault-plane fitting program `FPFIT` (Reasenberg & Oppenheimer 1985), and subsequently improved via waveform fitting with `ISOLA` (Sokos & Zahradnik 2008) using the 1-D wave speed model shown in Fig. 1(c), employing only the highest-SNR data within the frequency band 0.2–1.4 Hz, and with most solutions determined by at least four stations and from six to nine stations at offsets below 10 km. We refer to these double-couple mechanisms as the 'initial CMT solutions'—see Fig. 1(e). The majority of the fault plane solutions indicate thrust and strike-slip faulting, with thrust focal mechanisms oriented NS to NNE–SSW, and strike-slip faults dominantly in the NE–SE and NW–SE directions. This is consistent with the local tectonics, which is characterized by NNE–SSW directed thrust structures that dip to the SE, and an additional set of low-dip-angle thrust structures oriented NW–SE (see also Schurr *et al.* 2014; Kufner *et al.* 2018).

Both these modelling results require the updates that are the subject of this paper. Based on tomography under a paraxial ray tracing approximation (Beydoun & Keho 1987), the 3-D wave speed model does not take into account finite-frequency or full-waveform effects, and therefore insufficiently resolves complex structure in the surface. The CMT solutions, for their part, were determined using a local 1-D wave speed model and for small offsets, requiring further updates to become consistent with the 3-D wave speed model over the full range of station offsets.

3 METHODOLOGY

To prepare for full-waveform tomographic inversions for elastic wave speed structure underneath our area of interest, this paper describes our procedure for the full-waveform inversion for the CMT focal mechanisms that are an essential ingredient for wave speed modelling (Liu *et al.* 2004; Tape *et al.* 2009; Bozdağ *et al.* 2016; Lei *et al.* 2020).

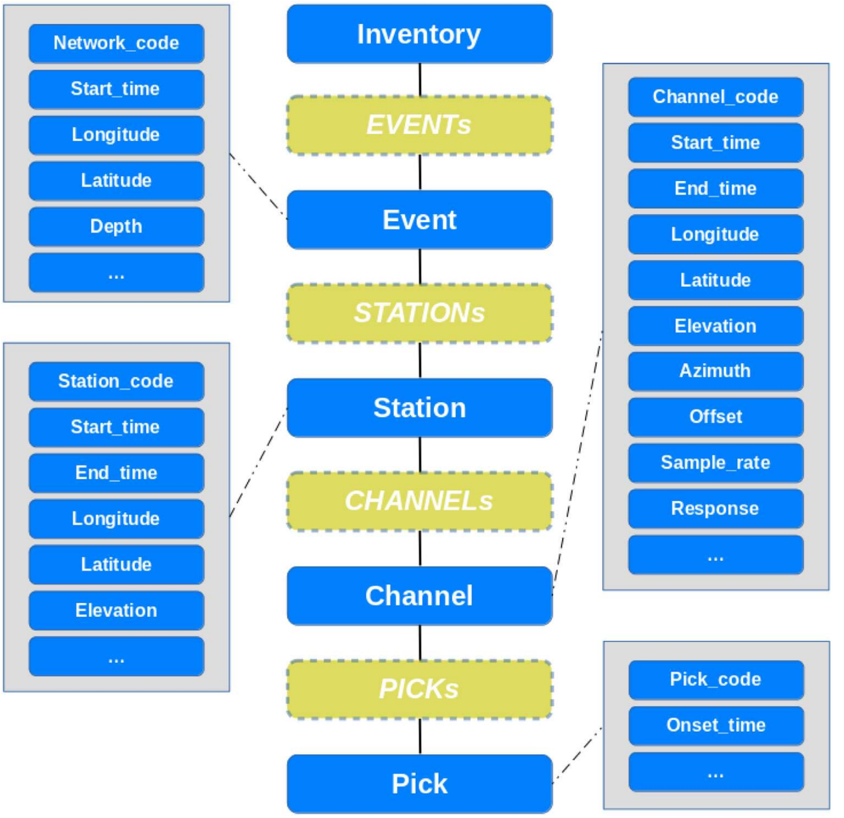


Figure 2. Data management structure used in our workflow.

The initial CMT solutions are to be updated and extended by increasing the station-event offsets, broadening the frequency band for the waveform data, and considering both amplitude and phase information. The pre-processing data management workflow is illustrated in Fig. 2, and the key modelling steps are introduced below.

3.1 Forward modelling

Under the point-source approximation for earthquakes, a CMT solution $\{\mathbf{M}, \mathbf{x}_s, S\}$ consists of a six-parameter moment tensor $\mathbf{M}$, a three-parameter set of spatial coordinates $\mathbf{x}_s$, and a source wavelet S. Here, we only focus on $\mathbf{M}$ and $\mathbf{x}_s$, taking S to be a band-limited Heaviside function for the wavefield (in particle displacement). Including the wave speed model in the notation, we will denote all model parameters collectively with the lower-case vector $\mathbf{m}$.

To enable the moment-tensor inversion of seismic traces, we must link the source and structural models to the data via forward modelling, which we can express in terms of the Green's functions of the wavefield operator. No such functions were stored or catalogued. Rather, the full-waveform simulation is carried out on the fly, based upon the spectral-element method (SEM; Komatitsch & Tromp 1999; Peter *et al.* 2011), which has been widely used in simulating seismic wave propagation across global (Bozdağ *et al.* 2016; Lei *et al.* 2020), regional (Zhu *et al.* 2012) and local (Borisov *et al.* 2018) scales.

Fig. 1(b) shows the *P*-wave speed model, which measures 144, 84 and 20 km in the horizontal (X and Y) and vertical (Z)-directions, respectively. The S wave speed model is similar to the P-wave model, but with small perturbations in the ratio of compressional to shear wave speeds, $V_P/V_S \sim 1.74$, and the density model is inferred from Gardner's equation (Gardner *et al.* 1974). To honour the topography, which is available on an evenly spaced grid at a resolution of 100 m, we construct a mesh of approximately regular hexahedra measuring about 500 m × 500 m × 500 m, containing 352, 208 and 48 Gauss–Lobatto–Legendre (GLL) nodes in each of the three coordinate directions, respectively. The top surface is a free surface, whereas we implement absorbing boundary conditions on all of the others. A detail of the meshed topography is shown in Fig. 1(d), without subsampling.

The central frequency of the Gaussian wavelet (in particle velocity) used in the simulation is 10 Hz. The time step is 5.0×10^{-3} s, in line with the Courant–Friedrichs–Lewy (CFL) conditions (Courant *et al.* 1967) corresponding to the local wave speeds and adapted to the mesh size intervals. The simulation of synthetics and the attendant calculation of Fréchet derivatives (Fichtner *et al.* 2006) for the source parameters are based on the GPU-accelerated code `SPECFEM3D`. The runtime to produce seismograms of 60 s recording duration is around 2 min per simulation, on a system with 16 Tesla K20 GPU cards. For scaling considerations with other hardware settings and resources, see Peter *et al.* (2011).

Fig. 3 shows two versions of an event gather for one particular earthquake source, event CMT001, sorted by offset. Fig. 3(a) shows synthetic seismograms calculated for this source mechanism in the initial wave speed model, whereas Fig. 3(b) shows a subset of the observations. In both panels we draw, as a solid line, a time–distance curve for a reference wave speed of 5.2 km s^{-1}, which we used to determine SNRs for the data at hand. Hence the culling of the observed data bottom panel, which only shows high SNR traces. In both panels the dotted lines mark the vertical average of the compressional and shear wave speeds in the initial model at the location

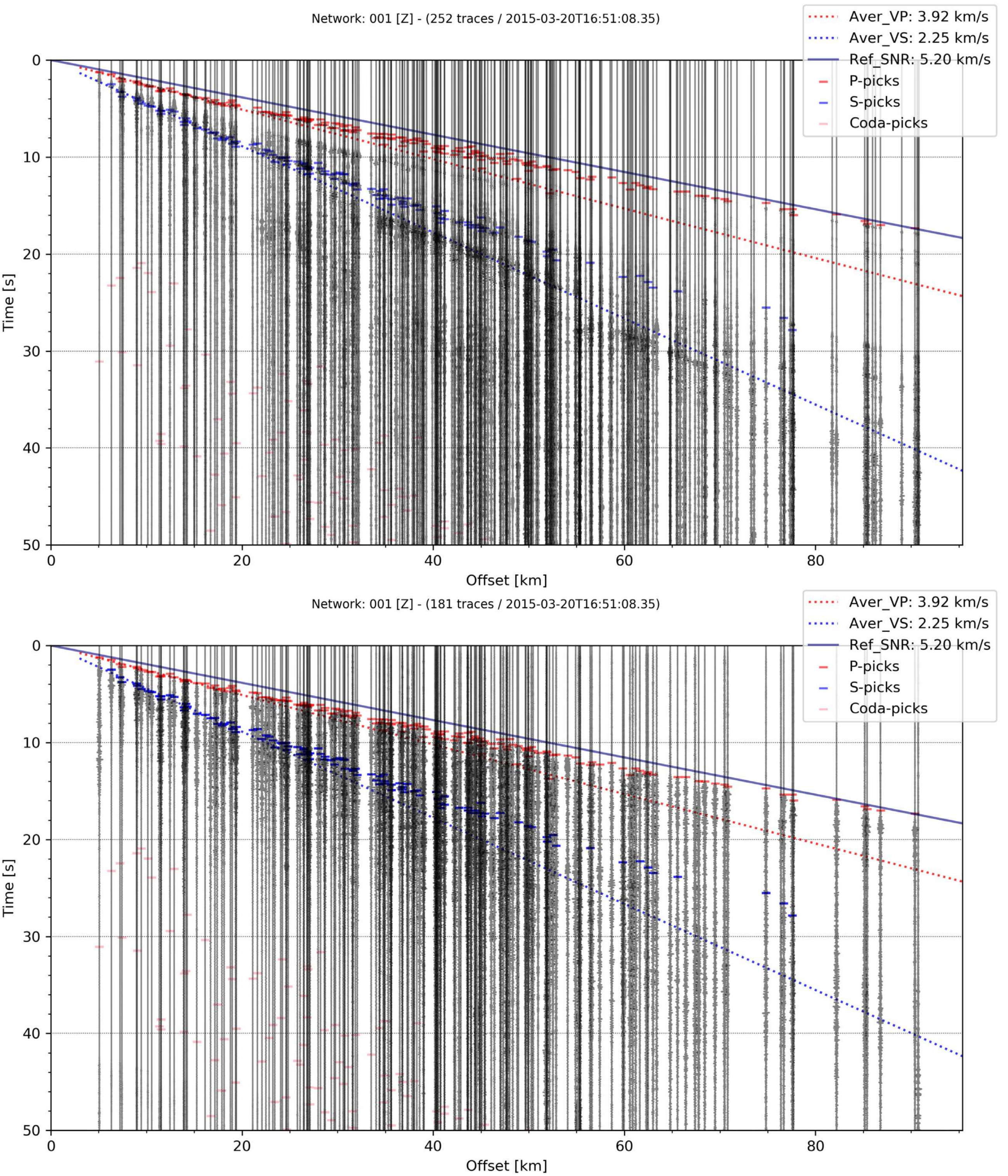


Figure 3. Time-offset vertical-component gathers for earthquake source CMT001. All traces are normalized by their maximum absolute value for clarity. (a) Synthetic seismograms calculated in the 3-D initial wave speed model shown in Fig. 1(b). (b) Subset of the observations deemed of high quality based on signal-to-noise considerations calculated with regards to the navy-blue reference line drawn at 5.2 km s^{-1}. The red dotted lines at 3.92 km s^{-1} and the blue dotted lines at 2.25 km s^{-1} correspond to the *P* and *S* wave speed, respectively, averaged over all depths in the 1-D layered model shown in Fig. 1(c) at the initial earthquake location. Red horizontal dashes are picked *P*-arrival times, blue ones are picked *S* arrivals. Pink horizontal dashes are coda-wave measures for which we have no further use.

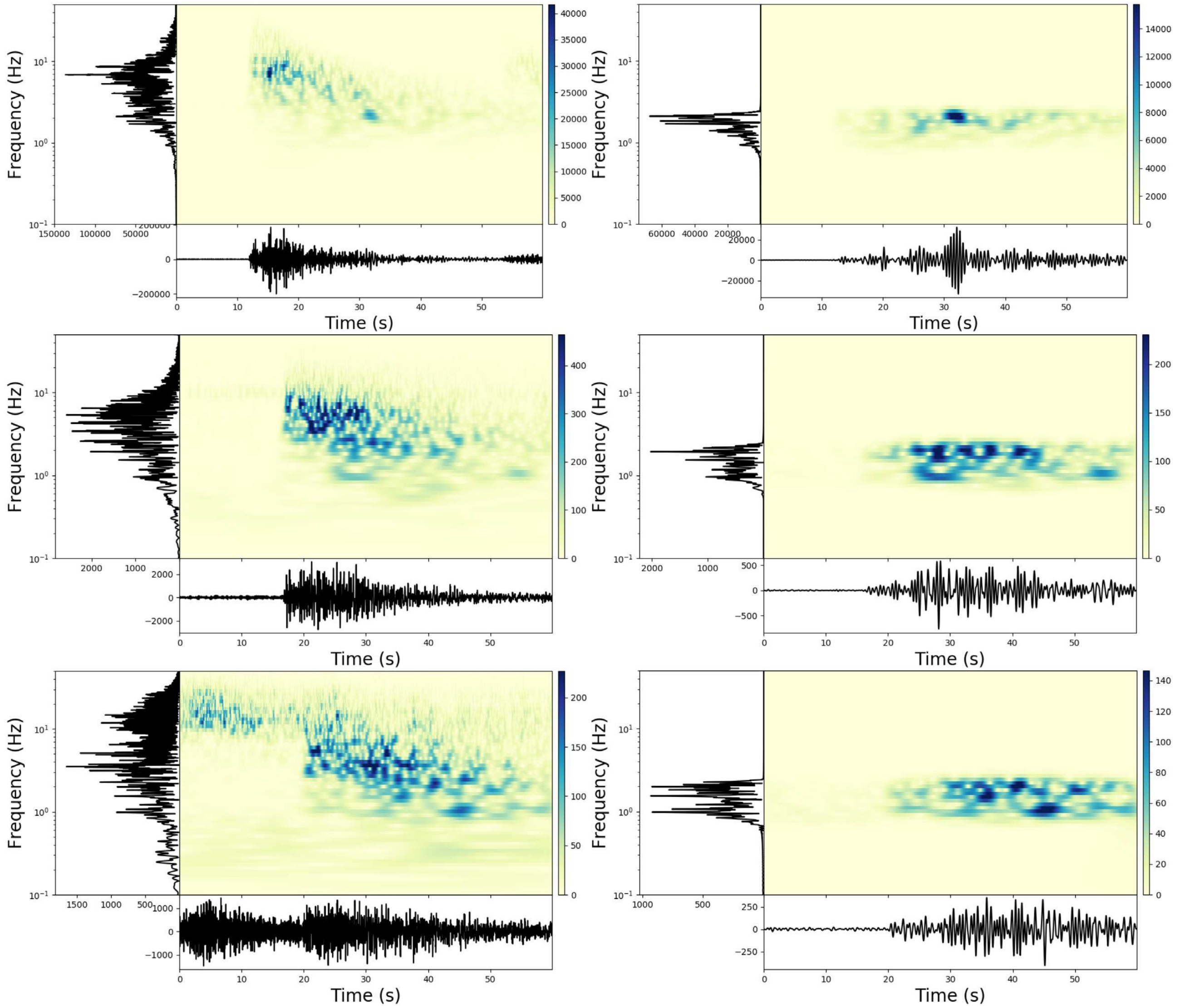


Figure 4. Spectrograms, spectral densities and seismic traces of (top) high-, (middle) intermediate- and (bottom) low-quality signals. The left-hand column shows the demeaned and detrended unfiltered data. Note the clear onsets of the seismic signals, despite the banded low-frequency noise below around 0.5 Hz in the middle left and middle right panels. The right-hand column shows the data after filtering within the passband 1–2 Hz. All traces from top to bottom now have similar SNRs.

of the epicentre. Short red and blue horizontal dashes mark the picked *P*- and *S*-arrival times. The pink dashes are additional 'coda' picks provided to us by SeismoTech, which were not incorporated into our modelling.

3.2 Data pre-processing and selection

Before commencing the inversion procedure we implement a number of quality-control measures on the data set, which serve to target those traces for which the SNR is sufficiently high. All seismograms are passed through the standard data processing procedures such as mean-removal, detrending, tapering, bandpass filtering, component rotation and, in the case of the observed data, instrument-response removal.

Fig. 4 illustrates the importance of targeting the right frequency band by plotting, in the left column, spectrograms of the broad-band observations considered to be of 'high', 'fair' and 'low' quality. The frequency range between 0.2 and 0.5 Hz is rather noisy, contaminated by microseismic ambient noise (Nakata *et al.* 2019). We begin by bandpassing the data using a filter with passband [0.5–1–2–2.5] Hz, whereby 0.5 and 2.5 Hz are the limit frequencies, and 1 and 2 Hz the corners. A Hann time window is applied to select ± 1 s around each arrival-time pick, both for the *P*- and the *S*-wave window. As the right column of Fig. 4 shows, the resulting seismograms have markedly higher SNRs.

The same observation can be made on the basis of Fig. 5, which again plots event gathers for the CMT001 source mechanism, as in Fig. 3, but now sorted by azimuth, aligned on the *P*-arrival pick, and highlighting the short data windows before and after filtering.

The following step is entirely data driven. We reject noisy seismograms trace-by-trace, based on two SNR criteria determined by comparing the amplitude and power of signal and noise. Sorted by epicentral distance, the seismic event gathers are divided in the time domain into a likely noise and a likely signal segment, demarcated by the reference time-distance curves drawn at 5.2 km s^{-1} in

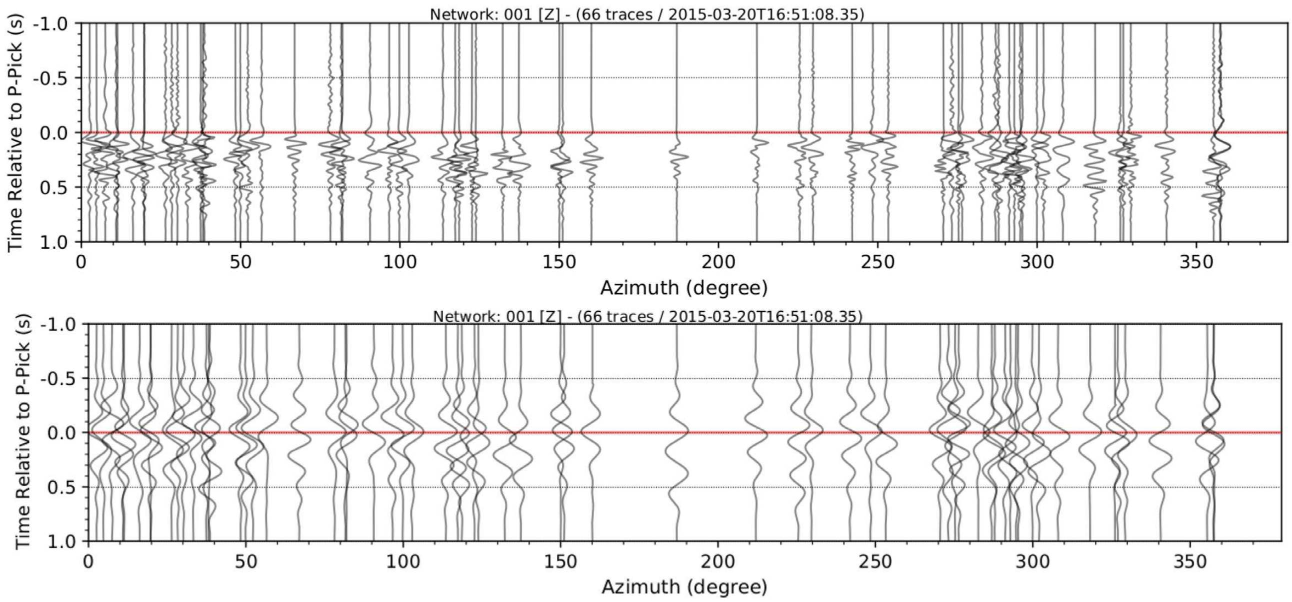


Figure 5. Time–azimuth vertical-component gathers for earthquake source CMT001. All traces are normalized by their maximum absolute value for clarity. Compared to Fig. 3, we are now only showing 2 s of every waveform, aligned on the picked P arrivals of the traces that passed quality control for offsets between 0 and 40 km. (a) Raw data and (b) data bandpassed between 1 and 2 Hz, where the agreement with the starting model is substantial.

Fig. 3. Amplitude and power ratios of the putative noise and signal segments are calculated according to

$$\Gamma = A_\mathrm{s}/A_\mathrm{n}, \tag{1}$$

$$Y = E_\mathrm{s}/E_\mathrm{n}, \tag{2}$$

where A_s and A_n are the maximum absolute values of the signal and noise segments, respectively, and E_s and E_n the mean-squared values. User-defined thresholds Γ_0 and Y_0 are applied to reject traces with low SNR.

The next step is model-derived. For the short measurement windows centred around the automatically-picked P- and S-arrival times, we formulate three different metrics that measure the similarity between the observed and the synthetic waveforms calculated in the initial model $\mathbf{m}$. The first two are correlative measures, namely the time lag, δt, between observations, $d(t)$, and synthetics, $s(t)$, at their maximum normalized cross-correlation value, and the value of their normalized cross-correlation, ξ, at that time lag. The third, η, is the energy ratio after aligning the traces by applying the time-shift δt. Without adding any more specifics for the moment, we thereby have the metrics

$$\delta t = \arg\max_{\delta t} \frac{\int s(t-\delta t, \mathbf{m})\, d(t)\, \mathrm{d}t}{\sqrt{\int s^2(t-\delta t, \mathbf{m})\, \mathrm{d}t}\sqrt{\int d^2(t)\, \mathrm{d}t}}, \tag{3}$$

$$\xi = \max \frac{\int s(t-\delta t, \mathbf{m})\, d(t)\, \mathrm{d}t}{\sqrt{\int s^2(t-\delta t, \mathbf{m})\, \mathrm{d}t}\sqrt{\int d^2(t)\, \mathrm{d}t}}, \tag{4}$$

$$\eta = 10 \log_{10} \frac{\int d^2(t)\, \mathrm{d}t}{\int s^2(t-\delta t, \mathbf{m})\, \mathrm{d}t}. \tag{5}$$

In the forthcoming figures we will use the labels *TS* (time-shift), *CC* (normalized cross-correlation) and $d\ln A$ (energy ratio). We only take candidate measurements that meet user-defined thresholds in each of the three metrics, that is, for which $\delta t < \delta t_0$, $\xi > \xi_0$, and $|\eta| < \eta_0$. Among them, δt_0 is frequency-dependent to avoid cycle-skipping, the normalized cross-correlation threshold empirically gets assigned the value $\xi_0 = 0.7$ (Maggi *et al.* 2009), and η_0 is estimated by numerical perturbation tests for a given CMT solution and wave speed model, and according to source-receiver geometry.

Applying these metrics individually is not always sufficiently selective for our challenging data set of small-magnitude events. As we show in the next section, the primary driver for our inversion is neither a multiplicative correlation nor a ratio-based measure of misfit, but rather a mean-squared waveform difference. Outliers in the metrics influence least-squares inversion strongly—quadratically. Within the thresholded subset of seismogram pairs with small time lags, we therefore take the time-lagged mean-squared waveform difference itself,

$$\chi = \int [d(t) - s(t - \delta t, \mathbf{m})]^2\, \mathrm{d}t, \tag{6}$$

as an additional quality-control metric. We accept waveforms for which $\chi < \chi_0$, with the threshold given in terms of the standard deviations of the distribution of χ over the data set of measurement windows.

In summary, the application of these quality-control measures ensures that the data fed to the optimization constitute a relatively well-fitting homogeneous set, and we monitor the evolution of all metrics over the course of the iterations. All metrics in eqs (3)–(6) apply independently over the P- and S-wave segments. In our study, we maintain $\delta t_0 = 0.35$ s, $\xi_0 = 0.7$, $\eta_0 = 40$ and $\chi_0 = \sigma_\chi$, whereby σ_χ is the standard deviation of the χ distributions per event gather, separately for the P and S windows.

Fig. 6 is a graphical rendition of the results of the selection procedure, in which the measurements selected for event CMT001 are shown highlighted, for the vertical (Z), radial (R) and tangential (T) components. Every row corresponds to a station, and all rows rendered in light blue correspond to stations with data that passed the crude quality metric of eqs (1)–(2). Specific P and S measurement windows that passed the quality control measures in eqs (3)–(6) are brightly coloured. While the P picks were made on the vertical and the S picks on the horizontal components, when the corresponding windows on the other components pass quality control, they become

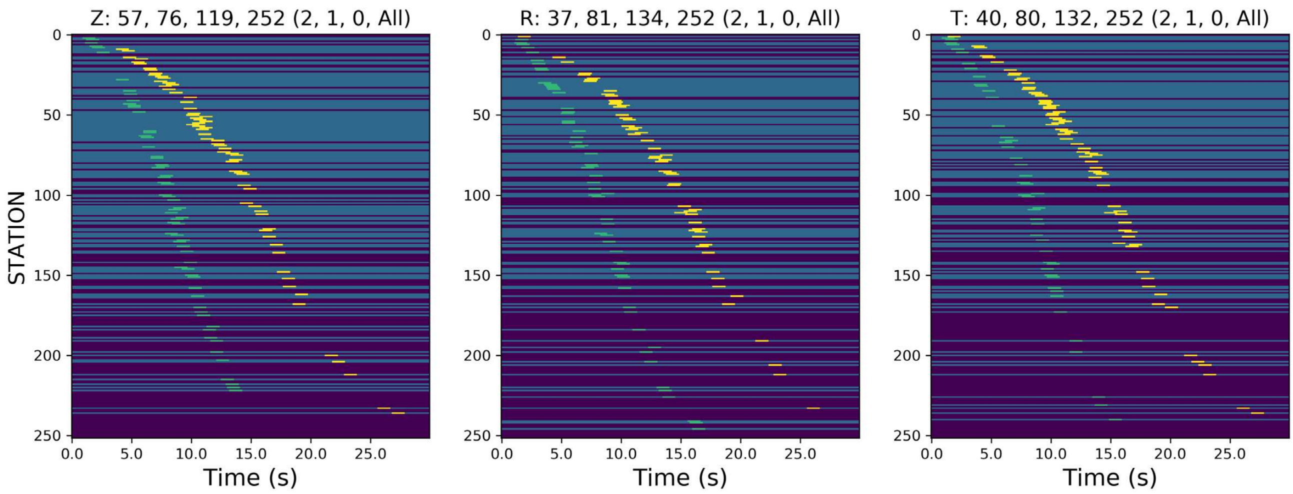


Figure 6. Windows of length 2 s containing the *P* wave (green) and *S* wave (yellow), for the vertical (Z), radial (R), transverse (T) components recorded for event CMT0001. Shown are all such windows that satisfy data-driven quality metrics and which show reasonable agreement with model-derived predictions using the starting source and wave speed models, as explained in the text. The original *P* arrivals were picked on the vertical, and the *S* arrivals only on the east and north components, but where the quality control measures allowed it, the windows were carried over across all components.

available for the inversion as well. Note, in particular, the presence of *P*-arrival windows on the transverse (T) components, a phenomenon ascribed to the scattering effect of the strong topography shown in Figs 1(b) and (d).

3.3 Centroid moment tensor inversion

The explicit model dependence on **m** for all the synthetic traces in eqs (3)–(6) includes both the full CMT solution, including the source-time function, and the elastic wave speed model. The focus in our present paper is limited to the further adjustment of the geometric CMT parameters (moment tensor and source location), and no further inversion for centroid time-shift, source-time function or 3-D wave speed structure is attempted here.

While we cannot altogether avoid the effects of errors in wave speed structure and source timing (Zhao & Helmberger 1994), for relatively long seismic periods, the differences between the true wave speed structure and a high-quality guess mainly affect the seismic arrival times and not as much their waveforms (Komatitsch *et al.* 2004). Moreover, we implemented a rather strict set of initial quality controls that guaranteed that the initial waveforms were indeed robustly compatible. Hence, allowing small time-shifts when matching the synthetic seismograms with the observations as Liu *et al.* (2004), we simply drive the inversion for the geometric source parameters using a time-lagged misfit between the observed and the synthetic traces at each iteration.

In summary, to seek a new moment-tensor solution and source location for a given data set, we minimize the collection of individual contributions to the full-waveform squared misfit between data and synthetics, over specific time portions T_i,

$$\chi_i = \int_{T_i} [d_i(t) - s_i(t - \delta t_i, \mathbf{m})]^2 \, \mathrm{d}t, \tag{7}$$

where we now use the index i to cycle over the *P*- and *S*-wave portions of the seismograms and the relevant components (Z, R, T) where those measurements are available. As before, d and s denote the appropriately filtered and tapered, rotated and instrument-corrected, observed and synthetic time-series, and δt the applicable time-shift. The overall penalty function is a weighted sum over all available segments:

$$\phi = \sum_{i}^{N} w_i \chi_i, \tag{8}$$

whereby w_i is a weight that can vary according to backazimuth, epicentral distance, component or any other categorical designation. In our case, we design the weighting functions as a product, with a for azimuth, d for epicentral distance and c for any other type of categorization,

$$w_i = w_i^a \, w_i^d \, w_i^c. \tag{9}$$

The specific forms of the factors contain normalized dependences of the form

$$w_i^a \propto 1/N_a \qquad w_i^d \propto \exp(-\Delta/\Delta_0), \qquad w_i^c \propto 1/N_c, \tag{10}$$

with N_a the number of measurements in a quadrant, Δ the epicentral distance for a given reference value Δ_0 and N_c the number of measurements in each category. The parameter w_i^a aims to balance contributions and control for uneven station coverage, w_i^d is designed to stress the contributions from the nearer epicentral distances, which improves the SNR of the observations and improves the fits between observations and synthetics, and w_i^c is used to mix and match contributions from different categories. Such a categorization can be based on seismogram component, compressional or shear arrival type, or frequency band.

The vector that we ultimately solve for, $\{\mathbf{M}, \mathbf{x}_s\}$, has nine parameters: six moment-tensor elements and three source-location coordinates. Given an initial wave speed model and an initial CMT solution for an earthquake, we can readily calculate the Fréchet derivatives of the synthetic seismograms, $\partial s_i/\partial M_j$ in each of the 9 model parameters $M_j, j = 1, 2,..., 9$, by performing suites of forward simulations in a perturbed model space, and then, as mentioned before, letting the time-shift vary freely within prior bounds.

3.4 Numerical implementation

We use the GPU accelerated spectral-element modelling package `SPECFEM3D` (Peter *et al.* 2011) for 3-D wave propagation calculations, and the package `PyCMT3D` (Liu *et al.* 2004; Lei 2020), slightly modified to handle our workflow. Both the observed and synthetic data are managed with their various metadata in the SAC format (Helffrich *et al.* 2013) as shown in Fig. 2, with the help of Obspy (Beyreuther *et al.* 2010). The topographic features are honoured by the mesher built into `SPECFEM3D`.

Throughout the inversion we maintain the data-driven data selection criteria espoused in eqs (1) and (2) for all events. Most notably, we verify that the wave speed reference 5.2 km/s shown in Fig. 3 remains applicable for the purpose of determining the SNR and use it as a basis for thresholded data selection, by comparing it with the picked arrivals shown in Figs 3 and 5. As to the model-derived data selection criteria, for every event we strictly follow the selection based on the expressions eqs (3)–(7), while we note that the use of different frequency-bands may lead to different data-selection results (Maggi *et al.* 2009). In particular, the δt_0 threshold value needs to be frequency-dependent to avoid cycle-skipping (Virieux & Operto 2009). As many details of the true 3-D wave speed structure model remain unknowable without conducting explicit wave speed inversions, we confine the CMT inversion to the relatively low frequencies of 1–2 Hz that remain insensitive to unmodelled small-scale heterogeneities.

The version of `PyCMT3D` that we modified to accommodate our workflow is used to perturb, via finite differences, the six focal mechanism parameters and the three location parameters, to evaluate the Fréchet derivatives $\partial s_i/\partial M_j$ of the synthetics with respect to the model parameters. Each Fréchet derivative calculation entails only forward simulations, conducted independently with respect to every one of the nine inversion parameters. A detailed set of accuracy tests for this procedure was presented by Liu *et al.* (2004). Here, we note that the Fréchet derivatives are linear functions of the six moment tensor parameters, while they are nonlinear functions in the hypocentre location parameters. We judiciously choose appropriate finite-difference intervals, especially for the hypocentre location, to ensure that the misfit function remains relatively linear within these intervals, as discussed at length by Liu *et al.* (2004).

The Fréchet derivatives are passed through the same preprocessing steps and share the same bandpass filtering and measurement windows as the synthetic data. We finally minimize the weighted objective function in eq. (8) by applying the calculated Fréchet derivatives. Some examples on the effects of weighting functions can be found in Figs A1–A4 in Appendix A.

4 RESULTS

To validate the performance of the proposed full-waveform CMT inversion workflow, we conducted tests with the data observed at the 252/247 stations shown in Fig. 1(a)—rotated, filtered, instrument-corrected, windowed, selected and weighted as discussed in Section 3, and using the initial 3-D wave speed model shown in Fig. 1(b), and the initial CMT solutions shown in Fig. 1(c). In this section we implement the full workflow pipeline and present the results of our full-waveform CMT inversion for our seismic data set of 86 events. We remind the reader that the two main challenges for our problem arise from the mountainous area with rugged topography differences of up to 3 km, and from the relatively low magnitudes, $1.3 \leq M \leq 3.0$, of the naturally occurring seismic events.

Fig. 7 is a gallery of comparisons made between the initial (blue) and newly inverted (red) CMT solutions, for 6 out of the 86 events. The first column shows the customary "beachball" projection of the initial and final moment-tensor solutions and lists the updates required by minimizing the weighted penalty function (ϕ, eq. 8), that is, the change to the scalar seismic moment, ΔM_{w}, the adjustments to the hypocentre location Δlon, Δlat and Δdep, and the relative change (initial minus final compared to initial) in the unweighted summed misfit criterion, $\Delta(\sum\chi)$, in per cent.

The second through last columns show the distributions of the various misfit criteria before (blue) and after (red) inversion, with the regions of overlap in the mixed colour. The annotations *TS*, *CC*, $d\ln A$ and χ identify the time-shift at maximum cross-correlation (δt, eq. 3), the cross-correlation after time-shift (ξ, eq. 4), the time-shifted energy ratio (η, eq. 5) and the time-shifted waveform-difference misfits (χ, eq. 6), respectively. Means (μ) and standard deviations (σ) of every histogram are listed inside the frames. In the last column, the relative percentage change in the unweighted penalty functions $\Delta(\sum\chi)$ is listed again, repeating the values quoted in the first column.

The benefits of full-waveform CMT inversion for the selection of individual events shown here can be inferred from the changes in those histograms, in particular for those of the waveform energy ratios, $d\ln A$, and the waveform-difference criterion, χ. After inversion, the histograms of the energy ratios tend to be more zero-centred with a smaller standard deviation. The better fit with respect to this particular measure, which is very sensitive to moment-tensor rotation, is also reflected in the waveform-difference misfit statistics. Even small changes in the CMT radiation pattern can result in dramatic misfit reduction on both counts.

The distributions of the time-shifts and the normalized cross-correlation values at those time-shifts stay largely within the bounds set for them at the outset. We elucidate these behaviours with the aid of another figure. We may encounter four types of relationships between observations and initial synthetics on the one hand, and between observations and final synthetics, on the other. Fig. 8(a) shows a situation where the time-shift increases in absolute value (i.e. the agreement worsens) and the normalized cross-correlation decreases (i.e. also worsens). Meanwhile, the logarithmic energy ratio moves closer to zero in absolute value (i.e. improves), and the mean-squared difference between observations and synthetics improves between the initial and the final set of model parameters. Fig. 8(b) illustrates a case in which the time-shift again gets worse, yet the cross-correlation improves, again while improving the amplitude and waveform matches. Fig. 8(c) showcases the scenario where the time-shift decreases in absolute value (i.e. the agreement improves), but the cross-correlation deteriorates, while again the energy ratio and waveform difference metrics improve. Finally, Fig. 8(d) has a case where the time-shift and the cross-correlation measures both show improvement, as do amplitude and waveform differences, although by smaller amounts. These scenarios show that by choosing to drive the CMT inversions by focusing on the reduction of a full-waveform difference measure, we typically achieve a closer agreement in waveform energy as well, while, on the other hand, it is not unusual, in the process, to increase the time-shift between data observations and synthetics, and reduce their normalized correlation after time-shifting.

As a final summary for this section, Fig. 9 renders the distributions of all metrics across all measurements over all 86 events, both prior (blue) to and after (red) inversion. Again, the histograms of energy ratios become markedly better centred and more peaked, and the waveform-differences reduce, for an overall relative $\Delta(\sum\chi)$ of

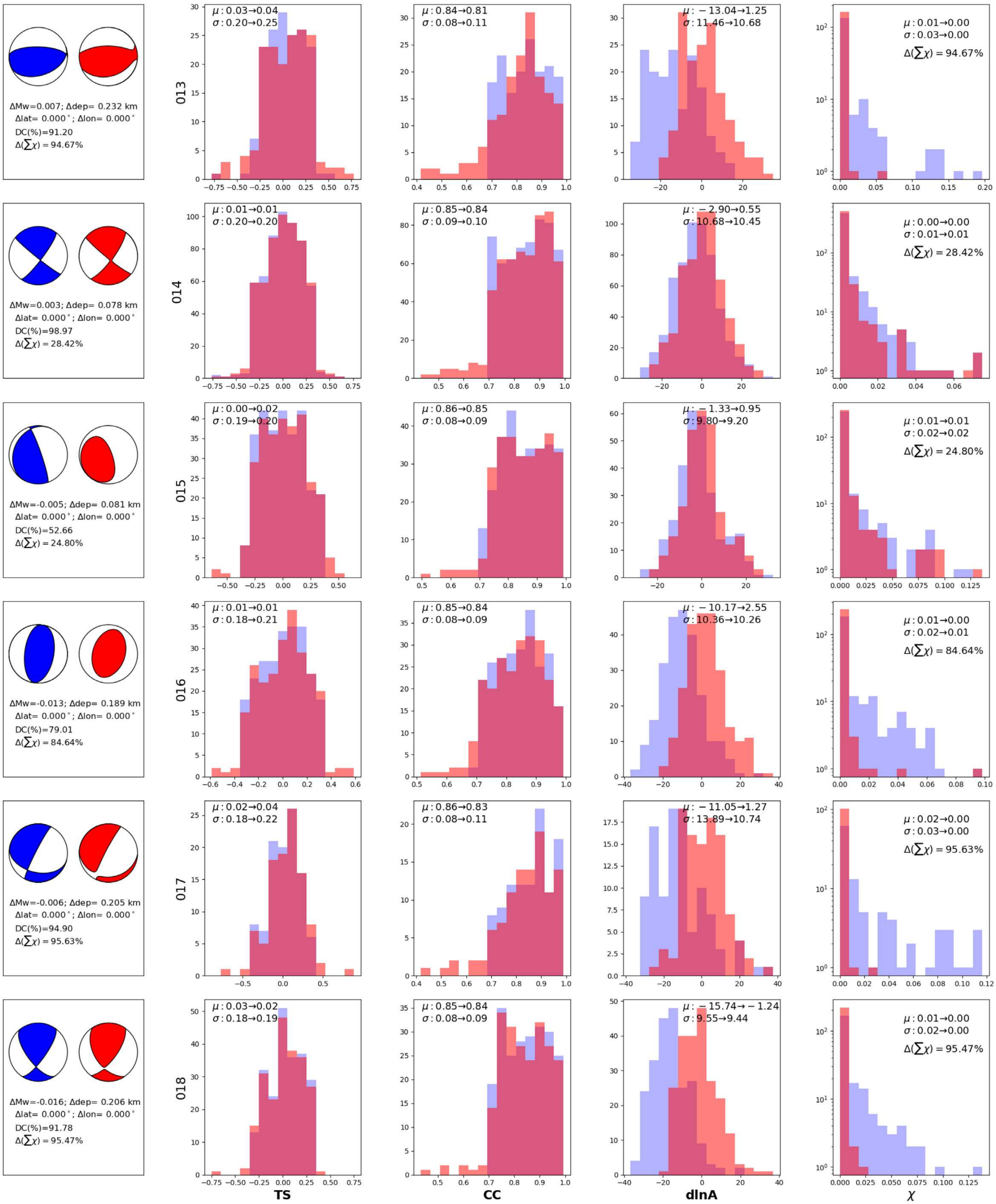


Figure 7. A selection of CMT inversion results. The left column shows initial and final focal mechanisms after inversion, in blue and red, respectively. The remaining columns show the distribution of misfit criteria: time-shift (*TS*), cross-correlation (*CC*), energy ratio ($d\ln A$) and waveform-difference misfit (χ). The model differences in scalar seismic moment and hypocentre coordinates, the double-couple percentages, and the change in overall data misfit are indicated, as are the means (μ) and standard deviations (σ) of the histograms of the misfit metrics. The distributions of the energy ratios and the waveform differences become more centred and narrower after inversion for focal-mechanism updates. Time-shift and cross-correlation values remain largely unchanged, with some leakage as discussed in the text accompanying Fig. 8.

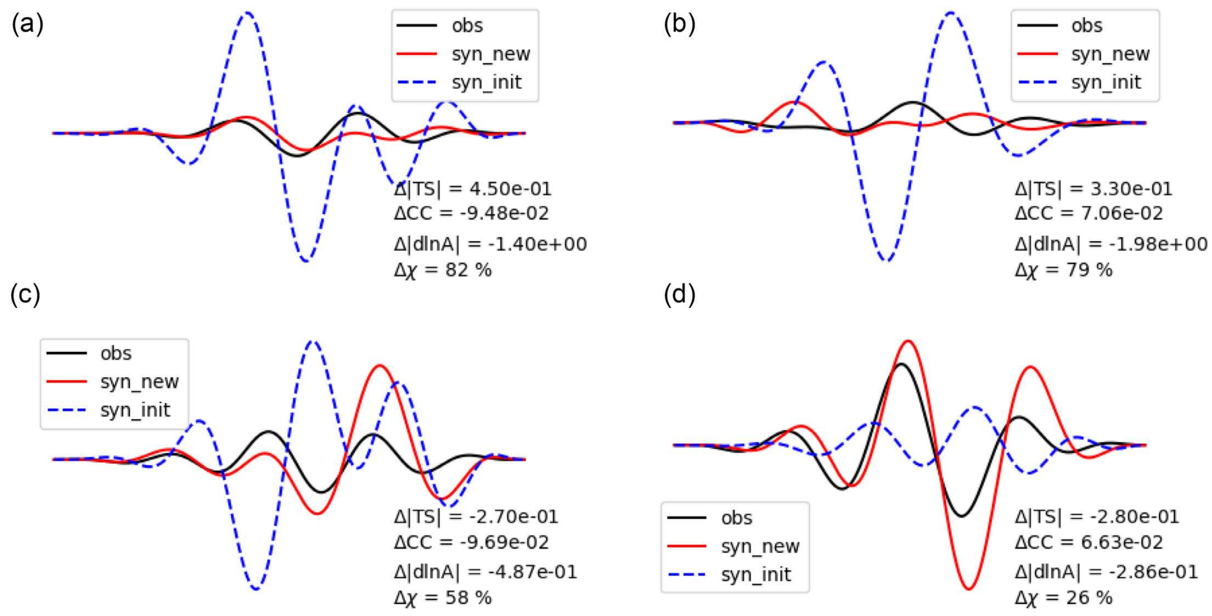


Figure 8. Comparisons between observed (black), initial (dashed blue) and final synthetic (red) waveforms, after time-shifting. Every segment shown is 2 s long. There are four scenarios, in which improving the agreement between the final synthetic and the observation, compared to the agreement between the initial synthetic and the observation, based on a waveform-difference criterion leads to (a) a worse time discrepancy and a diminished cross-correlation agreement; (b) a worse time discrepancy but an improved cross-correlation; (c) an improved time-shift but a diminished cross-correlation; and (d) an improved time fit and also an improved cross-correlation fit. In all four cases the inversion reduces the amplitude ($d\ln A$) and waveform-difference (χ) misfits. The measurements reported in Figs 7 and 9 show the full range of the four illustrated behaviours.

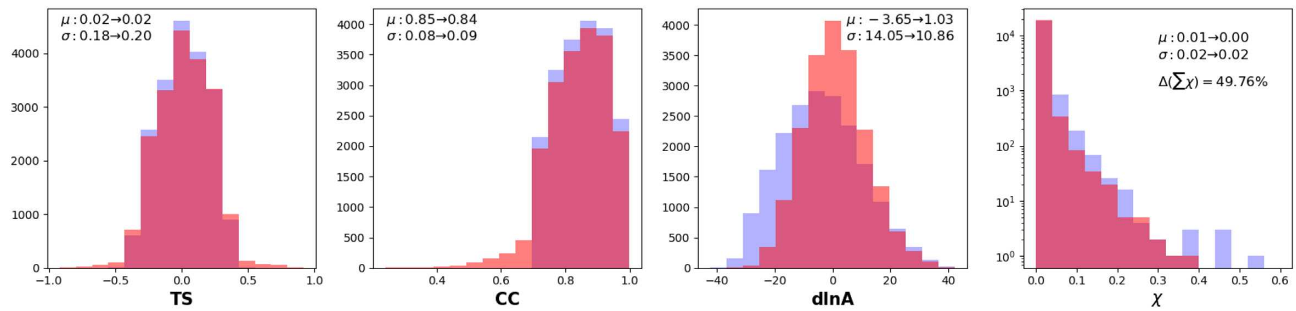


Figure 9. Summary of measures quantifying the quality of the inversion updates across all 86 events. As with columns 2 through 5 in Fig 7, every column shows the distribution of a different metric that captures the agreement between observed and synthetic traces: blue for those under the initial model and red for those in the final model, after inversion using the waveform-difference metric of eq (7). Energy-ratio ($d\ln A$) and waveform-difference (χ) most clearly show the improvement to the data fit by updating the CMT solutions. Traveltime shift (*TS*) and cross-correlation (*CC*) are not meaningfully affected, although there are trade-offs, as discussed in Fig. 8. Supporting Information Fig. S1 shows the joint behaviour of these various metrics as cross-plots.

49.76 per cent. Traveltime shifts remain virtually untouched from the initial set ±0.35 s, and the cross-correlation values largely stay within their initial quality-control bounds of 0.7. Supporting Information Fig. S1 renders the pairwise cross-plots between all metrics.

5 DISCUSSION

Fig. 10 shows the 86 initial (blue) and final (red) focal mechanisms as discussed in the previous section, but now in geographical context, and with the addition of known fault traces from the studies by Gągała *et al.* (2020), Abdulhameed *et al.* (2020) and Dedow *et al.* (2020). Note that Fig. 10(a) simply repeats the information formerly presented in Fig. 1.

Fig. 11 shows the distributions of the relationships between the initial and final 86 CMT solutions, focusing, in the first two columns, on the shifts in the moment magnitude ΔM_w and in hypocentre depth Δdep. Five outliers with $\Delta M_w > 0.25$ were omitted from the histogram, which corresponded to events located in the mountains close to the boundaries of the area of interest and for which the data coverage ultimately provided to be too poor to yield reasonable solutions. We do not show the horizontal relocation amounts, which were all very small compared to the depth changes. The last two columns list the relative change in waveform difference misfit $\Delta(\sum\chi)$, and the percentage of the double-couple contribution to the final solution, DC (per cent), obtained after tensor decomposition into isotropic, double-couple, and compensated linear vector dipole (CLVD) components (Jost & Herrmann 1989; Vavryčuk & Adamová 2020). No double-couple constraints were applied during the inversion (but see Appendix B). High fractional double-couple contributions are expected for tectonic earthquakes in this setting, and the large percentage with more than 50 per cent double-couple mechanism is highlighted in red in Fig. 10. See Supporting Information Fig. S2 for an alternative parametrization of this discrepancy, following Tape & Tape (2013).

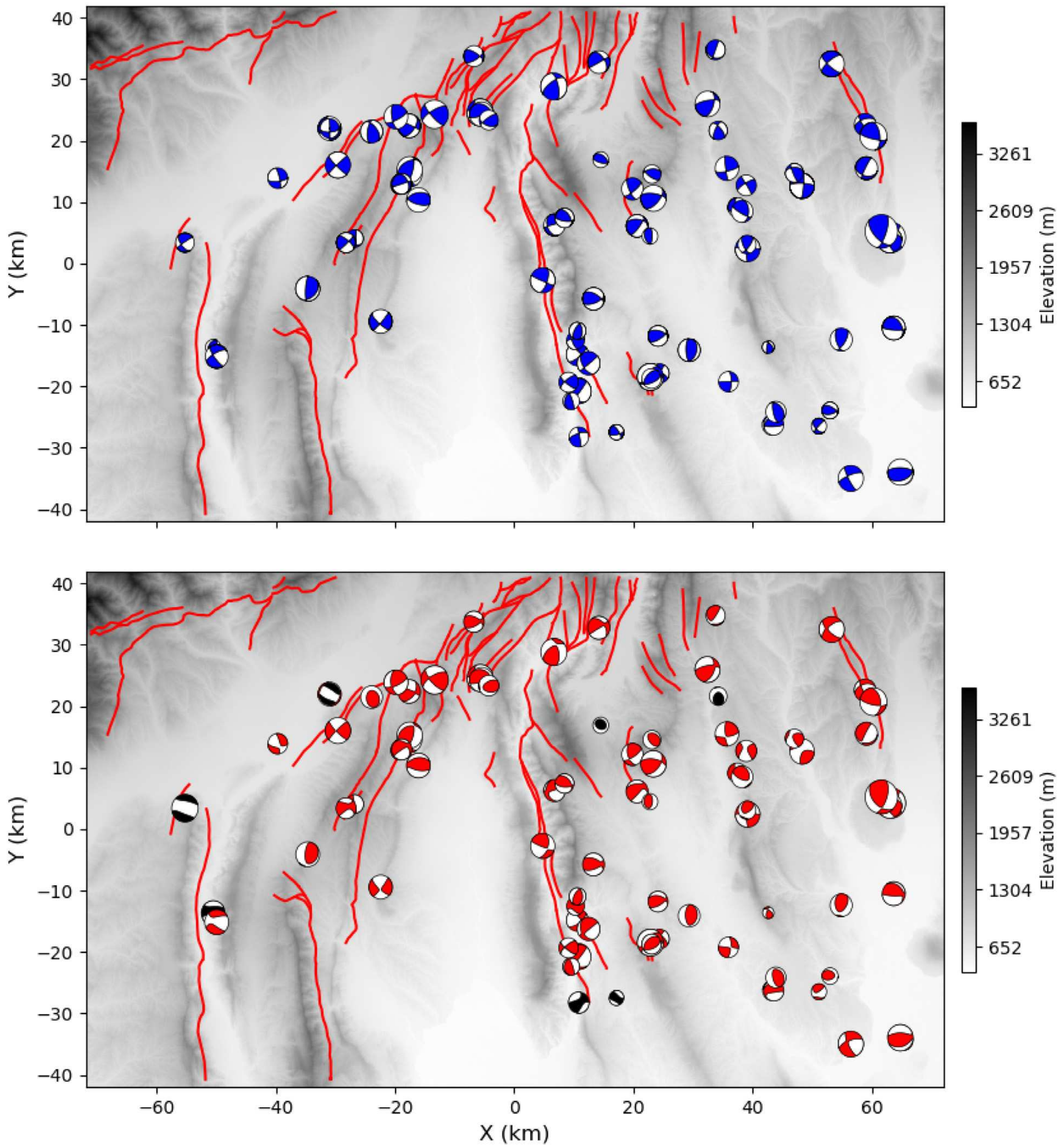


Figure 10. Initial (a, blue) and final (b, red) CMT solutions obtained in this study. The size of the beach balls is proportional to the event magnitudes. The solutions in (a) were previously shown in Fig. 1. In (b), focal mechanisms with double-couple components exceeding 50 per cent and magnitude changes below 0.25 are in red. The dominance of double-couple events is interpreted as being in line with their tectonic origin. Fault traces (Gągała *et al.* 2020) in red.

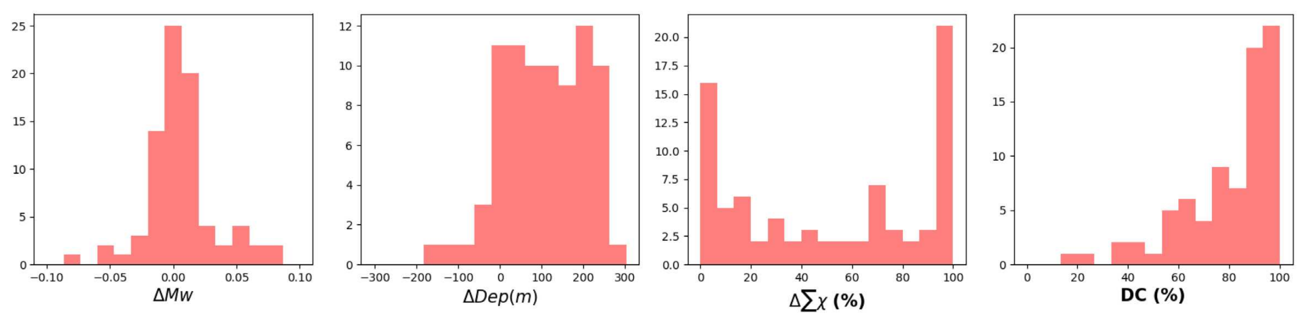


Figure 11. Distribution of the differences between the initial and the final 86 CMT solutions after our inversion. Shown are ΔM_{w}, the change in moment magnitude, and Δdep, the change in hypocentre depth. Also shown is $\Delta(\sum\chi)$ the relative change to the waveform-difference misfit metric, and *DC* (per cent), the double-couple percentage. The most robust solutions will display high values in those last two categories.

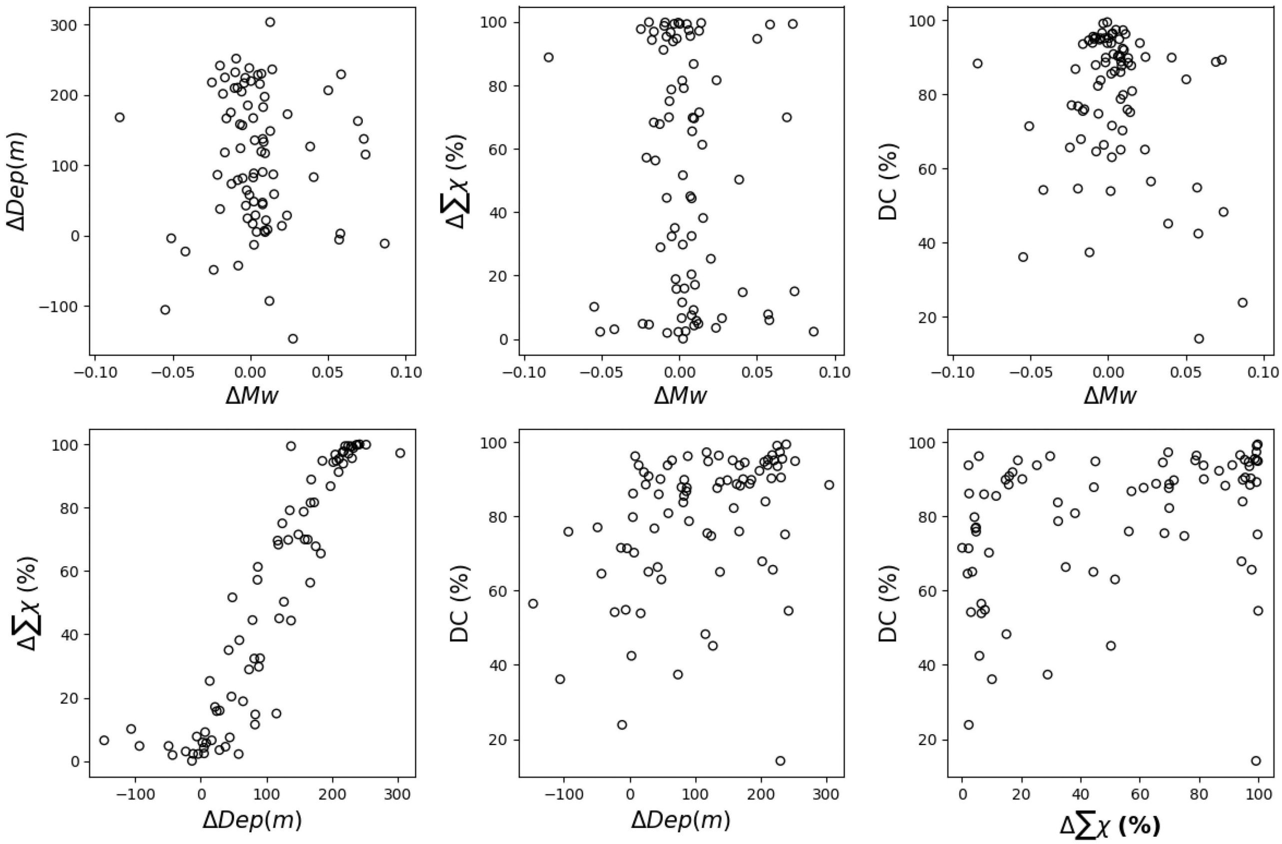


Figure 12. Cross-plot scatter diagrams for the four metrics shown in Fig. 11. These six subfigures are slices through the 4-D space spanned by ΔM_w, Δdep, $\Delta(\sum\chi)$ and *DC* (per cent). We conclude that $\Delta(\sum\chi)$ can be updated significantly without big changes in ΔM_w, and that, while the final solutions present significant $\Delta(\sum\chi)$ updates, most of the CMT solutions remain dominated by double-couple mechanisms, in agreement with their tectonic origin.

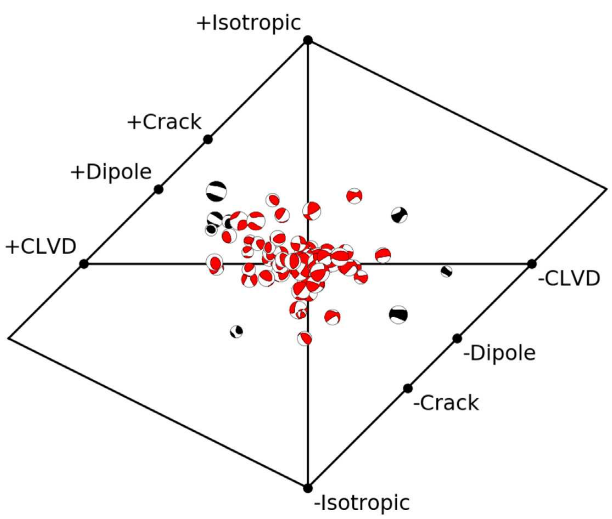


Figure 13. Hudson plot of the inverted CMT solutions. The colours and relative sizes of the beach balls are consistent with those rendered in Fig. 10.

Fig. 12 shows how the changes in parameters between the initial and final CMT solutions interrelate. The individual cross-plots are sections through the 4-D space spanned by ΔM_w, Δdep, $\Delta(\sum\chi)$ and DC (per cent). Most of the updated events moved to greater depths, without clear relation to the change in moment magnitude, whereas there is a positive relationship between the depth change and the improvement in waveform-difference misfit. Fig. 5 contained a clue to this behaviour for the event shown: many of the seismic arrivals appear to scatter behind the *P*-wave arrival picks, hence the need to move the sources deeper into the Earth to obtain better matches to the waveforms. The adjustments in the waveform-difference quality criterion required by the updates to the focal mechanism were large, while the changes in moment magnitude were, ultimately, small. Large, welcome, updates to the waveform fits notwithstanding, most of the final CMT solutions remain dominated by double-couple modes, as revealed by the Hudson plot in Fig. 13, despite the absence of inversion constraints, which is fitting for the tectonic background of these natural events.

We remind the reader that we performed a full-waveform CMT inversion for just 86 events in the magnitude range $1.3 \leq M \leq 3.0$, a small subset the total of 2607 events that were identified in the data set. Our subset of reinverted 86 CMT solutions is sparsely but relatively homogeneously distributed throughout the region. It is therefore implied that we might use these new mechanisms as initial solutions for neighbouring event clusters, provided sufficient computational resources (on our 16 GPU system, every `SPECFEM3D` simulation consumed about 2 min). In that sense our study is but an initial exploration of this rich passive data set obtained in this inhospitable mountainous region. Furthermore, our investigation has focused on a rather limited bandwidth of the data, analyses in short window lengths of two specific wave types, *P* and *S* waves, and without extensive exploration of possible changes in the source-time function. Our next step will be to conduct a full-waveform tomography on the basis of these new moment-tensor solutions, after which we should be able to extend the frequency bandwidth and time length of the data available for inversion, which will include more wave types, including surface waves, which can then be fed back into a next-round of full-waveform CMT inversion, possibly jointly with another wave speed inversion. We have sidestepped the role of wave speed errors by dealing with the waveform minimization of time-lagged waveform misfits. Future improvements of the wave speed model will lead to new 3-D wave propagation solutions that will ultimately weaken the role of time lags in CMT inversion of the kind conducted in this paper.

We judge the development and application of waveform-inversion methods to be a worthwhile step towards a future where the analysis of passive seismic events may routinely complement and possibly even replace active methods of exploration in such challenging terrains. Porting our automated data-processing procedures to a new region of focus will require access to a high-quality data set (automatically picked and/or hand-reviewed), a reasonable starting model of velocity structure and hypocentre locations (e.g. from a joint inversion), and a numerical mesh that appropriately captures the 3-D nature of the modelling domain (including topography).

In Appendix A, we show detailed examples of the various choices of weighting functions and their effects on the waveform matches across all three recorded components.

In Appendix B, we present CMT inversion results conducted by applying a double-couple constraint on the focal mechanism, to compare with the results presented in the main text, where we did not apply any such constraints.

6 CONCLUSION

We have presented a workflow for the regional-scale full-waveform CMT inversion of natural seismic events, and applied it to a passive seismic exploration project conducted using a high-density broad-band array operating for multiple months in a mountainous setting. The rugged topography and general weakness of the seismic events yielded data of variable SNRs, which required careful data selection involving both data- and model-driven approaches. Wavefield simulations were carried out in a 3-D tomographic initial wave speed model determined from hand-reviewed automatic *P*-wave traveltime measurements, using a 3-D Cartesian SEM on a high-resolution mesh that honours the topography. We determined updates to the initial set of CMT source parameters by minimizing the waveform misfits between the observed and synthetic data, using a global penalty function with weights intended to balance data coverage. The flexibility of our approach allows for the incorporation of information from many stations within a complex wave speed model including topography, without pre-calculating or storing Green's functions. We presented and discussed our results in comparison to the information based on more standard processing methods that was available to us from the outset, and showed our CMT mechanisms in their tectonic context.

DATA AND SOFTWARE AVAILABILITY

Our package `X-PyCMT3D` is available from GitHub at https://github.com/qcliu0/X-PyCMT3D. We accompany this paper by a complete list of our solutions and their uncertainties, such that they may be available for further study and interpretation by researchers interested in the particular geographic area of the Himalayan foothills that formed the focus of our exploration campaign. The Global CMT catalog is at https://www.globalcmt.org.

ACKNOWLEDGMENTS

We thank TotalEnergies SE for support and sharing the data, and SeismoTech Ltd for providing the initial CMT solutions and 3-D elastic models. We thank Lothar Ratschbacher for sharing the fault database. Computational support and resources were provided by the Princeton Institute for Computational Science and Engineering. We thank Wenjie Lei, Lucas Sawade and Jeroen Tromp for fruitful discussions. The insightful and constructive comments from associate editor Carl Tape, and from two anonymous reviewers, are gratefully acknowledged.

REFERENCES

Abdulhameed, S. *et al.*, 2020. Tajik basin and southwestern Tian Shan, northwestern India–Asia collision zone: 2. Timing of basin inversion, Tian Shan mountain building, and relation to Pamir-plateau advance and deep India–Asia indentation, *Tectonics,* **39**(5), e2019TC005873 , doi:10.1029/2019TC005873.

Aki, K. & Lee, W.H.K., 1976. Determination of three-dimensional velocity anomalies under a seismic array using first *P* arrival times from local earthquakes: 1. A homogeneous initial model, *J. geophys. Res.,* **81**(23), 4381–4399.

Beydoun, W.B. & Keho, T.H., 1987. The paraxial ray method, *Geophysics,* **52**(12), 1639–1653.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y. & Wassermann, J., 2010. Obspy: a Python toolbox for seismology, *Seismol. Res. Lett.,* **81**(3), 530–533.

Borisov, D., Modrak, R., Gao, F. & Tromp, J., 2018. 3D elastic full-waveform inversion of surface waves in the presence of irregular topography using an envelope-based misfit function, *Geophysics,* **83**(1), R1–R11.

Bozdağ, E., Peter, D., Lefebvre, M., Komatitsch, D., Tromp, J., Hill, J., Podhorszki, N. & Pugmire, D., 2016. Global adjoint tomography: first-generation model, *Geophys. J. Int.,* **207**(3), 1739–1766.

Convertito, V. & Herrero, A., 2004. Influence of focal mechanism in probabilistic seismic hazard analysis, *Bull. Seism. Soc. Am.,* **94**(6), 2124–2136.

Courant, R., Friedrichs, K. & Lewy, H., 1967. On the partial difference equations of mathematical physics, *IBM J. Res. Dev.,* **11**(2), 215–234.

Crosson, R.S., 1976. Crustal structure modeling of earthquake data: 1. Simultaneous least squares estimation of hypocenter and velocity parameters, *J. geophys. Res.,* **81**(17), 3036–3046.

Dedow, R. *et al.*, 2020. Tajik basin and southwestern Tian Shan, northwestern India–Asia collision zone: 3. Preorogenic to synorogenic retro-foreland basin evolution in the eastern Tajik depression and linkage to the Pamir hinterland, *Tectonics,* **39**(5), e2019TC005874 , doi:10.1029/2019TC005874.

Efron, B., 1979. Bootstrap methods: another look at the jackknife, *Ann. Stat.,* **7**(1), 1–26.

Ekström, G., Nettles, M. & Dziewonski, A.M., 2012. The global CMT project 2004–2010: centroid-moment tensors for 13,017 earthquakes, *Phys. Earth planet. Inter.,* **200–201,** 1–9.

Eyre, T.S. & van der Baan, M., 2017. The reliability of microseismic moment-tensor solutions: surface versus borehole monitoring, *Geophysics,* **82**(6), KS113–KS125.

Fichtner, A. & Simutė, S., 2018. Hamiltonian Monte Carlo inversion of seismic sources in complex media, *J. geophys. Res.,* **123**(4), 2984–2999.

Fichtner, A., Bunge, H.-P. & Igel, H., 2006. The adjoint method in seismology—I. Theory, *Phys. Earth planet. Inter.,* **157**(1–2), 86–104.

Gągała, Ł. *et al.*, 2020. Tajik basin and southwestern Tian Shan, northwestern India–Asia collision zone: 1. Structure, kinematics, and salt tectonics in the Tajik fold-and-thrust belt of the western foreland of the Pamir, *Tectonics,* **39**(5), e2019TC005871 , doi:10.1029/2019TC005871.

Gardner, G., Gardner, L. & Gregory, A., 1974. Formation velocity and density—the diagnostic basics for stratigraphic traps, *Geophysics,* **39**(6), 770–780.

Hejrani, B. & Tkalčić, H., 2020. Resolvability of the centroid-moment-tensors for shallow seismic sources and improvements from modeling high-frequency waveforms, *J. geophys. Res.,* **125**(7), e2020JB019643 , doi:10.1029/2020JB019643.

Helffrich, G., Wookey, J. & Bastow, I., 2013. *The Seismic Analysis Code. A Primer and User's Guide,* Cambridge Univ. Press.

Hingee, M., Tkalčić, H., Fichtner, A. & Sambridge, M., 2011. Seismic moment tensor inversion using a 3-D structural model: applications for the Australian region, *Geophys. J. Int.,* **184**(2), 949–964.

Jost, M.L. & Herrmann, R.B., 1989. A student's guide to and review of moment tensors, *Seismol. Res. Lett.,* **60**(2), 37–57.

Käufl, P., Valentine, A.P., O'Toole, T.B. & Trampert, J., 2014. A framework for fast probabilistic centroid-moment-tensor determination-inversion of regional static displacement measurements, *Geophys. J. Int.,* **196**(3), 1676–1693.

Käufl, P., Valentine, A., de Wit, R. & Trampert, J., 2015. Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition, *Bull. seism. Soc. Am.,* **105**(4), 2299–2312.

Kim, Y., Liu, Q. & Tromp, J., 2011. Adjoint centroid-moment tensor inversions, *Geophys. J. Int.,* **186**(1), 264–278.

Komatitsch, D. & Tromp, J., 1999. Introduction to the spectral element method for three-dimensional seismic wave propagation, *Geophys. J. Int.,* **139**(3), 806–822.

Komatitsch, D., Liu, Q., Tromp, J., Suss, P., Stidham, C. & Shaw, J.H., 2004. Simulations of ground motion in the Los Angeles basin based upon the spectral-element method, *Bull. seism. Soc. Am.,* **94**(1), 187–206.

Kufner, S.-K., Schurr, B., Ratschbacher, L., Murodkulov, S., Abdulhameed, S., Ischuk, A., Metzger, S. & Kakar, N., 2018. Seismotectonics of the Tajik Basin and surrounding mountain ranges, *Tectonics,* **37**(8), 2404–2424.

Lee, W. H.K. & Lahr, J.C., 1972. *HYPO71: A Computer Program for Determining Hypocenter, Magnitude, and First Motion Pattern of Local Earthquakes,* US Department of the Interior, Geological Survey National Center.

Lei, W., 2020. Global Seismic Full-Waveform Inversion, *PhD thesis*, Princeton University.

Lei, W. *et al.*, 2020. Global adjoint tomography—model GLAD-M25, *Geophys. J. Int.,* **223**(1), 1–21.

Li, J., Kuleli, H.S., Zhang, H. & Toksöz, M.N., 2011. Focal mechanism determination of induced microearthquakes in an oil field using full waveforms from shallow and deep seismic networks, *Geophysics,* **76**(6), WC87–WC101.

Liu, Q. & Gu, Y.J., 2012. Seismic imaging: from classical to adjoint tomography, *Tectonophysics,* **566–567,** 31–66.

Liu, Q., Polet, J., Komatitsch, D. & Tromp, J., 2004. Spectral-element moment tensor inversions for earthquakes in southern California, *Bull. seism. Soc. Am.,* **94**(5), 1748–1761.

Lois, A., Sokos, E., Martakis, N., Paraskevopoulos, P. & Tselentis, G.-A., 2013. A new automatic *S*-onset detection technique: application in local earthquake data, *Geophysics,* **78**(1), KS1–KS11.

Maggi, A., Tape, C., Chen, M., Chao, D. & Tromp, J., 2009. An automated time-window selection algorithm for seismic tomography, *Geophys. J. Int.,* **178**(1), 257–281.

McNamara, D.E. & Buland, R.P., 2004. Ambient noise levels in the continental United States, *Bull. seism. Soc. Am.,* **94**(4), 1517–1527.

Nakata, N., Gualtieri, L. & Fichtner, A., 2019. *Seismic Ambient Noise,* Cambridge Univ. Press.

Peter, D. *et al.*, 2011. Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes, *Geophys. J. Int.,* **186**(2), 721–739.

Reasenberg, P.A. & Oppenheimer, D.H., 1985. FPFIT, FPPLOT, and FPPAGE: Fortran computer programs for calculating and displaying earthquake fault-plane solutions, Open-file report 85–739, U.S. Dept.Interior, Geological Survey.

Ruan, Y., Lei, W., Modrak, R., Örsvuran, R., Bozdağ, E. & Tromp, J., 2019. Balancing unevenly distributed data in seismic tomography: a global adjoint tomography example, *Geophys. J. Int.,* **219**(2), 1225–1236.

Saltzer, R., Leahey, G., Schmedes, J., Roth, J. & Rumpfhuber, E., 2011. Earthquakes—a naturally occurring source of low-frequency data, in *SEG Technical Program Expanded Abstracts,* pp. 3689–3692, SEG, Denver, CO.

Saragiotis, C.D., Hadjileontiadis, L.J. & Panas, S.M., 2002. PAI-S/K: a robust automatic seismic *P* phase arrival identification scheme, *IEEE Trans. Geosci. Remote Sens.,* **40**(6), 1395–1404.

Schurr, B., Ratschbacher, L., Sippl, C., Gloaguen, R., Yuan, X. & Mechie, J., 2014. Seismotectonics of the Pamir, *Tectonics,* **33**(8), 1501–1518.

Sokos, E.N. & Zahradnik, J., 2008. ISOLA: a Fortran code and a Matlab GUI to perform multiple-point source inversion of seismic data, *Comput. Geosci.,* **34**(8), 967–977.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2009. Adjoint tomography of the southern California crust, *Science,* **325**(5943), 988–992.

Tape, W. & Tape, C., 2013. The classical model for moment tensors, *Geophys. J. Int.,* **195**(3), 1701–1720.

Thurber, C.H., 1983. Earthquake locations and three-dimensional crustal structure in the Coyote Lake area, central California, *J. geophys. Res.,* **88**(B10), 8226–8236.

Tichelaar, B.W. & Ruff, L.J., 1989. How good are our best models? Jackknifing, bootstrapping, and earthquake depth, *EOS, Trans. Am. geophys. Un.,* **70**(20), 593 & 605–606.

Triantafyllis, N., Venetis, I.E., Fountoulakis, I., Pikoulis, E., Sokos, E. & Evangelidis, C.P., 2021. Gisola: A high-performance computing application for real-time moment tensor inversion, *Seismol. Res. Lett.,* **93,**957–966.

Trnkoczy, A., 2012. Understanding and parameter setting of STA/LTA trigger algorithm, in *New Manual of Seismological Observatory Practice (NMSOP),* chap. IS 8.1, pp. 1–20, ed. Bormann, P., GeoForschungsZentrum.

Tromp, J., 2020. Seismic wavefield imaging of Earth's interior across scales, *Nat. Rev. Earth Environ.,* **1**(1), 40–53.

Tselentis, G.-A., Martakis, N., Paraskevopoulos, P. & Lois, A., 2011. High-resolution passive seismic tomography for 3D velocity, Poisson's ratio ν, and *P*-wave quality Q_P in the Delvina hydrocarbon field, southern Albania, *Geophysics,* **76**(3), B89–B112.

Um, J. & Thurber, C., 1987. A fast algorithm for two-point seismic ray tracing, *Bull. seism. Soc. Am.,* **77**(3), 972–986.

Vackář, J., Burjánek, J., Gallovič, F., Zahradník, J. & Clinton, J., 2017. Bayesian ISOLA: new tool for automated centroid moment tensor inversion, *Geophys. J. Int.,* **210**(2), 693–705.

Vavryčuk, V. & Adamová, P., 2020. Non-double-couple moment tensors of earthquakes calculated using empirical Green's functions, *Seismol. Res. Lett.,* **91**(1), 390–398.

Vicente, G.D., Cloetingh, S. A. P.L., Muñoz-Martín, A., Olaiz, A., Stich, D., Vegas, R., Galindo-Zaldívar, J. & Fernández-Lozano, J., 2008. Inversion of moment tensor focal mechanisms for active stresses around the microcontinent Iberia: tectonic implications, *Tectonics,* **27**(1), TC1009.

Virieux, J. & Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics, *Geophysics,* **74**(6), WCC1–WCC26.

Willacy, C., van Dedem, E., Minisini, S., Li, J., Blokland, J.-W., Das, I. & Droujinine, A., 2019. Full-waveform event location and moment tensor inversion for induced seismicity, *Geophysics,* **84**(2), KS39–KS57.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S. & Trujillo, J., 1998. A comparison of select trigger algorithms for automated global seismic phase and event detection, *Bull. seism. Soc. Am.,* **88**(1), 95–106.

Zhao, L.-S. & Helmberger, D.V., 1994. Source estimation from broadband regional seismograms, *Bull. seism. Soc. Am.,* **84**(1), 91–104.

Zhu, H., Bozdağ, E., Peter, D. & Tromp, J., 2012. Structure of the European upper mantle revealed by adjoint tomography, *Nat. Geosci.,* **5**(7), 493–498.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Figure S1. Cross-plot pairs between the metrics shown in Fig. 7, demonstrating their joint behaviour over the whole set of inverted traces. Details along the diagonals are shown in Fig. 9. As discussed with Fig. 8, the $d\ln A$ and $\ln\chi$ metrics may improve at the expense of the other metrics.

Figure S2. Relation between the double-couple percentages of the 86 events as quoted in the text, and the parametrization θ (Tape & Tape 2013).

APPENDIX A: WEIGHTING FUNCTIONS AND WAVEFORM-DIFFERENCE OUTLIER REJECTION

In this section we highlight the role of the weighting functions in eqs (8)–(10) applied to the penalty function in eq. (7), and comment on the initial rejection of outliers in the waveform-difference parameter χ defined in eq. (6). For the sake of simplicity, among those listed in eq. (10) we only illustrate the weights that involve azimuth (w^a) and epicentral distance (w^d).

Fig. A1 shows the case with neither w^a and w^d applied in the inversion. Fig. A2 uses only w^d, Fig. A3 only w^a, and Fig. A4 uses both, which is the preferred way, via which we obtained the results presented in the main text.

Fig. A5 shows the case that contains outliers in χ, while Fig. A6 shows the results obtained after excluding them. Outlier rejection effectively avoids their dominance in the least-squares inversion. In the inversion for most of the 86 CMT events, χ outlier rejection was a relatively inconsequential option, but to confidently tackle challenging data sets, it is recommended as a default protective step.

APPENDIX B: CENTROID MOMENT TENSOR INVERSION UNDER DOUBLE-COUPLE CONSTRAINT

In this section we discuss the difference between Centroid Moment Tensor inversions conducted with and without double-couple constraint. Figs B1 and B2 show output summaries of inversions for event CMT015 without double-couple constraint, and with double-couple constraint, respectively. Fig. B3 shows a selection of CMT results, similar to Fig. 7, but all of them inverted with the double-couple constraint. To further investigate the impact of double-couple constraint on the inversions, we estimate the uncertainties of the inverted CMTs, with and without constraints, by bootstrap resampling (Efron 1979; Tichelaar & Ruff 1989) over the data set. We randomly draw a percentage of the input data. The inverted CMTs are fit from the drawn samples and evaluated over the out-of-sample set. Repeating such procedures many more times returns means and standard deviations, allowing the uncertainty assessment of the inverted CMTs. Compared with those in Fig. 7, the distributions of the metrics look similar, but the inversion results have larger uncertainties, as revealed by Fig. B2.

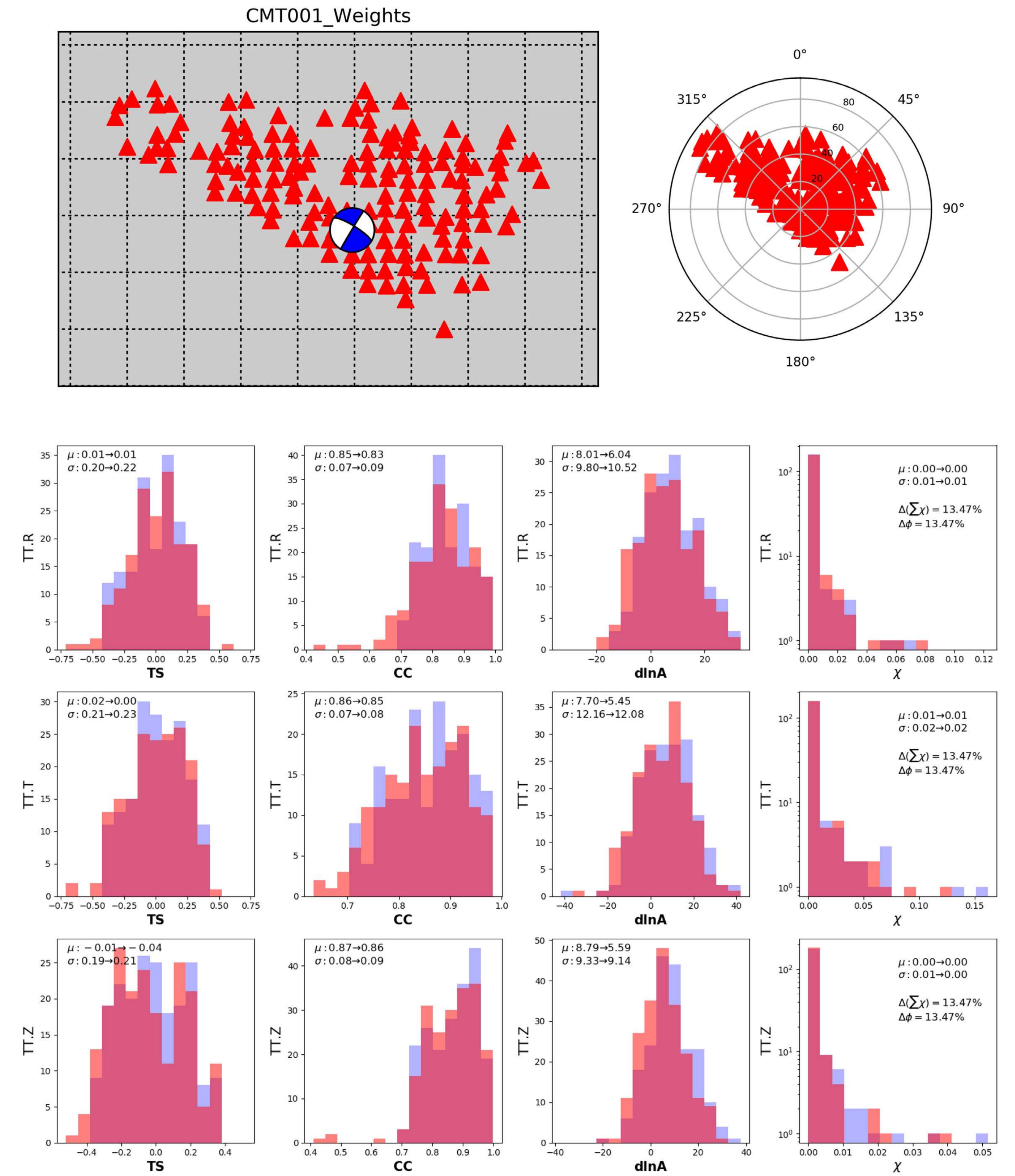


Figure A1. Waveform inversion for a CMT solution, without data weighting. (a) Source-receiver geometry in Cartesian and polar coordinates, respectively. (b) Distribution of misfit metrics between the observed data and the initial (blue) and final synthetics (red). Every row corresponds to a different seismogram component radial (R), transverse (T) and vertical (Z), respectively. Note that 'TT' denotes our network code. Every column corresponds to a misfit metric, labelled *TS* (time-shift), *CC* (cross-correlation), $d\ln A$ (energy ratio) and χ (waveform difference), as in Figs 7 and 9. Both scales on the graph with the histograms for χ are logarithmic scale. Means and standard deviations of the histograms are inset, as are the relative change in the unweighted sum of waveform-difference metrics, $\sum\chi$, and that of the weighted version, ϕ. Note that, without weighting, $\Delta(\sum\chi) = \Delta\phi$.

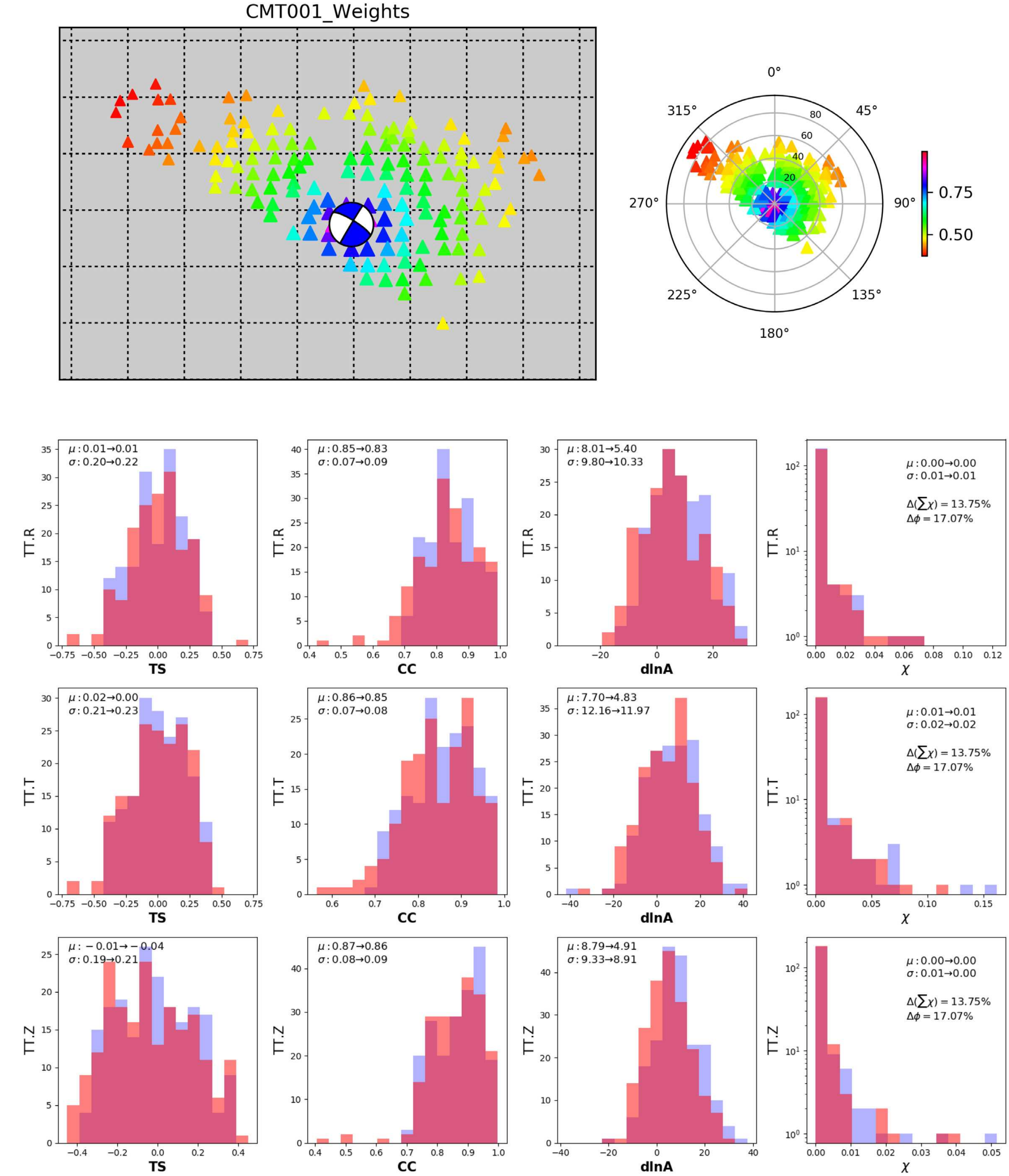


Figure A2. Waveform inversion for a CMT solution, with epicentral-distance weighting. (a) Source-receiver geometry in Cartesian and polar coordinates with weighting values. The values decrease with epicentral distance because we have more confidence for the near-offset data regarding data quality and data match. (b) Distribution of misfit metrics between the observed data and the initial (blue) and final synthetics (red), as in Fig. A1. The metrics μ and σ for $d\ln A$, and $\Delta(\sum\chi)$ for χ reveal that the results in this figure outperform those in Fig. A1 owing to the epicentral-distance weighting in the penalty function.

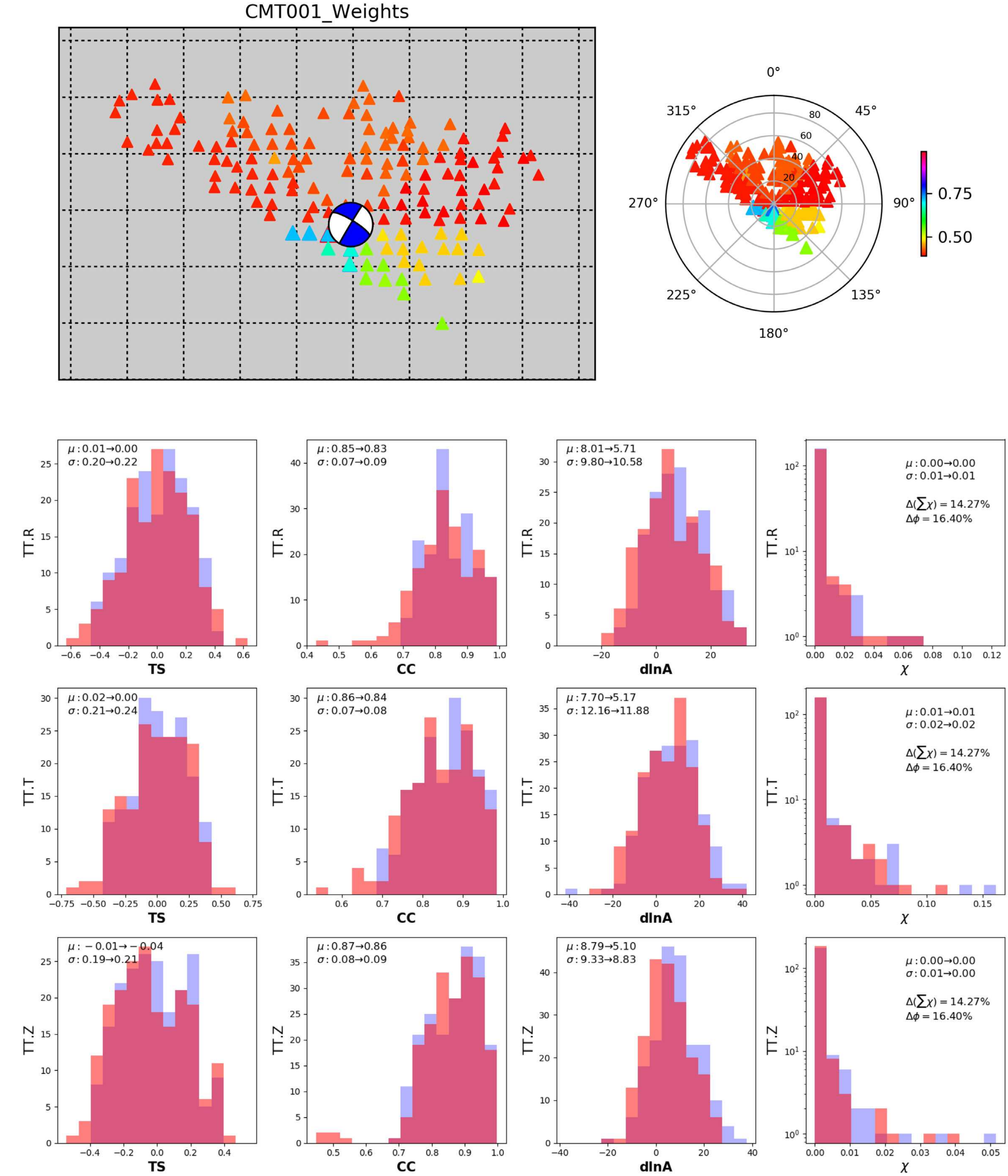


Figure A3. Waveform inversion for a CMT solution, with azimuthal weighting. As shown in the polar plot in (a), we divide the azimuth domain into eight equal-sized subquadrants. Each of them has the same contribution to the misfit function. The weightings are inversely proportional to the station number within quadrants, so as to balance the azimuthal coverage. (b) Distribution of misfit metrics between the observed data and the initial (blue) and final synthetics (red), as in Fig. A1. As in Fig. (A2), μ and σ for $d\ln A$, and $\Delta(\sum\chi)$ for χ reveal that the results in this figure outperform those in Fig. A1 owing to the azimuthal weighting in the penalty function.

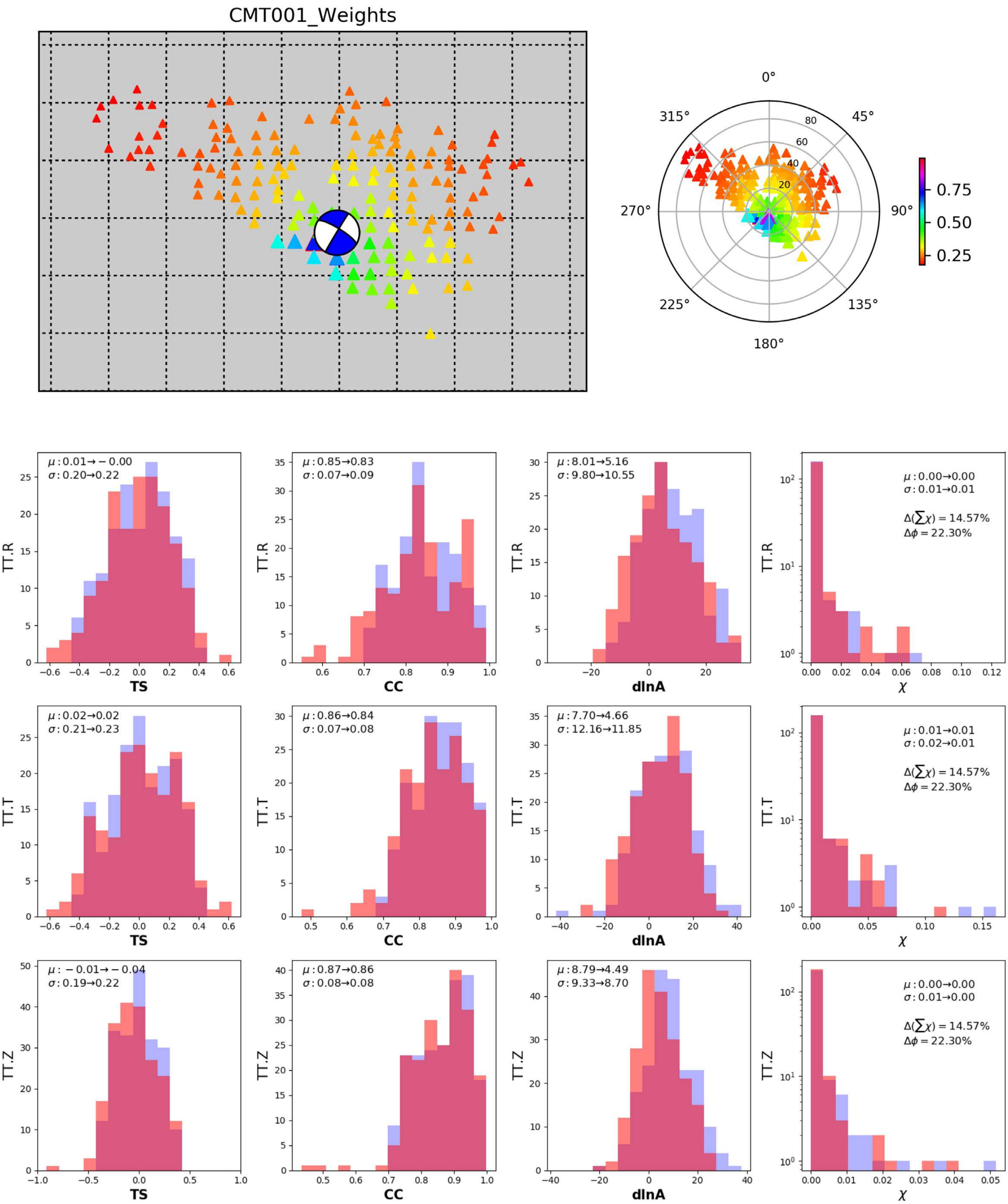
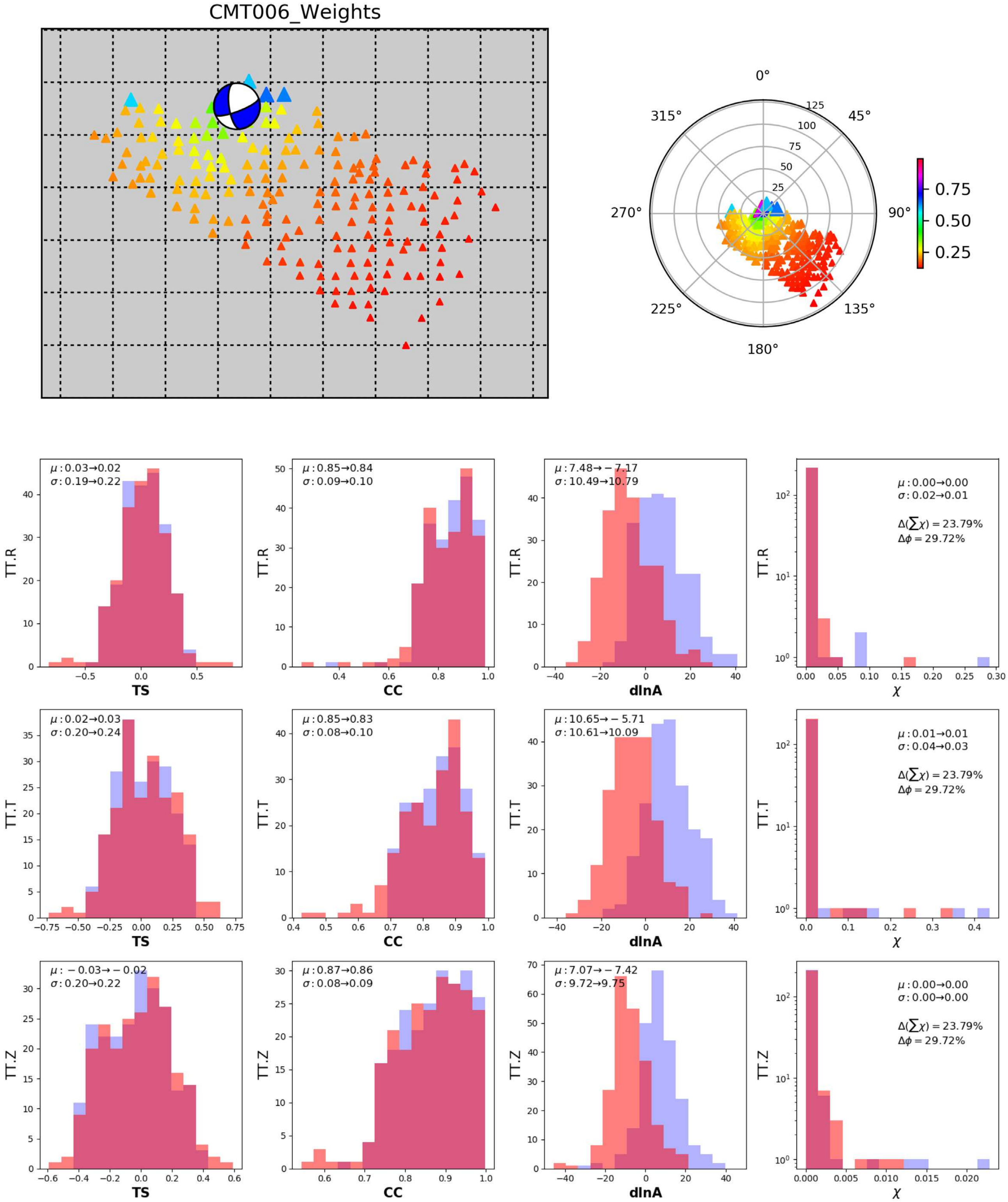


Figure A4. Waveform inversion for a CMT solution, with both azimuthal and epicentral-distance weighting. (a) Weights in Cartesian and polar coordinates, respectively. (b) Distribution of misfit metrics between the observed data and the initial (blue) and final synthetics (red), as in Fig. A1. The combination of weighting strategies illustrates how Fig. A4 outperforms both Figs A1–A3 in terms of $d\ln A$ and χ.



s

Figure A5. Waveform inversion for a CMT solution, without waveform-difference outlier rejection. Shown in (a) are weights identical to those in Fig. A4. (b) Distribution of misfit metrics between the observed data and the initial (blue) and final synthetics (red), as in Fig. A1. Although with significant data-misfit reductions, the histograms of *d*ln*A* are pushed away from being zero-centred.

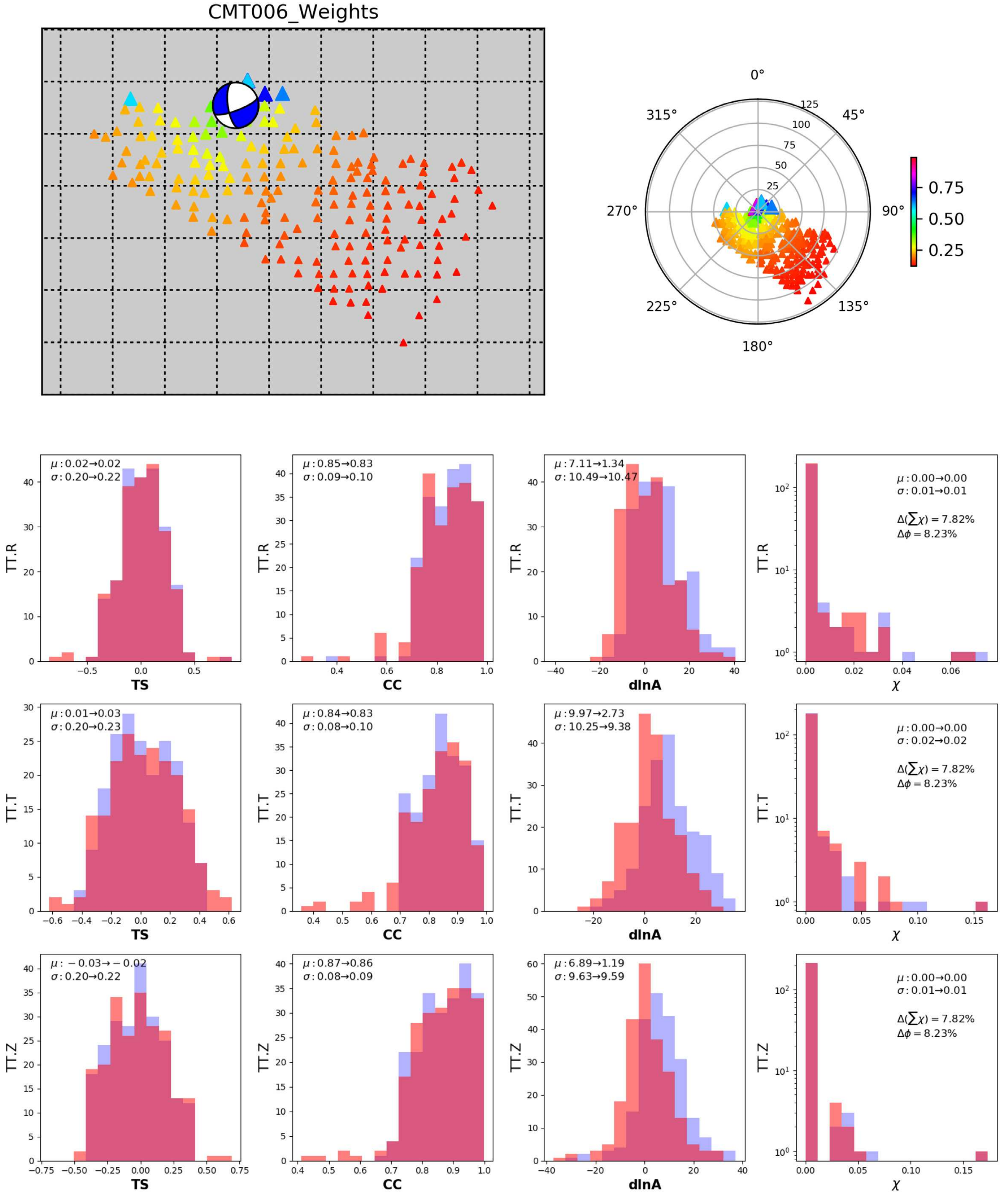


Figure A6. Waveform inversion for a CMT solution, with waveform-difference outlier rejection. The weighting functions also consider the number and measurements per station in an inversely proportional relation, and the station colours are dependent on the value of the weights. By comparison with the event-station map in Fig. A5, we notice a significant colour change for one station that turns from shallow to deep blue. Looking at the top station carefully, we also notice some changes hidden by this triangle. The weighting-value changes are caused by waveform-difference outlier rejection. (b) Metrics histograms, in which the $d\ln A$ histograms look to be more zero-centred and Gaussian-shaped. Notice the tick differences in χ with Fig. A5, caused by our normalization of χ by $\sum\chi$ before inversion. Waveform-difference outlier rejection changes the initial $\sum\chi$ value.

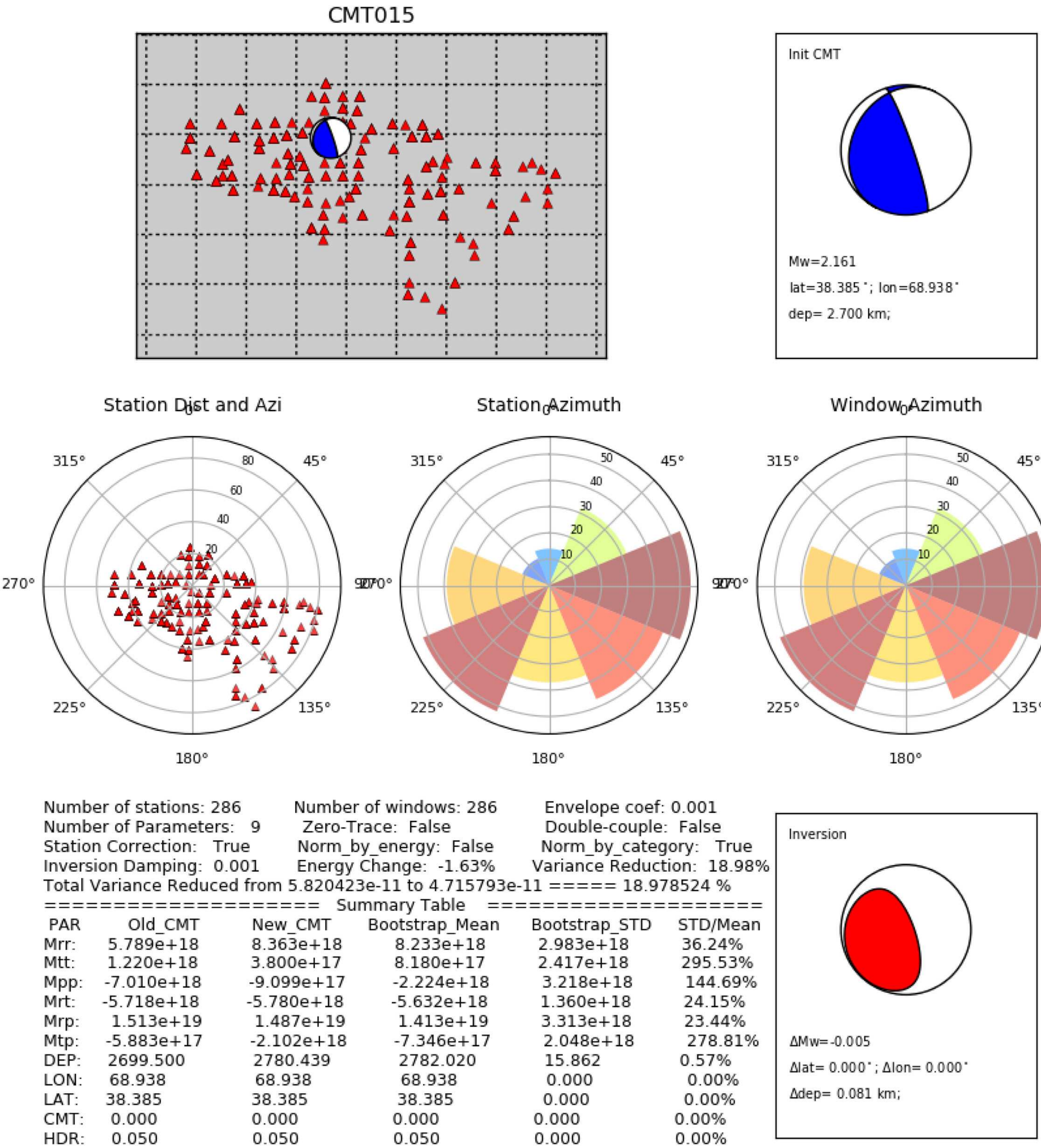


Figure B1. Output summary from a CMT inversion without double-couple constraint. Shown from top to bottom rows are the source-station constellation, the initial CMT solution, rose diagrams of the station distributions, weighting functions for stations and measurement windows in the azimuth domain, numerical details pertaining to our CMT inversion, and the inverted CMT solution.

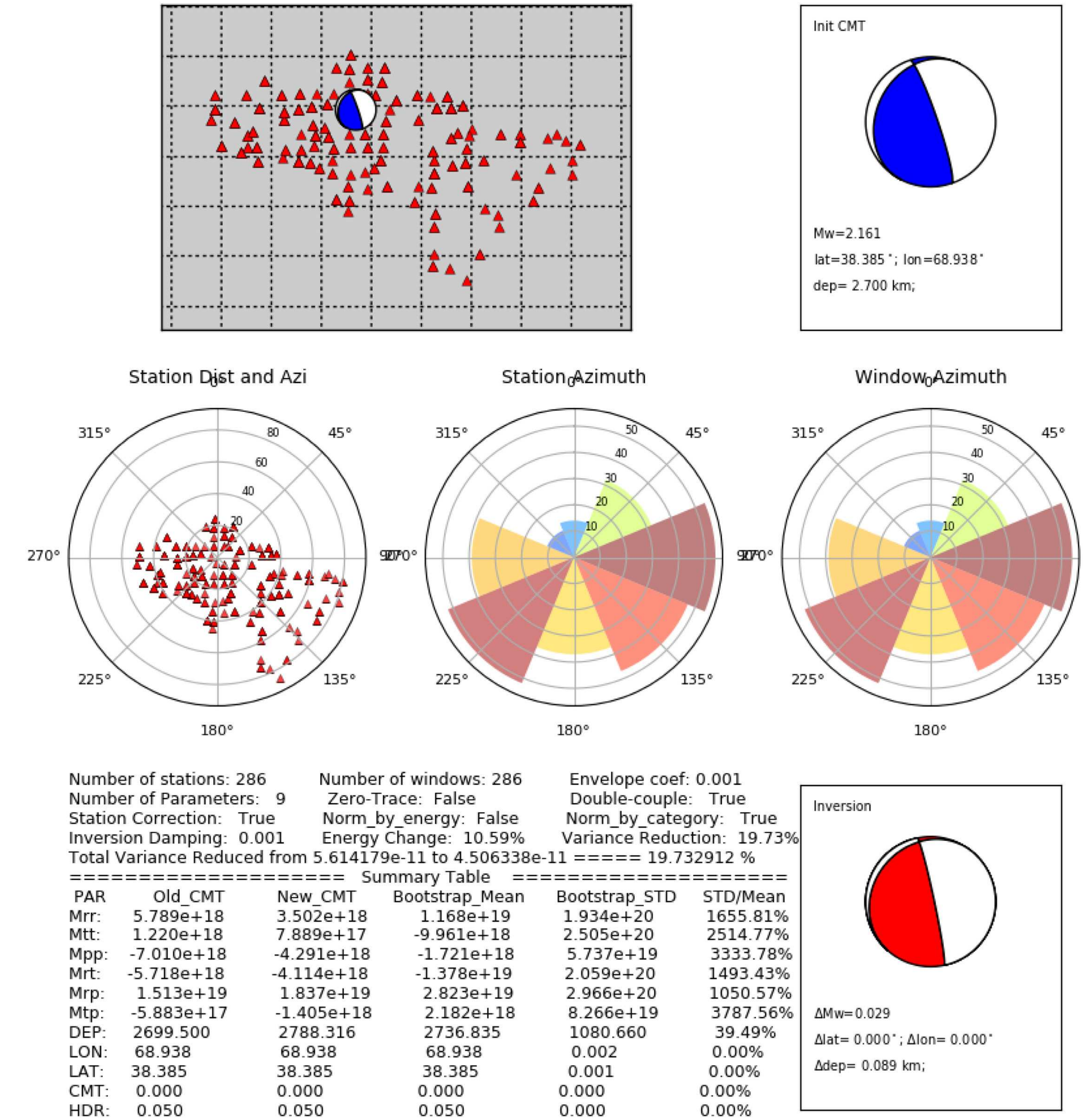


Figure B2. Output summary from a CMT inversion as in Fig. B1, but with double-couple constraint. Compared to the inversion without constraints, uncertainties are larger, as revealed by the 'Bootstrap_STD' columns.

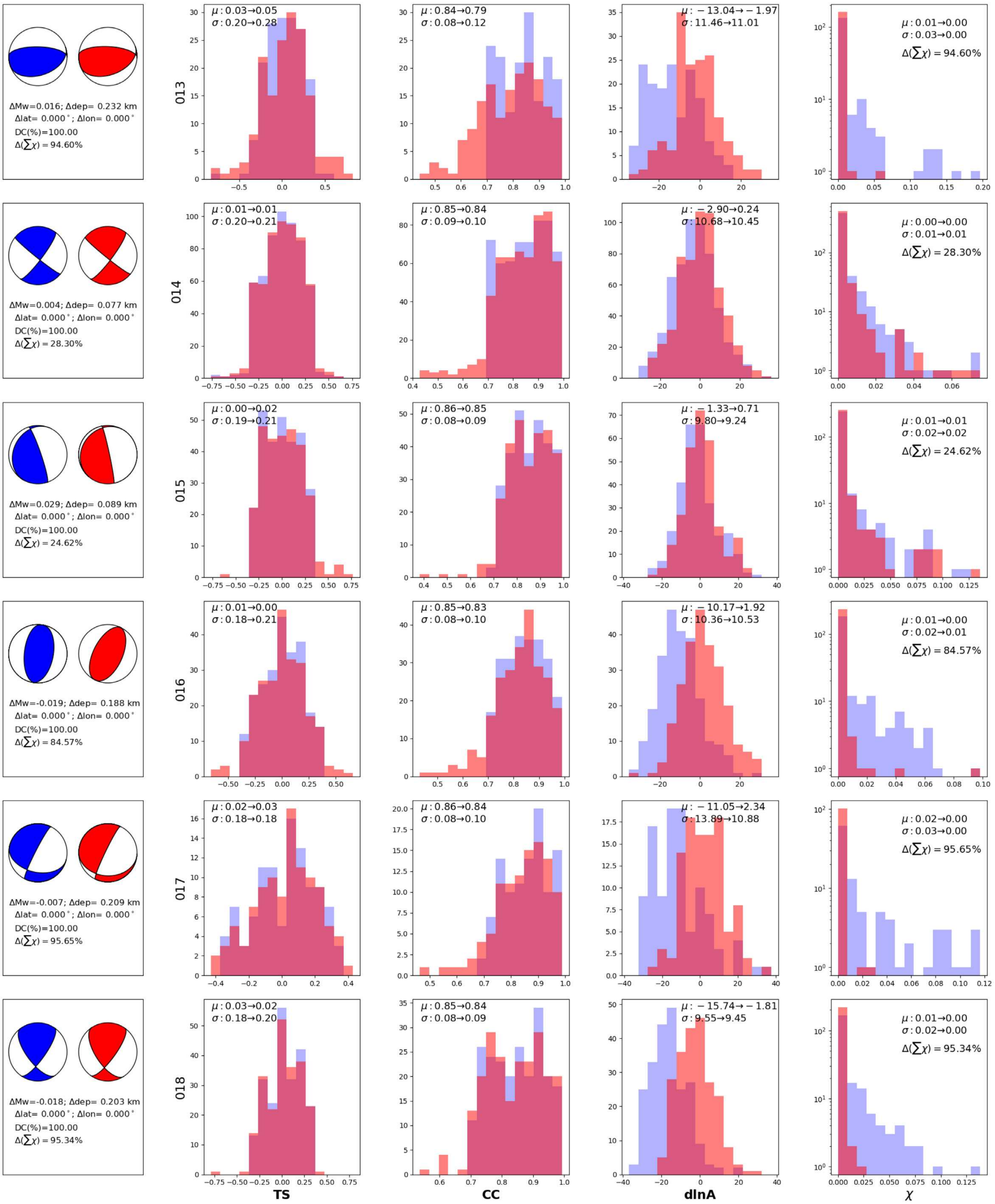


Figure B3. A selection of CMT inversion results, as in Fig. 7, but with the double-couple constraint applied during CMT inversion. The left-hand column shows initial and final focal mechanisms after inversion, in blue and red, respectively. The remaining columns show the distributions of misfit criteria: time-shift (*TS*), cross-correlation (*CC*), energy ratio ($d\ln A$) and waveform-difference misfit, χ. The model differences in scalar seismic moment and hypocentre coordinates, the double-couple percentages, and the change in overall data misfit are indicated, as are the means and standard deviations of the histograms of the misfit metrics. Compared with those in Fig. 7, the distributions of the metrics look similar, but the inversion results have larger uncertainties, as revealed by Fig. B2.

Geophys. J. Int. (0000) **000**, 000–000

Supplementary Materials for: Full-Waveform Centroid Moment Tensor Inversion of Passive Seismic Data Acquired at the Reservoir Scale

Qiancheng Liu[1,2], Umair bin Waheed[3], Dmitry Borisov[4], Frederik J. Simons[1], Fuchun Gao[5] and Paul Williamson[5]

[1] *Department of Geosciences, Princeton University, Princeton, NJ 08544, USA*
[2] *Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China*
[3] *Department of Geosciences, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia*
[4] *Kansas Geological Survey, Lawrence, KS 66047, USA*
[5] *Total E&P Research and Technology, Houston, Texas, USA*

1 April 2022

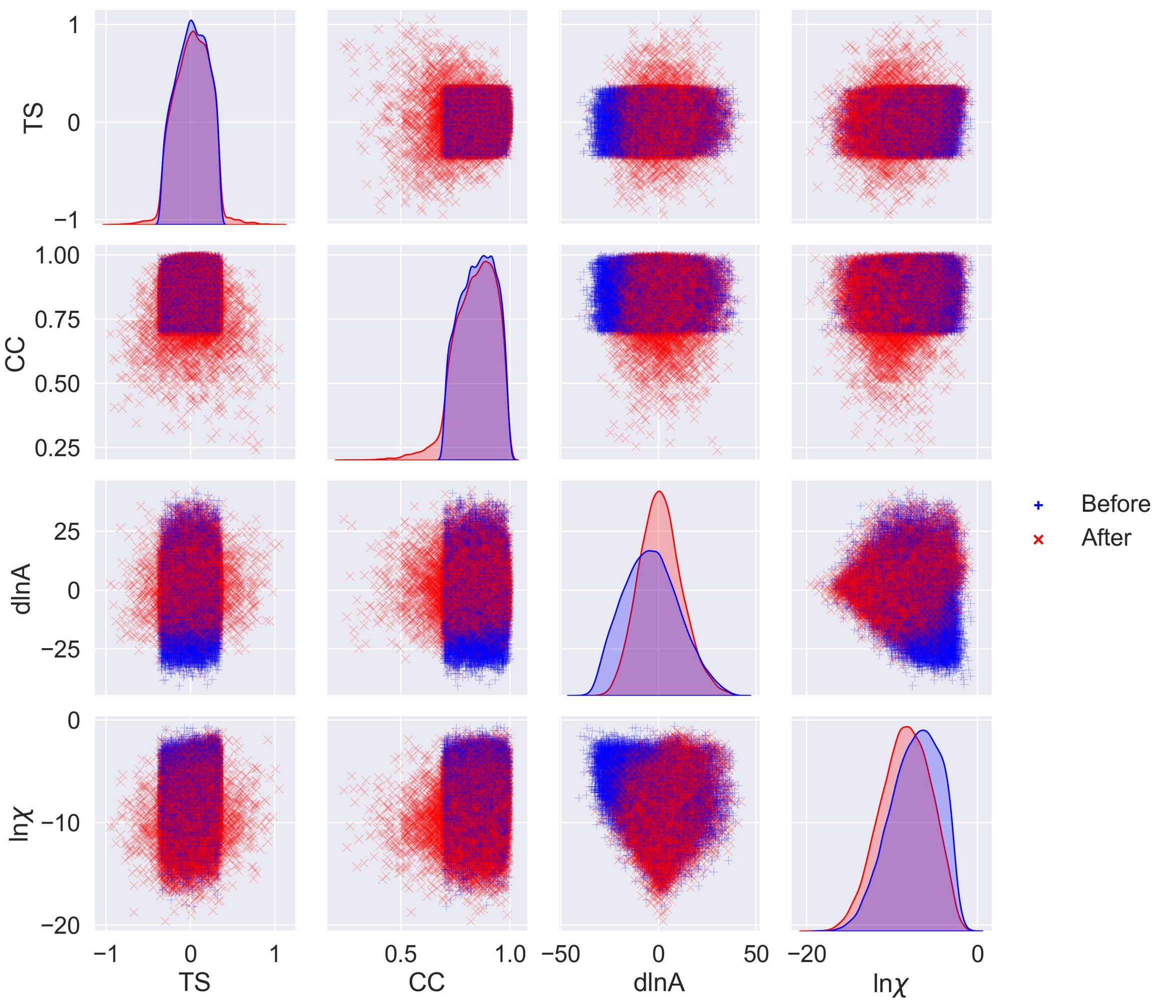


Figure S1. Cross-plot pairs between the metrics shown in Fig. 7, demonstrating their joint behavior over the whole set of inverted traces. Details along the diagonals are shown in Fig. 9. As discussed with Fig. 8, the $d\ln A$ and $\ln\chi$ metrics may improve at the expense of the other metrics.

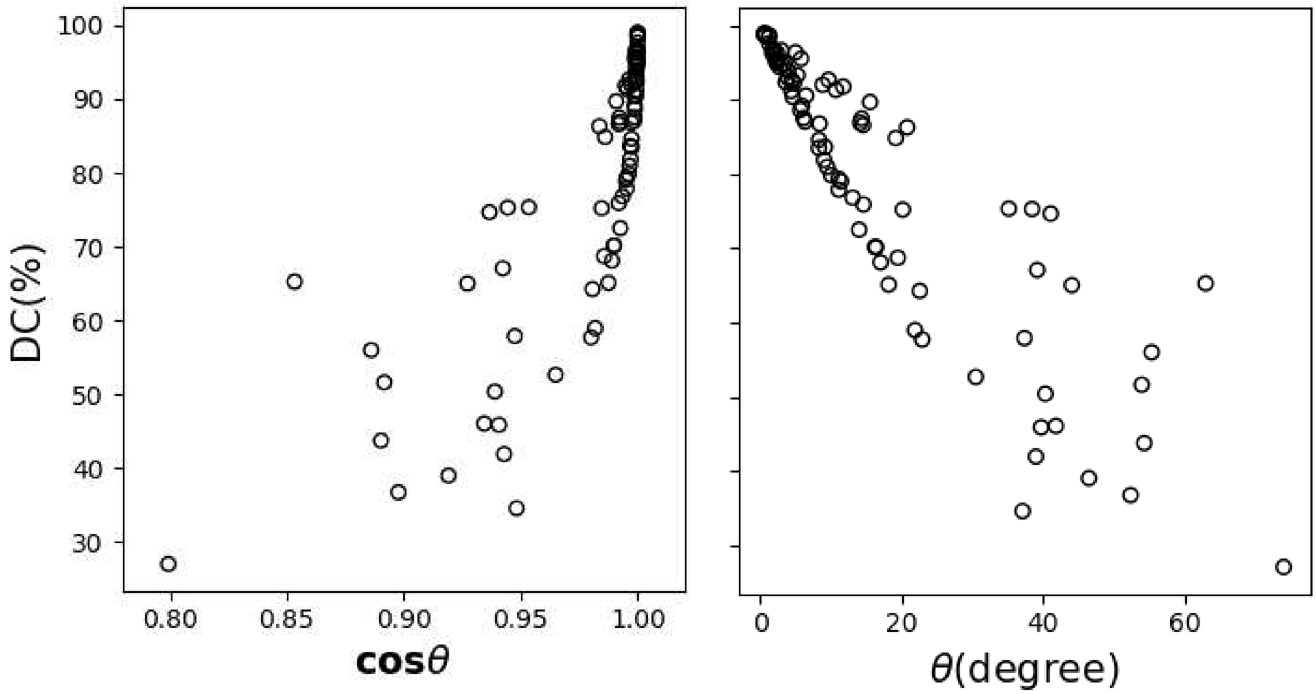


Figure S2. Relation between the double-couple percentages of the 86 events as quoted in the text, and the parameterization θ (Tape & Tape 2013).

.