Geophysical Journal International

Geophys. J. Int. (2019) **218,** 1044–1056
Advance Access publication 2019 May 16
GJI Seismology

doi: 10.1093/gji/ggz224

Analysis of surface and seismic sources in dense array data with match field processing and Markov chain Monte Carlo sampling

Chloé Gradon,[1] Ludovic Moreau,[1] Philippe Roux[1] and Yehuda Ben-Zion[2]

[1] *Université Grenoble Alpes, CNRS, IRD, IFSTTAR, ISTerre, Maison des Géosciences,* 38000 *Grenoble, France. E-mail: chloe.gradon@univ-grenoble-alpes.fr*
[2] *Department of Earth Sciences, University of Southern California, Los Angeles, CA* 90089*–*0740*, USA*

Accepted 2019 May 14. Received 2019 May 15; in original form 2019 March 28

SUMMARY

We introduce a methodology based on array processing to detect and locate weak seismic events in a complex fault zone environment. The method is illustrated using data recorded by a dense array of 1108 vertical component geophones in a 600 m × 600 m area on the Clark branch of the San Jacinto Fault. Because surface and atmospheric sources affect weak ground motion, it is necessary to discriminate them from weak seismic sources at depth. Source epicentral positions and associated apparent velocities are extracted from continuous seismic waveforms using Match Field Processing (MFP). We implement MFP at specific frequencies targeting surface and subsurface sources, using for computational efficiency a forward model of acoustic source in a homogenous medium and Markov Chain Monte Carlo sampling. Surface sources such as Betsy gun shots and a moving vehicle are successfully located. Weak seismic events are also detected outside of the array, and their backazimuth angle is retrieved and found to be consistent with the fault geometry. We also show that the homogeneous acoustic model does not yield satisfying results when extracting microseismic event depth, because of the ambiguity between depth and the apparent velocity based on surface data.

Key words: Computational seismology; Earthquake ground motion; Earthquake source observations; Wave propagation.

1 INTRODUCTION

With the improvement of sensor technology, such as the development of autonomous nodes, and the decrease in costs, the deployment of dense temporary arrays has become more popular in the context of academic research. These arrays are spatially denser than the permanent networks in place, and have already been successfully used in high-resolution imaging and monitoring of faults zone and volcanic structures. The improved spatial coherency at higher frequencies has allowed the use of noise-based tomography to image finer details of the Newport-Inglewood Fault (Lin *et al.* 2013) and the San Jacinto Fault (Roux *et al.* 2016; Zigone *et al.* 2019), with scales ranging from hundreds to tens of meters. Focal spot imaging (Hillers *et al.* 2016) has also been applied on the San Jacinto Fault.

The improved resolution provided by dense arrays can also be used to detect smaller earthquakes and other sources of radiation (e.g. tremor and anthropogenic sources) than is possible with regional networks (e.g. Inbal *et al.* 2016; Li *et al.* 2018; Meng & Ben-Zion 2018b). Many studies have successfully used arrays to locate weak subsurface sources for geophysical applications. However, no high-resolution study was performed to locate sources in the top few kilometers of fault zones. Such a study can provide fundamental information on the dynamics and properties of the shallow part of fault zones, their surrounding media and factors contributing to the observed ongoing ground motion.

Array methods commonly used to deal with low Signal-to-Noise Ratio (SNR) are generally 'stack-based' techniques, and in a lesser measure time-reversal. The former consists of time-shifting the signals at each station with an appropriate time-delay, so as to sum signals coherently to extract information hidden in noise. The appropriate time shifts depends on the distance between the dominant source and the array. A classic example of stack-based technique is beamforming. This involves scanning the angles of arrival of a wave, assuming it propagates with a plane wavefront. The phase shifts that correspond to each of the scanned angles are applied to the signals recorded at the array. The goal is to identify the arrival angle that maximizes the coherency in the shifted signals (e.g. Rost & Thomas 2002; Landès *et al.* 2010). Obviously, assuming plane wave propagation is valid only in the far-field of the source. To locate sources closer to the array, spherical waves can be considered instead when scanning for different source positions (Kao & Shan 2004; Langet *et al.* 2014; Cesca & Grigoli 2015; Grigoli *et al.* 2016; Poiata *et al.* 2016). Such methods are effective because the array gain classically compensates for low SNR, as long as spatial sampling meets the Nyquist's criterion.

Time reversal methods have been used to detect non-impulsive events occurring in complex media. This approach consists of back-propagating time-domain signals in a velocity model after they have been numerically time-reversed (Larmat *et al.* 2006, 2008; Artman *et al.* 2010). It is particularly effective on broad-band signals, where each frequency component adds up constructively when the velocity model matches the elastic properties of the medium. Due to the reciprocity of wave propagation, back-propagation focuses at the source, with resolution capabilities limited by the diffraction limit. However, applying time reversal methods to study seismic sources requires a detailed 3-D elastic velocity model including the shallow structure.

Localizing sources at shallow depth in a fault zone environment involves several challenges. For example, such sources are expected to have weak energy, high frequency content and short duration (Kwiatek & Ben-Zion 2016). These characteristics result in signals with poor SNR. In the San Jacinto Fault Zone (SJFZ) weak seismic events have peak energy between 0.5 and 20 Hz. These signals overlap with ambient seismicity, which is dominated by surface waves (Roux *et al.* 2016). This makes the detection of impulsive sources at depth difficult. Similarly, the existence of additional, non-seismic sources close to the array, is also an issue. For example, anthropogenic activities or atmospheric sources generate surface waves that exhibit coherency in the array (Riahi & Gerstoft 2015; Meng & Ben-Zion 2018a). This considerably complicates the interpretation of the signals. Discriminating between surface sources and sources at depth is therefore not trivial, as demonstrated by Inbal *et al.* (2018), and this is one key challenge of this study.

Another challenge is the high heterogeneity of the medium with strong velocity variations (e.g. Qin *et al.* 2018; Mordret *et al.* 2019) that can result in loss of coherency across the array. This would reduce the performance of stack-based methods, for which the recorded waveform must remain sufficiently coherent between sensors across the array in order to add constructively. Moreover, shallow sources may exist in the near-field of the array, hence the phases of surface and body waves may not be easy to separate. This raises additional issues in terms of interpretation.

This work is motivated by the need for a sensitive, yet robust way of localizing events at shallow depth in spite of these challenges. In the following, localizing refers to detecting and locating sources of radiation in continuous waveforms recordings. The performed analysis uses a very dense array and a statistical approach, and we rely on the high spatial resolution of array processing techniques to balance the poor *a priori* knowledge of sources characteristics and medium properties. Array processing methods benefit from the high number of sensors and can be automated using several days of continuous seismic noise. In order to localize short duration events, the use of short time windows results in high computational cost, making the development of an efficient localization method necessary. Consequently, time-reversal methods are unsuitable because they classically apply to high SNR seismic waveforms with 3-D elastic modelling, which generally results in time-consuming numerical computations (Larmat *et al.* 2006, 2008).

The study employs a 30-day long data set recorded by a dense array of 1108 vertical component geophones (Fig. 1a) covering an area of about 600 m × 600 m around the Clark branch of the SJFZ in the trifurcation area southeast of Anza, CA (Ben-Zion *et al.* 2015). The dense array was deployed in a relatively remote ranch, with structures and stationary machines at the surface, referred to as the Sage Brush Flat (SGB) site. The trifurcation area is the most seismically active portion of the SJFZ with tens to hundreds microseismic events per day (Hauksson *et al.* 2012; Ross *et al.* 2017; Meng & Ben-Zion 2018b). In addition to earthquakes, recorded waveforms in the area are affected by air-traffic events, wind and other non-seismic sources (Johnson *et al.* 2018; Meng & Ben-Zion 2018a) located mainly on the western side of the fault (Fig. 1c) around the position of the ranch. To provide benchmarks for various studies, Betsy gunshots were fired near the sensors marked with orange colours in Fig. 1(c). Finally, some sensors of the dense deployment overlap with six three-component seismometers and two sensors in shallow (∼100 m) boreholes (Ben-Zion *et al.* 2015) that were not used in this study.

In order to find shallow sources efficiently, we introduce a methodology based on Match Field Processing (MFP) combined with Markov Chain Monte Carlo (MCMC) sampling to infer the Probability Density Function (PDF) of the wavefield and medium parameters. MFP is equivalent to beamforming as a frequency-domain technique originally developed for ocean acoustics that takes advantage of the spatial coherency of a wavefield through an array with continuous recording (Baggeroer *et al.* 1993; Kuperman & Turek 1997). With the recent increase of dense seismic arrays, it has become possible to apply this approach also in the context of exploration (Corciulo *et al.* 2012) and hydrothermal geophysics (Cros *et al.* 2011; Vandemeulebrouck *et al.* 2013).

The combination of MFP with MCMC has been used in ocean acoustics to ensure convergence when extracting the medium parameters and to estimate source position in uncertain media (Richardson & Nolte 1991; Tollefsen & Dosso 2014). In this paper, we use this combination on seismic data to solve a complex optimization problem where classical grid searches are too time-consuming and gradient-descent like methods fail due to many local minima. The outputs of the MFP–MCMC processing are PDFs of the source position and average sub-surface velocity, which provide a quantitative measure for the confidence of the results.

The reminder of the paper is structured as follows. The outline of the method and its implementation are described in section 2, with a focus on source characterization that can discriminate between surface and shallow events. Section 3 presents representative example localization results for sources both at the surface and at depth, taken from the scan of the 26-day data set. The results are summarized and discussed in the final section.

2 MFP METHOD APPLIED TO GEOPHYSICAL DATA

MFP can be considered the frequency domain equivalent of a shift-and-stack technique. It is a model-based extraction method where the phase of a wavefield in array data is compared to the phase of a synthetic wavefield referred to as a replica. The best match between the data and the replica provides the best description of the seismic event given the model.

The main advantage of using replicas over stacking traces comes from the fact that, for a given frequency, they can be parametrized to describe the physical problem with any suitable degree of complexity, e.g. by including a search of the velocity profile in the medium for improved data fitting. Complex models and numerical solutions can be used to obtain a realistic representation of the data at a given frequency and yield superior results compared to stacking methods. Obviously, the efficiency of this approach is limited by the time required to compute the replicas, but this can be made very fast when a closed-form solution is available. In many geophysical studies, a simple homogeneous model of a spherical wave is considered (Cros *et al.* 2011; Corciulo *et al.* 2012; Vandemeulebrouck *et al.* 2013;

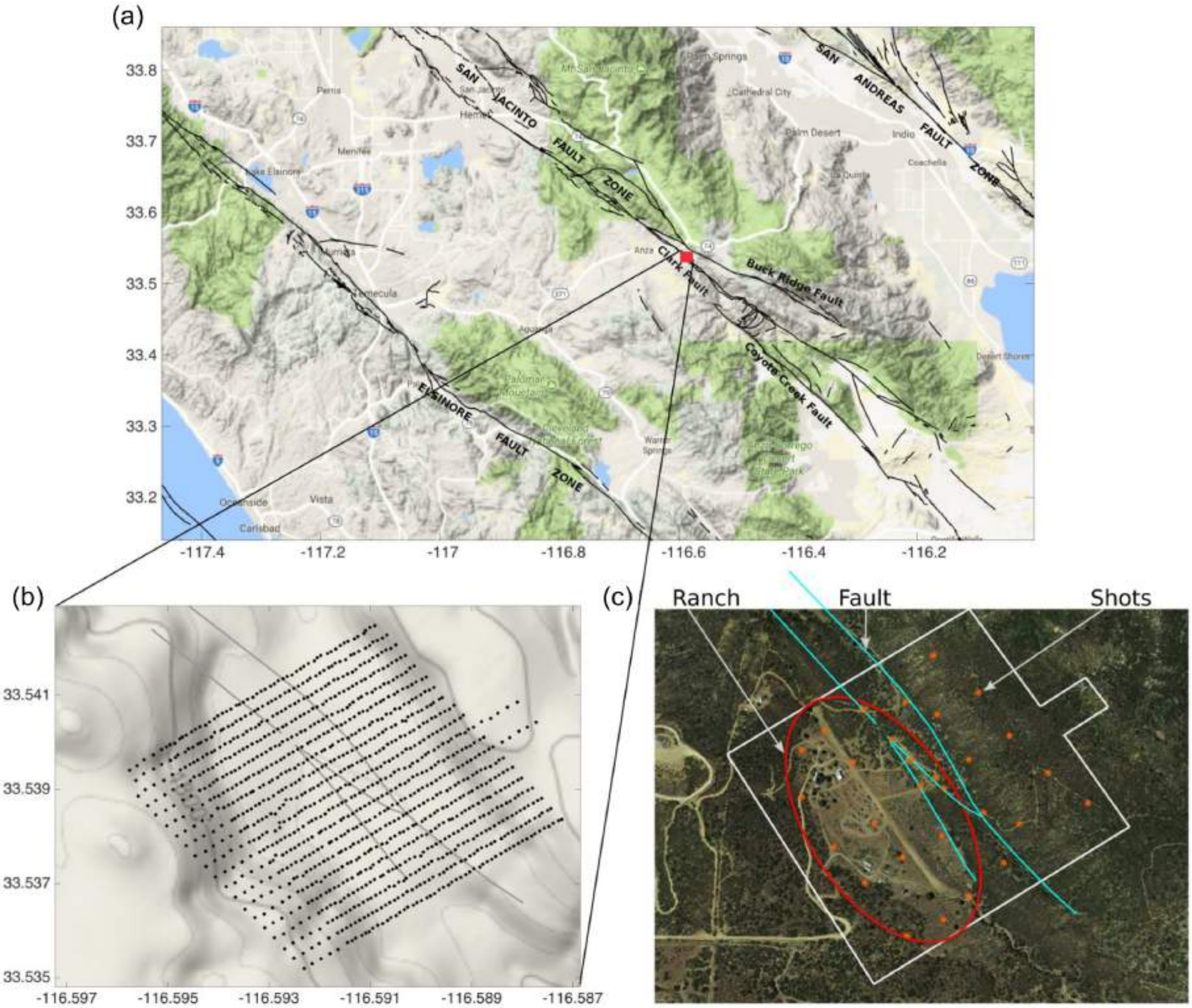

Figure 1. (a) Geographical situation of the Sage Bush Flat array on the San Jacinto Fault Zone (red square). (b) Geometry of the 1108 sensor array and situation with respect to fault traces (black lines) and local topography (Map data, Google 2017). Each black dot corresponds to a station. (c) Positions of expected main sources. Seismic source should mainly be located around the geological fault trace (marked with blue lines) and cultural sources from human activity and wind excitation are expected in the ranch area (within the red circle). The orange dots represent shots that were fired within the array.

Chmiel *et al.* 2016). In our case, the choice of such model can be questioned given the complexity of the medium and its possible impact on the spatial variation of the wavefield.

The primary goal of this study is to localize shallow seismic sources in the top 3 km of the crust. For very shallow events, the source-to-array distance is within one wavelength, which prevents the separation of *P*- and *S*-wave contributions in the wavefield recorded at the surface. The use of elastic models, requiring complex numerical computations with different wave velocities, is too computationally demanding. Tests using a library of pre-computed signals show that loading replica for more than 1000 sensors would also significantly increase computational time. For these reasons, we choose to use an acoustic model, which is far from representing the complexity of the medium under the array, but is nonetheless an appropriate model for initial testing of the methodology

In the following, we denote by **b** the replica computed from a candidate source. For a given frequency, **b** is a vector that contains the value of the wavefield at all sensor positions. For example, at sensor *j* the replica has the following expression:

$$b_j(\omega, \boldsymbol{a}) = A_j(\omega, \boldsymbol{a})\, e^{i\omega t_j(\boldsymbol{a})} \tag{1}$$

where A_j is an amplitude term, $\omega = 2\pi \mathrm{f}$ is the angular frequency and $\boldsymbol{a} = [\boldsymbol{X}\ V]$ is the set of parameters to be extracted consisting of source coordinates **X** and medium velocity V.

The traveltime $t_j(\boldsymbol{X})$ between the source and sensor *j* is,

$$t_j(\boldsymbol{a}) = \frac{r_j(\boldsymbol{X})}{V}, \tag{2}$$

with $r_\mathrm{j}(\boldsymbol{X})$ being the source–sensor distance.

In the study region, the SJFZ exhibits strong lateral variations of seismic velocities (Roux *et al.* 2016; Qin *et al.* 2018). This has two main consequences regarding the use of the spherical wave in a homogeneous medium approximation.

First, an average velocity model is needed to provide consistent source locations in spite of the medium heterogeneities. This means that four parameters, *x*, *y*, *z* and *v* should be used to determine the hypocentral position of a source. However, the depth, epicentral distance and velocity are linked through the apparent velocity measured at the surface. In practice, while the apparent velocity varies across the array due to variations of the epicentral distance, synthetic tests (not shown) confirm that the MFP method is only sensitive to the average apparent velocity. Consequently, for sources under the array the apparent velocity mostly depends on the average velocity of the medium and the source depth. We therefore use the apparent velocity V_{ap} as a proxy for both the velocity of the medium and the depth of the source, reducing the parameters to only x, y and V_{ap}.

The second issue brought by the lateral heterogeneities in the fault environment are amplitude variations in the data that are mainly a consequence of the structure of the medium. Therefore, using a simple geometrical decay in the replica is not sufficient to properly describe the amplitude variations of the wavefield. To avoid ambiguities when comparing the replica with the measurements, $A_j(\omega, \boldsymbol{a})$ is set to 1. This amplitude normalization means that every sensor on the array has the same weight in the MFP algorithm. The comparison between the data and the model is therefore only based on phase values. One advantage of MFP is that the origin time of an event does not need to be known, because the source localization only depends on the relative phase differences between sensors.

Phase differences are generally described in the form of the so-called Cross Spectral Density Matrix (CSDM). The CSDM, denoted by $\mathbf{K}$, is the frequency domain equivalent of the time-domain broad-band cross-correlations between all sensors. It contains the autocorrelation and intercorrelation between sensors in its diagonal and off-diagonal terms, respectively. This is calculated from the normalized data as:

$$\boldsymbol{K}(\omega) = \boldsymbol{d}^*(\omega) \, . \boldsymbol{d}(\omega) \, , \tag{3}$$

where $\boldsymbol{d}(\omega)$ is a complex vector obtained from the Fourier transform of windowed time-series records at each sensor and the star denotes complex conjugation. Fig. 2 illustrates values of $\mathbf{d}$ on the whole array for various sources and two different frequencies. Because we use normalized data, these examples correspond to spatial phase variations for a given windowed time-series. The time window is short enough to preserve the contribution from impulsive signals but long enough to ensure the stability of the Fourier transform. In practice at a given frequency, we use a time window of duration $T = 4$ periods. This means that we obtain a single localization for the dominant seismic source for each time interval T.

The comparison between the replica and CSDM can be performed using various operators, the choice of which depends on a trade-off between robustness and resolution. An operator with higher resolution, such as the Capon algorithm, is more sensitive to the accuracy of the model and requires an inversion of the CSDM. In our case, because of the complexity of the subsurface structure, the transient nature and low SNR of microseismic events, we use the more robust Bartlett operator (Cros *et al.* 2011; Corciulo *et al.* 2012), given by:

$$B(\boldsymbol{a}, \omega) = \sum_{\omega} \left| \boldsymbol{b}^*(\boldsymbol{a}, \omega) \, . \boldsymbol{K}(\omega) \, . \boldsymbol{b}(\boldsymbol{a}, \omega) \right| . \tag{4}$$

This operator is divided by the number of sensors squared (N^2) to scale its output between 0 and 1, where 1 signifies a perfect agreement between the data and the replica and $1/N$ represents a complete decorrelation of the two. When the Bartlett operator reaches a maximum, the physical parameters (source location and medium velocity) provide the best description of the dominant source given the model. MFP techniques are classically implemented with a grid search to find the solution that maximizes the Bartlett operator. However, the accuracy of grid searches depends on the density of the grid, and may require a prohibitive number of computations when the parameters space is multidimensional. This is especially true in our case were we process large amounts of data using short time windows.

Newton-Raphson-like methods are more efficient in this regard, but remain inadequate for problems where parameters follow a complex PDF with local minima. Moreover, none of these sampling methods provide the PDF of the parameters, which are essential for the interpretation of the solution. To tackle these issues, the MCMC sampling method is used here (Sambridge & Mosegaard 2002). The MCMC sampling algorithm is an iterative process that follows a Markov Chain to determine candidate solutions in multi-dimensional parameter space and evaluates their likelihood (PDF of parameters) given the data. Candidates solutions are subsequently accepted or rejected, based on a Bayesian criterion that compares this likelihood to that of the last accepted solution. This process results in an irregular sampling of the parameter space, where the area around the global optimum is evaluated more finely. In that case, the resolution of the final solution depends much less on the number of computations. However, the MCMC algorithm generally requires a burn-in period before reaching a stable state where the PDF is efficiently sampled. To speed up this burn-in process, we use simulated annealing global optimization as suggested in Moreau *et al.* (2014). Simulated annealing is an MCMC-like algorithm where the tolerance for unlikely candidates is gradually lowered. In practice, we first use simulated annealing to force the convergence and then use an MCMC process to sample the area around the global optimum. In the following only results of the MCMC process are plotted, since the output of the simulated annealing process is not a PDF.

The inclusion of the MCMC in the classical MFP process is illustrated in Fig. 3. The first step of the method is application of a narrow band filter with bandwidth of 4 Hz centred on a chosen frequency. Since MFP involves a frequency-domain phase-based optimization problem, the choice of frequency is a key step that is discussed below. Once the frequency is set, the filtered waveforms are segmented into successive time windows of length T. For each window, the CSDM is calculated using eq. (3) and the MFP outputs are obtained through the MCMC sampling process. The outputs are PDFs of the physical parameters associated with their corresponding Bartlett operator value. The value of the output maximum for candidate detections must be higher than an empirical threshold determined from the results of the 26-day scans presented in section 4.

The choice of frequency is critical in our process. To compensate for the fact that we do not use the depth of the source as a parameter, we take advantage of the capacity of the array to spatially sample the wavefield and use the frequency f as a filter that separates surface from deep sources. The geometry of the array sets the limits to the range of the apparent wavelengths $\lambda_{\mathrm{ap}} = V_{ap}/f$ that can be sampled. On one hand, a minimum aperture of about twice the apparent wavelength is necessary to allow a good spatial resolution. On the other hand, the Nyquist's criterion requires at least two sensors per wavelength to avoid spatial aliasing. In practice, we require four sensors per wavelength to properly sample the wavefield. This means that a given frequency sets the range of apparent velocities that can be detected by the array.

These limitations of the array detection capability can be turned into a practical way of separating between sources at depth and at the surface. The apparent velocity, V_{ap}, measured for surface sources is lower than for sources at depth. The reasons for this are twofold. First, velocities of P- and S- waves increase with depth and are larger than surface wave velocities. Second, the projection at the surface of a body wave propagating from sub-surface sources results in a higher apparent velocity. Consequently, the choice of frequencies translates directly into a filter for the measureable apparent velocities, and through that for depth.

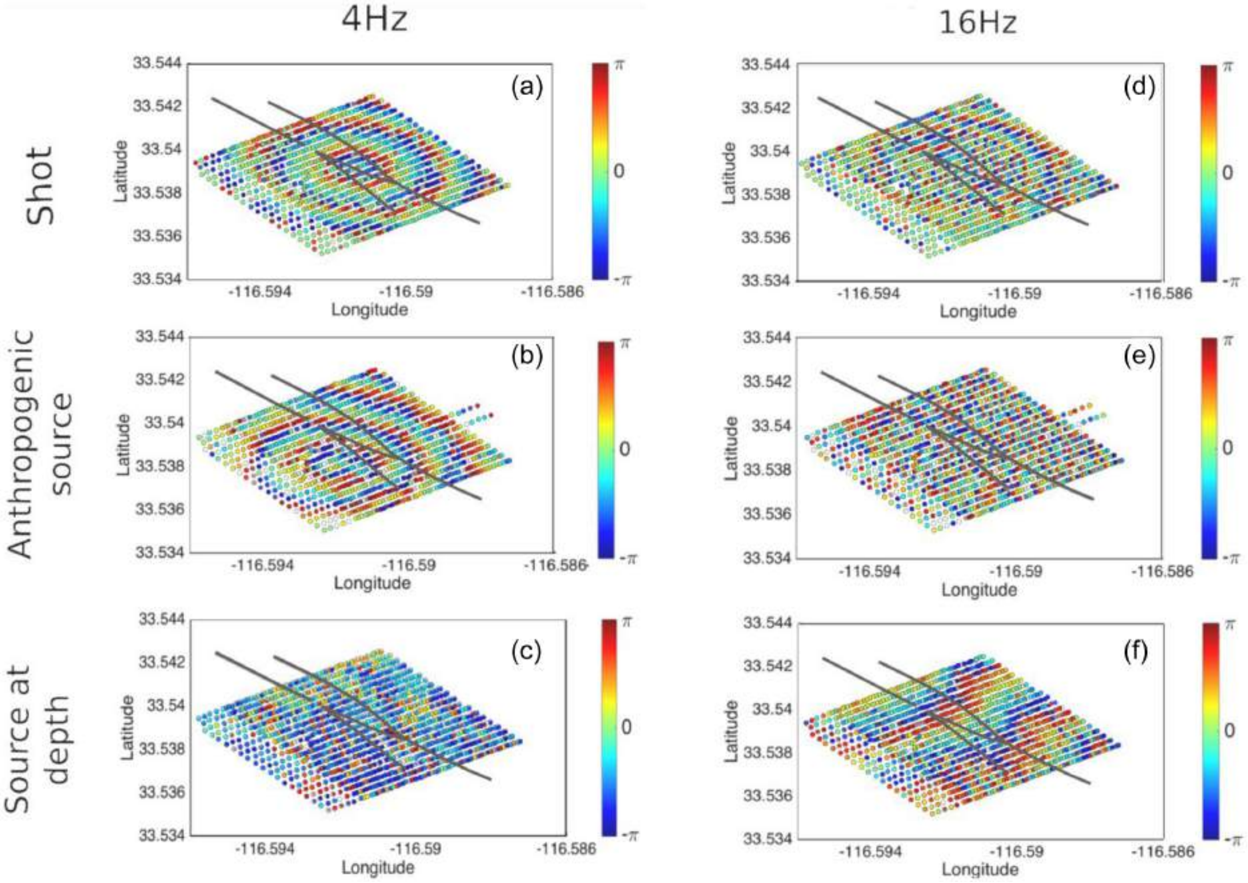

Figure 2. Phase patterns recorded at the array for 4 Hz (a, b, c) and 16 Hz (d, e, f). The patterns are obtained from the Fourier transform of 1-s (4 Hz) and 0.25-s (16 Hz) windows containing the events. The phases are highly coherent in space at 4 Hz for the first two sources (a and b) and at 16 Hz for the last one (f) and display clear propagation patterns. The signal from the source at depth shows little spatial variation at 4 Hz. The upper right-hand panels show spatial aliasing of the phase, inducing a loss of spatial coherency.

While the depth of source drives the choice of specific frequencies, the range of potential frequencies of study is limited by the medium and sensor characteristics. In theory, the highest frequency that can be used is only limited by the spatial sampling of the array, which sets the measurable phase coherency between sensors. However, initial tests show that above 20 Hz the phase measured across the array appears highly distorted due to medium heterogeneities and attenuation. At the other end of the spectrum, the lower frequency boundary is limited to 1 Hz due to both poor SNR (related to the 5 Hz corner frequency of the geophones) and the limited aperture of the array.

The employed dense array consists of 1108 geophones divided in 20 rows separated by 30 m, with an intersensor distance of around 10 m in each row (Fig. 1b). This translates to a diagonal aperture of 800 m and a mean interstation spacing of 25 m. In this configuration, all waves propagating with apparent wavelengths λ_{ap} between 75 and 400 m can be processed. The choice of central frequencies determines the range of apparent velocities V_{ap} for which detections are possible. Using 4 and 16 Hz provides a sufficient gap to distinguish between sub-surface and surface sources, while staying below the 20 Hz phase distortion limit. The apparent velocities V_{ap} that can be processed at 4 Hz are in the 300–1600 m s^{-1} range, while at 16 Hz they are in the 800–6400 m s^{-1} range.

Fig. 2 shows the phase distribution across the array for detected events with waveforms filtered around 4 and 16 Hz. We distinguish two surface sources and a third type involving sub-surface sources, associated with wavefields having estimated apparent velocities under 900 m s^{-1} for the first two and above 4000 m s^{-1} for the third type. Surface sources at 4 Hz (Figs 2a and b) exhibit a clear propagation pattern, whereas for a source at depth (Fig. 2c) the array is not wide enough to record the phase oscillation patterns. On the other hand, the wavelength of surface sources is too small at 16 Hz (Figs 2d and e), resulting in aliasing and loss of spatial coherency in the array, whereas the wavelength of deep sources is now reduced and can be seen within the array (Fig. 2f). This confirms the choice of the frequency as the range of apparent velocities that can be measured by the array, which directly translates into an effective discrimination between surface and deep sources.

3 APPLICATION TO DENSE ARRAY DATA

The array described in the previous section recorded continuous waveforms for 26 days at a 500 Hz sampling frequency, providing high-quality data at frequencies between 1 and 250 Hz. The duration of the recordings, large frequency range as well as the location and density of the array, are expected to provide a wide variety of sources to be detected and located. As mentioned, in addition to the seismic activity from the fault under the array, air-traffic, interaction of wind with obstacles above the surface, along with car-traffic and other activities produce complex ground motion (Fig. 1c). Small

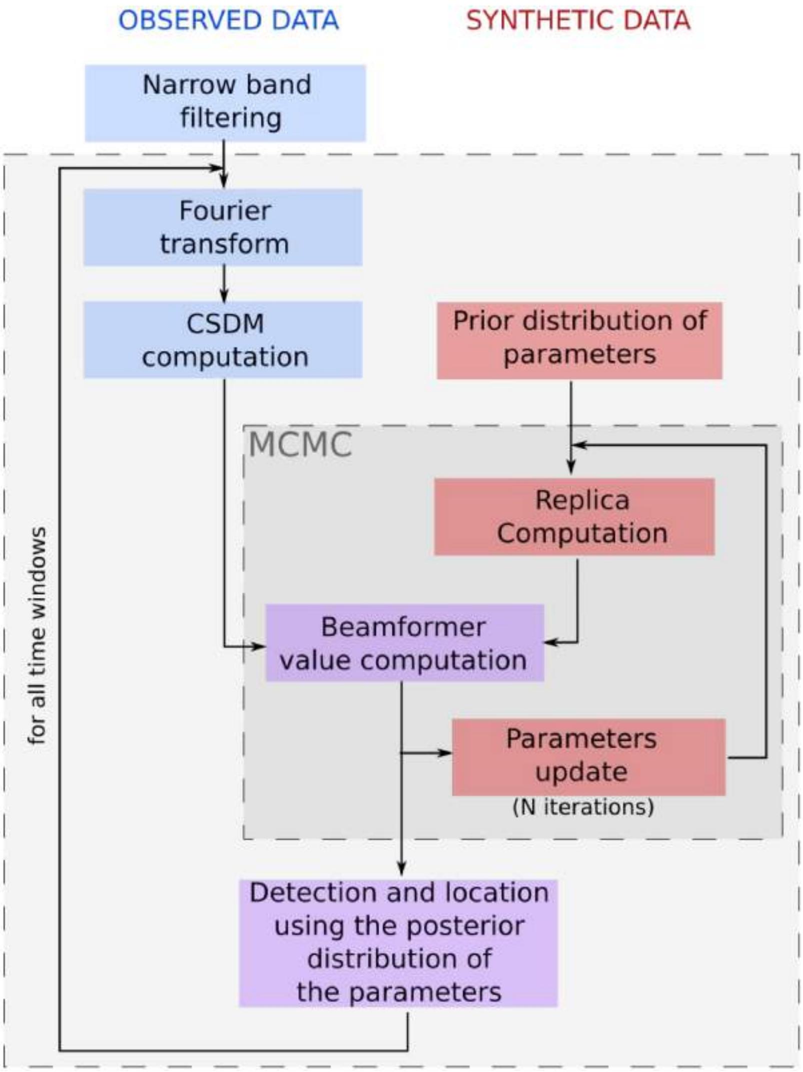

Figure 3. General workflow of the method, describing the implementation of the MFP method with a Markov Chain Monte Carlo scheme. The key steps of the latter are detailed in the darker grey box labelled MCMC.

Betsy gunshots fired at different locations of the array during the experiment (Ben-Zion *et al.* 2015) provide good benchmarks on surface source localization, as their position and time of occurrence is known.

Fig. 4 shows 10 s long time-series filtered between 0.5 and 20 Hz that correspond to three different events. For each event, the traces for a line of sensor are displayed. Some events, such as shots, are easily identifiable with a clear propagation across the array, even on these five traces only. However, times of arrival and wave propagation for events of longer duration such as anthropogenic sources are more difficult to distinguish in the traces, even with a good SNR. For weak sources at depth, the SNR is too low and events cannot be detected on single traces. Figs 4(d)–(f) show representations of the input to the MFP technique. Each panel is the wrapped phase of the Fourier transform of the time window in the corresponding time-series.

In these examples, the epicentral locations of the sources are readily identifiable because of the high array spatial density. The data representation in Figs 4(d)–(f) is equivalent to considering the spatial variations of the wavefield, which is then matched against different replica.

Two different sets of extraction parameters are used in the example data analysed. In the first set, only three parameters are taken into account: the source epicentral position x and y, and the apparent velocity at the array V_{ap}. This first set is used to scan the entire data set efficiently. Once the time-windowed signals that correspond to sources of interest are identified, a follow-up study that includes a higher number of parameter or a more elaborate model can be performed to better constrain the depth. In the second extraction, the parameters are the x, y and z coordinates of the source, and the velocity of the medium is set to 550 m s^{-1} found in previous studies to represent the very shallow crust (Roux *et al.* 2016; Meng & Ben-Zion 2018b). In the next section, we present example localization results obtained from examining 26 days of the data set.

4 EXPERIMENTAL RESULTS

A scan of the data over 26 days is performed at 4 and 16 Hz. We use the distribution of the maximum of the outputs at both frequencies, represented in Fig. 5, to determine thresholds (T_{exp4} and T_{exp16}) above which an event is detected. To reduce false detections, we use

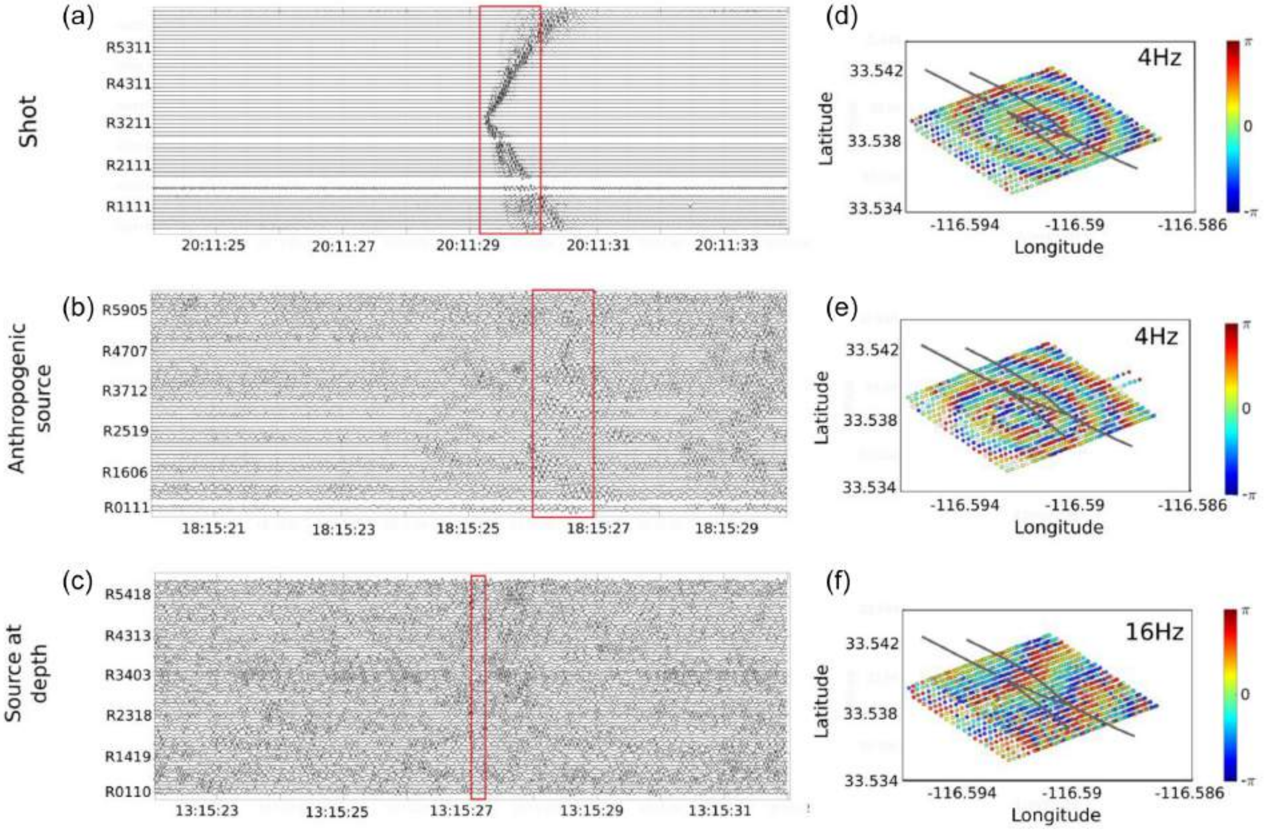

Figure 4. (a–c) show 10-s time signals measured by a line of sensors at the centre of the array. The traces are filtered between 0.5 and 50 Hz. Three different types of signal, coming from two surface sources (a and b) and a source at depth (c), are represented. The red boxes represent the 1 and 0.25-s time windows, used to obtain the phase patterns in the right-hand panels.

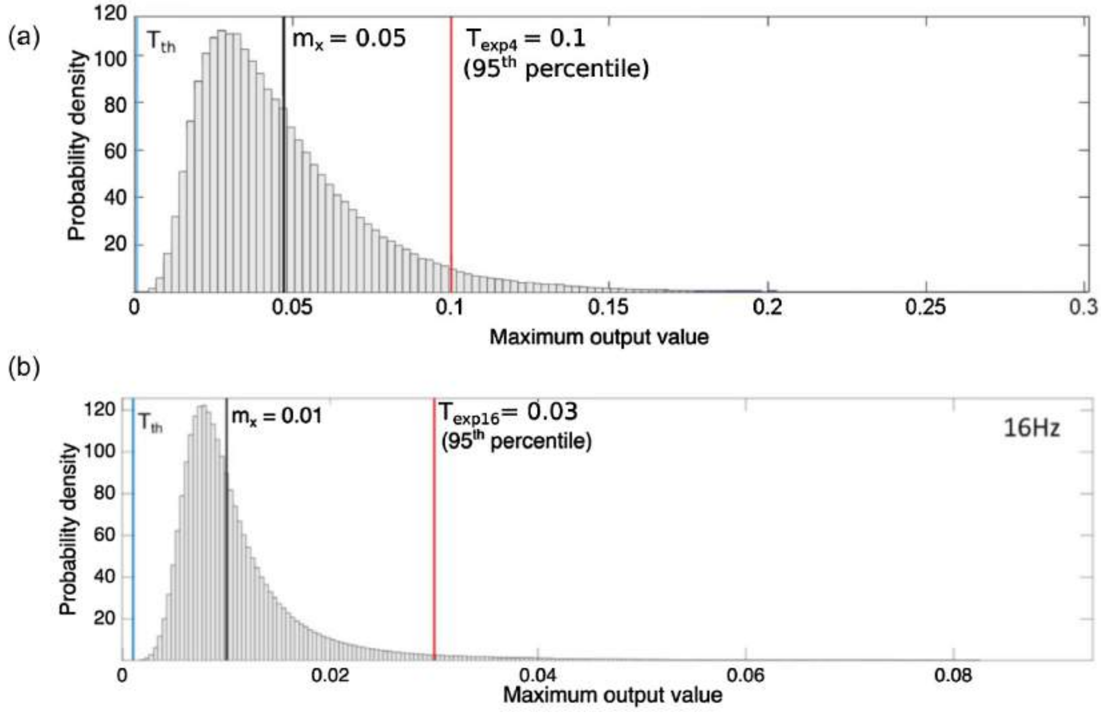

Figure 5. Distribution of the maximum values of the MCMC outputs of all windows for 26 d at 4 Hz (a) and 16 Hz (b). The theoretical threshold T_{th} and the mean of the distribution m_x are represented by the blue line and the black line, respectively. The chosen threshold T_{exp} (95th percentile) is the red line.

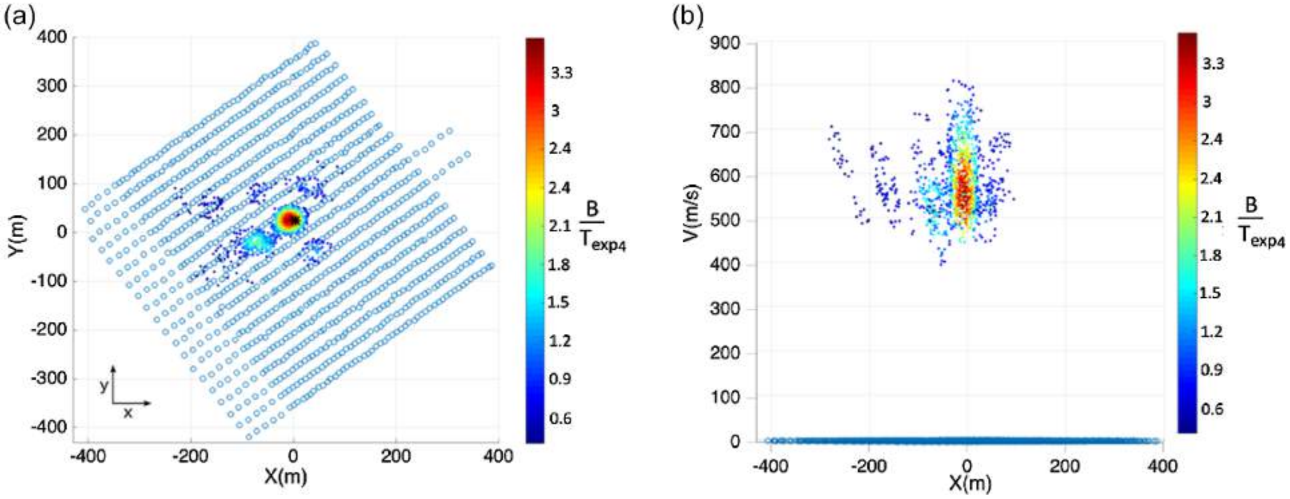

Figure 6. MCMC output for a shot according to X and Y positions an apparent velocity. The dots correspond to the candidate source position explored by the MCMC scheme after convergence of the simulated annealing. Colours represent the output value associated with each trial source position normalized by the detection threshold. The dots with the maximum values give the most probable source location and apparent velocity. (a) 2-D view showing the epicentral position of the output. The diameter at −3 dB is around 40 m. A black star represents the known position of the source. (b) 2-D view of the output according to apparent velocity and X direction. The colourbar corresponds to the values of the Bartlett operator B normalized by the experimental threshold T_{exp4}.

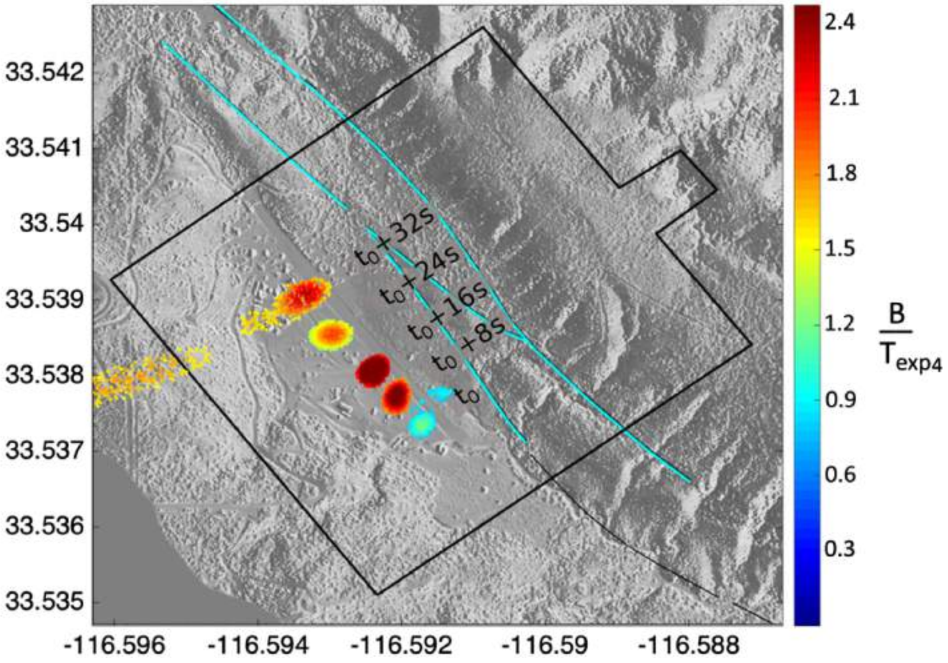

Figure 7. MCMC output for an anthropogenic source according to longitude and latitude positions for windows taken every 8 s. The colourbar correspond to the values of the Bartlett operator B normalized by the experimental threshold T_{exp4}.

the 95th percentile, ensuring that only 5 per cent of the windows are above the thresholds. In this case, $T_{\mathrm{exp4}} = 0.1$ and $T_{\mathrm{exp16}} = 0.03$. This results in around 55 200 and 167 000 localizations at 4 and 16 Hz, respectively. A more conservative approach can use higher threshold (e.g. 99th percentile). The large number of localizations is the main advantage of this scanning process since information about the subsurface geophysical structure can now be obtained from the statistics of the localizations and not only from the spatial accuracy of one single detection. Considering the short duration of each time window in the MFP process (1 s at 4 Hz and 0.25 s at 16 Hz), events with duration longer than the time window may sometimes be detected and located several times. Discriminating between localizations triggered by a long-duration event and localizations related to successive short duration events is not trivial. We choose to consider each localization separately, keeping in mind that they may originate from the same source.

It is possible to compare the empirical threshold to the theoretical limit of incoherent noise level for an array of N sensors. This corresponds to the power reduction associated with averaging N realizations of incoherent signals and is equal to $T_{\mathrm{th}} = 1/N$ (blue line in Fig. 5). This threshold is one or two order of magnitude below the mean of the 4 and 16 Hz MFP output distributions. This means that most of the noise recorded by neighbouring sensors is spatially coherent, which can be explained by the density of the array. This

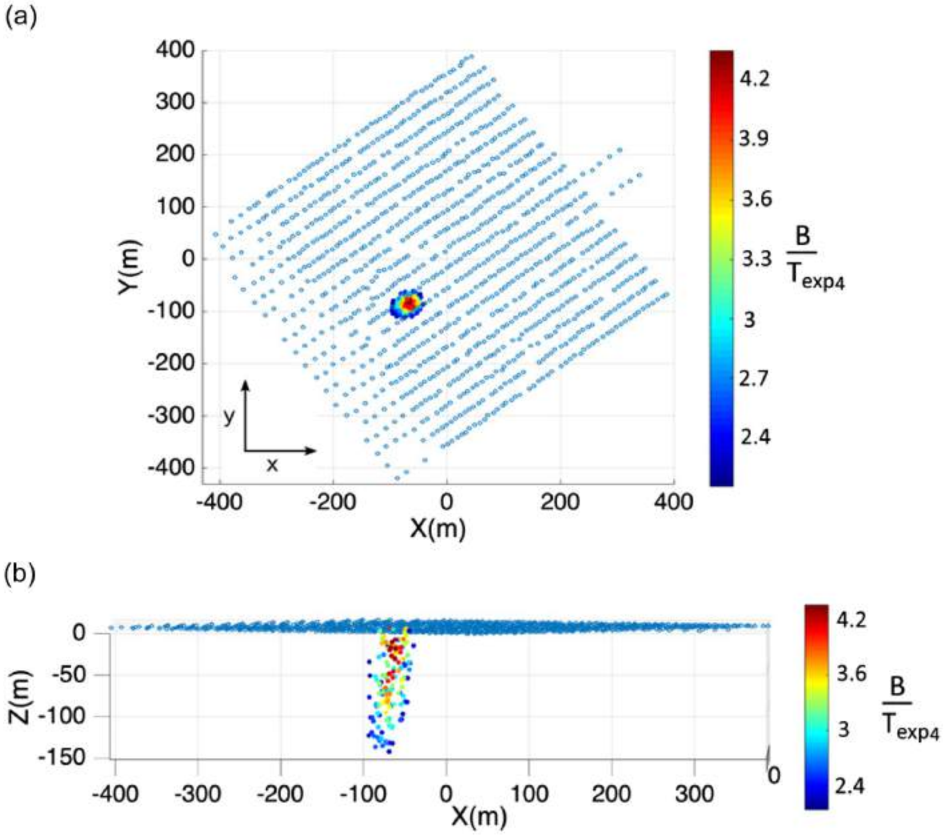

Figure 8. MCMC output for an anthropogenic source according to X and Y position and depth. The velocity of the model is fixed at 550 m s^{-1} (mean of the velocity of the XYV study). The output according to X and Y shown in (a) is of the same size as the previous study. (b) 2-D view according to X and Z. The uncertainty on the depth of the source displayed in (b) is greater than the uncertainties on X and Y, due to the position of the sensors at the surface and the simple used homogeneous model. The colourbar corresponds to the values of the Bartlett operator B normalized by the experimental threshold T_{exp4}.

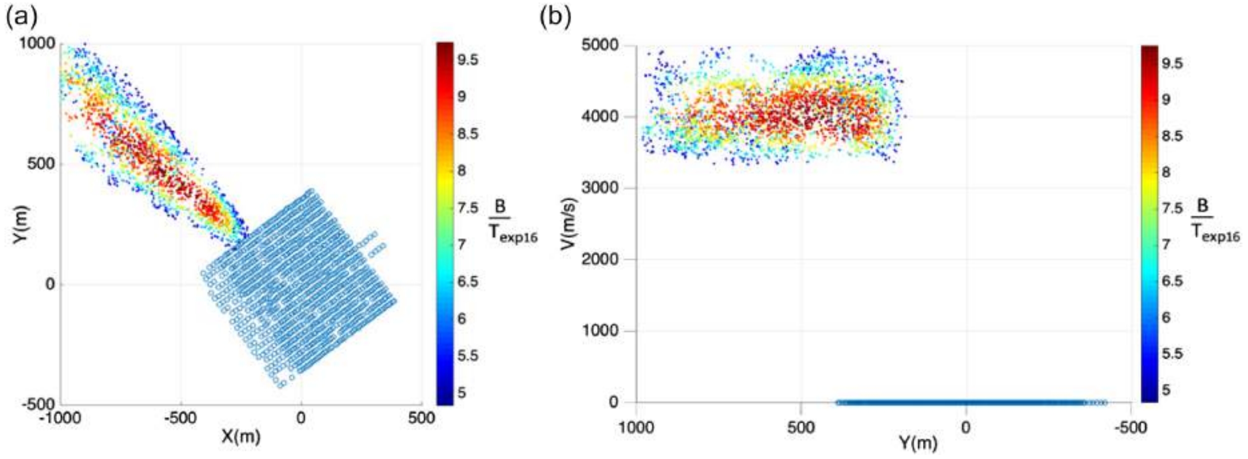

Figure 9. MCMC output for a deep source according to X and Y position and apparent velocity. The output according to X and Y presented in (a) is elongated in the radial direction, a consequence of the source being outside of the array. The apparent velocity displayed in (b) is much higher than for the two previous events, pointing at a source at depth. The colourbar corresponds to the values of the Bartlett operator B normalized by the experimental threshold T_{exp16}.

also explains the difference between the chosen thresholds at 4 and 16 Hz. At 4 Hz, the wavelength is longer and the noise will be more spatially coherent, resulting in higher MFP values.

In this section, three examples of localizations that are representative of the MFP results are investigated, with two sources at the surface and one source at depth. Further seismic interpretation of the statistics of the MFP outputs will be described in a future paper.

The first example is a Betsy shot, for which the positions (Fig. 1c) and time of occurrence are known. Fig. 6 presents results for a shot that was fired at the centre of the array, during Julian day 154 (see 1 s time windows in Fig. 3a). Because shots are surface sources, the frequency of investigation in this case is 4 Hz as discussed previously. MFP was developed for monochromatic source detection and successfully applied to geophysics to locate incoherent sources of

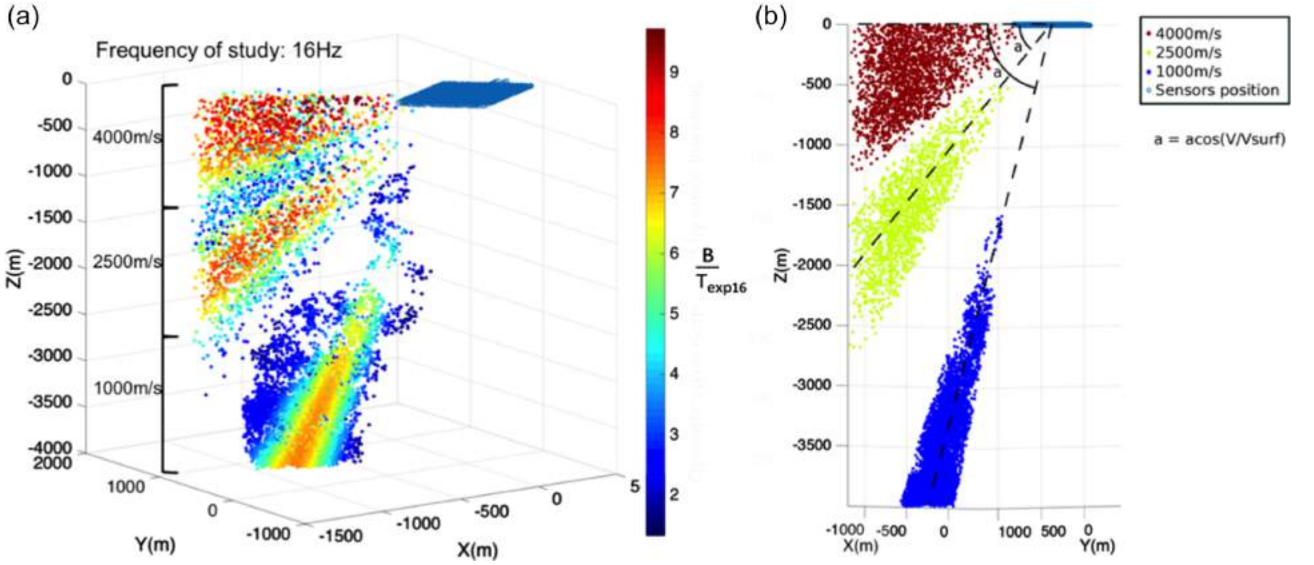

Figure 10. Dipping of the output according to the velocity of the medium and associated beamformer value. (a) Shows three clusters that correspond to three different outputs computed for a deep source. They were obtained by extracting for X, Y and Z position while successively using a fixed medium velocity of 1000, 2500 and 4000 m s^{-1}. The dipping angle depends on the ratio between the apparent velocity measured at the surface and the velocity of the medium, as is shown in the side view displayed in (b). The colourbar corresponds to the values of the Bartlett operator B normalized by the experimental threshold T_{exp16}.

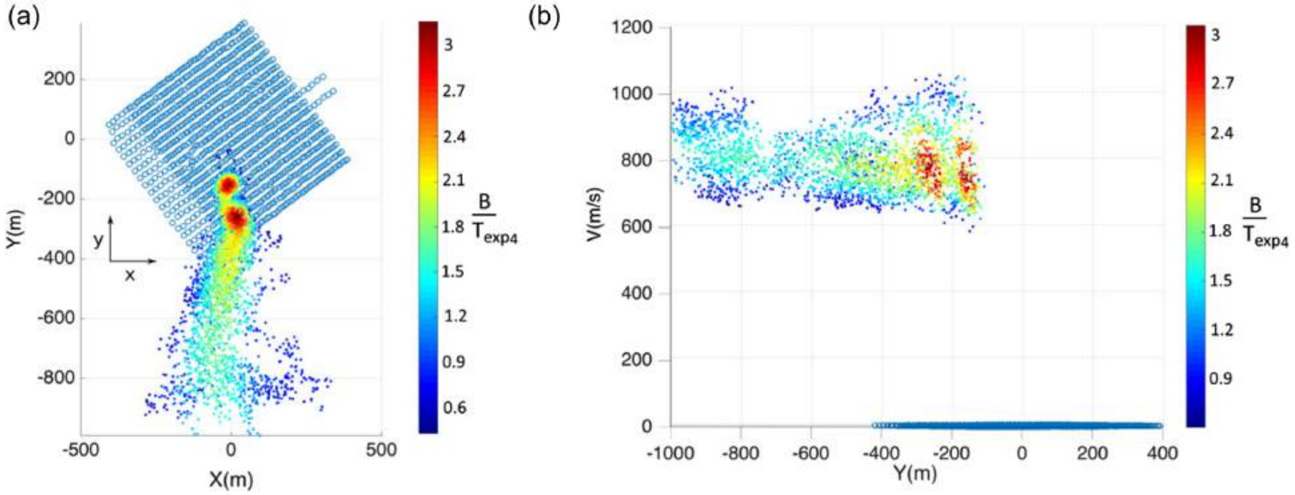

Figure 11. MCMC output for another window containing signal from the anthropogenic source according to X, Y and apparent velocity. (a and b) Show, respectively, the hypocentral position and a 2-D view of the output according to apparent velocity and Y. The results illustrate the limits of using the homogeneous model as a replica. Two distinct maxima are observed, which could either be interpreted as two distinct sources or as reflecting distortion of the wavefield by the heterogeneous medium. The colourbar corresponds to the values of the Bartlett operator B normalized by the experimental threshold T_{exp4}.

longer duration. Localizing a Betsy gunshot for which we know the position and time of the source, proves the ability of the method to detect impulsive events as well.

The maximum output is around three times higher than the noise threshold. This is a reasonable value considering the simple homogeneous model used in the calculations. The radius of the main spot gives an uncertainty of ±20 m on source position. The source localization is accurate within the uncertainty limit, the actual shot position being in the spot around the most probable source position. Given the uncertainties, the apparent velocity associated with the position of the maximum output, i.e. 580 m s^{-1}, is consistent with the phase velocity expected at the surface of the SGB site at 4 Hz (Roux *et al.* 2016; Mordret *et al.* 2019). Side lobes associated with the array response are also present. The method shows encouraging results for controlled sources such as shots.

When comparing MFP to time-domain stacking techniques, we find that the resolution and stability of the outputs are far better with MFP (Supporting Information Fig. S1). Two of the time domain stack results were obtained by shifting and stacking the time traces. The first one was filtered between 3 and 17 Hz (Supporting Information Fig. S1a) and the second between 3 and 5 Hz (Supporting Information Fig. S1c). The 3–17 Hz frequency bandwidth corresponds to the frequency range in our study, while the 3–5 Hz bandwidth is used to reproduce similar conditions to the MFP, which is calculated at 4 Hz (Supporting Information Fig. S1d). An additional comparison with an incoherent stacking, where the envelope

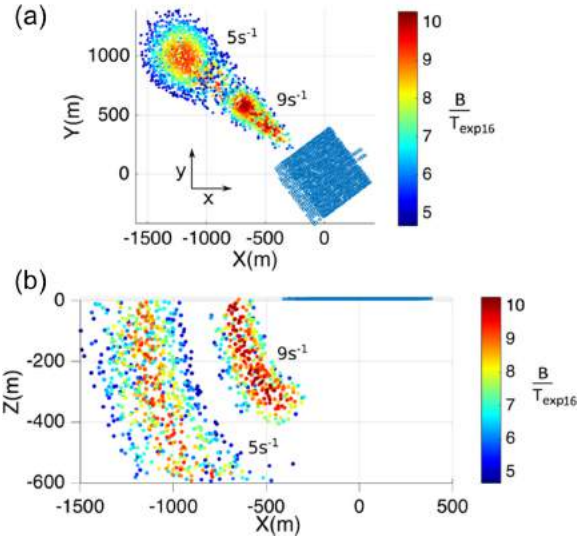

Figure 12. MCMC outputs for the deep source according to X, Y and Z for a medium with a fixed vertical velocity gradient. (a) 2-D view of the epicentral position of the output. (b) 2-D view according to X and Z. Two different outputs, for gradients of 5 and 9 s^{-1} are shown simultaneously in the two panels. In both cases the surface velocity is 800 m s^{-1}. The results show a substantial improvement in depth resolution. The colourbar corresponds to the values of the Bartlett operator B normalized by the experimental threshold T_{exp16}.

of the trace is used as an input, was also performed (Supporting Information Fig. S1b), indicating very poor resolution capability.

Fig. 7 presents a part of a series of events detected in successive 1-s time windows. The figure shows different outputs represented according to their epicentral position. Beside each output is the time relative to the first localization at time t_0. Localizations were performed with the first extraction parameters, and indicate a motion of the source in the selected time-windows with a mean velocity of 20 km h^{-1}. The source, possibly a moving vehicle, is considered to be anthropogenic. This signal is different than the shot discussed previously in terms of duration and frequency content. However, the uncertainties for the anthropogenic source are similar to those of the shot. This is expected, because we are working with narrow frequency bands and short time windows. The sizes of the spots around the most probable position remain comparable (around 40 m in diameter) and the associated apparent velocities are between 500 and 600 m s^{-1}.

A 3-D localization of this source for a window at $t_0 +$ 10 s was performed with the other extraction parameters to check consistency of our results. The velocity was fixed at 550 m s $^{-1}$, corresponding to the mean surface wave velocity at 4 Hz. The extraction output represented in Fig. 8, only for values higher than 50 per cent of the maximum beamformer for clarity, confirms that the source is located at the surface. Comparing the spots size for the three dimensions indicates that the uncertainty on depth is three times larger than the uncertainty on the epicentral location.

Fig. 9 shows results associated with a subsurface source analysed using a frequency of 16 Hz. The MFP output was computed for the 0.25-s time window in Fig. 4 (deep source) with the first extraction parameters. We observe a smearing of the PDF for the source position in the radial direction. This phenomenon is classical in array analysis for an event located outside of the array. The apparent velocity associated with the maximum output is 4000 m s^{-1}. This value is far too high to describe wave propagation near the surface, leading to the conclusion that the waves emitted by this source come from depth. This confirms that it is possible to detect and obtain an initial approximation of the source position without detailed analysis of depth.

Fig. 10 shows the MCMC outputs computed with X, Y and Z as extraction parameters together with a choice of different medium velocities V: 1000, 2500 and 4000 m s^{-1}. We observe a dipping of the PDF as the velocity V gets lower. This is consistent with Snell's law, which links the dipping angle of the MFP output, α, to the apparent velocity V_{ap}, and homogeneous velocity in the medium, such that $\alpha = \mathrm{acos}(V/V_{ap})$. The interdependence between MFP parameters is a classical issue when trying to extract depth information as well as medium velocity. Fig. 10 shows only a 25 per cent variation of the MFP output maximum between 0 and 4000 m depth, which is not sufficient to conclude on the source position.

As expected from a surface array, the resolution obtained when using a replica computed with a homogeneous velocity model does not allow determining the depth of a shallow seismic source and the velocity in the medium. This is essentially due to the trade-off between depth and velocity, but also because the medium under the array is complex. For the 3-D extraction, this makes the choice of a single velocity in the model difficult to determine. In light of this, the use of the apparent velocity as a proxy for depth and medium velocity is justified.

5 DISCUSSION AND CONCLUSIONS

This study of source localization in continuous seismic waveforms recorded by a dense array at the surface shows that it is possible to detect and locate both surface and sub-surface events by combining capabilities of the MFP and MCMC sampling. The MFP technique was developed originally for simpler applications of acoustic waves propagating in the ocean (e.g. Baggeroer Kuperman & Mikhalevsky 1993; Kuperman & Turek 1997). Application of the method for localization of near-surface sources recorded by seismic arrays around a fault zone is considerably more challenging because the velocity structure of the medium is more complex and a wide variety of noise sources contribute to the recorded ground motion (e.g. Riahi & Gerstoft 2016; Meng & Ben-Zion 2018a,b, Li *et al.* 2018). The 26-day continuous recording of the data also requires an efficient localization tool. To address these difficulties, we augmented the method for localization of sources at and below the surface of the earth by using the frequency of study as a 'filter' separating sources at the surface and at depth. This allows us to keep an analytic homogeneous model depending on only three parameters insuring the efficiency of the algorithm. The use of MCMC further reduces the computational time compared to grid search approaches and provides statistical information that can be useful when interpreting the data.

This methodology, introduced in section 2, gives access to the PDF of parameters that describe the results. It provides estimates of the source position and apparent velocity, as well as a confidence interval for the solution. In addition, our approach allows discrimination between surface and deep sources without determining the actual position of the source. Due to spatial sampling requirement, the choice of the frequency has an impact on the depth range of the detections: surface sources are detected at lower frequencies than deep sources, because the apparent velocity measured at the array increases with depth.

Surface sources such as shots and moving vehicles were successfully detected and located when extracting the epicentral position and the apparent velocity under the array. In those cases we obtain a good resolution for epicentral coordinates, when the events are occurring inside the array, and velocities that are consistent with previous studies. A potential shallow source located outside the array was also successfully detected based on the same model parameters. However, the use of a homogeneous velocity model to compute the replicas has limitations for source localization. First, the resolution of depth is not sufficient to determine the position of sources below the surface. The trade-off between depth and apparent velocity results in a strong ambiguity of the two parameters. Second, the simplicity of the replica cannot match complex phase patterns that can be produced by strong lateral and horizontal contrasts of seismic properties and wave interferences. In the case of anthropogenic sources, time windows may show outputs with distorted shapes or multiple spots (Fig. 11). It is not always possible to conclude on whether those outputs are due to multiple/moving sources, or the inability of MFP method with a simple replica to match data associated with one source.

A simple way to address the first issue and improve resolution at depth would be to use a 1-D velocity gradient. In such a case, it is still possible to compute the replica analytically and the model is closer to reality. Fig. 12 shows two outputs computed with velocity gradients of 5 and 9 s^{-1} and velocity at the surface of 800 $\mathrm{m\,s}^{-1}$. This value corresponds to the velocity at the elevation of the lowest sensor in the average 1-D velocity model below the array derived from data of the Betsy gunshots by Meng & Ben-Zion (2018b). As a simplification, when studying sources at depth, we consider the lowest elevation to be 0-m depth and do not extract the source positions in the volume above this boundary. We observe improved resolution and higher value for the maximum of the MFP output when using a model with 9 s^{-1} gradient. This is generally consistent to the model from Meng & Ben-Zion (2018b) and imaging results obtained in the area (Hillers *et al.* 2016; Roux *et al.* 2016). Another way to reduce depth uncertainties is to use data from the borehole sensor located within the array, to better separate sources at and below the surface and constrain the vertical position of sub-surface sources. This will be done in a follow up work, along with comprehensive seismic interpretation of the entire 26-day data set.

ACKNOWLEDGEMENTS

We thank Haoran Meng and Chris Johnson for discussions. YBZ acknowledges support from the National Science Foundation (grant EAR-1818589) and the Department of Energy (award DE-SC0016520). ISTerre is part of Labex OSUG@2020.

REFERENCES

Artman, B., Podladtchikov, I. & Witten, B., 2010. Source location using time-reverse imaging, *Geophys. Prospect.,* **58**(5), 861–873.

Baggeroer, A.B., Kuperman, W.A. & Mikhalevsky, P.N., 1993. An overview of matched field methods in ocean acoustics, *IEEE J. Ocean. Eng.,* **18**(4), 401–424.

Ben-Zion, Y. *et al.*, 2015. Basic data features and results from a spatially dense seismic array on the San Jacinto fault zone, *Geophys. J. Int.,* **202**(1), 370–380.

Cesca, S. & Grigoli, F., 2015. Full waveform seismological advances for microseismic monitoring, in *Advances in Geophysics,* Vol. **56,** pp. 169–228, Elsevier Ltd, doi:10.1016/bs.agph.2014.12.002.

Chmiel, M., Roux, P. & Bardainne, T., 2016. Extraction of phase and group velocities from ambient surface noise in a patch-array configuration, *Geophysics,* **81**(6), KS231–KS240.

Corciulo, M., Roux, P., Campillo, M., Dubucq, D. & Kuperman, W.A., 2012. Multiscale matched-field processing for noise-source localization in exploration geophysics, *Geophysics,* **77**(5), KS33–KS41.

Cros, E., Roux, P., Vandemeulebrouck, J. & Kedar, S., 2011. Locating hydrothermal acoustic sources at old faithful geyser using matched field processing, *Geophys. J. Int.,* **187**(1), 385–393.

Grigoli, F., Cesca, S., Krieger, L., Kriegerowski, M., Gammaldi, S., Horalek, J., Priolo, E. & Dahm, T., 2016. Automated microseismic event location using Master-Event Waveform Stacking, *Sci. Rep.,***6**(1), 25744, doi:10.1038/srep25744.

Hauksson, E., Yang, W. & Shearer, P.M., 2012. Waveform relocated earthquake catalog for Southern California (1981 to June 2011), *Bull. seism. Soc. Am.,* **102**(5), 2239–2244.

Hillers, G., Roux, P., Campillo, M. & Ben-Zion, Y., 2016. Focal spot imaging based on zero lag cross-correlation amplitude fields: application to dense array data at the San Jacinto fault zone, *J. geophys. Res.,* **121**(11), 8048–8067.

Inbal, A., Ampuero, J.P. & Clayton, R.W., 2016. Localized seismic deformation in the upper mantle revealed by dense seismic arrays, *Science,* **354**(6308), 88–92.

Inbal, A., Cristea-Platon, T., Ampuero, J., Hillers, G., Agnew, D. & Hough, S.E., 2018. Sources of long-range anthropogenic noise in Southern California and implications for tectonic tremor detection, *Bull. seism. Soc. Am.,* **108 ,** 3511–3527.

Johnson, C.W., Meng, H., Vernon, F.L., Nakata, N. & Ben-Zion, Y., 2018. Characteristics of ground motion generated by interaction of wind gusts with trees, structures and other obstacles above the surface, in *Abstract to the annual meeting of the Southern California Earthquake Center,* Southern California Earthquake Center.

Kao, H. & Shan, S.-J., 2004. The Source-Scanning Algorithm: mapping the distribution of seismic sources in time and space, *Geophys. J. Int.,* **157**(2), 589–594.

Kuperman, W.A. & Turek, G., 1997. Matched field acoustics, *J. Eng. Appl. Sci.,***11**(1), 141–148.

Kwiatek, G. & Ben-Zion, Y., 2016. Theoretical limits on detection and analysis of small earthquakes , *J. geophys. Res.,* **121**(8), 5898–5591.

Landès, M., Hubans, F., Shapiro, N.M., Paul, A. & Campillo, M., 2010. Origin of deep ocean microseisms by using teleseismic body waves, *J. geophys. Res.,* **115**(B5), B05302, doi:10.1029/2009JB006918.

Langet, N., Maggi, A., Michelini, A. & Brenguier, F., 2014. Continuous Kurtosis-based migration for seismic event detection and location, with application to piton de la Fournaise Volcano, La Reunion, *Bull. seism. Soc. Am.,* **104**(1), 229–246.

Larmat, C., Montagner, J.-P., Fink, M., Capdeville, Y., Tourin, A. & Clévédé, E., 2006. Time-reversal imaging of seismic sources and application to the great Sumatra earthquake, *Geophys. Res. Lett.,* **33**(19), L19312,

Larmat, C., Tromp, J., Liu, Q. & Montagner, J.-P., 2008. Time reversal location of glacial earthquakes, *J. geophys. Res.,* **113**(B9), B09314, .

.Lin, F, Li, D., Clayton, R. & .Hollis, D, 2013. High-Resolution 3D shallow crustal structure in Long Beach, California: Application of ambient noise tomography on a dense seismic array, *Geophysics,* **78**(4) Q45–Q56.

Li, Z., Peng, Z., Hollis, D., Zhu, L. & McClellan, J., 2018. High-resolution seismic event detection using local similarity for Large-N arrays, *Sci. Rep.,* **8**(1), 1646, doi:10.1038/s41598-018-19728-w.

Meng, H. & Ben-Zion, Y., 2018a. Characteristics of airplanes and helicopters recorded by a dense seismic array near Anza California, *J. geophys. Res.,* **123,** 4783–4797.

Meng, H. & Ben-Zion, Y., 2018b. Detection of small earthquakes with dense array data: example from the San Jacinto fault zone, southern California, *Geophys. J. Int.,* **212**(1), 442–457.

Mordret, A., Roux, P., Boué, P. & Ben-Zion, Y., 2019. Shallow 3-D structure of the San Jacinto Fault zone revealed from ambient noise imaging with a dense seismic array, *Geophys. J. Int.,* **216,** 896–905.

Moreau, L., Hunter, A., Velichko, A. & Wilcox, P., 2014. 3-D reconstruction of sub-wavelength scatterers from the measurement of scattered fields in elastic waveguides, *IEEE Trans. Ultrason. Ferroelectr. Freq. Control,***61**(11), 1864–1879.

Poiata, N., Satriano, C., Vilotte, J.P., Bernard, P. & Obara, K., 2016. Multi-band array detection and location of seismic sources recorded by dense seismic networks, *Geophys. J. Int.,* **205**(3), 1548–1573.

Qin, L., Ben-Zion, Y., Qiu, H., Share, P.-E., Ross, Z.E. & Vernon, F.L., 2018. Internal structure of the San Jacinto fault zone in the trifurcation area southeast of Anza, California, from data of dense seismic arrays, *Geophys. J. Int.,* **213**(1), 98–114.

Riahi, N. & Gerstoft, P., 2015. The seismic traffic footprint: tracking trains, aircraft, and cars seismically, *Geophys. Res. Lett.,* **42**(8), 2674–2681.

Riahi, N. & Gerstoft, P., 2016. Locating sources in a dense array through network-based clustering, in *2016 Information Theory and Applications Workshop (ITA),* pp. 1–8, IEEE, doi:10.1109/ITA.2016.7888149.

Richardson, A.M. & Nolte, L.W., 1991. A posteriori pobability source localization in an uncertain sound speed, deep ocean environment, *J. acoust. Soc. Am.,* **89**(5), 2280–2284.

Ross, Z.E., Hauksson, E. & Ben-Zion, Y., 2017. Abundant off-fault seismicity and orthogonal structures in the San Jacinto fault zone, *Sci. Adv.,* **3**(3), e1601946, doi:10.1126/sciadv.1601946.

Rost, S. & Thomas, C., 2002. Array seismology: methods and applications, *Rev. Geophys.,* **40**(3), 1008, doi:10.1029/2000RG000100.

Roux, P., Moreau, L., Lecointre, A., Hillers, G., Campillo, M., Ben-Zion, Y., Zigone, D. & Vernon, F., 2016. A methodological approach towards high-resolution surface wave imaging of the San Jacinto Fault Zone using ambient-noise recordings at a spatially dense array, *Geophys. J. Int.,* **206**(2), 980–992.

Sambridge, M. & Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems, *Rev. Geophys.,* **40**(3), 1009, doi:10.1029/2000RG000089.

Tollefsen, D. & Dosso, S.E., 2014. Three-dimensional source tracking in an uncertain environment, *J. acoust. Soc. Am.,* **125**(5), 2909–2917.

Vandemeulebrouck, J., Roux, P. & Cros, E., 2013. The plumbing of old faithful geyser revealed by hydrothermal tremor, *Geophys. Res. Lett.,* **40**(10), 1989–1993.

Zigone, D., Ben-Zion, Y., Lehujeur, M., Campillo, M., Hillers, G. & Vernon, F.L., 2019. Imaging subsurface structures in the San Jacinto fault zone with high frequency noise recorded by dense linear arrays, *Geophys. J. Int.,* **217,** 879–893.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Figure S1. MCMC output for a shot relocated with time-domain stacking. (a–c) and MFP (d). For the time-domain stacking, normalized time signals were shifted according to a time delay computed from the candidate time of origin, apparent velocity, and epicentral position. The shifted time traces were subsequently stacked and divided by the number of traces. The maximum of the stack is the output of the process. (a) 2-D view of the epicentral position of the time stack output. The time domain signals are filtered between 3 and 17 Hz and normalized. (b) 2-D view of the epicentral position of the incoherent time stack output. Here the normalized STA/LTA of the signals are stacked. (c) 2-D view of the epicentral position of the time-stack output. Time-domain signals are filtered between 3 and 5 Hz and normalized. (d) 2-D view of the epicentral position of the output of the MFP at 4 Hz.

Please note: Oxford University Press are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.