Petersen, G., Niemz, P., Cesca, S., Mouslopoulou, V., Bocchini, G. M. (2021): Clusty, the waveform-based network similarity clustering toolbox: concept and application to image complex faulting offshore Zakynthos (Greece). - Geophysical Journal International, 224, 3, 2044-2059.

https://doi.org/10.1093/gji/ggaa568

Geophysical Journal International

Geophys. J. Int. (2021) **224,** 2044–2059
Advance Access publication 2020 November 25
GJI Seismology

doi: 10.1093/gji/ggaa568

Clusty, the waveform-based network similarity clustering toolbox: concept and application to image complex faulting offshore Zakynthos (Greece)

G.M. Petersen,[1,2] P. Niemz,[1,2] S. Cesca,[1] V. Mouslopoulou[3] and G.M. Bocchini[4]

[1] *GFZ German Research Centre for Geosciences, Potsdam, Germany. E-mail: gesap@gfz-potsdam.de*
[2] *Institute of Geosciences, University of Potsdam, Potsdam, Germany*
[3] *National Observatory of Athens, Institute of Geodynamics, Athens* 11810*, Greece*
[4] *Ruhr University of Bochum, Institute of Geology, Mineralogy and Geophysics, Germany*

Accepted 2020 November 23. Received 2020 November 10; in original form 2020 July 30

SUMMARY

Clusty is a new open source toolbox dedicated to earthquake clustering based on waveforms recorded across a network of seismic stations. Its main application is the study of active faults and the detection and characterization of faults and fault networks. By using a density-based clustering approach, earthquakes pertaining to a common fault can be recognized even over long fault segments, and the first-order geometry and extent of active faults can be inferred. Clusty implements multiple techniques to compute a waveform based network similarity from maximum cross-correlation coefficients at multiple stations. The clustering procedure is designed to be transparent and parameters can be easily tuned. It is supported by a number of analysis visualization tools which help to assess the homogeneity within each cluster and the differences among distinct clusters. The toolbox returns graphical representations of the results. A list of representative events and stacked waveforms facilitate further analyses like moment tensor inversion. Results obtained in various frequency bands can be combined to account for large magnitude ranges. Thanks to the simple configuration, the toolbox is easily adaptable to new data sets and to large magnitude ranges. To show the potential of our new toolbox, we apply Clusty to the aftershock sequence of the M_w 6.9 25 October 2018 Zakynthos (Greece) Earthquake. Thanks to the complex tectonic setting at the western termination of the Hellenic Subduction System where multiple faults and faulting styles operate simultaneously, the Zakynthos data set provides an ideal case-study for our clustering analysis toolbox. Our results support the activation of several faults and provide insight into the geometry of faults or fault segments. We identify two large thrust faulting clusters in the vicinity of the main shock and multiple strike-slip clusters to the east, west and south of these clusters. Despite its location within the largest thrust cluster, the main shock does not show a high waveform similarity to any of the clusters. This is consistent with the results of other studies suggesting a complex failure mechanism for the main shock. We propose the existence of conjugated strike-slip faults in the south of the study area. Our waveform similarity based clustering toolbox is able to reveal distinct event clusters which cannot be discriminated based on locations and/or timing only. Additionally, the clustering results allows distinction between fault and auxiliary planes of focal mechanisms and to associate them to known active faults.

Key words: Persistence, memory, correlations, clustering; Seismicity and tectonics; Fractures, faults, and high strain deformation zones.

1 INTRODUCTION

The world-wide increasing number of seismic stations, even deployed in areas of moderate seismicity, significantly lowers earthquake detection thresholds. This enables seismologists to study spatial and temporal seismicity patterns in great detail. In general, earthquakes occur along pre-existing faults. Both, the extent and the stress state of seismogenic faults are of interest for structural studies and for seismic hazard assessment at local, regional or global scale. The association of seismic events to faults is a major

but also challenging task. Depending on the location of active faults, fault identification may involve field investigations (i.e. mapping, trenching, etc.), aerial investigations (analysis of satellite images or air-born lidar) or seismic-reflection/bathymetric data. Following large-magnitude earthquakes ($M_w > 6$), the geometry of the fault associated with the main rupture, as well as its slip distribution, is often estimated using seismological and geodetic tools (e.g. Koper *et al.* 2011; Yokota *et al.* 2011; Grandin *et al.* 2015; Cirella *et al.* 2020).

Moment tensor inversion represents a powerful tool to identify earthquake faulting mechanisms. Focal mechanisms obtained for seismic sequences are often used to obtain insight into the faulting style and the extent of an active fault (e.g. Örgülü & Aktar 2001; Serpetsidaki *et al.* 2010; Asano *et al.* 2011; Herrmann *et al.* 2011) or the geometry of multiple faults (e.g. Cesca *et al.* 2017). Although moment tensor inversion provides valuable insights, it has several limitations that complicate the identification of active faults. First, robust moment tensor inversions require a detailed knowledge of velocity structures and station instrumentation. Furthermore, the quality of moment tensor solutions strongly depends on the radiated frequencies: for lower magnitude events moment tensor inversion is often not feasible. In these cases, the signal to noise ratio is only sufficient at higher frequencies which cannot be modelled using simple 1-D velocity models. Finally, the causative fault plane cannot be distinguished from the auxiliary plane of the moment tensor (MT) without additional geological (e.g. fault geometry) or geophysical constraints (e.g. GPS displacements or aftershock distributions).

The clustering of earthquakes into groups of similar events is another approach to analyse the observed seismicity regarding underlying seismogenic processes. The clustering analysis can be based on various parameters such as: (1) spatial and/or temporal distributions (e.g. Frohlich 1987; Shearer *et al.* 2005; Ansari *et al.* 2009; Ouillon & Sornette 2011; Mouslopoulou & Hristopulos 2011; Mesimeri *et al.* 2019; Czecze & Bondár 2019); (2) the smallest rotation between moment tensors (e.g. Cesca 2020); (3) *P* and *S* polarities (e.g. Shelly *et al.* 2016) or (4) waveform similarities, as for example in Tsujiura (1983), Maurer & Deichmann (1995), Shearer *et al.* (2003), Barani *et al.* (2007), Trugman & Shearer (2017), Ruscic *et al.* (2019), Abramenkov *et al.* (2020) and in this study.

The clustering based on waveform similarity favours fault mapping by considering locations and mechanisms, since waveforms are inherently sensitive to both. Waveform similarity is generally assessed by cross-correlating waveforms of earthquakes at one or multiple stations. Very high waveform similarities (i.e. >0.9–0.95) are attributed to so-called repeaters (e.g. Geller & Mueller 1980; Igarashi *et al.* 2003; Baisch *et al.* 2008; Han *et al.* 2014). According to Geller & Mueller (1980) repeaters are located at distances smaller than a quarter of the dominant wavelength, however, also larger spatial separation was reported (e.g. Arrowsmith & Eisner 2006). Similar waveforms, observed at multiple stations, imply similar focal mechanisms and travel paths (locations and depths, e.g. Maurer & Deichmann 1995). Thus, the identification of clusters of similar events can shed light on the fault geometry and on the faulting style. In favourable conditions, waveform similarity studies can help to identify faults and map their geometries (e.g. Tsujiura 1983; Maurer & Deichmann 1995; Shearer *et al.* 2003). The waveform similarity based clustering approach is independent from the uncertainty of the hypocentral locations, therefore it can be applied even when hypocentral locations are poorly constraint. Only at a later stage of this study, when fault planes are inferred from the clusters, the location uncertainties are considered. Waveform similarity is also used to identify groups of events for relative relocation methods (e.g. Shearer *et al.* 2005; Trugman & Shearer 2017). High waveform similarities among a small magnitude foreshock or afterhock with a larger main shock with a known focal mechanism can be used to infer a similar mechanism for the weaker event. Such analyses can also be used for a more advanced declustering of a catalogue, not only relying on occurrence times (Barani *et al.* 2007), as well as for determining event pairs for an empirical green's function analysis.

Here we use a density-based clustering approach, which allows grouping earthquakes with a wide range of magnitudes, locations and focal mechanisms. End members of a density-cluster are not required to be as similar as neighbouring events, if they are connected via multiple events with gradually changing locations or mechanisms. Consequently, we are able to assign individual earthquakes assumed to be produced along an elongated fault into a single cluster.

Here, we introduce a new open-source, user-friendly and highly adaptable waveform clustering toolbox, named *Clusty*. The toolbox allows correlating and clustering hundreds to few thousands of events recorded by a network of stations based on what we refer to as the *network similarity* of the event pairs. We implemented different approaches to combine the waveform similarities computed for multiple stations across a network, allowing a comparison of the clustering methods and their results. In the development of the code we put emphasis not only on computational efficiency and the stability of results, but also on a broad range of analysis and plotting tools. Apart from the resulting catalogue of clustered events and accompanying plots, Clusty provides a list of representative events, i.e. one event for each cluster that is most similar to the rest of the cluster. The representative events can be used to perform moment tensor inversions aiming for a representative focal mechanism for each cluster.

In this study we apply the clustering toolbox to the aftershock sequence of the 25 October 2018 M_w 6.9 Zakynthos (Greece) earthquake (Chousianitis & Konca 2019; Cirella *et al.* 2020; Ganas *et al.* 2020; Karakostas *et al.* 2020; Mouslopoulou *et al.* 2020; Sokos *et al.* 2020). The data set includes >2300 events with $M \geq 2.8$ recorded at 33 stations from 25/10/2018 to 14/11/2019. The catalogue is available in Mouslopoulou *et al.* (2020). Zakynthos is located in the proximity of the western termination of the Hellenic subduction zone. The region is known for its high seismic activity and a great variety of faulting mechanisms (Mouslopoulou *et al.* 2020). Serpetsidaki *et al.* (2010) studied another seismic sequence offshore Zakynthos in April 2006 and emphasized the importance of the identification of active faults for regional seismic hazard assessment. Our waveform-based clustering analysis provides a better understanding of the geometries and kinematics of the faults involved in the 2018–2019 aftershock sequence. Further, we associate moment tensors inverted for representative events to the individual clusters. The identification of different waveforms excited by spatially clustered earthquakes provides evidence for the presence of various faulting styles on neighbouring faults, an outcome that is in agreement with the local geology (Mouslopoulou *et al.* 2020) and the regional stress field (Konstantinou *et al.* 2017).

We use the Zakynthos application to assess the stability of the clustering results, using different clustering settings, frequency ranges and discuss limits and opportunities of the toolbox. In Section 2, we describe the work-flow of the clustering toolbox Clusty. We applied our toolbox to the Zakynthos Earthquake aftershock case study and present the results of the clustering analysis in Section 3. We discuss both, the methods and the application with respect to the clustering results, inferred fault geometries and methodological limitations in Section 4.

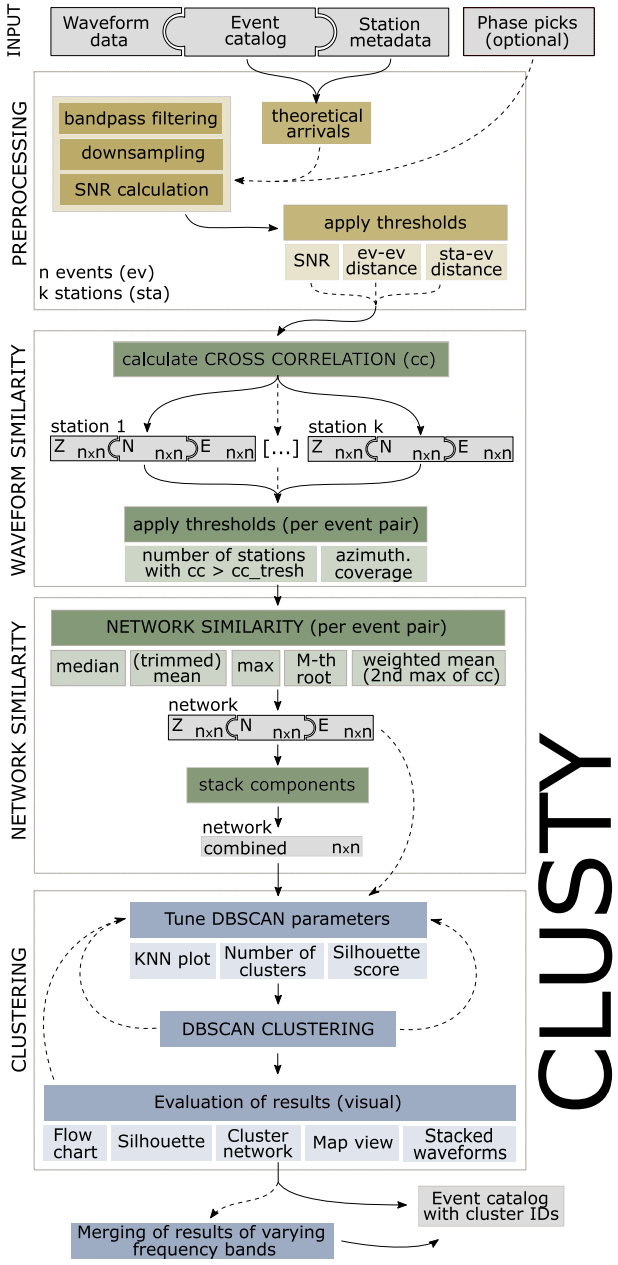

Figure 1. Schematic workflow diagram for the waveform-based network similarity clustering toolbox *Clusty*.

2 METHODOLOGY: THE CLUSTY NETWORK SIMILARITY CLUSTERING TOOLBOX

Clusty is a flexible, efficient and user-friendly *python* toolbox dedicated to seismic cluster analysis based on waveform similarity across a network of stations. It is based on the seismological *python* library Pyrocko (Heimann *et al.* 2017) and is running on Linux systems including desktop and server environments.

The general workflow is sketched in Fig. 1. As input Clusty requires an earthquake catalogue, waveform data and station metadata. If phase picks are not available, it is possible to compute theoretical arrival times using a chosen 1-D velocity model and *cake*, a tool implemented in *Pyrocko* to solve ray theory problems for layered earth models (Heimann *et al.* 2017). The user can either select a fixed time window for each phase or use our empirical relations (i.e. for surface waves: $[t_{\mathrm{onset}} - 10s,\ t_{\mathrm{onset}} + (3/f_{\mathrm{min}}) + 10s]$ and for body waves: P: $[t_P - 2,\ t_S]$, S: $[t_S - 2,\ 1.5(t_S - t_P)]$). Clusty preprocesses the waveforms, that is downsampling and bandpass filtering, and applies thresholds for inter-event distances (either epicentral (in this study) or hypocentral), event-to-station distances and signal-to-noise ratio (SNR). While none of these thresholds are strictly required, we recommend using them for computational efficiency. Distance-based thresholds should be set conservatively to avoid rejecting event pairs or station-event-pairs due to mislocated events. A minimum station–event distance is recommended as the clustering method assumes a station–event distance that is large compared to the event–event distance.

The workflow can be quickly adjusted to three channel or single channel data. For all event pairs passing the thresholds, Clusty computes the maximum cross-correlation coefficient (*cc*) at each station and for each component. For computational efficiency, this step runs in parallel on a user-defined number of cores on the CPU. Only event-pairs, that exceed an additional *cc* threshold (e.g. >0.7) at a minimum number of stations (e.g. >5), with a minimum azimuthal station coverage (e.g. $>60^{\circ}$) are considered in the subsequent analysis. However, it is important to notice that once these conditions are satisfied, all stations which passed the primary SNR and distance thresholds (and not only those passing the *cc* threshold) will be considered for the network similarity computation to assure that the statistics are not biased. The *cc* threshold does not represent a measure of the minimum similarity among events in the later applied clustering process. It only assures higher computational efficiency.

By applying the above mentioned thresholds, we reduced the number of calculated cross-correlations in our test data set from more than 378 million (45 stations, 3 components, 2367 events) to about 5 million. The pre-processing of the waveforms and the calculation of the *cc* values is the computationally most expensive step within Clusty. In the frequency band of 0.05–0.20 Hz (allowing a downsampling to 10 Hz) it takes about 4 hr on a cluster using 16 cores. All further steps within Clusty require only a few minutes on a single core. A memory saving option is available, so that Clusty can also be used on personal computers.

The *cc* values of the event pairs at each station are combined to a network similarity for each component using one of the methods described below. Subsequently, the components can be combined or analysed individually, for example to compare the results obtained from horizontal and vertical components. The network similarity matrix is then used as input for the clustering algorithm DBSCAN (Ester *et al.* 1996, see Section 2.2). The choice of appropriate clustering parameters is often difficult and sometimes subjective. To overcome these difficulties we implemented tools for testing various sets of clustering parameters and compare them using multiple analysis plots (see Section 2.3).

To analyse earthquakes with a broad range of magnitudes, the entire workflow can be repeated using different frequency bands. The resulting cluster labels can be harmonized with respect to a defined reference frequency band to create a joint cluster result catalogue. The user can run *Clusty* in one flow, but tuning the settings of the network similarity computation and the clustering parameters is a crucial point in this analysis. Therefore all steps can be run separately and repeated (e.g. computation of *cc*, network similarity, clustering, plotting of the results).

Settings for the toolbox like frequency filters, downsampling, SNR thresholds to retain or reject events as well as the choice of methods to compute the network similarity can be defined in a

configuration file. We provide an example configuration file in the git repository including information on the settings used for this study (https://git.pyrocko.org/clusty/clusty, last accessed October 2020). Clusty returns several figures to evaluate and present the results together with the output from the cluster analysis (i.e. clustering matrices and event-cluster identification). In addition, Clusty can provide stacked waveforms for each cluster as well as a list of representative events for subsequent studies.

2.1 Network similarity computation

The network similarity *nsim* of two events with index *i* and *j* (event pair *ij*) across a network of stations *s* with components *c* can be computed based on the maximum cross-correlation coefficients $cc_{ij,c,s}$ using a variety of methods implemented within the clustering toolbox to allow an easy comparison of different techniques. The network similarity of each event-pair is a value between 0 and 1, with 1 being the highest correlation.

For each pair of events *i*, *j* the maximum, the mean or the median of the $cc_{ij,c,s}$ value of all stations *s* (separate components *c*) can, among other methods, be used as a measure for network similarity. These three methods are computationally very efficient. However, the mean of the *cc* values of all stations is generally prone to outliers especially when calculated from a small sample of events, while the maximum of the *cc* values can be distorted in case of highly correlated monotonous noise or band-limited stations, for example due to high near-surface attenuation (Aster & Scott 1993). Moreover, the maximum-method is based on the *cc* value of a single station and cannot separate two different mechanisms which may radiate similar, highly correlated waveforms in the particular direction of the station. Therefore Aster & Scott (1993) suggest using the median of the *cc* values of all stations as best practice to determine the degree of similarity between two events. Consequently, the maximum value should only be used for testing, to adjust time windows and select appropriate bandpass filters or in cases where only single stations close to the epicentre have a sufficient SNR (Ruscic *et al.* 2019). For smaller magnitudes only the closest stations are expected to record an event, therefore it helps to use the mean or median of those stations that comply with the given SNR threshold.

For the same reason Maurer & Deichmann (1995) introduced an asymmetrically trimmed mean for the computation of the network similarity $nsim_{ij,c}$ across a total number of *M* stations: For each event pair the lowest *k* per cent of the *cc* values are removed before the mean is computed:

$$nsim_{ij,c} = \frac{1}{M - kM} \sum_{s=1}^{M-kM} cc_{ij,c,s}, \tag{1}$$

where $cc_{ij,c,s}$ is sorted by descending *cc* value. Lower *cc* values between events at some stations do not necessarily imply weaker correlation of the events in regard to mechanism and location but can also be caused by other influences, such as variable site effects or noise conditions (Akuhara & Mochizuki 2014).

The network similarity of an event pair can also be computed as a weighted sum of the *cc* values at all stations. The weights $w_{ij,c,s}$ are the absolute differences between the first and the second maximum of the according cross-correlation function (Shelly *et al.* 2016):

$$nsim_{ij,c} = \sum_{s=1}^{M} cc_{ij,c,s}\ w_{ij,c,s}. \tag{2}$$

The use of a weighted sum limits the influence of poorly correlated records from distant or noisy stations and stabilizes the computation. However, we recommend to use a threshold for a required *cc* value at a minimum number of stations. Further, the resulting weights should be analysed along with the network similarity to avoid that the result is dominated by few stations only.

Another approach to combine the *cc* values of all stations is a composite correlation measure computed as the *M*th root of the product of the *cc* values (Stuermer *et al.* 2011):

$$nsim_{ij,c} = [\prod_{s=1}^{M}(cc_{ij,c,s})]^{-M}. \tag{3}$$

Stuermer *et al.* (2011) combined *P* and *S* *cc* values extracted from the same single component trace in the product. When using three component data, we first compute the *M*th root of the product for each component *c* separately, and then combine the obtained network similarities in a consecutive step.

The network similarity matrices $nsim_{ij,c}$ are computed for the different components *c* (e.g. Z, N and E) separately and subsequently combined as a weighted sum:

$$nsim_{ij} = \sum_{c=1}^{C} nsim_{ij,c}\ \omega_c. \tag{4}$$

The weighting of the components ω_c is defined in the configuration file. A component-based weighting allows compensating site-effects, which can lead to complex horizontal traces. The weighting can also compensate for variations in waveforms originating from different mechanisms that affect horizontal and vertical components differently. By comparing the results of independent phases (e.g. *P* and *S* or Love and Rayleigh) and components (i.e. Z, N and E) one can learn about the sensitivity of the waveforms in regard to different faulting types.

2.2 Event clustering

For the clustering algorithm input, the network similarity matrix (with 1 being the highest correlation) is converted into a distance matrix (with 0 corresponding to identical events). To avoid confusion with the spatial distance we hereafter refer to it as the similarity-distance. At the current version of the clustering toolbox, the density-based DBSCAN algorithm (Ester *et al.* 1996), as implemented in the python package *scikit-learn* (Pedregosa *et al.* 2011), is used for clustering. Other clustering algorithms, such as OPTICS (Ankerst *et al.* 1999) or *k*-means (Lloyd 1982) can be added by the user depending on the clustering targets.

Clusters derived using the DBSCAN algorithm can have any shape and the number of clusters is not predefined. Further, the algorithm allows for unclustered events (a noise class). Following the definitions of Ester *et al.* (1996), events belonging to one cluster are either core events or border events. Core events have at least a minimum number of neighbouring events (*MinPts*) within the similarity-distance *Eps*. Events at the border of the cluster (border events) are connected to at least one core point, but have less then *MinPts* neighbouring events within the similarity-distance *Eps*. Clusters are formed based on the concept of density reachability. An event *i* is considered directly density-reachable from a core event *j*, if it is within the similarity-distance *Eps*. Further the events *i* and *j* are density-connected if they are density-reachable through one or more density-connected core events.

The DBSCAN clustering procedure starts with the selection of an arbitrary event of the data set. All events that are density-reachable from this very first event (with respect to *Eps* and *MinPts*) are

retrieved. A cluster is only formed if there is at least one core event. If not, DBSCAN visits the next point of the database. This process is continued until all points have been processed (Ester *et al.* 1996). Events not lying within similarity-distance *Eps* of any other event are assigned to a noise class (unclustered). *Eps* and *MinPts* need to be tuned by the user according to the data set.

2.3 Tuning of the clustering parameters and graphical analysis of clustering results

Our clustering toolbox provides several analysis plots that facilitate the tuning of the clustering parameters and the evaluation of the stability of the clusters. Further, these plots provide detailed insight into the clustering results. The plotting tools can also be used to analyse, compare and choose multiple target frequency bands to include surface waves for larger, distant events and body waves for smaller, local events. The graphical output is generated using GMT (Wessel *et al.* 2013) and the python plotting packages *matplotlib* (Hunter 2007) and *plotly* (Plotly Technologies Inc. 2015). The plots presented in this section illustrate the analysis that was performed to obtain optimal clustering settings and stable results for the application to the aftershock sequence of the 25 October 2018 Zakynthos M_w 6.9 earthquake in Greece.

Clusty allows the user to run the entire clustering process for different DBSCAN parameters (*Eps* and *MinPts*) in parallel to compare the results. As mentioned above, the input similarity-distance matrix for DBSCAN is computed from the cross-correlations of waveforms. Therefore the similarity-distance radius (*Eps*) is directly related to the underlying physical process and a rough first estimate of *Eps* can be made based on expected similarities. However, the expected *cc* values, and consequently, the optimal *Eps* value may vary depending on the length of the considered waveform time windows, the frequency content as well as on site and noise conditions at the stations. An *Eps* of 0.1 implies that a pair of connected events has at least a network similarity of 0.9. Depending on the chosen method for the network similarity computation, waveform cross-correlation values at single stations can be smaller if other stations with higher values compensate for it. *Eps* needs to be adjusted to the purpose of the clustering. Using a small *Eps* value allows finding very similar events or repeaters. However, in this case other events are omitted, which would still be considered similar when clustering is performed using a higher *Eps* for fault identification and tracing.

Ester *et al.* (1996) suggested a *k*-nearest neighbour (*k*-NN) plot (Fig. 2a) to choose the *Eps* parameter. Therein, the average similarity-distance of every sample to its *k* nearest neighbours (here corresponding to the *MinPts* parameter) is calculated and plotted in an ascending order to visually find a 'knee', that corresponds to the optimal *Eps* value for the given data set (Ester *et al.* 1996). In Fig. 2(a), the sorted similarity-distances of the *k*th nearest neighbours are shown for *MinPts* values from 5 to 8. For *MinPts* 5, significant gradient changes are seen for *Eps* values of 0.16 and 0.21 (red arrows in Fig. 2a). For increased *MinPts* values these gradient changes are observed for larger *Eps* values. However, we prefer smaller *Eps* values, because otherwise we observe rather unstable and heterogeneous clusters in our application (Fig. 2b and following paragraphs). Therefore, we suggest three additional metrics to constrain a range of appropriate DBSCAN clustering parameters for fault tracing purposes: (1) the silhouette score, (2) the number of clusters and (3) the total number of clustered events. These metrics reflect the ensemble of all clusters, while the influence of different parameter sets onto single clusters can be analysed using more sophisticated analysis tools, introduced hereafter. Fig. 2(b) shows these metrics for *Eps* values between 0.01 and 0.30 and *MinPtS* values of 5 and 8. The trends of the three curves are similar for both *MinPts* values. The silhouette score is a measure of the homogeneity of all clusters (Rousseeuw 1987), here neglecting the unclustered events. It is the mean of the silhouette coefficients of all clustered events. The silhouette coefficient of a single event expresses how similar that event is compared to the other events within the same cluster and compared to the events of the nearest other cluster. The silhouette coefficient is defined as:

$$s = (\overline{icd} - \overline{ncd})/max(\overline{icd}, \overline{ncd}), \quad (5)$$

where $\overline{ncd}$ is the mean nearest cluster similarity-distance for each event and $\overline{icd}$ is the mean intracluster similarity-distance (Rousseeuw 1987). The silhouette coefficient ranges between –1 and 1, where 1 corresponds to a cluster of identical events, that are completely different from events belonging to other clusters. Coefficients between –1 and 0 indicate that the similarity-distance of an event with respect to events of other clusters is smaller than the average similarity-distance to events of its own cluster. We use the implementation of *scikit-learn* (Pedregosa *et al.* 2011) to calculate the silhouette coefficients. The silhouette score in Fig. 2(b) is largest at very low *Eps* values, when only highly similar earthquakes are assigned to one or few clusters. Thereafter, the silhouette score decreases with increasing *Eps* value, so in fact we are visually searching for local maxima or changes in the gradient of the curve, but not for the global maximum. The shift between the two lines for *MinPts* 5 and 8 (grey and black, respectively) in Fig. 2(b) is the result of an increased number of earthquakes required per cluster for higher *MinPts* values. The first cluster can therefore only be found for a slightly higher *Eps* value in case of a higher *MinPts* value. In both curves a local maximum is seen at an *Eps* value of 0.06. Several minor changes in the gradient are observed between 0.10 and 0.14, followed by a major gradient change at 0.15 (black arrow in Fig. 2b). Below *Eps* 0.15, the silhouette score is relatively stable on a low level. By decreasing the *Eps* only by 0.01 or 0.02 we obtain significantly higher silhouette scores, thus more homogeneous clusters (Fig. 2b), a prerequisite for a reliable identification of active faults. The total number of clusters and the number of clustered events decreases with increasing *MinPts*, resulting from a higher required number of earthquakes to form a cluster. The number of clustered events increases rapidly until an *Eps* value of 0.13 (*MinPts* 5, green arrow in Fig. 2b) and 0.16 (*MinPts* 8) and shows a smaller gradient afterwards. Local maxima of the number of clusters are found for 0.10 and 0.13 (blue arrow in Fig. 2b) for *MinPts* 5 and 0.13 for *MinPts* 8. For larger *Eps* values single clusters collapse into larger, more heterogeneous ones, as can be seen in the flow diagram (Fig. 3).

The flow diagram helps to assess the stability of the clustering results. It allows a comprehensive comparison of clustering results obtained using different clustering parameters (Fig. 3) or waveform frequency filters (Fig. S1). Fig. 3 shows the flow diagram of clustering results obtained for an *Eps* range of 0.06–0.30. The width of the connecting bands between two clusters obtained with two sets of parameters is proportional to the number of common events. In this way, the diagram reflects conserved quantities as well as the splitting or merging of clusters (Fig. 3). In Fig. 3 the small clusters in the lower part of the diagram remain stable over a wide range of *Eps* values. The two largest and distinct green and light blue (3 and 5) clusters collapse into one heterogeneous cluster when *Eps*

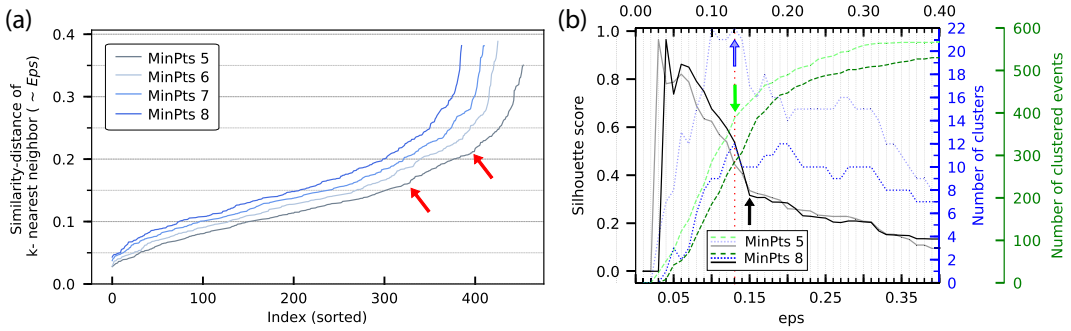

Figure 2. Selection of the DBSCAN clustering parameters. (a) The *k*-nearest neighbour plot (here for *MinPts* of 5–8) helps selecting *Eps* by identify a 'knee' (gradient changes, red arrows). (b) Silhouette score (black solid lines), number of clusters (blue dotted lines) and total number of clustered events (green dashed lines) for *Eps* values between 0.01 and 0.4 and *MinPts* values of 5 (lighter colours) and 8 (darker colours). Arrows mark features discussed in the text for *MinPts* 5. Blue: max. number of clusters, green: gradient change in number of clustered events, black: gradient change in silhouette score. Red dotted line indicates the *Eps* value used in the application to the Zakynthos data set (0.13).

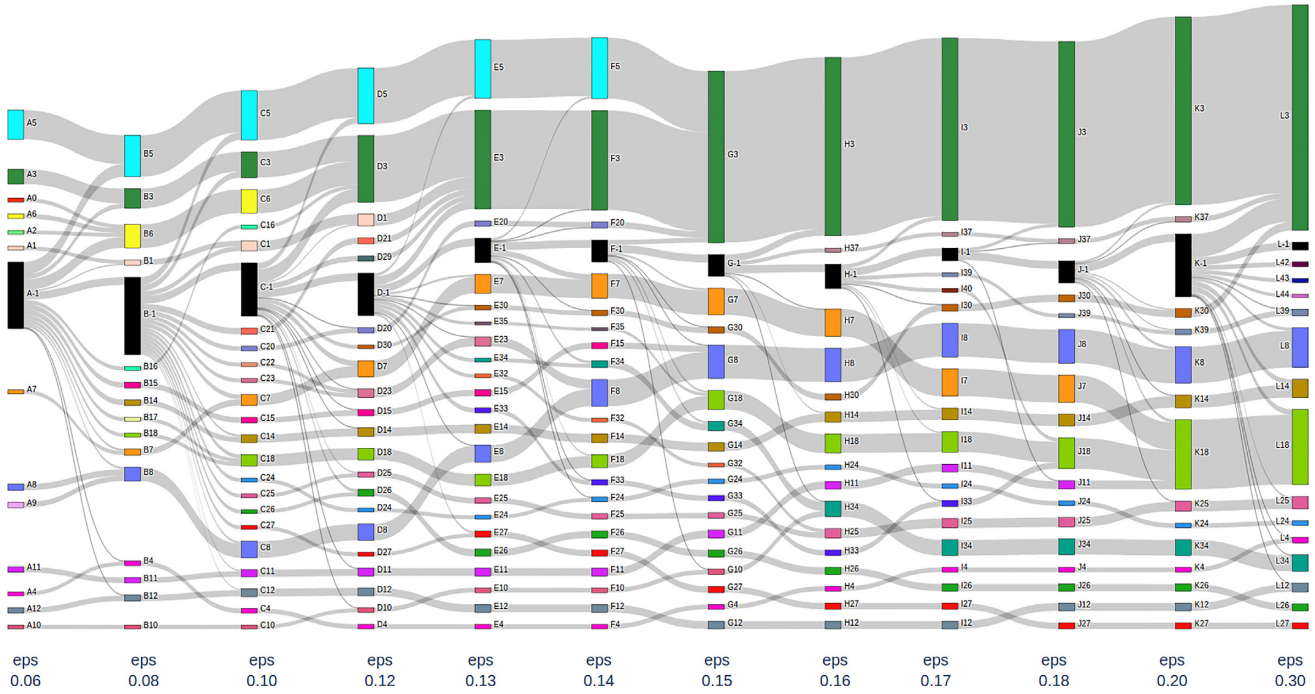

Figure 3. Screenshot of the interactive flow diagram for a comparison of clustering results obtained with *Eps* values between 0.06 and 0.30, *MinPts* 5 (0.05–0.2 Hz). Black represents the noise cluster (here reduced in size to emphasize the clustered events). Same colours represent the same cluster. The size of the clusters varies depending on the required similarity (Eps value), the block size is proportional to the number of events within the cluster. The grey bands connect clusters obtained with different clustering parameters which share at least one event. The thickness of the grey bands is proportional to the number of shared events. This representation allows evaluating the stability of cluster results when changing clustering parameters.

increases from 0.14 to 0.15. For *Eps* values as small as 0.06 only few small clusters are found. For *Eps* = 0.30 large clusters with clearly distinguishable event types (when using a smaller *Eps*) collapse into one cluster.

When using multiple clustering settings, the resulting clusters as well as their labels will differ. Therefore, we implemented a function that provides harmonized cluster labels across the different clustering results. The harmonization of labels can lead to a discontinuous cluster label numbering but assures that labels are persistent.

Finally, we introduce two more visualization tools to analyse the clustering parameters and control the clustering results: the silhouette coefficient plot (Figs 4a–c) and the event-connectivity plot (Figs 4d–f). Both depict the homogeneity and the connectivity within each cluster or among different clusters, respectively.

The silhouette coefficient plot (Figs 4a–c) shows how similar each event is to the events in its own cluster compared to the events of the most similar cluster (Rousseeuw 1987). Each coloured block represents the events of one cluster, sorted by their silhouette coefficient. The silhouette plot helps to find appropriate *Eps* and *MinPts* settings and the optimal number of clusters by evaluating the similarity of events within each cluster. The connectivity plot (Figs 4d–f) provides a complementary visualization of the similarity between events as well as between clusters. Within this force-directed projection (Fruchterman & Reingold 1991) the relative distance between events or clusters of events represent their similarity. When choosing *Eps*=0.06 (Figs 4a and d) only a few, very similar events are clustered. However, the visualization of the connectivity (Fig. 4d) shows that there are many more clusters of similar events. This

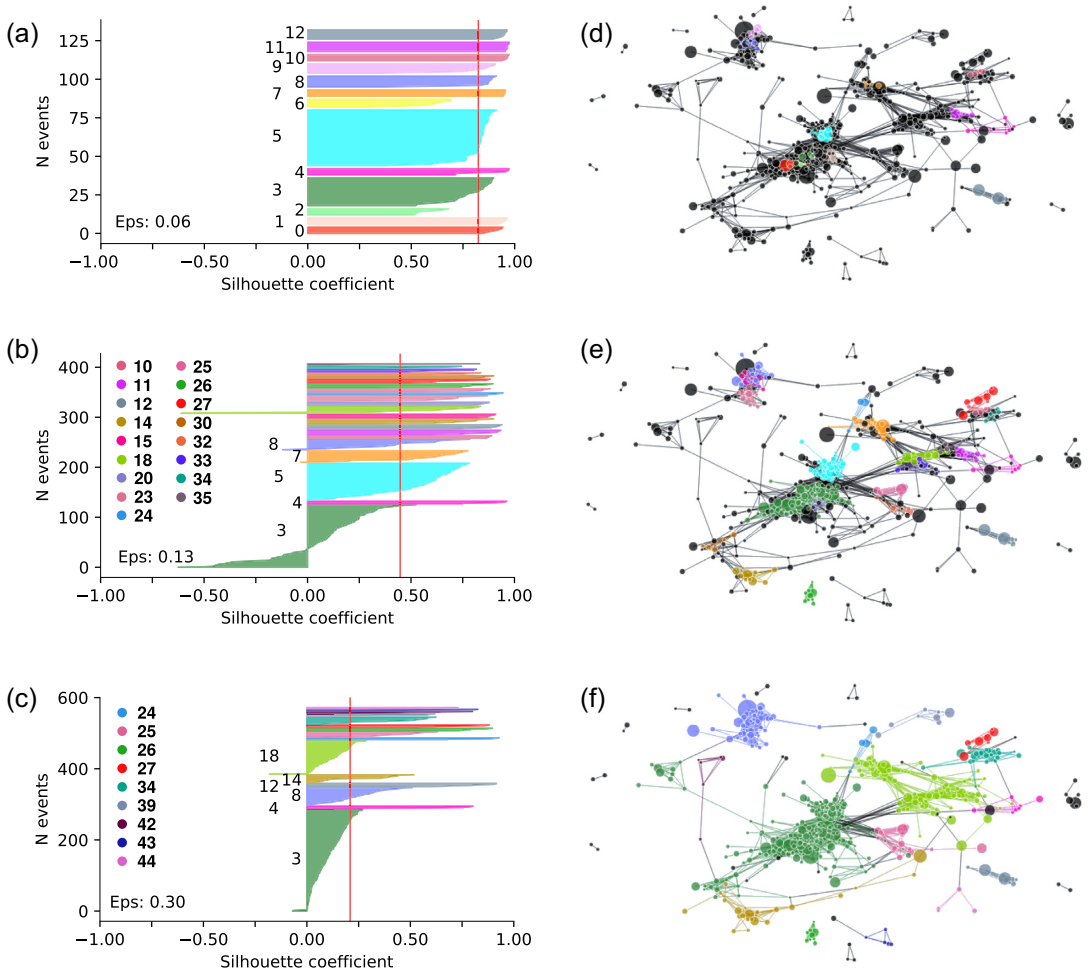

Figure 4. Example for silhouette coefficient plots (a–c) and connectivity plots (d–f), obtained for the Zakynthos data set with *Eps* values 0.06, 0.13 and 0.3, *MinPts* = 5. The numbers next to the clusters in (a–c) indicate the cluster label. The red vertical line is the silhouette score (mean of silhouette coefficients) of the clustered events. The connectivity plots (d–f) provide an additional visualization of the similarity of events within the same cluster as well as among different clusters. Events are coloured according to their cluster as in (a–c). Unclustered events are shown in black. In this projection the relative distance between events or clusters represents their similarity. The absolute locations within the projection have no meaning.

cannot be seen from the silhouette plot alone (Fig. 4a). By increasing *Eps* to 0.13 (Figs 4b and e) all clusters are well separated and homogeneous, except for cluster 3 and 18. For *Eps*=0.30 (Figs 4c and f) the central clusters collapse into one. In the latter case, the silhouette plot indicates that the clusters are generally more heterogeneous and larger. We want to stress the importance of the analysis of both, the similarity between separated clusters as well as among the events that belong to the same cluster, before interpreting the results. The user can get insights into the quality of the performed clustering analysis by comparing the presented plots for a range of parameters.

Clusty provides maps and waveform plots as final graphical output along with the catalogue of clustered events (for examples, see also Section 3). Station-wise waveform plots display all aligned waveforms per cluster and component. The waveform plots provide another direct visualization to evaluate the similarity of waveforms. We would like to point out that the clustering algorithm, the plots to evaluate the stability of the results (flow diagram, silhouette and connectivity plots) and the final maps may be used independently with any other distance matrix provided by the user. For example, these distance matrices could be based on Kagan angles or spatial distances.

3 APPLICATION: OFFSHORE AFTERSHOCK SEQUENCE OF THE *M* 6.9 ZAKYNTHOS EARTHQUAKE, GREECE

3.1 Study area

The study area is located at the western margin of the Hellenic Subduction System (HSS), along which the oceanic lithosphere of the African Plate is subducted beneath the continental lithosphere of the Eurasian Plate with a NE dipping slab (Fig. 5). Within the study area, faults of varying geometries and slip movements (Mouslopoulou *et al.* 2020) accommodate the northward kinematic transition from convergence to strike-slip (i.e. Pérouse *et al.* 2017; Sachpazi *et al.*

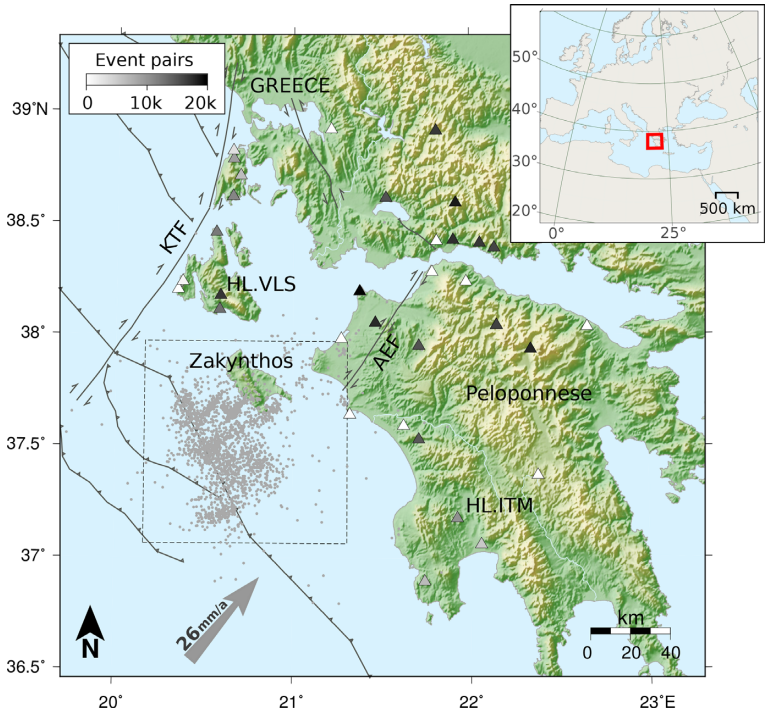

Figure 5. Study area and station network. Triangles indicate seismic stations. Colour intensity depicts the contribution of each station to the clustering result (number of clustered event pairs for which the station was used). The dashed square outlines the extent of Figs 6(a) and 7. The grey arrow indicates the relative movement of the African Plate with respect to stable Eurasia (Pérouse *et al.* 2017). Active regional faults from Basilic *et al.* (2013), topography from SRTM (Farr *et al.* 2007). KTF, Kefalonia Transform Fault; AEF, Achaia-Elia Fault.

2000). Although all types of faulting (thrust, normal and strike-slip) may occur (Konstantinou *et al.* 2017; Mouslopoulou *et al.* 2020), thrust faulting appears to prevail south and southwest of Zakynthos (Papadimitriou *et al.* 2013; Wardell *et al.* 2014), while strike-slip faulting is dominant to the northwest (Louvari *et al.* 1999; Sachpazi *et al.* 2000), onshore Peloponnese (Feng *et al.* 2010; Stiros *et al.* 2013) and in the offshore area between Zakynthos and Peloponnese (Kokkalas *et al.* 2013; Haddad *et al.* 2020; Mouslopoulou *et al.* 2020). Normal faulting is accommodated in the shallower sections of the crust (above 15 km), often at high angles to the prevailing strike of the mapped thrust/strike-slip faults (Mouslopoulou *et al.* 2020).

The study region is characterized by intense seismic activity and strong main shocks ($M > 6$, Papazachos & Papazachou 2003). On 25 October 2018, a magnitude M_w 6.9 earthquake struck southwest of Zakynthos (Chousianitis & Konca 2019; Cirella *et al.* 2020; Ganas *et al.* 2020; Karakostas *et al.* 2020; Mouslopoulou *et al.* 2020; Sokos *et al.* 2020). It occurred after a 4-yr-long phase of seismic unrest which was probably triggered by a slow-slip event (Mouslopoulou *et al.* 2020) and was followed by strong aftershock activity, which is still ongoing (Mouslopoulou *et al.* 2020; Sokos *et al.* 2020). The complex moment tensor of the main shock, with a significant non-double couple component, was attributed to subevents of thrust faulting and moderately dipping right lateral strike-slip faulting, in accordance with the African–Eurasian Plate motion (Cirella *et al.* 2020; Mouslopoulou *et al.* 2020; Sokos *et al.* 2020). While seismological results alone cannot clearly discriminate between a splay-thrust and a subduction-thrust fault scenario for the main candidate earthquake fault, the scenario of a splay-thrust fault is supported by published seismic-reflection and bathymetric data (Mouslopoulou *et al.* 2020) and the recording of a minor tsunami that suggests rupture of the sea-bed (Cirella *et al.* 2020).

Our study is based on the catalogue of the aftershock sequence reported by Mouslopoulou *et al.* (2020). It consists of $\geq$2300 events ($M_w \geq 2.8$), including about 80 double-couple solutions showing a large variability of thrust, normal and strike-slip mechanisms. This hints at the activation of a complex fault system, in accordance with local fault diversity (e.g. Konstantinou *et al.* 2017; Mouslopoulou *et al.* 2020). Thanks to the complex tectonic setting, together with the multitude of activated faults, we consider the Zakynthos data set an ideal case study for our clustering analysis tool. The waveform data of the networks HA, HC, HL, HP, HT, MN (University Of Athens 2008; Technological Educational Institute Of Crete 2006; National Observatory Of Athens, I. O. G. 1997; University Of Patras, G. D. 2000; Aristotle University Of Thessaloniki Seismological Network 1981; MedNet Project Partner Institutions 1990) used in this study was obtained using the pyrocko fdsn client to access the databases of the National Observatory of Athens Seismic Network (NOA, http://www.gein.noa.gr/en/), GEOFOrschungsNetz (GEOFON; https://geofon.gfz-potsdam.de/), Observatories and Research Facilities for European Seismology (ORFEUS; https://www.orfeus-eu.org/), Incorporated Research Institutions for Seismology (IRIS; https://www.iris.edu/hq/) and Instituto Nazionale di Geofisica e Vulcanologia (INGV; http://webservices.ingv.it).

3.2 Results

Here, we present the clustering results for the Zakynthos data set obtained using a 30 per cent-trimmed mean for the calculation of the network similarity from waveforms of the seismic stations presented in Fig. 5. Vertical (HHZ) and horizontal (HHN and HHE) components were combined with weightings of 0.4, 0.3 and 0.3, respectively. Only event-pairs with *cc* values $\geq$0.7 and an SNR $\geq$ 2

at more than five stations, covering a minimum azimuthal range of 60°, are considered. We discuss the choice of the network similarity computation method and the DBSCAN clustering parameters (here: Eps 0.13, MinPts 5, primary frequency band 0.05–0.20 Hz, time window 80 s) in Section 4.

We used four different frequency bands to account for surface waves (0.02–0.15 Hz and 0.05–0.20 Hz) and body waves (0.1–0.5 Hz and 0.2–1.0 Hz). The overall patterns of clustered events are similar in all four frequency bands. Considering the stability and homogeneity as well as the total number of clustered events, the frequency band 0.05–0.20 Hz provides the best results. Using this frequency band, the clustering toolbox grouped 387 of 2361 (16 per cent) earthquakes with $M_w > 2.8$ into 22 clusters (Fig. 6a). 75 per cent of the events in the catalogue were rejected because they did not meet the quality thresholds (SNR, min. number of available traces) described above. Despite the small number of events compared to the total number of events in the catalogue, we consider the clustered events representative for the entire aftershock sequence as they cover 70 per cent of the cumulative moment.

The results of the primary frequency band are combined with the other, secondary frequency bands, mainly to account for smaller events with a low SNR at low frequencies. About 50 events were added to the clustering results, resulting in a total of approximately 430 clustered events. For each cluster, we computed deviatoric MTs (Fig. 6a) for one representative event using the probabilistic full waveform inversion framework Grond (Heimann *et al.* 2018), following the approach described in Mouslopoulou *et al.* (2020). The inversion includes 101 bootstrap chains with different weightings of the station-component-based misfits. The ten best MTs of each bootstrap chain, referred to as the ensemble of solutions, are used to analyse the uncertainties of the best solution obtained in the inversion.

Fig. 7 shows the temporal activity and moment release of the clusters. 84 per cent of the cumulative seismic moment of the aftershocks is released within the first month of the sequence. The central clusters (3, green; 4, pink; 5, light blue and 20, blue) are activated soon after the main shock. Our representative mechanisms (Fig. 6a) as well as the MT solutions of Mouslopoulou *et al.* (2020) for events belonging to the central clusters (3, 4, 5, 20 in Fig. S2) indicate predominantly thrust faulting. The thrust clusters 3 and 5 release 50 per cent of the cumulative seismic moment of the 1-yr aftershocks sequence or 70 per cent of the cumulative seismic moment of the clustered events (Fig. 7, inset). The proximity to the main shock, the time of initiation and the thrust nature of these events collectively suggest that they may be directly triggered by slip during the main shock. This is further supported by the representative mechanisms of cluster 3 and 5, which resemble the geometry resolved for the thrust part of the main shock by Sokos *et al.* (2020) and Mouslopoulou *et al.* (2020), possibly reflecting slip on the same (or neighbouring) fault. The representative mechanism of cluster 4 (Fig. 6a) shows a shallowly dipping (<10°) fault plane, however, its slip mechanism cannot be resolved unambiguously.

Following the rupture of the thrust clusters 3 and 5, and within hours of the main event, several strike-slip faults were activated north and south of the main shock's epicentre (Fig. 7), contributing significantly (12 per cent) to the total moment release of the aftershock sequence (Fig. 7, inset). West of the island of Zakynthos, we observe two NE–SW elongated seismicity clusters, which are associated with strike-slip faulting: overlapping clusters 11, 18 and 35 and the isolated cluster 7 (Fig. 6a). The small strike-slip cluster 24 abuts against the cluster 5 which is associated with thrust faulting. Due to the vicinity of these two clusters, the smaller strike-slip cluster cannot be detected based on spatio-temporal clustering. While the strike-slip clusters 7 and 24, which are located to the east and to the west of cluster 5, respectively, are active within the first ten days after the main shock, the activity of the overlapping strike-slip clusters 11, 18 and 35 starts two months later (Fig. 7). Cluster 33, which overlaps spatially with cluster 11, 18 and 35, in contrast is active within the first days after the main shock and shows a more oblique mechanism (Fig. 6). South of the main shock, three spatially and temporally overlapping strike-slip clusters (8, 15, 23) show a similar elongation in NE–SW direction (Fig. 6a). There, the activity starts within the first week after the main shock. The mechanisms of the three representative events of these clusters, together with the solutions from Mouslopoulou *et al.* (2020) (Fig. S2) indicate strike-slip on NS or EW striking fault planes, incompatible with the distribution of hypocentres. The latest cluster in the aftershock sequence is cluster 27 (red cluster in Figs 6a and 7), located to the south of the main shock. It is associated with a thrust slip on a NW–SE striking plane. The cluster consists of eight highly similar events.

4 DISCUSSION

In the introduction we briefly described the problem of the identification of active seismic faults and two related seismic methods, MT inversions and clustering of earthquakes based on selected precomputed features. Unlike the clustering of seismic events by their moment tensor, the clustering based on waveform similarities, which we propose here, is able to resolve closely located faults of different mechanisms without the limitation to larger magnitudes. Spatial clustering analysis is not limited by the magnitude, either, but is not able to resolve differences in the faulting mechanism. The clustering approach upon waveform similarity reflects the sensitivity of mechanism, location and depth, thus, providing a tool for the identification of active faults. Following a discussion on the methodological implementation, we review how a joint analysis of the clustering results and MT solutions for representative events can help to identify and describe active faults.

4.1 Discussion Part I: On the methodological implementation

The clustering toolbox presented here is dedicated to the study of active faults based on the waveform similarities of event pairs across a network of seismic stations. Compared to a single station approach, the network similarity has several advantages. By taking into account spatially distributed stations, a larger portion of the seismic radiation pattern is considered. Therefore a network similarity allows distinguishing mechanisms which cannot be distinguished in single station approaches. Especially in narrow frequency bands, it is possible to achieve high correlations at single stations that are excited by different faulting mechanisms. Further, even for a single high quality station noise conditions may vary temporarily and data gaps are likely. Using multiple stations in our network similarity approach assures the most efficient use of the available data.

We tested all methods for the network similarity computation (Section 2) by applying them to the aftershock sequence of the 25 October 2018, Zakynthos Earthquake. We observe that the network similarity based on the highest *cc* value across the network cannot resolve small differences between clusters of similar location and mechanism. The other methods implemented in the toolbox, that is median, mean, trimmed mean, the weighted sum and the composite

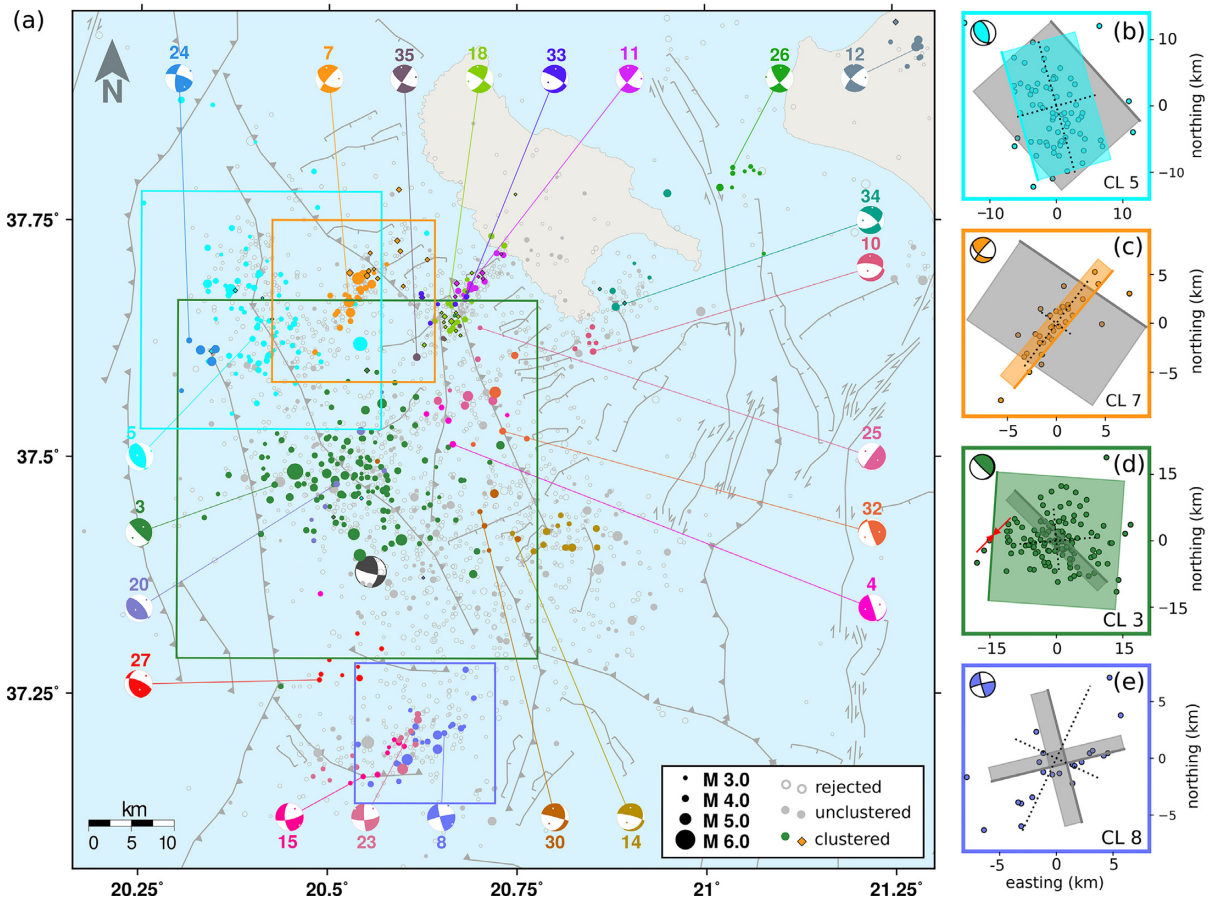

Figure 6. (a) Combined waveform-based clustering results for the aftershock sequence of the 25 October 2018, M_w 6.9 earthquake offshore Zakynthos, Greece (black MT). Clusters and representative MTs are colour-coded. Cluster label numbers are discontinuous due to harmonization of different *Eps* values and frequency bands (see Figs 3 and 4). Open grey circles represent events rejected from the clustering analysis due to selection criteria. The primary frequency band results (0.05–0.20 Hz) are shown as dots, diamonds refer events added to the clusters using the secondary frequency ranges. For the four largest clusters surface projections of the nodal planes of the representative MTs are shown in (b)–(e). Causative planes are coloured. For cluster 3 the strike angle is poorly resolved in the MT solution. The red arrows show the slip direction on the shallow nodal plane (green rectangle). Dashed lines depict the principle axes of principle component analyses of epicentres (see Section 4). Offshore faults are compiled and reinterpreted by Mouslopoulou *et al.* (2020) from bathymetric data and seismic-reflection profiles provided by Kokkalas *et al.* (2013), Wardell *et al.* (2014) and EMODnet Bathymetry Consortium (2018).

correlation measure (see Section 2.1), return comparable results after slightly adjusting the clustering parameters. For the sake of clarity, we only refer to the 30 per cent-trimmed mean network similarity when discussing the choice of the clustering parameters and the clustering results.

In the methodological section, we introduced the density-based DBSCAN clustering algorithm (Ester *et al.* 1996; Pedregosa *et al.* 2011). DBSCAN does not require that all events within one cluster are (highly) similar to all other events. Instead, it is sufficient that events within one cluster are connected by more similar events. Events with small differences in waveforms due to gradual changes in site effects, faulting mechanism or the travel path (location) can still belong to the same cluster if there are other connecting events in between them. Consequently, this approach is not only able to identify repeaters (e.g. Figs 4a, d and 8) (Geller & Mueller 1980), but allows grouping of events located on elongated faults. In our clustering toolbox we allow for unclustered events: If an event is not exhibiting a high similarity to any cluster of events, it is assigned to the noise class. In contrast, for instance the *k*-means clustering algorithm assigns every event to one of the given clusters without allowing for a noise class (Lloyd 1982). Therefore, we do not recommend using *k*-means for fault mapping purposes. Furthermore, in contrast to density-based clustering algorithms like DBSCAN, centroid-based clustering like *k*-means require a predefined number of clusters. Another common density-based clustering algorithm is OPTICS (Ankerst *et al.* 1999). Contrary to DBSCAN, which has a fixed radius *Eps*, OPTICS can handle varying cluster densities. However, for the fault tracing, we intend to have fixed criteria in regard to the required similarity of events and therefore use a fixed search radius. Thus, we rely on DBSCAN that assures that the event similarities, which result from physical processes and interevent distances, are comparable between the clusters. However, since our toolbox is set-up in a modular fashion, more methods can easily be implemented.

Clusty is applicable to different seismological scales since it directly uses waveforms and does not require precomputed features, such as moment tensors, characteristic functions, polarities or amplitude ratios. Potential applications range from acoustic emissions in laboratory or mining experiments to sequences of regional seismicity. Days to weeks long swarm activity as well as yearlong seismic sequences can be analysed. The flexibility in combining results from different frequency bands allows to investigate events with a large range of magnitudes. Thanks to the output of representative events of each cluster and stacked waveforms (optional),

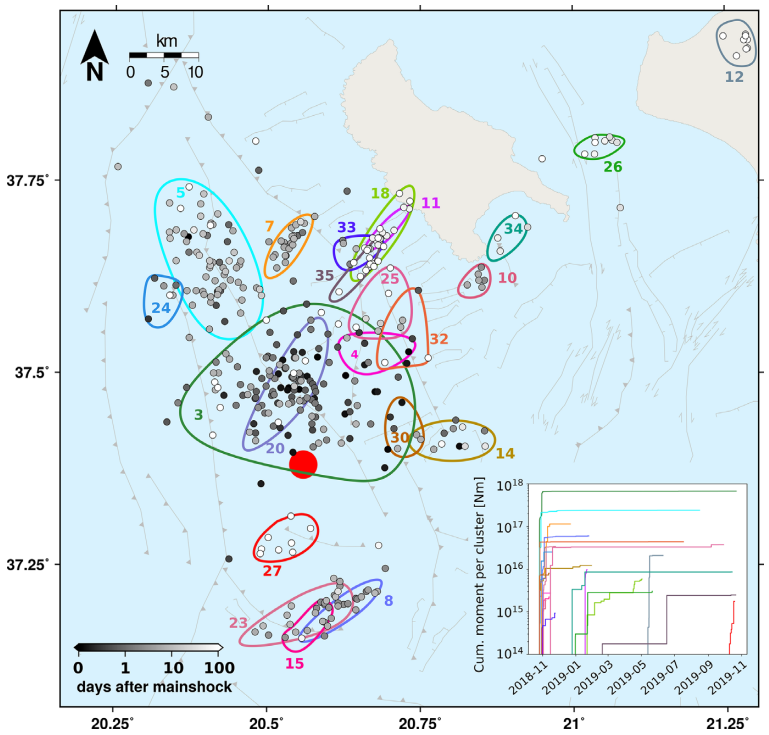

Figure 7. Spatial and temporal distribution of earthquake clusters during the Zakynthos aftershock sequence. The origin time of the clustered earthquakes are relative to the main shock. Contours of clusters and cluster labels for orientation and comparison to Fig. 6(a). The red dot indicates the main shock epicentre. Offshore faults are adopted from Mouslopoulou *et al*. (2020). Inset: Cumulative seismic moment of clusters over time.

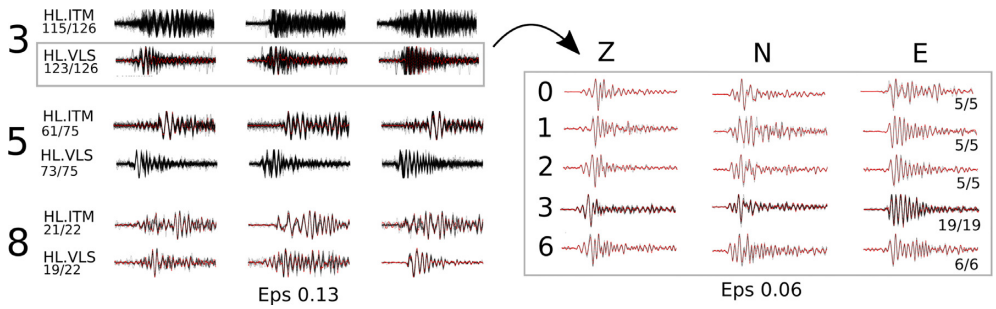

Figure 8. Aligned waveforms of the events in the two thrust clusters 3 and 5 and in the strike-slip cluster 8 at stations HL.ITM and HL.VLS (left-hand panel). See Fig. 5 for station locations. Numbers below the station name indicate the number of stacked traces versus the number of events within the clusters. Differences arise from waveform quality thresholds or missing data. When lowering the *Eps* to 0.06, cluster 3 splits up into smaller, more homogeneous subclusters (right-hand panel).

further analyses, such as subsequent moment tensor inversion, is facilitated.

The clustering toolbox returns several analysis plots to calibrate the settings for each study and to avoid black-box like usage. The DBSCAN parameter *Eps* should always be carefully adjusted. Larger values result in larger and more heterogeneous clusters. In contrast, low *Eps* values result in a higher similarity within the single clusters at the cost of a smaller number of clustered events, eventually losing information on the fault orientation. This trade-off needs to be considered when choosing an *Eps* value. We recommend testing different *Eps* values in parallel, for example from 0.05 to 0.30, and inspect what can be learned with respect to event clusters. We chose the *Eps* value of 0.13 for the cluster analysis after the joint consideration of the analysis tools as presented in Section 2. By testing different *MinPts* values, we find that the parameter does not significantly influence the observed pattern of earthquake clustering. To allow for smaller clusters to be included in the results, we set *MinPts* to 5.

Fig. 8 (left-hand panel) shows the aligned waveforms of clusters 3, 5 and 8 at two stations located north and east of the epicentral region (Fig. 5). The stacked waveforms of cluster 3 are clearly more diffuse than those of the other clusters. When lowering the *Eps* to 0.06 the large cluster 3 splits up into multiple homogeneous subclusters (right-hand part of Fig. 8 and see also Fig. 3). While the homogeneity within the subclusters is much higher, this approach substantially reduces the total number of clustered events (40 in 5 subclusters versus 126 in one cluster), showing a trade-off which was previously described.

4.2 Discussion Part II: On the application to the Zakynthos sequence

By applying the clustering toolbox to the aftershock sequence of the Zakynthos Earthquake, we are able to assign about 430 events to 22 distinct clusters. This is five times the number of aftershocks (∼80) that were clustered using the Kagan angle in Mouslopoulou *et al.* (2020). The increased number of clustered events enables a more precise characterisation of seismic patterns. Contrary to other clustering approaches that use event locations and/or times (e.g. Mouslopoulou & Hristopulos 2011; Ouillon & Sornette 2011; Karakostas *et al.* 2020), here we are able to distinguish events that are located close to each other but have different focal mechanisms and, thus, are expected to excite different waveforms, as seen for thrust cluster 5 and strike-slip cluster 24. Karakostas *et al.* (2020) identify 8 clusters for the same aftershock sequence based on event locations. We identify several additional small clusters (e.g. 24, 14 and 30), extending the insight into the complex fault system. Spatial or temporal clustering cannot separate events in complex faulting patterns as seen for the southernmost strike-slip clusters 8, 15, and 23. Cluster 33 overlaps spatially with clusters 11, 18 and 35, but the waveform similarity clearly separates these event groups, which are also separated temporally by 2 months.

Location errors need to be taken into account in the analysis of structures inferred from the clustering results. Earthquake (re-)location offshore Zakynthos is challenging due to the effects of a non-homogeneous velocity model, large azimuthal gaps and a sparse station coverage (e.g. Karastathis *et al.* 2015; Sachpazi *et al.* 2016). The locations and their uncertainties in the catalogue of Mouslopoulou *et al.* (2020) that we use here, were obtained using NonLinLoc (Lomax *et al.* 2000, 2009). The clustered events in this study have median uncertainties of 1.8 km and 2.7 km in horizontal and vertical direction, respectively (95 per cent confidence interval: 3.8 and 4.5 km). Due to the location errors, we do not consider any small structures (<10 km) in our analysis of fault planes.

Moment tensor solutions for representative events or for stacked waveforms of all events in a cluster enable the interpretation of the seismicity clusters with respect to faulting styles and fault orientations. Using stacked waveforms for each cluster can facilitate moment tensor inversion if the SNR of single events is too low otherwise. However, stacked waveforms need a particularly careful checking due to possible artefacts. For the application to the Zakynthos aftershock sequence the magnitudes of the representative events (M_w 3.5–4.6) allow for full waveform moment tensor inversions. In case of cluster 3, we report the moment tensor for the event with the second highest silhouette coefficient, because its magnitude is significantly larger (M_w 4.6 versus 3.9), providing a more stable MT result. Since Clusty is based on the Pyrocko python package, subsequent MT inversions using the probabilistic inversion framework Grond (Heimann *et al.* 2018) are facilitated.

The representative MTs show a wide variety of faulting types including thrusts, strike-slips and few normal faults. In general, our results are consistent with the MT solutions from Mouslopoulou *et al.* (2020) (Fig. 6a and Fig. S2). The *P* axes of their MTs are oriented NE–SW in general agreement with the regional compression.

For 15 of our 22 clusters, there is at least one MT available from Mouslopoulou *et al.* (2020, Fig. S2). Our representative MT solutions and MTs from Mouslopoulou *et al.* (2020) for events that belong to our clusters differ by a Kagan angle <30° for >50 per cent of the clusters and by <40° for 12 of our 15 clusters, respectively, implying homogeneous clusters. A Kagan angle of 30° is often used as a threshold for similar focal mechanisms in literature (e.g. Lee *et al.* 2014), while an angle ≪60° still indicates a good correspondence (Pondrelli *et al.* 2006; d'Amico *et al.* 2011). The large clusters 3 and 5 have increased mean Kagan angles of 40 and 55°, respectively. The variations of mechanisms in clusters 3 and 5 might primarily reflect varying dips of thrust fault planes.

For cluster 3, Mouslopoulou *et al.* (2020) report oblique strike-slip MTs for four earthquakes besides the predominant thrust mechanism. The oblique strike-slip and thrust mechanisms within this cluster cannot be distinguished, even when the *Eps* value is as low as 0.06. In our clustering approach we use a broader frequency band (0.05–0.20 Hz) compared to the MT inversions by Mouslopoulou *et al.* (2020). Repeating the MT inversion for all events for which solutions are reported by Mouslopoulou *et al.* (2020) in a broader frequency band (0.02–0.07 Hz), we obtain thrust mechanisms with minor oblique components. Kagan angles between our representative event for cluster 3 and our own MT solutions for the events that were also reported by Mouslopoulou *et al.* (2020) result in a mean angle of 25°, indicating similar event mechanisms. We assume that the narrower bandwidth used by Mouslopoulou *et al.* (2020), along with the unfavourable station distribution along the coast, results in two possible mechanisms that could not be distinguished in the MT inversion in the case of these four events.

Varying fault plane dip angles could be attributed to listric thrust faults offshore Zakynthos (Kokinou *et al.* 2005; Papoulia & Makris 2010; Kokkalas *et al.* 2013). Cluster 3 was active immediately after the main shock (Fig. 7). As the cluster is located within the main shock rupture area (Sokos *et al.* 2020), its heterogeneity may be linked to the complexity of the main shock, which possibly involves two overlapping events (Mouslopoulou *et al.* 2020; Sokos *et al.* 2020). The main shock itself does not belong to any of the clusters. Its waveforms are different probably because of its larger magnitude (and lower corner frequency) and/or because of its rupture complexity.

Strike-slip clusters are activated to the south and to the north of the M_w 6.9 epicentre few days after the main shock (and after the activation of the dominant thrust clusters 3 and 5). Activity on these distinct faults may have been triggered by stress perturbations imposed by the main shock and aftershock activity. The overlapping strike-slip clusters 11, 18, 35 southwest of the island of Zakynthos are activated two months after the main shock.

The identification and tracing of fault planes cannot be automated in this study. Increased vertical uncertainties of the hypocentres in the studied catalogue inhibit the direct fitting of fault planes into clouds of clustered events. Instead, we use their epicentral locations, geological constraints, and the projection of the two nodal planes of the representative MTs onto the surface (Figs 6b–e) to distinguish the fault plane from the auxiliary plane. The nodal planes are centred at the mean cluster location. The area of the nodal planes is estimated using an empirical relation to the cumulative seismic moment magnitude of the cluster following Wells & Coppersmith (1994). Local magnitudes are converted to moment magnitudes using an empirical relation derived from the MT solutions of Mouslopoulou *et al.* (2020). Since we only analyse event clusters with a moderate cumulative moment, we assume that a square-shaped fault model is representative for the fault plane (Delouis *et al.* 2009). We compare the projected planes and the epicentre distribution to distinguish the fault plane from the auxiliary plane. Additionally, we apply a Principle Component Analysis (PCA, Jolliffe 2002; Shearer *et al.* 2003) to each cluster by determining the eigenvectors ($\boldsymbol{v}_1$, $\boldsymbol{v}_2$) and eigenvalues (λ_1, λ_2) of the covariance matrix of the epicentres within each cluster. The length of the axes (dotted lines in Figs 6b–e) is

Table 1. Results of the joint interpretation of the clustering analysis and the representative MTs (nodal plane orientation) for clusters containing ≥10 earthquakes (n_{ev}). Time period is activity in days after the main shock. Interpreted fault planes in bold type. TF, thrust fault; SSF, strike-slip fault; NF, normal fault.

Cluster label	n_{ev}	Time period	ML_{max}	cum. moment (Nm)	Nodal planes (strike/dip/rake)	Mechanism
3	126	0–23	5.9	1.3e+18	134/84/83, **4/8/138**	TF
5	75	0–47	5.7	4.7e+17	138/37/68, **344/55/105**	TF
14	11	0–74	4.6	2.6e+16	284/80/-70, 39/22/-152	NF
7	24	3–20	5.3	2.3e+17	124/51/-9, **219/82/-140**	SSF
8	22	5–67	5.2	1.2e+17	256/84/7, 165/82/173	SSF
23	12	7–27	5.1	6.9e+16	84/72/-8, 176/81/-162	SSF
11	10	83–87	4.7	2.1e+16	131/63/3, **40/87/153**	SSF
18	15	87–187	4.1	1.3e+16	302/83/-23, **35/66/-173**	SSF
12	10	199–226	4.9	4.3e+16	128/70/-14, 223/75/-159	SSF

estimated from the 95 per cent confidence interval. When using epicentres (2-D case), the eigenvector of the largest eigenvalue can be interpreted as the strike direction of the fault, while the smaller eigenvalue can provide insight into the dip of the fault. For steep faults we expect $\lambda_1 > > \lambda_2$, while $\lambda_1 \approx \lambda_2$ represents a low-angle dip.

We test the results of Clusty against structural fault mapping offshore Zakynthos by focusing on prominent seismicity patterns: (1) the central thrust clusters 3 and 5 north of the M_w 6.9 epicentral area and (2) the large strike-slip clusters 7 and 8 north and south of the epicentral area. Figs 6(b) and (d) show the central clusters 3 and 5. The representative MT for cluster 3 has one steeply dipping (84°) SE–NW striking nodal plane and one shallowly dipping (<10°), NS striking plane. Due to the shallow dip angle of the latter and the unfavourable station distribution, it is difficult to resolve the strike angle of this nodal plane: among the ensemble of MT solutions having a small misfit the strike direction of the low-angle dipping nodal plane varies between NNW and N. Despite the limited resolution of the strike angle, both, the orientation of the *P*- and *T*-axis and the NE-ward slip of the shallow nodal plane (red arrow in Fig. 6d), are distinct and in agreement with the tectonic setting. Considering the broad scatter of the epicentres and the regional tectonic setting we identify the shallowly dipping nodal plane as the fault plane (coloured nodal plane in Fig. 6d). It has a similar orientation as the fault planes that Cirella *et al.* (2020) and Ganas *et al.* (2020) inferred for the main shock by jointly considering geodetic and seismic data. However, in contrast to our representative MT of cluster 3, the main shock mechanism has a large strike-slip component (Cirella *et al.* 2020; Ganas *et al.* 2020; Mouslopoulou *et al.* 2020; Sokos *et al.* 2020). The PCA supports the identification of the causative plane, as both PCA axes have a similar length, which indicates a low dip angle. However, in this case of two axes of similar length the PCA cannot resolve the strike direction of the fault plane. For cluster 5, both nodal planes (striking NNW and SE) could explain the scatter of events. A mapped thrust fault coincides with the NNW strike direction and ENE dip direction of the first nodal plane as well as with the major axis from the PCA (Fig. 6b). The identification of the causative plane is further supported by the regional tectonic setting. (Fig. 6a). The lack of seismic activity between the two large thrust clusters (3 and 5) may reflect a locked patch on an otherwise creeping fault plane (Moreno *et al.* 2011) or a rupture on two fault segments with deviating geometries, as seen in Fig. 6(a). However, it may also reflect a bias from the short observational period (∼1 yr).

For cluster 7, the SW striking nodal plane of the representative MT clearly coincides with the elongated epicentre distribution. This makes the selection of the fault plane unambiguous, also when relying on the PCA results (Fig. 6c). Cluster 8 is an example of a more complex fault system. The two steeply dipping nodal planes strike in SSE and WSW direction while the cluster is elongated in NE–SW direction, as indicated by the PCA (Fig. 6e). Consequently, an unambiguous identification of the fault plane is not possible. Considering similar observations for the clusters 15 and 23 we propose that this deviation is not caused by limitations in the analysis or a systematic bias in epicentre locations. Instead, it may be attributed to multiple strike-slip faults forming a bookshelf structure (en-echelon). Sokos *et al.* (2020) decompose their main shock moment tensor into one major strike-slip segment and a thrust segment. Similar to our finding for clusters 8, 15 and 23, they describe that the N10°E striking nodal plane of the strike-slip subevent is not in accordance with the alignment of the aftershocks.

The strike-slip faults associated with clusters 7 and 8 (north and south of the epicentral area) have not been constrained geologically using the available bathymetric data, possibly due to their subtle signature on the seabed and the limited resolution of the bathymetric data. Seismic-reflection data are not available for these regions. In addition to the mapped normal faults south of Zakynthos, we identify strike-slip clusters (11, 18, 35) which are oriented parallel to cluster 7 (Fig. 6a). This possibly implies the presence of NE–SW trending strike-slip faults north and south of the M_w 6.9 epicentral area, providing an example of how the toolbox Clusty can enhance or complement available tectonic information on active faults which might be of importance for seismic hazard scenarios.

Additional identified fault planes from the cluster analysis are presented in Table 1. The analysis demonstrated here depends on the availability of both, a sufficient number of cluster members with reliable event locations, and representative MTs. This prevents a fault plane identification for small clusters. If MTs are not available or cannot be computed, the clusters of earthquakes can still be compared to mapped faults, possibly providing additional information on activated faults and on the faulting style.

In summary, we show how our new waveform-based network similarity clustering toolbox Clusty helps to better constrain the geometries and kinematics of earthquake sequences that rupture multiple faults. We applied our toolbox onto the western-end of the HSS in the eastern Mediterranean, a complex tectonic setting where all types of faulting occur simultaneously. However, Clusty is applicable on other multifault systems globally, including subduction terminations (e.g. Mouslopoulou *et al.* 2019), closely spaced faults in narrow rift basins (Nicol *et al.* 2006) or fault intersections (Mouslopoulou *et al.* 2007). In ongoing studies we use it to analyse acoustic emission activity from a mine-scale experiment as well as

low magnitude seismicity in areas where moment tensor inversion meets its methodological limits.

5 CONCLUSIONS

The open source toolbox Clusty clusters earthquakes based on the waveform similarity across a seismic network. Thanks to comprehensive analysis tools, and the flexible choice of methods, the toolbox provides an easily tunable workflow and produces transparent results. Based on the analysis plots (i.e. flow diagram, silhouette score plot, connectivity plots), the user can visually inspect the results and select the most appropriate settings such as frequency bands, the quality thresholds and DBSCAN clustering parameters (*Eps* and *MinPts*). Besides the clustered catalogue and its graphical representation, Clusty provides a list of representative events and optionally stacked waveforms for each cluster to facilitate subsequent analyses such as moment tensor inversions and fault plane identifications. The modular setup of the toolbox allows an easy adaption to a broad range of applications e.g. local swarm like activity or regional long-term seismic patterns. The toolbox is open-source and can be downloaded at https://git.pyrocko.org/clusty/clusty. We applied the clustering toolbox to a seismic sequence following the magnitude M_w 6.9 Zakynthos Earthquake, Greece. We show how clustering parameters can be selected using the analysis plots provided by the toolbox. As a result we identify 22 clusters comprising more than 430 events that represent >70 per cent of the cumulative seismic moment released during the investigated time period. Relying on full waveform analysis, we can distinguish closely located events with different faulting styles. Moment tensor inversions for representative events of each cluster complement the clustering analysis of the seismic sequence. We show how our waveform-based clustering approach can be used to discriminate the fault plane from the auxiliary planes. Using 1 yr of seismic activity, we are able to associate clusters of events to individual faults and shed light onto the complex fault system in the study region. Thrust faulting is observed in two large clusters that are activated immediately after the main shock and remain active during the entire observation period, although the largest portion of the seismic moment from these clusters is released within the first days after the main shock. We suggest that these events are closely related to the M_w 6.9 earthquake and possibly occur on the same fault plane, accommodating subduction-related strain. However, the main shock itself does not show a high waveform similarity compared to these clusters. Clusty suggests the presence of strike-slip faults north and south of the main shock, in areas which are poorly resolved by seismic-reflection data. The results are broadly compatible with the geometry and kinematics of offshore faults mapped using seismic-reflection profiles and bathymetric data.

ACKNOWLEDGEMENTS

We like to thank the Editor, an anonymous reviewer and J. Zahradnik for constructive comments that helped to improve the manuscript. GMP is funded by DFG project 'From Top to Bottom - Seismicity, Motion Patterns and Stress Distribution in the Alpine Crust' (Project Number 362440331), a subproject of 'SPP 2017: Mountain Building Processes in 4D' (Project Number 313806092). PN is currently funded by the BMBF (German Federal Ministry of Education and Research) project SECURE (grant agreement No. 03G0872A).

6 DATA AVAILABILITY

The code for the toolbox is open-source and can be accessed at https://git.pyrocko.org/clusty/clusty. The event catalog used in this study can also be downloaded from the git repository (database last accessed in November 2019). The waveform data is freely available via the FDSN services (see section 3.1).

REFERENCES

Abramenkov, S., Shapiro, N.M., Koulakov, I. & Abkadyrov, I., 2020. Clustering of long-period earthquakes beneath Gorely Volcano (Kamchatka) during a degassing episode in 2013, *Geosciences,* **10**(6), 230.

Akuhara, T. & Mochizuki, K., 2014. Application of cluster analysis based on waveform cross-correlation coefficients to data recorded by ocean-bottom seismometers: results from off the Kii Peninsula, *Earth, Planets Space,* **66**(1), 80.

Ankerst, M., Breunig, M.M., Kriegel, H.-P. & Sander, J., 1999. Optics: ordering points to identify the clustering structure, in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data,* SIGMOD '99, pp. 49–60, Association for Computing Machinery, New York, NY, USA.

Ansari, A., Noorzad, A. & Zafarani, H., 2009. Clustering analysis of the seismic catalog of Iran, *Comput. Geosci.,* **35**(3), 475–486.

Aristotle University Of Thessaloniki Seismological Network, 1981. Permanent regional seismological network operated by the Aristotle University of Thessaloniki. International Federation of Digital Seismograph Networks, https://doi.org/10.7914/SN/HT.

Arrowsmith, S.J. & Eisner, L., 2006. A technique for identifying microseismic multiplets and application to the valhall field, north sea, *Geophysics,* **71,** V31–V40.

Asano, Y., *et al.*, 2011. Spatial distribution and focal mechanisms of aftershocks of the 2011 off the Pacific coast of Tohoku Earthquake, *Earth, Planets Space,* **63**(7), 669–673.

Aster, R.C. & Scott, J., 1993. Comprehensive characterization of waveform similarity in microearthquake data sets, *Bull. seism. Soc. Am.,* **83**(4), 1307–1314.

Baisch, S., Ceranna, L. & Harjes, H.-P., 2008. Earthquake cluster: what can we learn from waveform similarity?, *Bull. seism. Soc. Am.,* **98**(6), 2806–2814.

Barani, S., Ferretti, G., Massa, M. & Spallarossa, D., 2007. The waveform similarity approach to identify dependent events in instrumental seismic catalogues, *Geophys. J. Int.,* **168**(1), 100–108.

Basilic, R., *et al.*, 2013. The European Database of Seismogenic Faults (EDSF) compiled in the framework of the Project SHARE. http://diss.rm.ingv.it/share-edsf/, doi: 10.6092/INGV.IT-SHARE-EDSF.

Cesca, S., 2020. Seiscloud, a tool for density-based seismicity clustering and visualization, *J. Seismol.,* **24**(1), doi:10.1007/s10950-020-09921-8.

Cesca, S., *et al.*, 2017. Complex rupture process of the Mw 7.8, 2016, Kaikoura earthquake, New Zealand, and its aftershock sequence, *Earth planet. Sci. Lett.,* **478,** 110–120.

Chousianitis, K. & Konca, A.O., 2019. Intraslab deformation and rupture of the entire subducting crust during the 25 October 2018 Mw 6.8 Zakynthos earthquake, *Geophys. Res. Lett.,* **46**(24), 14 358–14 367.

Cirella, A., *et al.*, 2020. The 2018 Mw 6.8 Zakynthos (Ionian Sea, Greece) earthquake: seismic source and local tsunami characterization, *Geophys. J. Int.,* **221**(2), 1043–1054.

Czecze, B. & Bondár, I., 2019. Hierachical cluster analysis and multiple event relocation of seismic event clusters in Hungary between 2000 and 2016, *J. Seismol.,* **23,** 1313–1326.

d'Amico, S., Orecchio, B., Presti, D., Gervasi, A., Zhu, L., Guerra, I., Neri, G. & Herrmann, R., 2011. Testing the stability of moment tensor solutions for small earthquakes in the Calabro-Peloritan Arc region (southern Italy), *Boll. Geof. Teorica. Appl.,* **52**(2), doi:10.4430/bgta0009.

Delouis, B., Charlety, J. & Vallée, M., 2009. A method for rapid determination of moment magnitude Mw for moderate to large earthquakes from the near-field spectra of strong-motion records (MWSYNTH), *Bull. seism. Soc. Am.,* **99**(3), 1827–1840.

EMODnet Bathymetry Consortium, 2018. EMODnet Digital Bathymetry (DTM 2018), Tech. rep., EMODnet Bathymetry Consortium.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al.*, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,* August 1996, pp. 226–231.

Farr, T.G., *et al.*, 2007. The shuttle radar topography mission, *Rev. Geophys.,* **45**(2), doi:10.1029/2005RG000183.

Feng, L., Newman, A.V., Farmer, G.T., Psimoulis, P. & Stiros, S.C., 2010. Energetic rupture, coseismic and post-seismic response of the 2008 MW 6.4 Achaia-Elia Earthquake in northwestern Peloponnese, Greece: an indicator of an immature transform fault zone, *Geophys. J. Int.,* **183**(1), 103–110.

Frohlich, C., 1987. Aftershocks and temporal clustering of deep earthquakes, *J. geophys. Res.,* **92**(B13), 13 944–13 956.

Fruchterman, T.M. & Reingold, E.M., 1991. Graph drawing by force-directed placement, *Softw.: Pract. Experience,* **21**(11), 1129–1164.

Ganas, A., *et al.*, 2020. The 25 October 2018 Mw= 6.7 Zakynthos earthquake (Ionian Sea, Greece): a low-angle fault model based on GNSS data, relocated seismicity, small tsunami and implications for the seismic hazard in the west Hellenic Arc, *J. Geodyn.,* **137,**.

Geller, R.J. & Mueller, C.S., 1980. Four similar earthquakes in central California, *Geophys. Res. Lett.,* **7**(10), 821–824.

Grandin, R., Vallée, M., Satriano, C., Lacassin, R., Klinger, Y., Simoes, M. & Bollinger, L., 2015. Rupture process of the Mw= 7.9 2015 Gorkha earthquake (Nepal): insights into Himalayan megathrust segmentation, *Geophys. Res. Lett.,* **42**(20), 8373–8382.

Haddad, A., Ganas, A., Kassaras, I. & Lupi, M., 2020. Seismicity and geodynamics of western Peloponnese and central Ionian Islands: insights from a local seismic deployment, *Tectonophysics,* **778,** 228353.

Han, L., Wu, Z., Li, Y. & Jiang, C., 2014. Cross-correlation coefficients for the study of repeating earthquakes: an investigation of two empirical assumptions/conventions in seismological interpretation practice, *Pure appl. Geophys.,* **171**(3–5), 425–437.

Heimann, S., *et al.*, 2017. Pyrocko - an open-source seismology toolbox and library, V. 0.3. GFZ Data Services. https://doi.org/10.5880/GFZ.2.1.2017.001.

Heimann, S., Isken, M., Kühn, D., Sudhaus, H., Steinberg, A., Daout, S., Cesca, S., Bathke, H. & Dahm, T., 2018. Grond: a probabilistic earthquake source inversion framework. V. 1.0. GFZ Data Services. https://doi.org/10.5880/GFZ.2.1.2018.003.

Herrmann, R.B., Malagnini, L. & Munafó, I., 2011. Regional moment tensors of the 2009 L'Aquila earthquake sequence, *Bull. seism. Soc. Am.,* **101**(3), 975–993.

Hunter, J.D., 2007. Matplotlib: a 2D graphics environment, *Comput. Sci. Eng.,* **9**(3), 90–95.

Igarashi, T., Matsuzawa, T. & Hasegawa, A., 2003. Repeating earthquakes and interplate aseismic slip in the northeastern Japan subduction zone, *J. geophys. Res.,* **108**(B5), doi:10.1029/2002JB001920.

Jolliffe, I.T., 2002. *Principal Component Analysis, Springer Series in Statistics,* 2nd edn, Springer-Verlag.

Karakostas, V., Kostoglou, A., Chorozoglou, D. & Papadimitriou, E., 2020. Relocation of the 2018 Zakynthos, Greece, aftershock sequence: spatiotemporal analysis deciphering mechanism diversity and aftershock statistics, *Acta Geophys.,* **68,** 1263–1294.

Karastathis, V., Mouzakiotis, E., Ganas, A. & Papadopoulos, G., 2015. High-precision relocation of seismic sequences above a dipping Moho: the case of the January-February 2014 seismic sequence on Cephalonia island (Greece), *Solid Earth,* **6**(1), 173.

Kokinou, E., Kamberis, E., Vafidis, A., Monopolis, D., Ananiadis, G. & Zelilidis, A., 2005. Deep seismic reflection data from offshore western Greece: a new crustal model for the Ionian Sea, *J. Petrol. Geol.,* **28**(2), 185–202.

Kokkalas, S., Kamberis, E., Xypolias, P., Sotiropoulos, S. & Koukouvelas, I., 2013. Coexistence of thin-and thick-skinned tectonics in Zakynthos area (western Greece): Insights from seismic sections and regional seismicity, *Tectonophysics,* **597,** 73–84.

Konstantinou, K., Mouslopoulou, V., Liang, W.-T., Heidbach, O., Oncken, O. & Suppe, J., 2017. Present-day crustal stress field in Greece inferred from regional-scale damped inversion of earthquake focal mechanisms, *J. geophys. Res.,* **122**(1), 506–523.

Koper, K.D., Hutko, A.R., Lay, T., Ammon, C.J. & Kanamori, H., 2011. Frequency-dependent rupture process of the 2011 Mw 9.0 Tohoku earthquake: comparison of short-period p wave backprojection images and broadband seismic rupture models, *Earth, Planets Space,* **63**(7), 16.

Lee, S.-J., Liu, Q., Tromp, J., Komatitsch, D., Liang, W.-T. & Huang, B.-S., 2014. Toward real-time regional earthquake simulation II: real-time online earthquake simulation (ROS) of Taiwan earthquakes, *J. Asian Earth Sci.,* **87,** 56–68.

Lloyd, S., 1982. Least squares quantization in PCM, *IEEE Trans. Inform. Theory,* **28**(2), 129–137.

Lomax, A., Virieux, J., Volant, P. & Berge-Thierry, C., 2000. Probabilistic earthquake location in 3D and layered models, in *Advances in Seismic Event Location,* pp. 101–134, Springer.

Lomax, A., Michelini, A. & Curtis, A., 2009. Earthquake location, direct, global-search methods, in *Encyclopedia of Complexity and System Science,* pp. 1–33, Springer.

Louvari, E., Kiratzi, A. & Papazachos, B., 1999. The Cephalonia transform fault and its extension to western Lefkada Island (Greece), *Tectonophysics,* **308**(1–2), 223–236.

Maurer, H. & Deichmann, N., 1995. Microearthquake cluster detection based on waveform similarities, with an application to the western Swiss Alps, *Geophys. J. Int.,* **123**(2), 588–600.

MedNet Project Partner Institutions, 1990. Mediterranean very broadband seismographic network (MedNet). Istituto Nazionale di Geofisica e Vulcanologia (INGV), https://doi.org/10.13127/SD/fBBBtDtd6q.

Mesimeri, M., Karakostas, V., Papadimitriou, E. & Tsaklidis, G., 2019. Characteristics of earthquake clusters: application to western Corinth Gulf (Greece), *Tectonophysics,* **767,** 228160.

Moreno, M., *et al.*, 2011. Heterogeneous plate locking in the South-Central Chile subduction zone: building up the next great earthquake, *Earth Planet. Sci. Lett.,* **305**(3–4), 413–424.

Mouslopoulou, V. & Hristopulos, D.T., 2011. Patterns of tectonic fault interactions captured through geostatistical analysis of microearthquakes, *J. geophys. Res.,* **116**(B7), doi:10.1029/2010JB007804.

Mouslopoulou, V., Nicol, A., Little, T. & Walsh, J., 2007. Displacement transfer between intersecting regional strike-slip and extensional fault systems, *J. Struct. Geol.,* **29**(1), 100–116.

Mouslopoulou, V., Saltogianni, V., Nicol, A., Oncken, O., Begg, J., Babeyko, A., Cesca, S. & Moreno, M., 2019. Breaking a subduction-termination from top to bottom: the large 2016 Kaikōura Earthquake, New Zealand, *Earth planet. Sci. Lett.,* **506,** 221–230.

Mouslopoulou, V., Bocchini, G.M., Cesca, S., Saltogianni, V., Bedford, J.R., Petersen, G.M., Gianniou, M. & Oncken, O., 2020. Earthquake-swarms, slow-slip and fault-interactions at the western-end of the Hellenic Subduction System precede the Mw 6.9 Zakynthos Earthquake, Greece, *Geochem., Geophys., Geosyst.,* doi:10.1029/2020GC009243.

National Observatory Of Athens, I. O. G., 1997. National Observatory of Athens Seismic Network. International Federation of Digital Seismograph Networks, https://doi.org/10.7914/SN/HL.

Nicol, A., Walsh, J., Berryman, K. & Villamor, P., 2006. Interdependence of fault displacement rates and paleoearthquakes in an active rift, *Geology,* **34**(10), 865–868.

Örgülü, G. & Aktar, M., 2001. Regional moment tensor inversion for strong aftershocks of the August 17, 1999 Izmit earthquake (Mw= 7.4), *Geophys. Res. Lett.,* **28**(2), 371–374.

Ouillon, G. & Sornette, D., 2011. Segmentation of fault networks determined from spatial clustering of earthquakes, *J. geophys. Res.,* **116**(B2), doi:10.1029/2010JB007752.

Papadimitriou, E., Gospodinov, D., Karakostas, V. & Astiopoulos, A., 2013. Evolution of the vigorous 2006 swarm in Zakynthos (Greece) and probabilities for strong aftershocks occurrence, *J. Seismol.,* **17**(2), 735–752.

Papazachos, B. & Papazachou, C., 2003, *The Earthquakes of Greece,* Ziti Publication (In Greek), pp. 356.

Papoulia, J. & Makris, J., 2010. Tectonic processes and crustal evolution on/offshore western Peloponnese derived from active and passive seismics, *Bull. Geol. Soc. Greece,* **43**(1), 357–367.

Pedregosa, F., *et al.*, 2011. Scikit-learn: machine learning in Python, *J. Mach. Learn. Res.,* **12,** 2825–2830.

Pérouse, E., *et al.*, 2017. Transition from collision to subduction in Western Greece: the Katouna–Stamna active fault system and regional kinematics, *Int. J. Earth Sci.,* **106**(3), 967–989.

Plotly Technologies Inc., 2015, *Collaborative Data Science,* Plotly Technologies Inc., https://plot.ly.

Pondrelli, S., Salimbeni, S., Ekström, G., Morelli, A., Gasperini, P. & Vannucci, G., 2006. The Italian CMT dataset from 1977 to the present, *Phys. Earth planet. Inter.,* **159**(3–4), 286–303.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.,* **20,** 53–65.

Ruscic, M., Bocchini, G.M., Becker, D., Meier, T. & van Keken, P., 2019. Variable spatio-temporal clustering of microseismicity in the Hellenic Subduction Zone as possible indicator for fluid migration, *Lithos,* **346,** 105154.

Sachpazi, M., *et al.*, 2000. Western Hellenic subduction and Cephalonia Transform: local earthquakes and plate transport and strain, *Tectonophysics,* **319**(4), 301–319.

Sachpazi, M., *et al.*, 2016. Slab segmentation controls the interplate slip motion in the SW Hellenic subduction: new insight from the 2008 Mw 6.8 Methoni interplate earthquake, *Geophys. Res. Lett.,* **43**(18), 9619–9626.

Serpetsidaki, A., Sokos, E., Tselentis, G.-A. & Zahradnik, J., 2010. Seismic sequence near Zakynthos Island, Greece, April 2006: identification of the activated fault plane, *Tectonophysics,* **480**(1–4), 23–32.

Shearer, P., Hauksson, E. & Lin, G., 2005. Southern California hypocenter relocation with waveform cross-correlation, Part 2: results using source-specific station terms and cluster analysis, *Bull. seism. Soc. Am.,* **95**(3), 904–915.

Shearer, P.M., Hardebeck, J.L., Astiz, L. & Richards-Dinger, K.B., 2003. Analysis of similar event clusters in aftershocks of the 1994 Northridge, California, earthquake, *J. geophys. Res.,* **108**(B1), doi:10.1029/2001JB000685.

Shelly, D.R., Hardebeck, J.L., Ellsworth, W.L. & Hill, D.P., 2016. A new strategy for earthquake focal mechanisms using waveform-correlation-derived relative polarities and cluster analysis: Application to the 2014 Long Valley Caldera earthquake swarm: a new strategy for focal mechanisms, *J. geophys. Res.,* **121**(12), 8622–8641.

Sokos, E., Gallovič, F., Evangelidis, C.P., Serpetsidaki, A., Plicka, V., Kostelecký, J. & Zahradník, J., 2020. The 2018 Mw 6.8 Zakynthos, Greece, earthquake: dominant strike-slip faulting near subducting slab, *Seismol. Res. Lett.,* **91**(2A), 721–732.

Stiros, S., Moschas, F., Feng, L. & Newman, A., 2013. Long-term versus short-term deformation of the meizoseismal area of the 2008 Achaia–Elia (MW 6.4) earthquake in NW Peloponnese, Greece: evidence from historical triangulation and morphotectonic data, *Tectonophysics,* **592,** 150–158.

Stuermer, K., Kummerow, J. & Shapiro, S.A., 2011. Waveform similarity analysis at Cotton Valley, Texas, in *Proceedings of the SEG Technical Program Expanded Abstracts 2011,* San Antonio , Society of Exploration Geophysicists, pp. 1669–1673.

Technological Educational Institute Of Crete, 2006. Seismological Network of Crete. International Federation of Digital Seismograph Networks, https://doi.org/10.7914/SN/HC.

Trugman, D.T. & Shearer, P.M., 2017. GrowClust: a hierarchical clustering algorithm for relative earthquake relocation, with application to the Spanish Springs and Sheldon, Nevada, earthquake sequences, *Seismo. Res. Lett.,* **88**(2A), 379–391.

Tsujiura, M., 1983. Characteristic frequencies for earthquake families and their tectonic implications: evidence from earthquake swarms in the Kanto District, Japan, *Pure Appl. Geophys.,* **121**(4), 573–600.

University Of Athens, 2008. Seismological Laboratory. International Federation of Digital Seismograph Networks, https://doi.org/10.7914/SN/HA.

University Of Patras, G. D., 2000. PSLNET, permanent seismic network operated by the University of Patras, Greece. International Federation of Digital Seismograph Networks, https://doi.org/10.7914/SN/HP.

Wardell, N., Camera, L., Mascle, J., Nicolich, R., Marchi, M. & Barison, E., 2014. The structural framework of the Peloponnese continental margin from Zakynthos to Pylos from seismic reflection and morpho-bathymetric data, *Boll. Geof. Teorica Appl.,* **55**(2), 343–367.

Wells, D.L. & Coppersmith, K.J., 1994. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement, *Bull. seism. Soc. Am.,* **84**(4), 974–1002.

Wessel, P., Smith, W.H., Scharroo, R., Luis, J. & Wobbe, F., 2013. Generic mapping tools: improved version released, *EOS, Trans. Am. geophys. Un.,* **94**(45), 409–410.

Yokota, Y., Koketsu, K., Fujii, Y., Satake, K., Sakai, S., Shinohara, M. & Kanazawa, T., 2011. Joint inversion of strong motion, teleseismic, geodetic, and tsunami datasets for the rupture process of the 2011 Tohoku earthquake, *Geophys. Res. Lett.,* **38**(7), doi:10.1029/2011GL050098.

SUPPORTING INFORMATION

Supplementary data available at *GJI* online:

Figure S1 Screenshot of interactive flow diagram for a comparison of clustering results obtained in four different frequency bands, targeting surface waves and body waves. The diagram was produced using the waveform-similarity based clustering toolbox Clusty and data from the 2018/2019 Zakynthos aftershock sequence. Each colour refers to one cluster of earthquakes. The width of connecting bands is proportional to the number of shared events between the results obtained in the different frequency bands.

Figure S2 Moment tensor solutions from Mouslopoulou *et al.* (2020) along with the estimated extent of our clusters, for comparison. MTs of events that were grouped into clusters in our study are coloured according to the cluster they belong to. Offshore faults are adopted from Mouslopoulou *et al.* (2020).

Please note: Oxford University Press is 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.