ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/175288/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Alves, Tiago M. , Taylor, Joshua, Corcoran, Padraig and Ze, Tao 2025. Data sampling strategies for accurate fault analyses: A scale-independent test based on a Machine Learning approach. Journal of Structural Geology , 105342. 10.1016/j.jsg.2025.105342

Publishers page: https://doi.org/10.1016/j.jsg.2025.105342

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

Journal Pre-proof

Data sampling strategies for accurate fault analyses: A scale-independent test based on a Machine Learning approach

Tiago M. Alves, Joshua Taylor, Padraig Corcorant, Tao Ze

PII: S0191-8141(25)00006-9

DOI: https://doi.org/10.1016/j.jsg.2025.105342

Reference: SG 105342

To appear in: *Journal of Structural Geology*

Received Date: 11 July 2024

Revised Date: 8 January 2025

Accepted Date: 9 January 2025

Please cite this article as: Alves, T.M., Taylor, J., Corcorant, P., Ze, T., Data sampling strategies for accurate fault analyses: A scale-independent test based on a Machine Learning approach, *Journal of Structural Geology*, https://doi.org/10.1016/j.jsg.2025.105342.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2025 Published by Elsevier Ltd.

Journal Pre-proof

Data sampling strategies for accurate fault analyses: A scale-independent test based on a Machine Learning approach

Tiago M. Alves[1], Joshua Taylor[1,2], Padraig Corcorant[2]; Tao Ze[3]

[1]3D Seismic Lab, School of Earth and Environmental Sciences, Cardiff University, Main Building, Park Place, Cardiff, CF10 3AT, United Kingdom (email: alvest@cardiff.ac.uk)

[2] School of Computer Science and Informatics, Cardiff University, Abacws Building, Senghennydd Road, Cathays, Cardiff, CF24 4AG, United Kingdom

[e] Key Laboratory of Tectonics and Petroleum Resources, Ministry of Education, China University of Geosciences, Wuhan, 430074, China

Abstract

Seismic and outcrop data from SE Brazil, Greece and SW England are used to develop a new method to correctly identify tectonic fault segments – either active or quiescent - using a machine learning approach. Three-dimensional (3D) analyses of tectonic faults are often based on the mapping of throw values (T) along their full length (D) or depth (Z) using a wide range of data. Yet, the collection of these throw values using geophysical or outcrop data is often time-consuming and onerous. In contrast to many empirical measurements of T/D and T/Z, our new method supports the mapping of active (or potentially active) fault segments and limits data undersampling, a caveat that results in the grouping of faults as single zones, systematically overlooking their natural segmentation. The new method is scale-independent and resulted in the definition of a minimum sampling ratio necessary for accurate fault segment mapping. Determined through the gradual downsampling of T/D and

T/Z data to a critical point of information loss, the minimum sampling interval (δ) in T/D and T/Z data, expressed as a percentage of fault length, or height, is: a) $1.02\% \pm 0.02$ for faults that are longer or higher than 3.5 km; b) $4.167\% \pm 0.18$ for isolated faults that are shorter than 3.5 km in either length or height. This work is therefore important as it shows that one should never acquire T/D and T/Z data above a threshold δ value of 4% to identify successive, linked fault segments, whatever their scale. Total accuracy in fault-segment detection is only assured for δ values of 1% when in the presence of fault zones with segments longer than 3.5 km. As a corollary, we confirm that T/D and T/Z data are often undersampled in the published literature, leading to a significant bias of subsequent interpretations towards coherent constant-length growth models when analyzing both active and old, quiescent fault systems.

Keywords: Data sampling; Machine Learning; Tectonic faults; fault growth; Sampling errors; Fault propagation models

1. Introduction

The mapping and geometrical characterization of faults and joints at varied scales of observation are vital to geological, structural and earthquake-risk analyses. Recognizing faults and joints is also important in hydrocarbon and geothermal energy production, in engineering works, and to the implementation of sub-surface storage solutions (Gudmundsson et al., 2002; Misra and Mukherjee, 2018; Trippetta et al., 2019; Torabi et al., 2023). Measurements of both active and quiescent tectonic faults need to be accurate because: a) the trapping and accumulation of subsurface fluid often depend on the geometry and interaction styles of faults and joints (Yielding, 2015), b) drilling-related hazards are

Journal Pre-proof

frequent in highly-faulted areas, as well as in prospects where reservoir quality is much reduced by joint systems (Saeidi et al., 2014; Kozłowska et al., 2017), c) the migration and preservation of sub-surface fluid is, in many a prospect, associated with the timing of formation, growth, and sizes of tectonic faults and joints (Ferrill et al., 2017; 2020). The size of tectonic faults, their inherent geometry, and the location of their intersection (linkage) points are important for a safe and sustainable production of geological resources (Jentsch et al., 2020; Purba et al. 2019; Moska et al., 2021; Huenges et al., 2013).

Another characteristic of tectonic faults is that their size is a predictor of earthquake magnitude, i.e. faults over a certain length are capable of generating destructive earthquakes and associated geohazards (Trippeta et al., 2019). Importantly, large magnitude earthquakes can be generated in seemingly discrete fault segments that are connected to form a single large fault zone, at depth, whereas relatively isolated, smaller fault segments present a much lower seismic risk (Cloetingh et al. 2010; He et al. 2019; Alves, 2024). Generally speaking, fault intersections, geotechnically unstable fault zones and active faults capable of generating earthquakes must be avoided in engineering projects. The systematic undersampling of tectonic faults' geometries can result in a rapid degradation of infrastructure after the completion of construction works, or in unexpected cost increases (Aydin et al., 2004; Wang et al. 2022).

Journal Pre-proof

Notwithstanding the fact that accurate fault analyses are crucial in structural geology, producing a high-resolution image of fault structures is time-consuming, expensive and not always practical. Automatic methods to extract faults using remote sensing data have been developed by authors such as Gloaguen et al. (2007), but these types of data are not always available or concern the scales of observation necessary for a particular aim, or analysis. In most instances, the solution followed by both academia and industry is to reduce data collection to the minimum required level, just enough to understand where the loci of fault

Journal Pre-proof

interactions are. Unfortunately, such an approach results in coarse and often random data sampling ratios and techniques, as one can easily verify in most scientific articles published in the past 20-30 years. Purposely anonymous examples from the published literature, include: a) fault-throw values measured every 600 m for a single fault segment that is 4 km-long, in which only 6.6 data points were acquired for such a segment (for an average of 15 measurements per fault in the same article), b) fault throws measured every 3 cm for faults that are 1.0 m-long, returning 33 measurements per fault, on average, c) a third example in which fault throws are measured every 100-200 m for a fault zone that is 4 km-long, returning an average of less than 20 data points per fault segment. The coarse, and often random samplings of throw/distance (T/D) and throw/depth (T/Z) data are common in the literature, and surprisingly not always deriving from the use of seismic and geophysical data of relatively poor resolution.

Such erratic sampling strategies can lead to a generic failure in recognizing that fault segments are components of a larger fault zone (Walsh et al., 2003). In fact, Tao and Alves (2019) have shown the systematic undersampling of fault throws in seismic, remote sensing and outcrop data will inevitably lead to an over-reliance of models reflecting coherent ‘fast propagation’ styles of fault growth (Walsh et al., 2003; Nicol et al., 2020). In other words, naturally segmented faults, or fault zones, will appear as single long structures if fault throws are undersampled. This caveat is compounded when interpreters overlook map-view geometries and concentrate only on collecting throw values without an accurate structural mapping accompanying their workflows.

The aim of this work is to produce reliable predictions of fault segmentation in an automated manner, without human bias, and avoiding any under- or overfitting of data to emphasize a particular fault growth model. Overfitting in this case would involve finding more faults than exist through misinterpretation of signal noise and height undulation caused

Journal Pre-proof

by erosion or poor exposure, for instance. Underfitting would be to exaggerate fault throw so that multiple segments appear coherent in their growth and part of a single fault zone (Torabi and Berg, 2011; Tao and Alves, 2019). Our approach is scale-independent and works for both active and quiescent (ancient) tectonic faults that may or may not reactivate by anthropogenic means. In summary, the main research questions addressed in the work include:

a) What mathematical methods can be applied to Machine Learning tools to avoid interpretative errors when identifying tectonic faults?

b) What are the implications of misrepresenting fault segmentation in terms of understanding their growth modes?

c) What are the threshold fault-throw (or displacement) values necessary for a correct identification of fault segmentation in nature?

2. Theoretical aspects concerning fault-segment recognition

2.1 Coherent vs. isolated growth modes and scale variance in structural observations

Tectonic faults and joints, universally named as ‘rock fractures’ in the published literature, comprise sets of related segments, or strands, that can be kinematically and spatially related (Pollard and Segall, 1987; Gudmundsson, 2012) (Fig. 1). They represent continuous, brittle breaks in rocks, be it crustal-scale stress in the case of tectonic faults or smaller localized stresses that hardly offset rocks in the case of polygonal faults and joints (Peacock et al., 2017; Laubach et al., 2018). The largest of faults, those documenting a clear vertical or horizontal offset in strata or rocks, are often part of system of related fault segments that interact and link - and are restricted to a relatively narrow band or volume -

Journal Pre-proof

also called a Fault Zone (Peacock et al., 2000; 2017). Fault zones are formed by the 3D linkage of multiple segments in a broad region of deformation, leaving behind fault segments not frequently affected by such a strain (Rotevatn et al., 2019).

Fault segments may show geometries that are indicative of a 'fault-linkage' and 'coherent' growth (Kim and Sanderson, 2005) or, instead, develop individually to obey an 'isolated' growth mode (Walsh et al., 2003) (**Fig. 2**). In practice, many 'linked' or 'coherent' faults are part of a larger zone of deformation, while isolated faults show growth histories and throw distributions that are independent or disparate from nearby faults segments (Nicol et al., 2020) (**Fig. 2**). The recognition of such fault growth modes in geophysical or outcrop data relies on the correct mapping of fault throws (T) against fault zone length (D) and depth (Z) to produce T/D and T/Z plots (Cartwright et al., 1998; Baudon et al., 2008) (**Figs. 1 and 2**). Multiple examples of how throw data can be used to understand fault growth modes are given in the literature for Norway (Tvedt et al., 2013; King and Cartwright, 2020), SE Brazil (Varela and Mohriak, 2013; Plawiak et al., 2024), Gulf of Mexico (Cartwright et al., 1998; Shen et al., 2018) and for tectonically active areas in the Gulf of Corinth (Fernández-Blanco et al., 2019; Robertson et al., 2020; Nixon et al., 2024), offshore Crete (Caputo et al., 2010; Nicol et al., 2020; Mechernich et al., 2023) or the Basin and Range, where topographic information has been combined with local tectonic analyses (Lee et al., 2023). Whenever available, fault displacement should be used instead of throw (see **Fig. 3**), but the acquisition of such data is time-consuming in practice when analysing outcrop or geophysical data - as a result, fault throw (T) is more frequently measured (e.g. Cartwright et al., 1998). Fault throw is a measure of the vertical distance between the footwall tip of a fault and its corresponding hanging-wall tip (Mukherjee, 2019) (**Fig. 3**). Fault displacement concerns the total movement of two fault blocks along a fault plane, measured in any specified direction. It represents the distance between two separated pieces of a marker layer on both sides of a fault. The time

Journal Pre-proof

and effort needed to collect such fault data is often the source of 'censorship' and 'truncation' in data (Torabi and Berg, 2011), leading to incorrect assumptions regarding the relative timing of fault activity.

A caveat often overlooked by structural interpreters is that recognizing fault segments depends on the distinction of meaningful throw gradients that represent segment linkages on T/D (or D_{max}/L) plots, accompanied by their analysis on vertical sections and map view (Walsh and Watterson, 1991; Walsh et al., 2002, 2003; Kim and Sanderson, 2005) (**Fig. 1**). With lower resolution images, or remote-sensing data of lower quality, comes a high level of uncertainty over the linkage points of discrete fault segments when acquiring such T/D or T/Z data. The lack of chronostratigraphic markers can also result in the misinterpretation of important gaps between faults, and multiple small segments may appear as a single large fault when a low-resolution dataset obscures lows, or minima, in throw (Tao and Alves, 2019).

2.2 Use of T/D and T/Z data in fault-segment recognition

Fault throw/distance (T/D) and throw/depth (T/Z) data are often measured on a seismic section, or exhumed fault plane, in order to identify distinct fault segments and interpret a fault propagation mode (Torabi and Berg, 2011). Fault throw (T) is often used instead of displacement as it is an easier variable to define, and quantify, in geophysical and field data, regardless if a fault is planar or listric. Throw measurements in listric faults will overlook their horizontal component (heave) but can still be used to identify discrete fault segments. In parallel, throw/distance (T/D) plots measure throw distributions along a fault's length and can be complemented by Throw-Depth (T/Z) measurements. While T/D data help recognizing distinct, linked fault segments, T/Z data indicate the areas where the mechanical

Journal Pre-proof

properties of rocks may vary across a fault, at the same highlighting any evidence for vertical fault linkage (Cartwright et al., 1998; Baudon et al., 2008).

Distinct faults, and also their constituting segments, show distinct orientations and curvatures in map view (Kim and Sanderson, 2005). On T/D profiles, steep decreases in throw values relate to the existence of an intersection (a 'hard' or 'soft' linkage point) between two fault segments or, instead, points out to a fault's lateral tip (**Figs. 1 and 2**). Two linking fault segments will also be recorded as sudden gradient changes in T/D and T/Z plots (**Figs. 1 and 2**). Conversely, variations in fault height caused by erosion and local sediment deposition will be seen as high-frequency, low-magnitude undulations that resemble a noise-like pattern of throw distributions (Torabi et al., 2019). Throw and displacement can be particularly affected by erosion of a fault scarp, as both are measured from a defined height at the immediate footwall block of a fault (**Fig. 3**).

3. Data and Machine Learning methods

3.1 Fault-throw data

Measurements of fault throw used in this work were taken from distinct parts of the world (**Fig. 4a**). T/D and T/Z measurements for 415 faults were used in our analysis - they were collected at regular intervals and used to test the sampling distance necessary to correctly interpreted fault linkages and their growth modes. The primary source of data comes from the Southern North Sea and Southeast Brazil (Alves et al. 2022; Zhang et al., 2022; Tao and Alves, 2016; 2019). Outcrop data were gathered in various locations in Crete and Somerset (Tao and Alves 2019; Gaki-Papanastassiou et al. 2009; Caputo et al. 2010; Alves and Cupkovic 2018).

3.1.1 3D Seismic data

Seismic data in this work comprises two high-quality seismic volumes from the SE Brazil (**Figs. 4a,b and 5a**). The volume was stacked with a bin (or trace) spacing of 12.5 m and a vertical sampling rate of 2 ms. The vertical resolution of the investigated seismic data varies from 5 to 8 m near the seafloor, and c. 12 m at the maximum depth of faults investigated in this work (**Fig. 5a**). Fifty-nine (59) faults, including crestal faults, radial faults and low-angle normal faults flanking salt diapirs were interpreted every 1, 3, 5, 10 and 20 inlines and crosslines (**Fig. 5a**). Composite lines were also used, when needed, to collect data perpendicularly to fault-plane dip. Interpreted faults are 225 m to 5,000 m long and show throw values varying from 6 ms to 73 ms two-way time (twt). These faults are still active at present as some offset strata that are very close to the modern seafloor due to on-going salt tectonics in SE Brazil (**Fig. 5a**).

Journal Pre-proof

3.1.2 Ierapetra Fault Zone (SE Crete)

The modern Ierapetra Fault Zone is located in SE Crete and is >25 km long (**Fig. 4a,c and 5b**). It has been active since, at least, the Late Miocene and is one of the most prominent structures on the island (Caputo et al., 2010; Gaki-Papanastassiou et al., 2009). Several fault segments striking NNE–SSW and dipping to the WNW played a crucial role in the evolution of the fault zone, namely the Kavousi, Ha and Ierapetra segments (Gaki-Papanastassiou et al., 2009) (**Fig. 5b**). Each of these segments has its own characteristic geometry (**Fig. 5b**). Due to

Journal Pre-proof

its activity, thick sediments cover the fault zone's hanging-wall, while the immediate footwalls are barren of marine sediment and feed adjacent basins at present (**Fig. 5b**).

Throw/distance (T/D) data reveal that the fault segments are 0.5 to 7.1 km long, show maximum throw values between 250 and 1000 m. Nevertheless, a synchronous Holocene reference horizon was identified in the study area and used as a marker to compile T/D plots for outcropping fault segments (**Fig. 3**). During the collection of fault-throw data, the following were performed:

(i) Fault scarps were mapped in detail in the field and projected on 1:50,000 maps from the Hellenic Mapping and Cadastral Organization – the maps with the highest resolution in the region. The present-day height of footwall tips and any associated erosional and depositional features were taken into consideration in our throw measurements of active tectonic faults,

(ii) Throw data were collected at a regular interval of 50 m along the fault segments observed in the field. Throw measurements were gathered where the geometry of the faults is clear on the maps and in panoramic photos (**Fig. 5b**).

3.1.3 Sub-seismic scale faults from SW England (Kilve)

The Bristol Channel Basin records four distinct stages of faulting: 1) N-S extension and associated normal faulting in the Mesozoic, accompanying the development of the Bristol Channel Basin, 2) reactivation of some of the normal faults formed during the first stage, 3) reverse reactivation of Mesozoic and older structures during the Alpine orogenic pulses (Underhill and Paterson, 1998), 4) reverse-reactivation of normal faults that were

Journal Pre-proof

subsequently cut by conjugate strike-slip faults (Dart et al., 1995), 5) jointing of strata after Alpine-related fault reactivation (Rawnsley et al., 1998).

A certain degree of tectonic reactivation thus occurred in the Bristol Channel Basin during the Cenozoic and was of an enough magnitude to generate: a) structures formed by N-S contraction - chiefly reverse reactivated planar normal faults, b) structures formed by east–west contraction, c) intersecting N- to NNW-trending and NE-trending faults (Glen et al., 2005). Importantly, the faults analysed in this paper were formed by N-S extension, record no apparent tectonic reactivation, and only occur in Liassic limestones and shales (Peacock et al., 2017).

Thirteen (13) faults with lengths varying from 1.65 m to 7.55 m, and maximum throw values ranging from 3 cm to 29 cm, were measured and interpreted in the field (**Figs. 4 and 5c,d**). Fault-throw measurements in the field depended on how clear they were exposed at the surface. Throw values were measured where the hanging-wall and footwall were totally exposed on the two sides of the fault trace. The throw-distance data were acquired along the exposed fault trace every 5 cm. T/D plots were also computed and analyzed for these faults considering different sampling spacings as exemplified in the Supplementary Materials in Tao and Alves (2019).

3.2 Machine Learning and mathematical algorithms

Machine Learning algorithms were implemented using the Python programming language applied on NumPy (Harris et al., 2020), PyWavelets/Pywt V1.4.1 (Lee et al., 2023) and SciPy 1.0 (Virtanen et al., 2020) software libraries.

Journal Pre-proof

3.2.1 Wavelet transforms for fault-segment detection

The main advantage of using Wavelet Transforms to detect discrete fault segments is that they permit the analysis of features that vary in character over different scales (Kalbermatten et al., 2012; Shen et al., 2022). For acoustic or optical signals, such features are often frequencies varying over time. In image data, features of interest include edges and textures, as is the case of throw maxima and minima in T/D and T/Z curves (Shen et al., 2022), or object-based classes of images recorded after segmenting remote sensing data into homogeneous regions (Gloaguen et al., 2007).

In mathematical terms, Wavelet Transforms allow for the decomposition of an input signal into the intensity of individual frequency bands. The advantage of the Wavelet Transform method over the Fast Fourier Transform is the former's ability to identify both the frequency and spatial position of frequencies in the data. Fast Fourier Transforms only provide frequency information over a fixed range, with no location value along that range (Sifuzzaman et al., 2009). A wavelet can thus be convolved with a signal and the resulting signal gives the intensity of the wavelet at each point along the signal. The wavelet size can be changed to give an intensity for each frequency band.

In order to have a successful Wavelet Transform, a wavelet must follow a set of criteria, namely the wavelet function ψ needs to return a zero average:

$$\int_{-\infty}^{+\infty} \psi\,(t)dt = 0. \qquad \text{(Eq. 1)}$$

The wavelet is then multiplied by a scale parameter *s* and translated by *u* such as:

Journal Pre-proof

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right) \qquad \text{(Eq. 2)}$$

The Wavelet Transform of *f*, at a scale *s* and position *u*, is finally computed by correlating *f* with a wavelet atom:

$$Wf(u,s) = \int_{-\infty}^{+\infty} f\,(t)\,\frac{1}{\sqrt{s}}\psi^{*}\left(\frac{t-u}{s}\right)dt \qquad \text{(Eq. 3)}$$

The above equations use *t* as the measure of displacement across the signal, as wavelets are most often related to signals represented as a function of time. In our particular case the Wavelet Transform will not be reprocessed as a function of time; it will be estimated along a measured distance, making no difference to the mathematics used. Time (t) will be replaced by distance (D) in Wavelet Transforms, this parameter D being the distance along a fault plane considered earlier in the paper, with frequency and wavelength being also be processed in reference to distance. By convention, t is used in signal processing, but for our case study distance (D) is used where t is seen in Equations 1 to 3.

Particular wavelet types are more often used in signal processing, and is thus best to choose one of these common types when performing a Wavelet Transform. A wavelet that follows a similar shape to the expected signal is required to get the best results (Mallat, 2009). In this work we used the so-called *Ricker wavelet* (see **Fig. 6**). Such a wave shape allows for the isolation of peaks, or throw maxima, in a fault segment, with throw minima being mathematically defined as the wavelet boundaries – as with distinct fault segments that

Journal Pre-proof

are part of a fault zone (Mallat, 2009). Hence, the *Ricker wavelet* closely matches the linkage behaviour of fault segments (see Wang, 2015a; 2015b), i.e. it better identifies sharp throw minima, which are known to indicate the places where distinct fault segments were originally linked (**Fig. 6**). Such an approach results in the strongest correlation possible whenever the *Ricker wavelet* is equal to the fault size.

3.2.2 Polynomial regressions as a complementary method for fault-segment detection

Polynomial regressions follow a similar process to linear regressions whereby a line with a minimum average distance to the data points is found. Such a distance is quantified by a Sum of Square Errors (Heiberger and Neuwirth, 2009). The advantage of such a polynomial regression relates to the ability of using a higher order equation to define the line of best fit to T/D and T/Z data, respectively where the x-axis is distance (D) and the y-axis corresponds to depth (Z). In the case of a third order polynomial, the key values are the coefficients, and a polynomial regression model can be simplified to these coefficient terms, i.e. the terms can be used as predictors for the values in the 'real' field data. This simplification to a single equation is important in our work, as it allows for data comparisons for the same fault whenever the T/D and T/Z measurements are correctly sampled vs. when data are downsampled.

To avoid data overfitting, we used a lower degree polynomial of degree 3. In practical terms, a three-parameter polynomial equation is first generated for each of the identified fault segments. The absolute minimum number of sample points used to generate this first model of fault shape is three (3), so the sampling space is so low that only the two tips and the point of maximum throw of a fault are identified (e.g. **Fig. 7**). The purpose of this method is to allow a comparison of fault detection approaches, using different sampling ratios, by reducing them all to the same dimensions. The low complexity of this method also helps to ensure that the model does not overfit the T/D and T/Z data in this work. We verified that the above findings could be generalized to our specific data by verifying that the error between the model and the T/D and T/Z data in question was small. A more detailed explanation of the polynomial regression process can be found in Ostertagová (2012) and James et al. (2013) and Section 4.4 in this work.

The modelling of faults via polynomials works well due to the process of fault creation, itself the result of forces, or stresses, developing and growing fractures in a volume of rock. Over geological periods of time, such forces change in terms of their direction and magnitude, and multiple factors can cause local variations in space and time (4D) in stress-strain relationships (Kim and Sanderson, 2005). At a single point in time, a skewed polynomial shape can accurately follow the shapes of faults and joints in nature, as the forces acting on a volume of rock result in a fault following a path of least resistance. This promotes the formation, in nature, of Gaussian T/D and T/Z curve shapes in faults and joints. The various (unpredictable) factors acting on these same structures, and altering their T/D and T/Z profiles, can thus be simplified as skewed Gaussian curves. Goff (1991) found that a skewed Gaussian curve provides a model of low complexity that accurately fits our type of data.

4. Results

As a summary, the workflow used in this work is shown in **Fig. 8** to highlight the different steps of the proposed machine learning methodology.

Journal Pre-proof

Journal Pre-proof

4.1 Step 1 – Application of discrete Continuous Wavelet Transforms (CWT) to resolve faults at different scales

Theoretically, Continuous Wavelet Transforms (CWTs) can produce a 2D plot of frequency band strength. For the purposes of this work, these band strengths correspond to variations in fault throw (T) when this throw is interpreted as a part of a wave. Hence, fault segments in the field or in seismic data can be represented as wavelets.

In this work, computed CWTs were visualized against T/D plots, with a clear correlation being observed between frequency band strength and the throw maxima recorded for each fault segment (**Figs. 7 and 8**). In fault zones containing multiple throw maxima, the largest fault segments correlate with a peak in low frequency wavelet amplitude (**Fig. 7**).

When performing a CWT, the wavelets of various frequencies are compared across the input signal. The correlation of the signal with that wavelet is measured at each point. Therefore, when reaching the throw maxima of fault segments with a similar frequency, the accuracy behind correlating wavelets with T/D (and T/Z) plots (i.e. *correlation strength*) reaches a maximum. Such an approach simplifies the analysis of fault segments by splitting the throw measurements made in the field, or in seismic data, into frequency bands. This allows a computer algorithm to pick out certain frequencies that are likely to correspond to fault segments. Wavelets that are most similar in shape and size to fault segments, will result in a higher correlation between the CWTs and real T/D and T/Z data after convolution. This means the peak in convolution output will give the 'best match' from possible wavelet sizes and locations along a fault. Peaks in the CWTs' output can then be assumed to be the 'top' (i.e. the point of maximum throw) of a fault of a particular size.

Modifications were made in our analysis to the CWT technique so that discrete fault segments could be found. The main modification consisted in changing how the frequency of a fault is decided. Initially the frequency of discrete faults was found by identifying the point of greatest throw amplitude across a fault, or a fault zone, for each frequency. We then visually confirmed which of these amplitude maxima coincides with the throw maxima of fault segments by comparing then with acquired throw data, acquired at maximum resolution. However, such an approach was deemed unreliable when: a) multiple segments in a fault zone show similar lengths, and b) an entire fault zone follows a shape similar to the wavelet, in which case a very low frequency will be used, spanning the entire fault zone. This caveat results in the smallest segments being ignored by the algorithm. An example can be found in **Fig. 7**, where some obvious local fault segments were missed.

A successful solution was found using an approach that required the application of a computational step to remove the highest frequencies representing a ‘noisy’ signal. Peaks and Troughs in the computed wavelets were found by a comparison of points to their immediate neighbours (see Section 4.2). This is called mathematically as calculating *prominence*. Prominence is calculated by finding the minimum between a peak and the next higher peak, so the comparison happens over a range around a peak, not just the immediate neighbours - a full mathematical explanation of *prominence* is given in https://www.mathworks.com/help/signal/ug/prominence.html. If a point was found to be higher in value than its adjacent points and had sufficient *prominence* in the whole of the fault zone, it was taken as a Peak by our algorithm. Troughs were found in the same way, using an inverse algorithm so the same function can be used.

In a second stage of this process, a wavelet band was chosen by removing wavelets that are not considered relevant, as they mostly represent noise (**Fig. 8**). The highest wavelet frequency band remaining in the dataset was then considered to be ready for fault scanning. The use of the highest wavelet frequency avoided locating the longest faults early in the process, as the scale (and wavelet range) of these long faults usually overprint the smallest

Journal Pre-proof

fault segments before these are found. For instance, in **Fig. 7** we can identify discrete throw maxima relating to the presence of small fault segments that were overlooked by the algorithm that, in Step 1, was focused on picking the greatest throw maxima. This means that the identification of relatively small throw maxima need to be prioritized in a Machine Learning approach.

In summary, the maximum value in the wavelet band that is not interpreted by the algorithm as a discrete fault segment was defined as the maximum throw value of a new segment. Conversely, the throw minima on each side of this maximum were taken as comprising the lateral tips of a fault segment. Such a method could be applied to a map all maxima and minima in the produced CWT matrix. This method allows for a rigorous definition of fault segment distribution and their linkage points. To avoid errors in our analysis a cross-validation was used to select the most suitable frequency. We split the data into training and test cases. The frequency was selected using the training cases and this was determined to be a suitable frequency through evaluation of the test cases. A subset of the dataset was chosen randomly to use for validation of the frequency constant. A ground-truthed set of fault locations was then marked on the dataset. The constant that came closest to this ground-truthed data was taken and modified by smaller amounts for a different subset of the data, repeating the same process to address any bias introduced.

4.2 Step 2 – Detection of throw gradients from the point of throw minima

Step 2 in this work consisted in the application of a gradient descent from the point of frequency minima. The aim was to find the nearest throw minimum representing the linkage point between two fault segments. If no frequency minima are found before reaching the end

Journal Pre-proof

of the dataset, the last value picked by the algorithm is taken as the end of the segment (**Figs. 8 and 9**).

The method consisted in the scanning of every wavelet frequency for their Peaks and Troughs, which are then reduced down to frequencies that contain enough Peaks and Troughs to form at least one discrete fault segment. A second reduction is completed by removing the frequencies that result in too many Peaks per meter. Such a step is important for removing frequencies that reflect irrelevant, spurious throw maxima, usually comprising measurement errors and resolution issues when measuring throw data in seismic and at outcrop (truncation and censoring cf. Torabi and Berg, 2011). The threshold Peak values can be changed, with a stricter threshold resulting in the identification of only the larger fault segments (see **Fig. 9**), and a looser threshold resulting in multiple fault segments being found. Naturally, if it is set too loose, unwanted segments may appear in one's fault tracing.

4.3 Step 3 – Integration of Continuous Wavelet Transforms (CWTs) with a threshold Peak rate

To improve the accuracy of our results, a re-sampling was applied as a third step before undertaking a CWT. The sample count was scaled to 1,000 times the longest wavelet length, which resulted in a less unusual behavior whereby segments are too large to be detected by any of the wavelets. This allows the wavelet bands to be kept the same for all tests, even while the dataset sample sizes vary. After all processing was done, an optional process allowed for the joining of the segments, to remove gaps between them. Step 3 returned more accurate results when undertaken on a series of faults where no gaps are expected, i.e. the approach also meant the lowest throw between any two segments was always considered as the linkage point of successive fault segments, regardless of their scale in nature.

Journal Pre-proof

A value between 0.03 and 0.04 for the threshold Peak rate (number of peaks per sample) was found to provide good results in the datasets tested in this work. This value was decided by plotting Peak rate values against frequency and finding where the graph **in Fig. 10** begins to level out. In this graph, the rapid descent recorded with increasing wavelength sizes represents the reduction in noise occurring as the small changes in throw are filtered out by the algorithm. Once this noise is filtered out, and the curve approaches a flat, we can be confident that the remaining data is accurate. Cross-validation can then be used to select the above values for the threshold Peak rate.

Selecting a threshold Peak rate must be consistently applied across all tests to make one's results comparable later on. However, in practical terms, the threshold can be changed based on the smallest fault sizes one has to find in a dataset, although using a threshold too high results in the detection of throw maxima that are the result of random noise or constitute irrelevant changes in fault height in a discrete segment. The adoption of a 0.04 Peak rate returned positive results in this work - all Peaks that are clearly not part of faults were ignored, without overlooking any possible faults, examples of which can be seen in **Fig. 11**. Cross-validation against the ground-truthed throw data was again used to obtain a value of 0.04. If a different dataset with different properties is used, then cross-validation is also performed with respect to that dataset to select the most appropriate value. In practice, 0.04 was chosen by validating it across a large dataset and should be considered a 'default' value to use but can also be changed depending on whether its use causes false positive or false negative faults. This approach allowed us to use the previously defined method of Wavelet-Transform scanning described in Section 4.1, starting with the highest frequency, as we have now removed noisy wavelet bands that could hinder such a Machine Learning approach.

Journal Pre-proof

4.4 Step 4 – Throw-profile fitting via a cubic model

The computational steps so far described are successful in identifying the tips and throw maxima for each fault segment, but fault shape is often not accurately depicted. To best represent fault segment shape, a third order polynomial regression needs to be applied individually to each fault segment (**Fig. 12**).

In our database, fault shape approaches a cubic equation in almost all cases; the evolution of fault shape is a result of stress and ruptures in the lithosphere that can be interpreted using the same models that dictate the geometry of failure in the smaller scale and in other materials (Scholz and Aviles, 2013). A discrete fracture developing in a rheological uniform material usually produces a parabolic fault in 2D (Walsh et al., 2002; 2003; Kim and Sanderson, 2005). However, in nature the interaction with varying rock types, adjacent faults and other irregularities within the crust add an order of complexity to fault shapes. This is correctly accounted for with the use of a cubic model (Goff, 1991; Ostertagová, 2012). A second order polynomial is only capable of modelling a curve with a single peak or a single trough. Since the dataset used in this work contains multiple peaks and troughs, such a model is unsuitable; using a third order polynomial overcomes this limitation whereby it can model curves with multiple peaks and troughs.

In this fourth step, the regression model developed for fault segment detection is provided with throw data at the maximum resolution possible. However, a set weighting was added for the minima, maxima and Peak throw values, thus ensuring the final curve passes through each of these points. In addition, a lower weighting is given to the peak to prevent the detection of unusual shapes due to other points being ignored by the model. A regression was then applied through the implementation of the python software library Scikit-learn (Grisel et al., 2023), which implements a simple and effective regression algorithm that

Journal Pre-proof

allows for quick implementation into the code used in previous steps (Grisel et al., 2023; Raschka and Mirjalili, 2018). An advantage of this step is that the resulting curves can model each fault using a single equation, and the computation of such equation simplifies any further analysis needed for a fault (**Fig. 12**). It should be noted that the resulting equation will only give an accurate model of fault shape within the range of the fault's predicted length. Outside this range the cubic equation does not fit with the real fault shape (**Fig. 12**).

5. Critical mathematical tests of minimum sampling rates for T/D and T/Z analyses

A minimum sampling rate for T/D and T/Z analyses was previously estimated by Tao and Alves (2019) as a percentage of the smallest fault segment. They approached the detection of fault linkage points to the mapping of a fault's total area, and geometry, in the 2D space. Hence, a downsampling method was gradually applied by Tao and Alves (2019) to data collected at maximum resolutions so to highlight fault linkage points in T/D and T/Z plots. Fault-segment linkage points were detected for each iteration. The number of fault segments was then measured and, once this number was reduced, fault segments could not be detected below a specific sampling rate δ.

Mathematically speaking, the standard approach to downsampling a dataset is through decimation, which involves the application of an integer decimation factor M. The new decimated data are obtained by simply selecting every M^{th} value of a signal $x(n)$, a step that returns a new sample rate of:

$$n' = \frac{n}{M} \qquad \text{(Eq. 4)}$$

Decimation methods most commonly used involve the application of a low-pass filter prior to decimation so that aliasing is avoided. However, to most accurately simulate the degradation of data that derives from a lower sampling of field measurements, we decided to avoid the application of a low-pass filter to our data. In fact, the decimation approach in Equation 4 is the most basic and allows only for quick tests of the effect of sampling reductions on the shape of T/D and T/Z data.

In our analysis, decimation was found to introduce a bias to downsampled data. The results were often determined by the locations of the decimated samples relative to the 'real' linkage points of discrete fault segments. Hence, to allow for a consistent approach to downsampling, an interpolation algorithm was used whereby an interpolation function was generated and followed the input data. This interpolation function was then applied to a new set of sample points, an approach most closely following what happens when acquiring T/D and T/Z data in the field, or in seismic data, as we can choose – by using this interpolation function - a completely new set of sampling points independently of how the original data was acquired.

A linear interpolation was therefore followed in our approach by computing two adjacent samples, with the desired sampling point falling between these adjacent sample locations along the throw axis. The normalized spacing between these two samples is *1/U*. If the distance of the first sample comes before the desired sample distance by x_m, then the sampling distance of the second sample leads the desired sampling distance by $(1/U) - x_m$. If we designate the two samples as $y_1(m)$ and $y_2(m)$, and use a linear interpolation, the approximation of the desired sample becomes (Proakis, 1992):

Journal Pre-proof

Journal Pre-proof

$$\begin{aligned} y(m) &= (1 - a_m) y_1(m) + a_m y_2(m) \\ \text{where} \quad a_m &= U x_m \quad \text{and} \quad 0 \leq a_m \leq 1 \end{aligned}$$

(Eq. 5)

Through this resampling approach, a sampling ratio can be increased and decreased while retaining the original fault shape. The position of sampling points can also be tweaked to find possible sampling intervals that cause information loss.

5.1 Integral Error test

The main result of performing a polynomial regression fit is that an interpreted can obtain a discrete equation for each perceived fault. Building upon the method of Modulus Error analysis in Tao and Alves (2019), we created a measure of the scale of changes caused by a reduction in throw sampling. We subtracted the equations of faults measured at different sampling spaces and took the absolute value of the resultant equation, where x is distance:

$$\begin{aligned} \text{Total error} &= \frac{\sum_{i=0}^{n} \int_{p_i}^{p_{i+1}} |f_i(x) - g_i(x)|\, dx}{\sum_{i=0}^{n} \int_{p_i}^{p_{i+1}} f_i\,(x) dx} \\ p &= \text{Intersection points} \end{aligned}$$

(Eq. 6)

For a single fault, Equation 6 can be simplified to:

Journal Pre-proof

$$\text{Fault error} \quad = \frac{\int_p^q |f(x) - g(x)|\, dx}{\int_p^q f\,(x) dx}$$

$$p = \text{Fault start} \qquad \text{and} \quad q = \text{Fault end}$$

(Eq. 7)

Performing an Integral Error calculation on each step of a sampling reduction test reveals some of the effects imposed on the identification of fault-linkage points when one randomizes data (throw) sampling (**Fig. 13**). As the sampling is reduced, the Integral Error increases, responding to the fact that the sampled locations may miss the fault linkage points if the sampling is too coarse. The error will reach a maximum value and then decrease over smaller changes in sampling. This means that coarse and random sampling techniques can drastically change the results, leading to erroneous estimations of fault segments' shape, hindering their subsequent identification. In other words, it is certain that one is overlooking the presence of discrete fault segments when the error starts to decrease in its magnitude (**Fig. 13**). In addition, when the sampled points are being incrementally reduced along a fault, the distance to the nearest sample may also vary with some degree of randomness. An interpreter may thus be fortunate enough (or not) to collect data near a point where fault segments are linked solely by chance. The influence this has on the error value means that sometimes, but also randomly, error will decrease for a lower number of samples.

5.2 Modulus Error test

The approach in Section 5.1 resulted in the calculation of a ratio resolving the size of the error relative to the size of the fault $f(x)$. As we mostly recorded an increase in Integral Error up to the point where a fault is no longer detected, the variation in Integral Error

Journal Pre-proof

became a good indication of the reliability of predictions made at decreasing sampling space. The similarity of this equation to the Modulus Error equation in Tao and Alves (2019) allows for a direct comparison between different error-calculation methods as a function of sampling space:

$$\text{Modulus Error} = \frac{\sum_1^n |A_m - A'_m|}{\sum_1^n A_m}$$

(Eq. 8)

Taking the integral of Equation 8 will give a value for the area between the two faults, which can be used as a measure of error between two measurements of the same fault zone. In our case it was used to compare the downsampled datasets to the original ground-truth ones as the sampling space is being tested.

5.3 Intersection Error test

The lateral tips of discrete fault segments can sometimes change in their relative position (as identified by our algorithm) if data decimation is too ‘coarse’. Once again, an interpreter may be fortunate enough (or not) to collect data near a point where fault segments are linked solely by chance. As a result, information is lost; when fault linkage points are not identified in their accurate location, any resulting interpretations of a fault’s geometry may be inaccurate. Small changes in the location of fault segments’ linkage points may not indicate issues with their identification, so a threshold value needs to be defined if a particular sample strategy is inaccurate.

Journal Pre-proof

We devised a way to measure the change in intersection points, i.e. the difference between the lateral tip of a fault in one case is compared to the closest lateral tip of a fault in another measurement of the same fault zone. This distance is divided by the length of the fault to give an error value. The average of all the faults' errors gives a final *Intersection Error* for the comparison.

$$\text{Intersection Error} = (n_{max} - n_{min}) + \sum_{i=1}^{n_{min}} \frac{\min\left(\left|a_{i,0} - b_{(1\to n_{max}),0}\right|\right)}{a_{i,2} - a_{i,0}}$$

$$\begin{array}{rl} n = \text{number of faults} & a_{c,d}, b_{c,d} = \text{sequence of faults} \\ c = \text{fault number} & d = \text{fault start, peak and end} \end{array}$$

(Eq. 9)

The Intersection Error returns similar values to the Integral Error. However, it will more clearly identify situations where a fault segment has been overlooked. Other measures of error also prioritize changes in the general shape of faults, while in many cases the more important aspect of the faults we analyzed is where they lateral tips are, i.e. where they begin and end laterally.

6. Discussion

6.1 Downsampling techniques to highlight interpretation errors

A comparison of error percentages when reducing the sampling spacing in T/D and T/Z data reveals some interesting trends (**Fig. 14**). In most cases the error gradually increases when sampling decreases, but there are some examples of minima in Integral and Intersect errors occurring due to a sample coinciding exactly with a fault segment linkage point (see low error percentages in **Fig. 14**). In other words, by simple coincidence, one can select a

Journal Pre-proof

sample that coincides exactly with, or be very close to, a fault intersection point. This finding constitutes an important addition to the analysis of Tao and Alves (2019); it provides further confirmation that obeying a minimum threshold sampling ratio is paramount to analyzing fault segmentation in nature.

We applied an iterative downsampling approach to all the fault data available to find a minimum sample ratio as a percentage of fault length. Three (3) approaches were followed to measure minimum sampling ratios from the strictest to the most lenient:

a) Strict - sampling considers a percentage of the total data input range, i.e. the total length of a fault zone that is composed of multiple segments,

b) Moderate - sampling is calculated considering the longest segment found in a fault zone, and,

c) Lenient - sampling only considers the very first fault lost as a result of reducing throw sampling rate.

The use of multiple minimum sampling definitions allowed us to identify what are the upper and lower limits for the required sampling ratio in order to map discrete fault segments with accuracy. Our datasets often include a wider range of fault geometries, with faults varying in size along a fault zone. Results are shown in **Fig. 15**.

The results show that, with relatively short fault zones, in which only a few faults need to be found and modelled, relatively lenient sample ratios are sufficient when compared to long fault zones. The main caveat of analyzing fault zones is that they may contain long and short fault segments, and the shortest segments need to be accurately identified using strict sampling ratios. This means some fault zone geometries require a much higher sample ratio than, for instance, two-three linked segments with relatively constant sizes.

Journal Pre-proof

Journal Pre-proof

6.2 Minimum sampling ratios in T/D and T/Z analyses

Figure 16 illustrates the relationship between each error-testing approach and the critical sampling ratio, with detailed information being provided in **Table 1**. The purpose of quantifying error is to understand how much information is lost by a reduction in the sampling ratio, or distance. The larger the percentage error observed in **Fig. 16**, at a critical sample ratio, the better the measure of the accuracy of fault predictions is. The critical sample ratio is the point at which important fault information is lost.

Modulus Error works independently of any fault shape data, so it results in a smaller distribution error - it cannot reliably tell an interpreter how much information is lost in terms of fault shapes and their linkage points. In comparison, Integral Error reflects a compromise between the Modulus and the Intersection errors, though it only returns information on the accuracy of lateral tips (start and end points) of fault segments. In spite of this, Integral Error has a much higher average result for error, meaning the changes in fault shape are relatively greater than the change in position of faults' linkage points.

From these results, and also via the successful visualization of fault shape, we demonstrate that Integral Error is a superior tool to gauge the loss in information when comparing variable sampling ratios for faults. The high correlation with Intersection Error also tells us that there is little use for combining the two error-defining methods (Intersection and Interval errors) in individual cases, as they are heavily dependent. The Intersection Error can therefore be used separately to Interval Error as a good indication for how trustworthy the identification of fault linkage points will be.

After establishing a relationship between error values and critical sample ratios, we could reach a conclusion on the minimum sample ratios necessary for accurate fault analyses.

We found a minimum sample ratio that would be appropriate for various cases, with a 95% success rate (**Table 2**). The success rate measured in these cases is based on the use of a fully automated wavelet method (**Fig. 17** and **Table 2**). With the use of other tools, as well as a human input (fault segment and curvature mapping sensu Kim and Sanderson, 2005) a higher success rate will be likely achieved.

The critical values in **Fig. 17** show the minimum sampling ratios calculated for the three downsampling approaches considered in Section 6.1. The Strict approach can be taken as reflecting the minimum sampling length/fault length ratio (δ) for large fault zones comprising fault segments of varied dimensions (see also **Table 2**). These were commonly observed in the datasets gathered in SE Crete where fault-segment length is variable, but with some segments >3.5 km long. Based on these constraints, the point of data loss over a wide range of data sets was calculated in this work and resulted in the estimate of the following δ values:

a) A δ of 1.02% ± 0.02 if one uses a Strict approach for the sampling of throw data. This value is particularly important when in the presence of fault zones that are >3.5 km long,

b) A δ of 4.167% ± 0.18 for a Moderate approach, in which the choice of sampling ratio prioritizes the identification of the longest segment in a fault zone,

c) A minimum δ of 5.882% ± 1.26 is necessary to identify segments in a fault zone using a Lenient sampling approach.

For a typical fault zone that is longer than 20 km, such as Ierapetra's with its largest segments c. 3.5 km long, the results above indicate that the collection of throw values every 35 m is the minimum sampling ratio one should use. In 3D seismic data, this translates into mapping

Journal Pre-proof

fault throws every 3 lines for a typical volume with a bin spacing of 12.5 m. Moderate and Lenient approaches will respectively translate into the collection of throw data every 140 m and 200 m along the Ierapetra Fault, i.e. every 11 and 16 lines for a similar fault in a 3D seismic volume processed with a bin spacing of 12.5 m. In SW England, sub-seismic faults are 1.65 m to 7.55 m long, and that results in a Strict sampling that varies from 1.68 cm to 7.7 cm. A more Lenient sampling would require throws sampled every 9.7 cm and 44.39 cm for such structures.

It is worth noting these are not prescriptive sampling distances as, recognizing, the minimum sampling length/fault length ratio (δ) is a function on fault length. Moreover, this same rule also applies to the collection of throw data for T/Z (throw-depth) plots so to prevent the grouping of distinct segments into a single unlinked (coherent) fault.

6.3 Implications for T/D and T/Z analyses

Ze and Alves (2019) recognized that depositional rates near active normal faults vary significantly on their hanging-wall and footwall blocks, as well as recording variable sediment pathways. This renders the use of expansion indexes and layer-by-layer interpretations of throw troublesome in seismic data imaging relatively old, buried basins. The Strict approach to using a δ of 1.02% will compensate for any of the issues indicated in Tao and Alves (2019), helping in the identification of early-stage fault segmentation. It will prevent the tendency, in the published literature, of considering the constant-length model as predominant in nature. In order to reduce risk of important data loss in the interpretation of short, minor faults, we recommend the use of a δ value of 1.0% preventing the loss of important fault information. Taking the smallest fault in the area as the reference point for a δ value also gives less room for interpretation error.

A limitation concerning the use of T/D and T/Z data in fault analyses is that the scale at which structural geologists acquire and interpret fault throw (or displacement) data is variable. It depends on the inherent scale of the structures of interest, and the aims of the survey or study in question. The chosen scale of observation is also dependent on data resolution and pre-defined structural criteria (e.g. Walsh and Watterson, 1991, Walsh et al., 2002, Walsh et al., 2003, Kim and Sanderson, 2005, Torabi and Berg, 2011). Therefore, to acquire data at a scale that is several orders of magnitude greater than that in which fault segmentation likely occurred, e.g. interpreting deeply buried faults using seismic data of poorer quality persuades interpreters to readily recognize coherent fault-growth models to the detriment of the isolated growth model. This is particularly the case when faults crossing sedimentary basins, but not rooted into basement units (and, therefore, not developed at a crustal scale), are interpreted in seismic data. At what temporal scale is the 'fast-propagation', coherent fault model applicable is another important caveat in many of these models – the time-dependent growth and ultimate linkage of small faults is not easily resolved in seismic data, nor are stratigraphic (age) constraints often accurate enough. For these reasons, we consider that fault segmentation can be systematically overlooked by interpreters when adopting of broad, one-fits-all, attitude to data sampling, against which the Strict δ values suggested in this work should be used in fault analyses, but rarely are.

7. Conclusions

This work shows that the application of a Wavelet-Transform detection system in fault analysis is useful to automate fault mapping and remove human bias from interpretation workflows. With human oversight and adjustments, this system improves the productivity of interpreters analyzing complex fault arrays. As a corollary, this work proves the need to consider a threshold sampling ratio (δ) in T/D and T/Z data as necessary, based on the following results:

a) A lower sampling ratio (δ) is required when interpreting long, segmented fault zones composed of faults of multiple lengths and heights. This is important as the linkage points between fault segments often coincide with regions of throw minima that are much smaller than the throw maxima of adjacent faults. The adoption of a low sampling ratio is independent of the style of linkage between discrete fault segment, e.g. hard-linkage, soft-linkage, or relay ramps. It is also independent of the type of fault one considers (normal, reverse or strike-slip).

b) This work suggests a minimum sampling ratio (δ) of 4.167% for faults that are relatively short, and clearly isolated. This is, however, a rough guideline, as faults in nature can have some unpredictable geometries and a Strict approach (δ of 1.02% ± 0.02) may still be the appropriate, in many instances, if recognizing fault segmentation is the main aim of a study.

Journal Pre-proof

c) For the fault zones we analyzed in the field and at outcrop, a Strict sampling ratio of 1.02% will translate into throw data collected every 35 m if a fault zone contains segments greater than 3.5 km. In 3D seismic data, this translates into mapping fault throws every 3 lines for a typical volumes with a bin spacing of 12.5 m. Moderate and Lenient approaches to fault measurements will respectively translate into the collection of throw data every 140 m and 200 m for such a fault zone geometry. The smaller sub-seismic faults of SE England require a sampling every 1.65 cm (Strict approach) to 44.39 cm (Lenient approach).

d) The final decision regarding the use of Strict sampling ratios of 1.02% ± 0.02 should be based on all geological information available on the fault zone, or region, being analyzed.

If there is any major uncertainty around fault size, one should follow a Strict approach and consider a δ of 1.02% ± 0.02.

e) Mathematically speaking, a combination of Continuous Wavelet Transforms and Polynomial Regressions allows for an accurate mapping of fault segmentation from T/D and T/Z data. The Continuous Wavelet Transform is used to define fault ranges. A cubic (polynomial) regression model is later applied on these ranges to obtain fault shape in a separate stage. The high reliability of this technique allows for its systematic application using Machine Learning tools.

The results in this work are based on mathematical methods tested on a large dataset comprising 415 faults. The method we propose are applied with minimal human intervention, meaning results can be directly linked to the mathematical equations. The results also demonstrate the significant impact data sampling techniques can have on the resulting interpretation of fault location, and growth modes, particularly whenever small faults are quickly lost due to sub-scale imaging or incorrect measuring approaches. Significant changes to the perception of the entire fault zone can be seen when a single fault becomes indistinguishable. For these reasons, we recognize that fault segmentation is systematically overlooked in the published literature when adopting a broad, one-fits-all, attitude to data sampling, against which the δ values suggested in this work should be used.

Journal Pre-proof

Acknowledgements

Schlumberger supported this study and the 3D Seismic Lab at Cardiff University. CGG is acknowledged for the provision of new, reprocessed seismic data from SE Brazil. Total Energies, Petrobras, Husky Energy (now Cenovus) and the Royal Society of London are acknowledged for their support to field work campaigns in SE Crete. We thank A. Nicol

(Univ. Canterbury, NZ) for the provision of an editable version of Fig. 2 and M. Glen for providing the map of Kilve and Watchet in Fig. 4d. The authors also thank editors I. Alsop and S. Mukherjee, together with two anonymous reviewers, for their constructive comments to an earlier draft of this paper.

Data statement

The data that support the findings of this study are available on request from the corresponding author, Tiago M. Alves.

References

Alves, T.M., 2024. Networks of geometrically coherent faults accommodate Alpine tectonic inversion offshore SW Iberia. Solid Earth, 15, 39–62. https://doi.org/10.5194/se-15-39-2024

Alves, T.M., and Cupkovic. T., 2018. Footwall Degradation Styles and Associated Sedimentary Facies Distribution in SE Crete: Insights into Tilt-Block Extensional Basins on Continental Margins. Sedimentary Geology, 367, 1–19. https://doi.org/10.1016/j.sedgeo.2018.02.001

Alves, T.M., Mattos, N.H., Newnes, S., and Goodall, S., 2022. Analysis of a Basement Fault Zone with Geothermal Potential in the Southern North Sea. Geothermics 102, 102398. https://doi.org/10.1016/j.geothermics.2022.102398

Aydin, A., Ozbek, A., and Cobanoglu, I., 2004. Tunnelling in Difficult Ground: A Case Study from Dranaz Tunnel, Sinop, Turkey. Engineering Geology, 74, 293–301. https://doi.org/10.1016/j.enggeo.2004.04.003

Journal Pre-proof

Baudon, C., and Cartwright, J., 2008. The kinematics of reactivation of normal faults using high resolution throw mapping. Journal of Structural Geology, 30, 1072-1084. https://doi.org/10.1016/j.jsg.2008.04.008

Caputo, R., Catalano, S., Monaco, C., Romagnoli, G., Tortorici, G., and Tortorici, L., 2010. Active Faulting on the Island of Crete (Greece). Geophysical Journal International, 183, 111–26. https://doi.org/10.1111/j.1365-246X.2010.04749.x

Cartwright, J., Bouroullec, R., James, D., and Johnson, H., 1998. Polycyclic motion history of some Gulf Coast growth faults from high-resolution displacement analysis. Geology, 26, 819-822. https://doi.org/10.1130/0091-7613(1998)026<0819:PMHOSG>2.3.CO;2

Cloetingh, S., Van Wees, J.D., Ziegler, P.A., Lenkey, L., Beekman, F., Tesauro, M., Förster, A., Norden, B., Kaban, M., and Hardebol, N., 2010. Lithosphere Tectonics and Thermo-Mechanical Properties: An Integrated Modelling Approach for Enhanced Geothermal Systems Exploration in Europe. Earth-Science Reviews, 102, 159–206. https://doi.org/10.1016/j.earscirev.2010.05.003

Dart, C.J., McClay, K., and Hollings, P.N., 1995. 3D analysis of inverted extensional fault systems, southern Bristol Channel basin, UK. In: Buchanan, J.G. and Buchanan, P.G. (Eds). *Basin Inversion*, Geological Society of London, Special Publications, 88, pp. 393-413. https://doi.org/10.1144/GSL.SP.1995.088.01.21

Fernández-Blanco, D., de Gelder, G., Lacassin, R., and Armijo, R., 2019. Geometry of flexural uplift by continental rifting in Corinth, Greece. Tectonics, 39, e2019TC005685. https://doi.org/10.1029/2019TC005685

Ferrill, D.A., Morris, A.P., McGinnis, R.N., Smart, K.J., Wigginton, S.S., and Hill, N.J., 2017. Mechanical Stratigraphy and Normal Faulting. Journal of Structural Geology, 94, 275–302. https://doi.org/10.1016/j.jsg.2016.11.010

Journal Pre-proof

Ferrill, D.A., Samrt, K.J., and Morris, A.P., 2020. Fault failure modes, deformation mechanisms, dilation tendency, slip tendency, and conduits v. Seals. *In:* Ogilvie, S.R., Dee, S.J., Wilson, R.W., Baileu, W.R. (Eds), Integrated Fault Seal Analysis. Geological Society, London, Special Publications, 496, pp. 75-98. https://doi.org/10.1144/SP496-2019-7

Gaki-Papanastassiou, K., Karymbalis, E., Papanastassiou, D., and Maroukian, H. 2009. Quaternary Marine Terraces as Indicators of Neotectonic Activity of the Ierapetra Normal Fault SE Crete (Greece). Geomorphology, 104, 38–46. https://doi.org/10.1016/j.geomorph.2008.05.037

Glen, R.A., Hancock, P.L., and Whittaker, A., 2005. Basin inversion by distributed deformation: the southern margin of the Bristol Channel Basin, England. Journal of Structural Geology, 27, 2113-2134. https://doi.org/10.1016/j.jsg.2005.08.006

Gloaguen, R., Marpu, P.R., and Niemeyer, I., 2007. Automatic extraction of faults and fractal analysis from remote sensing data. Nonlinear Processes in Geophysics, 14, 131–138, https://doi.org/10.5194/npg-14-131-2007

Goff, J.A., 1991. A Global and Regional Stochastic Analysis of Near-Ridge Abyssal Hill Morphology. Journal of Geophysical Research: Solid Earth, 96, 21713–21737. https://doi.org/10.1029/91JB02275

Grisel, O., Mueller, A., Lars, Gramfort, A., Louppe, G., Fan, T.J., Prettenhofer, P., et al., 2023. Scikit-Learn/Scikit-Learn: Scikit-learn 1.3.0. Zenodo. https://doi.org/10.5281/ZENODO.8098905

Gudmundsson, A., 2012. *Rock Fractures in Geological Processes*. Cambridge University Press, Cambridge, 578 pp. https://doi.org/10.1017/CBO9780511975684

Journal Pre-proof

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., et al., 2020. Array Programming with NumPy. Nature, 585, 357–62. https://doi.org/10.1038/s41586-020-2649-2

He, D., Lu, R., Huang, H., Wang, X., Jiang, H., and Zhang, W., 2019. Tectonic and Geological Setting of the Earthquake Hazards in the Changning Shale Gas Development Zone, Sichuan Basin, SW China. Petroleum Exploration and Development, 46, 1051–64. https://doi.org/10.1016/S1876-3804(19)60262-4

Heiberger, R.M., and Neuwirth, E., 2009. Chapter 11 – Polynomial Regression. In: Heiberger, R.M. and Neuwirth, E. (Eds), R Thorugh Excel – A spreadsheet interface for statistics, data analysis, and graphics. Springer Verlag New York, pp. 269-284. https://doi.org/10.1007/978-1-4419-0052-4

Huenges, E., Kohl, T., Kolditz, O., Bremer, J., Scheck-Wenderoth, M., and Vienken, T., 2013. Geothermal Energy Systems: Research Perspective for Domestic Energy Provision. Environmental Earth Sciences, 70, 3927–33. https://doi.org/10.1007/s12665-013-2881-2

James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. Chapter 7 - Moving Beyond Linearity. In: James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), An Introduction to Statistical Learning: With Applications in R, Springer Texts in Statistics, 103, New York: Springer, pp. 265–301.

Jentsch, A., Jolie, E., Jones, D.G., Taylor-Curran, H., Peiffer, L., Zimmer, M., and Lister, B., 2020. Magmatic Volatiles to Assess Permeable Volcano-Tectonic Structures in the Los Humeros Geothermal Field, Mexico. Journal of Volcanology and Geothermal Research, 394, 106820, https://doi.org/10.1016/j.jvolgeores.2020.106820

Kalbermatten, M., van de Ville, D., Turberg, P., Tuia, D. and Joost. S. 2012. Multiscale analysis of geomorphological and geological features in high resolution digital

Journal Pre-proof

elevation models using the wavelet transform. Geomorphology, 138, 352-363. https://doi.org/10.1016/j.geomorph.2011.09.023

Kim, Y.-S, and Sanderson, D.J., 2005. The Relationship Between Displacement and Length of Faults: A Review. Earth-Science Reviews, 68, 317–334. https://doi.org/10.1016/j.earscirev.2004.06.003

King, J.J., and Cartwright, J.A., 2020. Ultra-slow throw rates of polygonal fault systems. Geology, 48, 473-477. https://doi.org/10.1130/G47221.1

Kozłowska, M., Brundzinski, M.R., Friberg, P., and Currie, B.S., 2017. Maturity of nearby faults influences seismic hazard from hydraulic fracturing. Proceedings of the National Academy of Sciences, 115, E1720-E1729. https://doi.org/10.1073/pnas.171528411

Laubach, S.E., Lamarche, J., Gauthier, B.D.M., Dunne, W.M., Sanderson, D.J., 2018. Spatial arrangement of faults and opening-mode fractures. Journal of Structural Geology, 108, 2-15. https://doi.org/10.1016/j.jsg.2017.08.008

Lee, G., Gommers, R., Wohlfahrt, K., Wasilewski, F., O'Leary, A., Nahrstaedt, H., Sauvé, A. et al. 2022. PyWavelets/Pywt: V1.4.1. Zenodo. https://doi.org/10.5281/ZENODO.1407171

Lee, J., Stockli, D.F., and Blythe, A.E., 2023. Cenozoic slip along the southern Sierra Nevada normal fault, California (USA): A long-lived stable western boundary of the Basin and Range. Geosphere, 19, 878-899. https://doi.org/10.1130/GES02574.1

Mallat, S. G. 2009. *A Wavelet Tour of Signal Processing*. 2nd ed. San Diego: Academic Press, 895 pp. https://doi.org/10.1016/B978-0-12-374370-1.X0001-8

Mechernich, S., Reicherter, K., Deligiannakis, G., and Papanikolaou, I., 2023. Tectonic geomorphology of active faults in Eastern Crete (Greece) with slip rates and earthquake history from cosmogenic ^{36}Cl dating of the Lastros and Orno faults. Quaternary International, 652, 77-91. https://doi.org/10.1016/j.quaint.2022.04.007

Journal Pre-proof

Misra AA, and Mukherjee S. 2018. Atlas of Structural Geological Interpretation from Seismic Images. Wiley Blackwell. ISBN: 978-1-119-15832-5.

Moska, R., Labus, K., and Kasza, P., 2021. Hydraulic Fracturing in Enhanced Geothermal Systems—Field, Tectonic and Rock Mechanics Conditions—A Review. Energies, 14, 5725. https://doi.org/10.3390/en14185725

Mukherjee S., 2019. Particle tracking in ideal faulted blocks using 3D co-ordinate geometry. Marine and Petroleum Geology 107, 508-514. https://doi.org/10.1016/j.marpetgeo.2019.05.037

Nicol, A., Walsh, J., Childs, C, and Manzocchi, T., 2020. Chapter 6 - The growth of faults. *In:* Understanding Faults – Detecting, Dating and Modeling (Eds: Tanner, D, Brandes), Elsevier, pp. 221-255. https://doi.org/10.1016/B978-0-12-815985-9.00006-0

Nixon, C.W., McNeill, L.C., Gawthorpe, R.L., Shillington, D.J., Michas, G., Bell, R.E., Moyle A., Ford, M., Zakharova, N.V., Bull, J.M., and de Gelder, G., 2024. Increasing fault slip rates within the Corinth Rift, Greece: A rapidly localising active rift fault network. Earth and Planetary Science Letters, 636, 118716. https://doi.org/10.1016/j.epsl.2024.118716

Ostertagová, E., 2012. Modelling Using Polynomial Regression. Procedia Engineering, 48, 500–506. https://doi.org/10.1016/j.proeng.2012.09.545

Peacock, D.C.P., Nixon, C.W., Rotevatn, A., Sanderson, D.J., and Zuluaga, L.F., 2017. Interacting faults. Journal of Structural Geology, 97, 1-22. https://doi.org/10.1016/j.jsg.2017.02.008

Plawiak, R.A.B., Carvalho, M.J., Sombra, C.L., Brandão, D.R., Mepen, M., Ferrari, A.L., and Gambôa, L.A.P., 2024. Structural controls of the migration of mantle-derived CO_2 offshore in the Santos Basin (Southeastern Brazil). Frontier in Earth Sciences, 11, https://doi.org/10.3389/feart.2023.1284151

Journal Pre-proof

Pollard, D.D. and Segall, P., 1987. Theoretical displacements and stresses near fractures in rock: with applications to faults, joints, veins, dikes, and solution surfaces. In: Atkinson, B.K. (Ed.*), Fracture Mechanics of Rock*, Academic Press, London, pp. 277-349.

Proakis, J.G., 1992. Multirate Digital Signal. In: Proakis, J.G. (Ed.), Advanced Digital Signal Processing, Macmillan, 141–204. https://books.google.co.uk/books?id=4_RSAAAACAAJ

Purba, D.P, Adityatama, D.W., Umam, M.F., and Muhammad, F., 2019. Key Considerations in Developing Strategy for Geothermal Exploration Drilling Project in Indonesia. Proceedings, 44th Geothermal Reservoir Engineering, Stanford University, Stanford, California, SGP-TR-214.

Raschka, S., and Mirjalili, V., 2018. Predicting Continuous Target Variables with Regression Analysis." In: Raschka, S., Mirjalili, V. (Eds.), Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow, Packt Publishing, Second edition, fourth release, pp. 309–346.

Rawnsley, K.D., Peacock, D.C.P., Rives, T., and Petit, J.P., 1998. Joints in the Mesozoic sediments around the Bristol Channel Basin. Journal of Structural Geology, 20, 1641-1661. https://doi.org/10.1016/S0191-8141(98)00070-4

Robertson, J., Roberts, G.P., Iezzi, F., Meschis, M., Gheorghiu, D.M., Sahy, D., Bristow, C., and Sgambato, C., 2020. Distributed normal faulting in the tip zone of the South Alkyonides Fault System, Gulf of Corinth, constrained using ^{36}Cl exposure dating of late-Quaternary wave-cut platforms. Journal of Structural Geology, 136, 104063. https://doi.org/10.1016/j.jsg.2020.104063

Journal Pre-proof

Rotevatn, A., Jackson, C.A.-L., Tvedt, A.B.M., Bell, R.E., and Blækkan, I., 2019. How do normal faults grow? Journal of Structural Geology, 125, 174-184. https://doi.org/10.1016/j.jsg.2018.08.005

Saeidi, O., Rasouli, V., Vaneghi, R.G., Gholami, R., and Torabi, S.R., 2014. A modified failure criterion for transversely isotropic rocks. Geoscience Frontiers, 5, 215-225. https://doi.org/10.1016/j.gsf.2013.05.005

Shen, S., Li, H., Chen, W., Wang X., and Huang, B., 2022. Seismic Fault Interpretation Using 3-D Scattering Wavelet Transform CNN, IEEE Geoscience and Remote Sensing Letters, 19, 1-5, Art no. 8028505. doi: 10.1109/LGRS.2022.3183495

Scholz, C.H., and Aviles, C.A., 2013. The Fractal Geometry of Faults and Faulting. In: Das, S., Boatwright, J., Scholz, C.H. (Eds), Geophysical Monograph Series, American Geophysical Union, Washington, D.C., 147–155. https://doi.org/10.1029/GM037p0147

Sifuzzaman, M., Islam, M.R., and Ali, M.Z., 2009. Application of Wavelet Transform and its advantages compared to Fourier Transform. Journal of Physical Sciences, 13, 121-134. ISSN: 0972-8791

Tao, Z., and Alves, T.M., 2019. Impacts of Data Sampling on the Interpretation of Normal Fault Propagation and Segment Linkage. Tectonophysics, 762, 79–96. https://doi.org/10.1016/j.tecto.2019.03.013

Torabi, A., Alaei, B., and Libak. A., 2019. Normal Fault 3D Geometry and Displacement Revisited: Insights from Faults in the Norwegian Barents Sea. Marine and Petroleum Geology, 99, 135–55. https://doi.org/10.1016/j.marpetgeo.2018.09.032

Torabi, A., and Berg, S.S., 2011. Scaling of fault attributes: A review. Marine and Petroleum Geology, 28, 1444-1460. https://doi.org/10.1016/j.marpetgeo.2011.04.003

Journal Pre-proof

Torabi, A., Rudnicki, J., Alaei, B., and Buscarnera, G., 2023. Envisioning faults beyond the framework of fracture mechanics. Earth-Science Reviews, 238, 104358. https://doi.org/10.1016/j.earscirev.2023.104358

Trippetta, F., Petricca, P., Billi, A., Collettini, C., Cuffaro, M., Lombardi, A. M., Scrocca, D., Ventura, G., Morgante, A., and Doglioni, C., 2019. From mapped faults to fault-length earthquake magnitude (FLEM): a test on Italy with methodological implications. Solid Earth, 10, 1555–1579. https://doi.org/10.5194/se-10-1555-2019

Tvedt, A.B.M., Rotevatn, A., Jackson, C.A.-L., Fossen, H., and Gawthorpe, R.L., 2013. Growth of normal faults in multilayer sequences: A 3D seismic case study from the Egersund Basin, Norwegian North Sea. Journal of Structural Geology, 55, 1-20. https://doi.org/10.1016/j.jsg.2013.08.002

Underhill, J.R. and Paterson, S. 1998. Genesis of tectonic inversion structures: seismic evidence for the development of key structures along the Purbeck–Isle of Wight Disturbance. Journal of Geological Society, London, 155, 975-992. https://doi.org/10.1144/gsjgs.155.6.0975

Varela. C.L., and Mohriak, W.U., 2013. Halokinetic rotating faults, salt intrusions, and seismic pitfalls in the petroleum exploration of divergent margins. American Association of Petroleum Geologists Bulletin, 97, 1421-1446. https://doi.org/10.1306/02261312164

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., et al., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, 261–72. https://doi.org/10.1038/s41592-019-0686-2

Walsh, J., Nicol, A., Childs, C., 2002. An alternative model for the growth of faults, Journal of Structural Geology, 24, 1669-1675. https://doi.org/10.1016/S0191-8141(01)00165-1

Journal Pre-proof

Journal Pre-proof

Walsh, J.J., Bailey, W.R., Childs, C., Nicol, A., and Bonson, C.G., 2003. Formation of Segmented Normal Faults: A 3-D Perspective. Journal of Structural Geology, 25, 1251–1262. https://doi.org/10.1016/S0191-8141(02)00161-X

Walsh, J.J., and Watterson, J., 1991. Geometric and Kinematic Coherence and Scale Effects in Normal Fault Systems. Geological Society, London, Special Publications, 56, 193–203. https://doi.org/10.1144/GSL.SP.1991.056.01.13

Wang, Y., 2015a. Frequencies of the Ricker wavelet. Geophysics, 80, 1MA-Z50. https://doi.org/10.1190/geo2014-0441.1

Wang, Y., 2015b. Generalized seismic wavelets. Geophysical Journal International, 203, 1172-1178. https://doi.org/10.1093/gji/ggv346

Wang, Y., Li, J., Wang, Z.-F., and Chang, H., 2022. Structural Failures and Geohazards Caused by Mountain Tunnel Construction in Fault Zone and Its Treatment Measures: A Case Study in Shaanxi. Engineering Failure Analysis, 138, 106386. https://doi.org/10.1016/j.engfailanal.2022.106386

Yielding, G., 2015. Trapping of buoyant fluids in fault-bound structures. *In:* Industrial Structural Geology: Principles, Techniques and Integration (Eds: Richards, F.L., Richardson, N.J., Rippington, S.J., Wilson, R.W., Bond, C.E.). Geological Society, London, Special Publications, 421, 29-39. https://doi.org/10.1144/SP421

Ze, T., and Alves, T.M., 2016. The Role of Gravitational Collapse in Controlling the Evolution of Crestal Fault Systems (Espírito Santo Basin, SE Brazil). Journal of Structural Geology, 92, 79–98. https://doi.org/10.1016/j.jsg.2016.09.011

Zhang, Q., Alves, T.M., and Martins-Ferreira, M.A.C., 2022. Fault Analysis of a Salt Minibasin Offshore Espírito Santo, SE Brazil: Implications for Fluid Flow, Carbon and Energy Storage in Regions Dominated by Salt Tectonics. Marine and Petroleum Geology, 143, 105805. https://doi.org/10.1016/j.marpetgeo.2022.105805

Journal Pre-proof

Figure and table captions

Figure 1 – Schematic representation of how tectonic faults interact and link in nature. Faults evolve from isolated to interacting faults by linking vertically and laterally. The ratio of d_{max}/L (maximum displacement vs. length) increases as lateral propagation occurs in a fault. Stage 1 corresponds to the formation of isolated, non-interacting fault segments. Stage 2 relates to the start of fault interaction, overlap and joint growth. Stage 3 represents a fully linked pair of faults that grow together from that moment onwards. Figure is modified from Kim and Sanderson (2005).

Figure 2 – Schematic representation of normal-fault evolution. Isolated propagating faults (left) consist of isolated segments that coalesce to form long, interlinked fault strands. The coherent constant-length growth model (right) assumes that lateral fault propagation is rapid but vertical propagation is limited. Figure modified from Nicol et al. (2020).

Figure 3 – Diagram summarizing the way fault-throw data are measured at outcrop, or using stratigraphic markers in seismic data. The diagram is modified from Ze and Alves (2019) and based on the Ierapetra Fault Zone, SE Crete, one of the faults analyzed in this work. Throw measurements are usually taken relative to a correlative surface that is present on the footwall and hanging-wall blocks of faults. However, this can be made difficult by fault scarp erosion, and by the covering of the immediate hanging-wall depocentre to the fault by strata. Heave corresponds to the lateral displacement accommodated by a fault during its movement. Fault displacement is the resultant vector of throw and heave.

Journal Pre-proof

Figure 4 – a) World map indicating the location of the regions where T/D and T/Z data were acquired for this study. b) Location of the seismic surveys interpreted in SE Brazil from which fault-throw data were acquired. c) Location of the Ierapetra Fault relative to other fault families, local sedimentary basins and regional basement terrains. d) Map of SW England's coast highlighting the locations where fault-throw data were acquired at a sub-seismic scale (see black squares on the map). Figure 4b is modified from Alves and Cupkovic (2018). Figure 4d is modified from Glen et al. (2005).

Figure 5 – Examples of faults analyzed in this work, from where throw measurements were acquired. a) Some of the salt-related faults at the scale of industry seismic data acquired from a high-resolution seismic survey shot in SE Brazil. b) Panoramic view of the central part of the Ierapetra Fault Zone and its constituting fault segments. In the parentheses are shown the height of footwall blocks associated with what is a > 25 km long normal fault zone. c) and d) Faults in the SW England (Bristol Channel) at the sub-seismic scale.

Figure 6 – Normalized Ricker wavelet, a symmetrical wavelet used to represent signal changes in the time domain Wang, 2015a, 2015b). In this work, the time domain was replaced with by a spatial component (length or height) in order to apply the *Ricker* wavelet theory to the identification of fault segments.

Figure 7 – Graphical example of the Continuous Wavelet Transform technique used to identify discrete fault segments (Step 1 in this work, Section 4.1) at the lower polynomial degree 3. Note the obvious correlation between frequency band strength and the throw maxima recorded for each fault segment. Note that fault segmentation using this technique

Journal Pre-proof

Journal Pre-proof

results in the smallest segments being ignored by the algorithm. This figure thus stress the fact that a Continuous Wavelet Transform cannot identify throw maxima in the smaller fault segments – it is focused on picking the greatest throw maxima in a given T/D and T/Z dataset.

Figure 8 – Workflow suggested in this paper for the identification of fault segments using a Machine Learning approach.

Figure 9 – Example of the improved fault recognition resulting from applying gradient measurements from the point of threshold minima (Step 2 in this work, Section 4.2). Step 2 focused on finding the nearest throw minimum representing the linkage point between two fault segments. Threshold values can be changed in the algorithm, with a stricter threshold resulting in the identification of only the larger fault segments, and a looser threshold resulting in multiple fault segments being found. If no frequency minimum is found before reaching the end of the dataset, the last value picked by the algorithm is taken as the end of the segment. In Step 2, some of the smallest fault segments were still overlooked by the algorithm but not on such a scale as revealed in Step 1 (see **Figs. 7 and 8**).

Figure 10 – Graph used to estimate noise floor in the data used in this work. The rapid descent recorded with increasing wavelength sizes represents the reduction in noise occurring as a result, as small changes in throw are filtered out by the algorithm. Once this noise is filtered out, and the curve approaches a flat, we can be confident that the remaining data is accurate. Peak rate values, when plotted against the frequency of data, show that adopting a threshold peak rate of 0.04 is a valid approach.

Journal Pre-proof

Figure 11 – Example of the improved fault recognition after applying a peak rate threshold to a Continuous Wavelet Transform (Step 3 in this work, Section 4.3). a) Fault R2_H3 interpreted in high-resolution seismic data from SE Brazil. b) Fault L2 H4-1 from onshore SE Brazil. Note the improved results in Step 3 when compared with Step 2, but with some smaller peaks being still overlooked in parts of the fault segments analyzed. The adoption of a 0.04 peak rate (see **Fig. 10**) returned positive results in Step 3 - all Peaks that are clearly not part of discrete segments were ignored, without overlooking any possible faults.

Figure 12 - Examples of regression curves modelling fault shape in T/D and T/Z data using a cubic model (Approach 4). Overall, this was the method that returned a better correlation between the fault segments identified in our dataset and the segments identified by the algorithm used. a) Fault R2 H2 analyzed from high-resolution 3D seismic data from SE Brazil. b) Segmented fault zone R2 H3 interpreted in SE Brazil using high-resolution seismic data. c) Fault L2 H4-1 from offshore SE Brazil.

Journal Pre-proof

Figure 13 – a) Visualization of T/D plots before and after a critical sampling ratio is applied. b) Example of the changes in fault shape when sampling ratio is reduced to an Integral Error of 11.6%.

Figure 14 – Change in error rate observed while the number of samples is reduced. a) In keystone fault 6-11, Modulus Error increases at a constant rate, whereas integral and intersect errors vary erratically due to loss of fault intersection points. b) Fault C24 records a rapid oscillation of error values is recorded. In most cases the error gradually increases when

decreasing the sampling ratio, but there are some examples of minima in Integral and Intersect errors occurring due to a sample coinciding exactly with a fault segment linkage point (see Section 5.1 in this article).

Figure 15 – Total distribution of minimum sample ratios (δ) for all datasets in this work. Results are shown separately for three different downsampling approaches: Strict, Moderate, and Lenient (see Section 6.2 in this article).

Figure 16 – Error distribution after a critical sampling ratio is applied to the throw data in this work.

Figure 17 – Graph showing the minimum sampling ratio calculated for Strict, Moderate and Lenient approaches to T/D and T/Z sampling. The sampling ratio (δ) values corresponding to a 95% success rate in fault-segment recognition are highlighted, with each data point represented by a vertical line.

Table 1 - Key statistics concerning the box plot in Fig. 16.

Table 2 – Minimum sampling ratios (δ) calculated based on a 95% success rate in fault-segment recognition for each downsampling approach: Strict, Moderate, and Lenient. See **Fig. 17** for a graphical representation of these values.

Journal Pre-proof

Table 1 - Key statistics concerning the box plot in Fig. 16.

Error Type	Min.	Q1	Median	Mean	Q3	Upper	Max.
Integral	2.48%	55.5%	89.5%	71.8%	96.8%	99.97%	99.97%
Modulus	0.401%	4.50%	6.72%	11111%	10.1%	17.7%	49.1%
Intersection	0%	21.0%	40.4%	36.6%	59.7%	86.0%	86.0%
Reduction (%)	10%	19.5%	27.3%	31.7%	52.9%	90.4%	90.4%

Journal Pre-proof

Table 2 – Minimum sampling ratios (δ) calculated based on a 95% success rate in fault-segment recognition for each downsampling approach: Strict, Moderate, and Lenient. See Fig. 17 for a graphical representation of these values.

Method	Critical Value	Uncertainty
Lenient	5.882%	±0.37%
Moderate	4.167%	±0.18%
Strict	1.020%	±0.02%

Journal Pre-proof

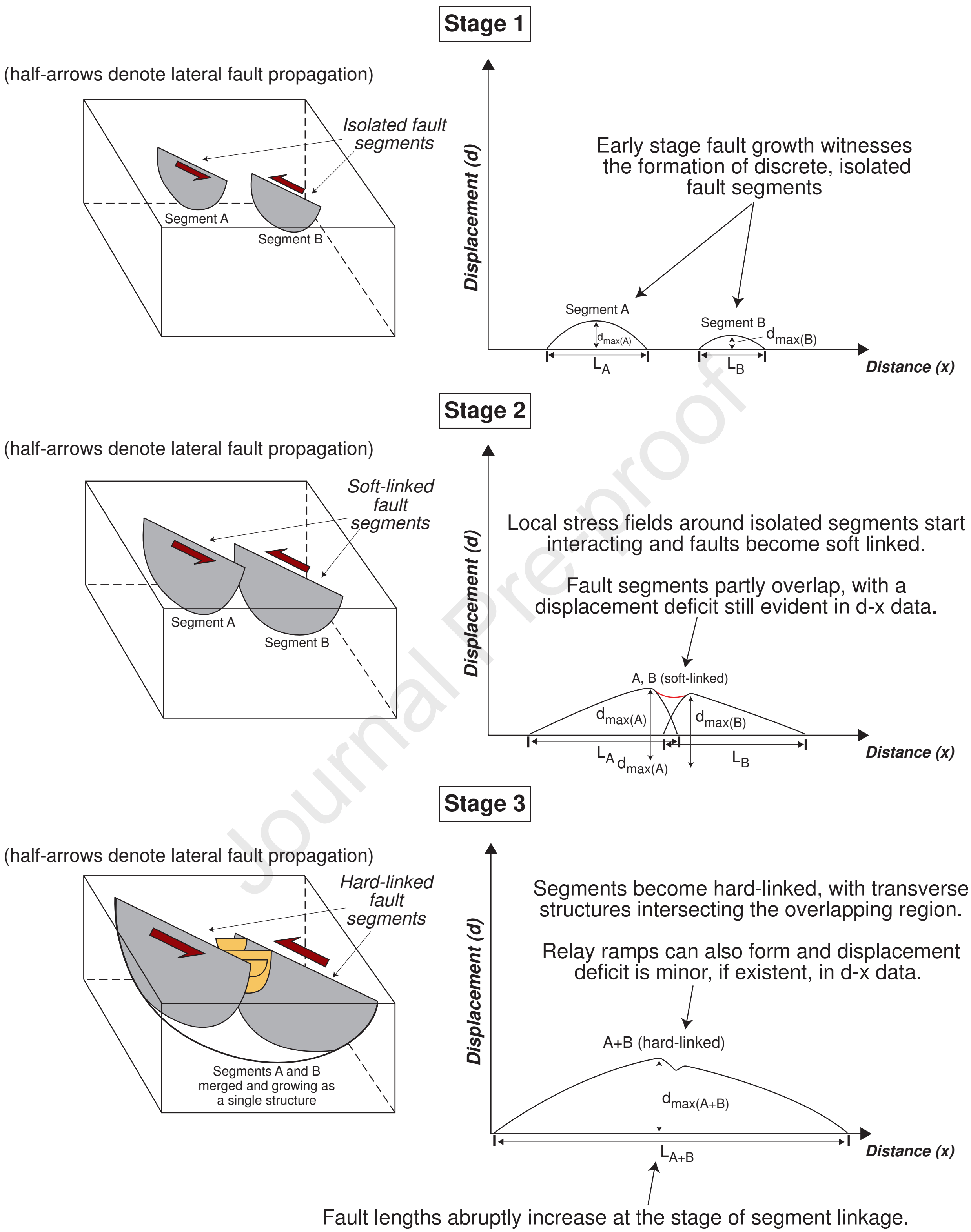


Figure 1

Journal Pre-proof

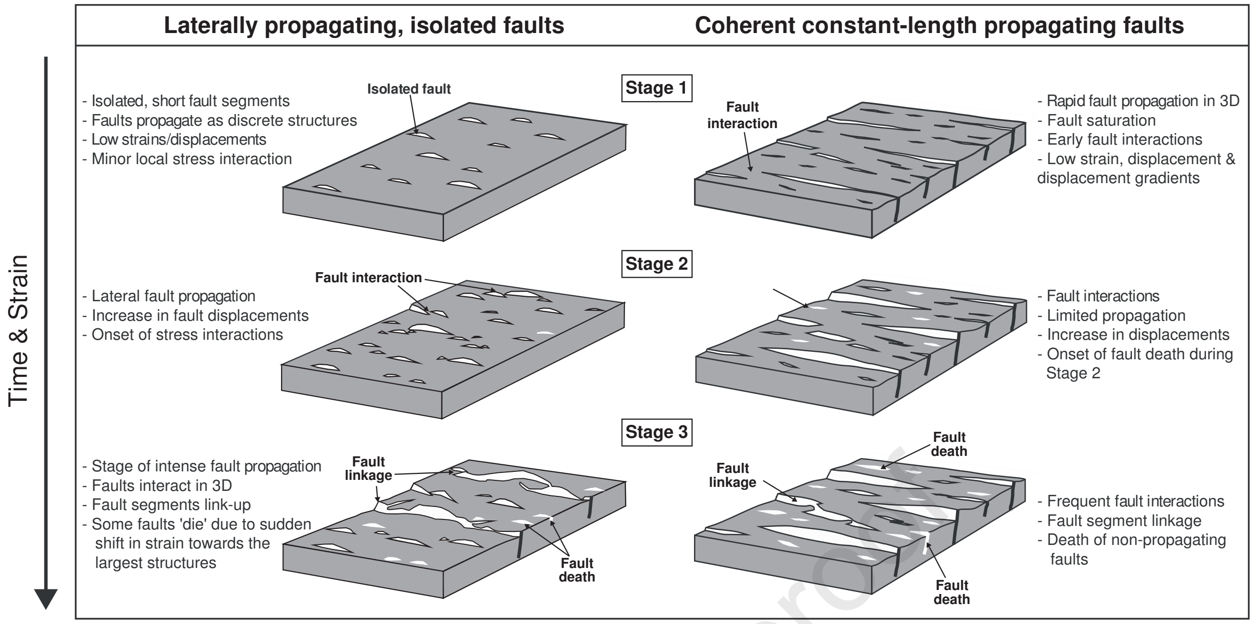


Figure 2

Journal Pre-proof

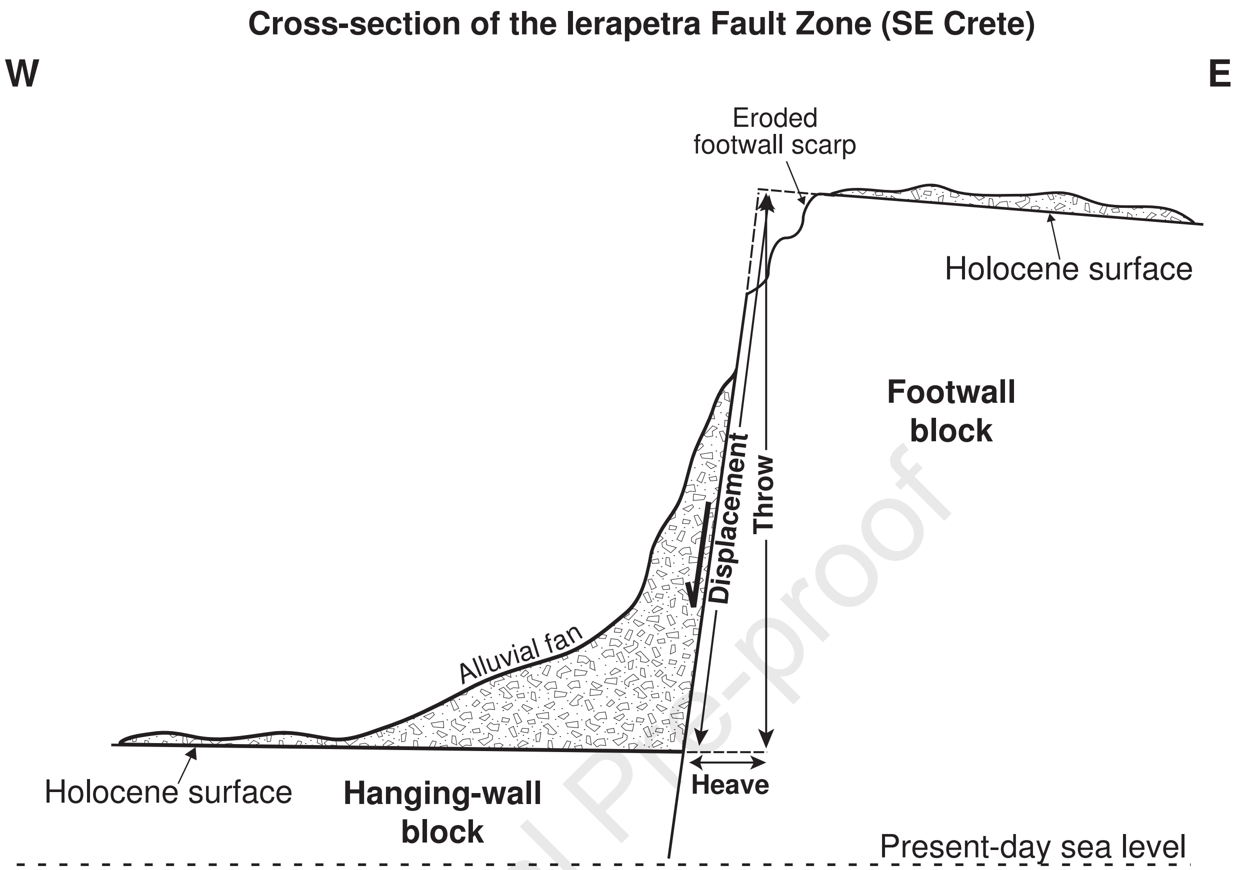


Figure 3

Journal Pre-proof

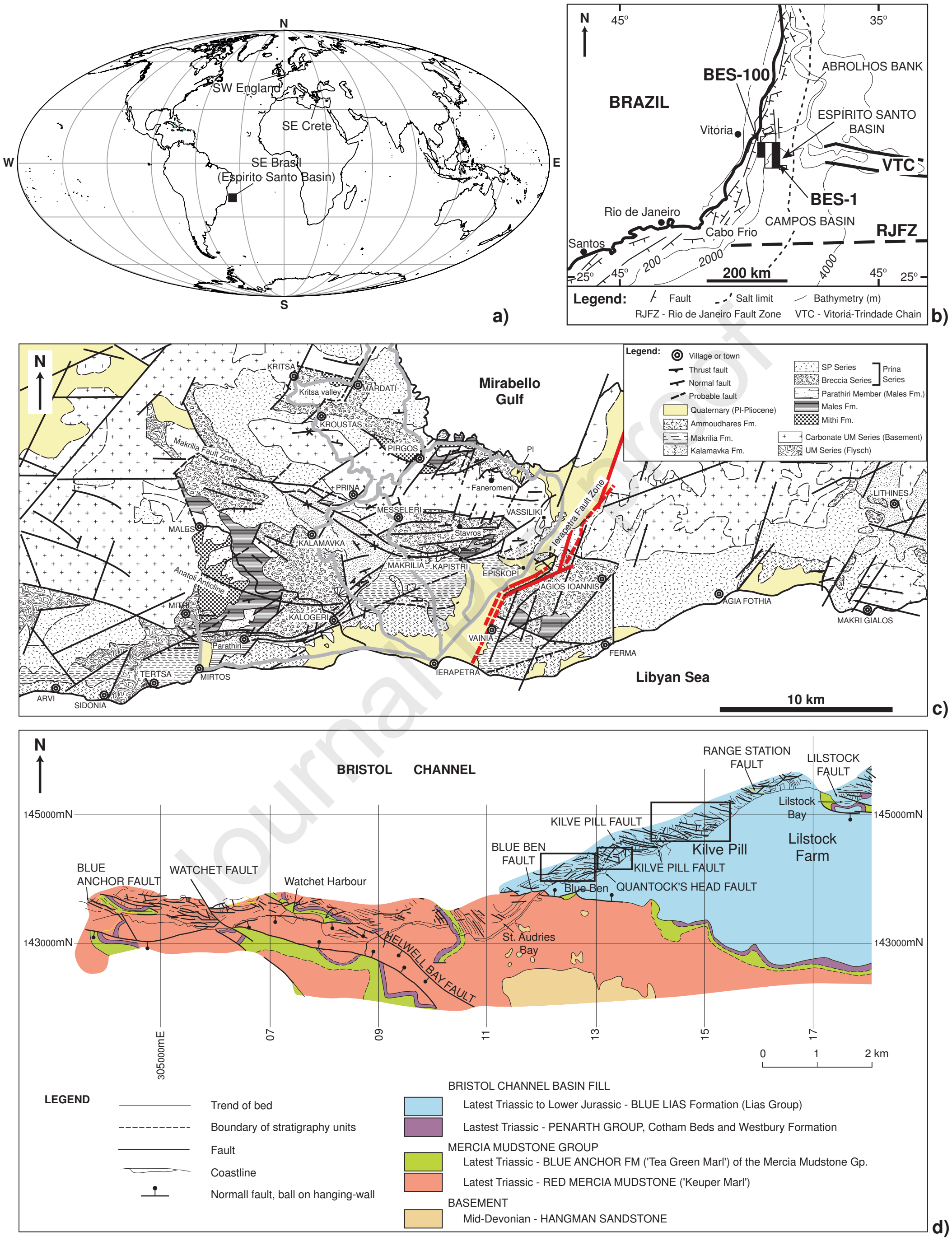
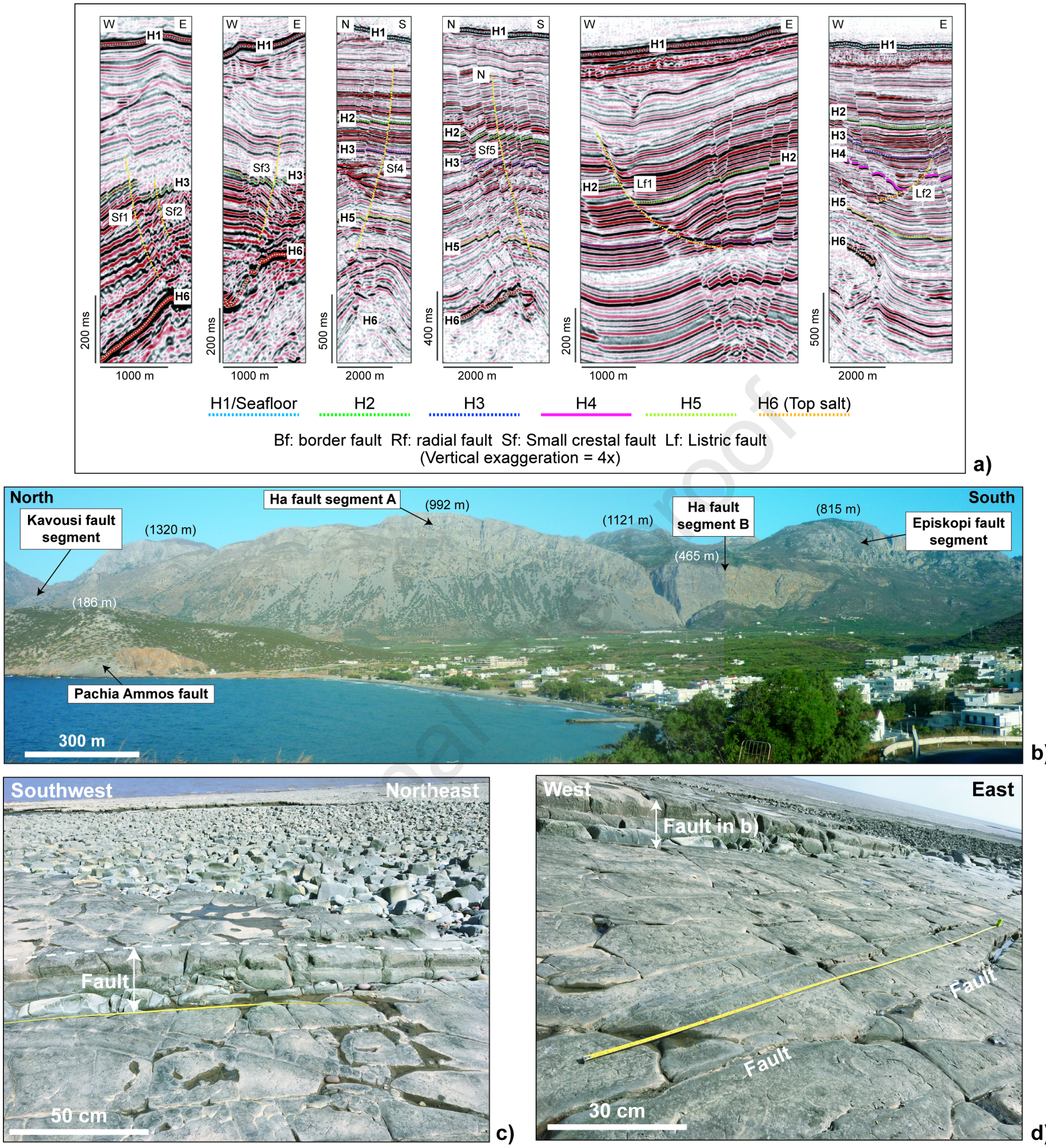


Figure 4

Journal Pre-proof

Journal Pre-proof



W E H1 H3 Sf1 Sf2 H6
W E H1 Sf3 H3 H6
N S H1 H2 H3 Sf4 H5 H6
N S H1 N H2 Sf5 H3 H5 H6
W E H1 H2 Lf1 H2
W E H1 H2 H3 H4 Lf2 H5 H6
200 ms 200 ms 500 ms 400 ms 200 ms 500 ms
1000 m 1000 m 2000 m 2000 m 1000 m 2000 m
H1/Seafloor H2 H3 H4 H5 H6 (Top salt)
Bf: border fault Rf: radial fault Sf: Small crestal fault Lf: Listric fault
(Vertical exaggeration = 4x)
a)
North
South
Kavousi fault segment
Ha fault segment A
(992 m)
(1320 m)
(1121 m)
Ha fault segment B
(815 m)
Episkopi fault segment
(465 m)
(186 m)
Pachia Ammos fault
300 m
b)
Southwest
Northeast
Fault
50 cm
c)
West
East
Fault in b)
Fault
Fault
30 cm
d)

Figure 5

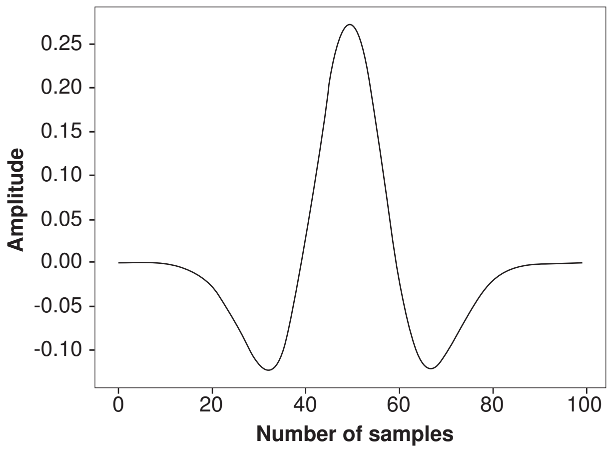


Figure 6

Journal Pre-proof

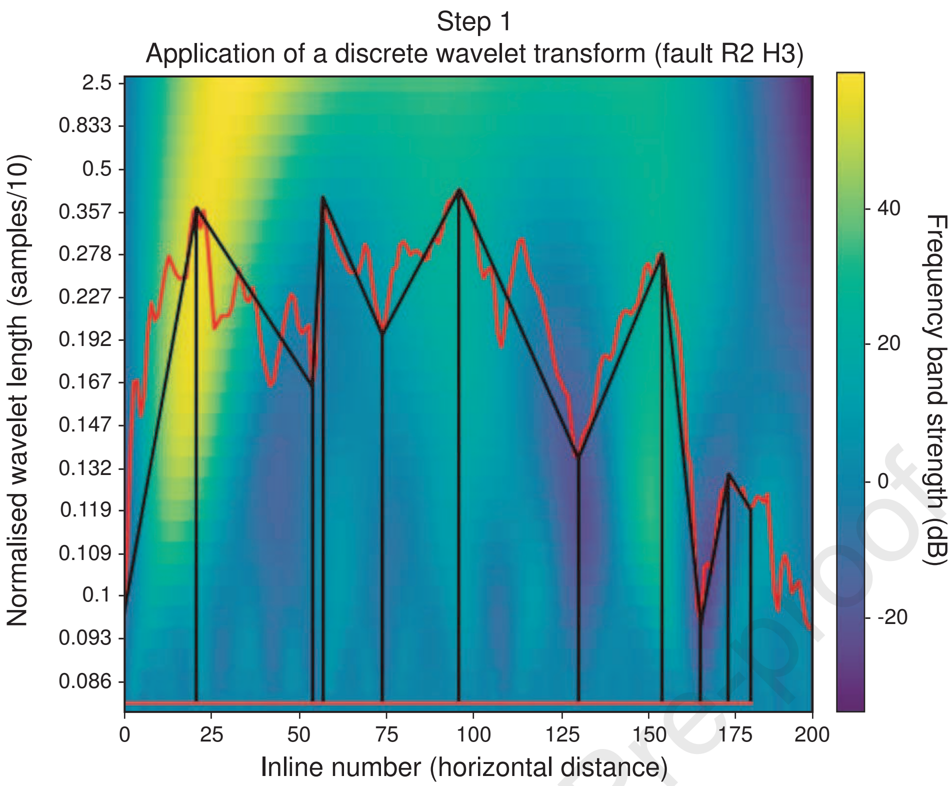


Figure 7

Journal Pre-proof

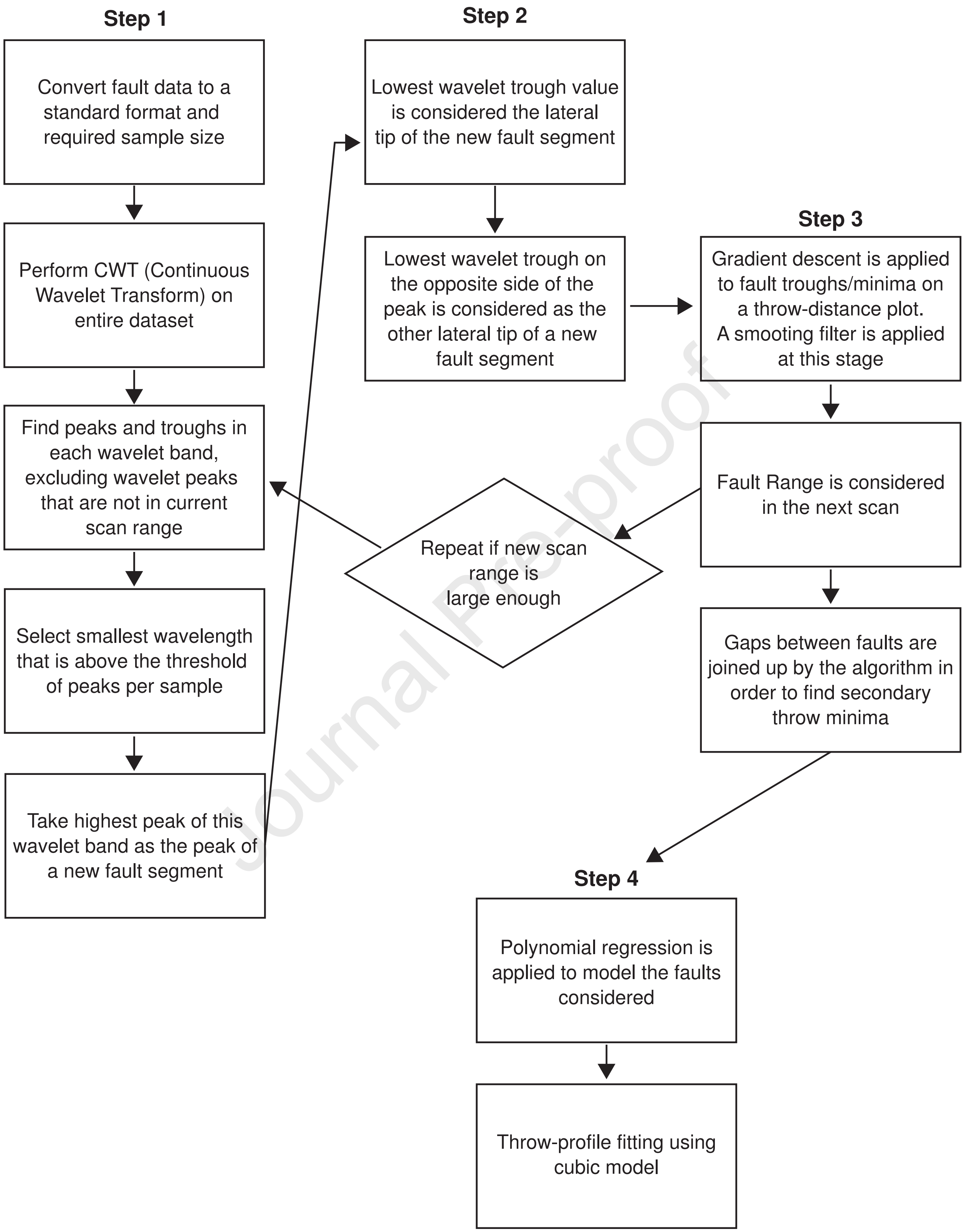


Figure 8

Journal Pre-proof

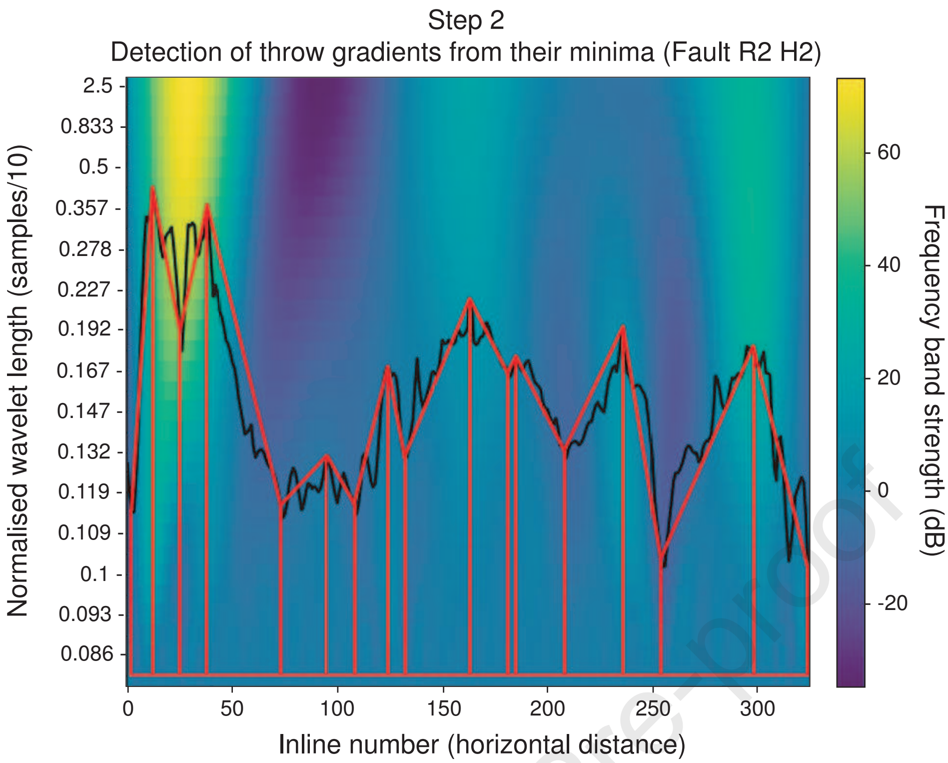


Figure 9

Journal Pre-proof

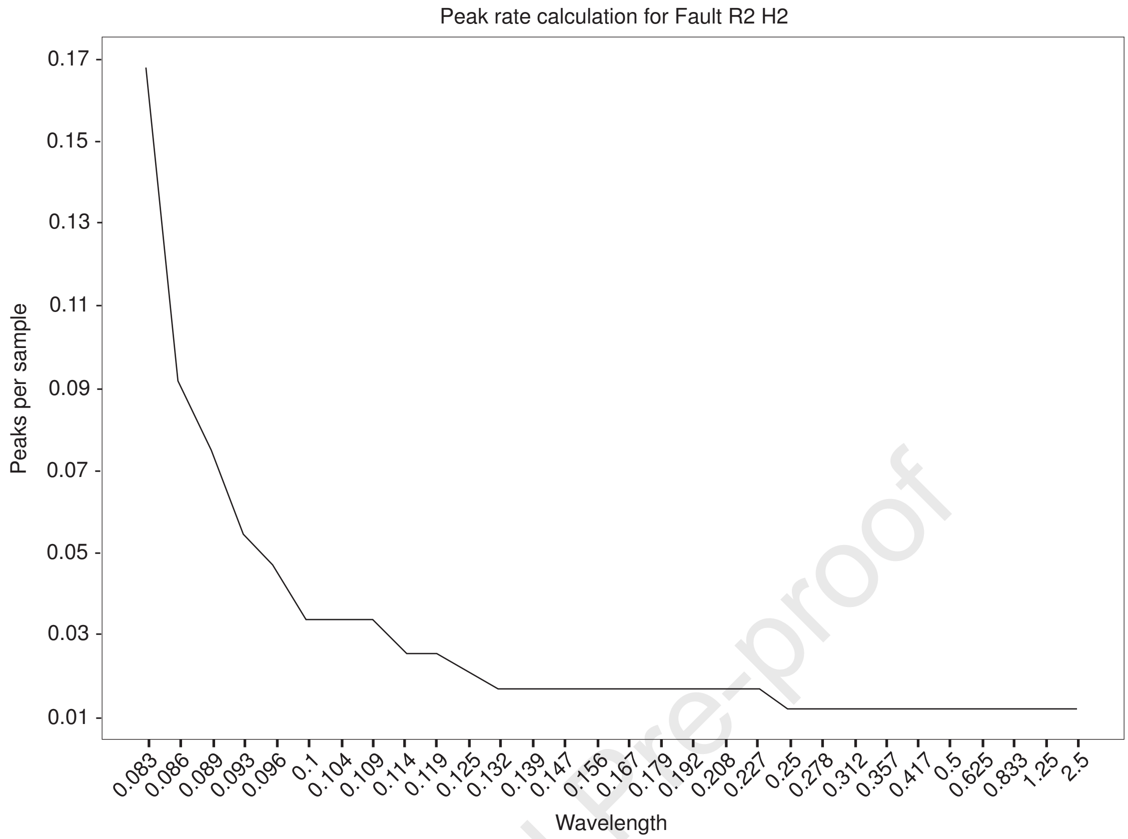


Figure 10

Journal Pre-proof

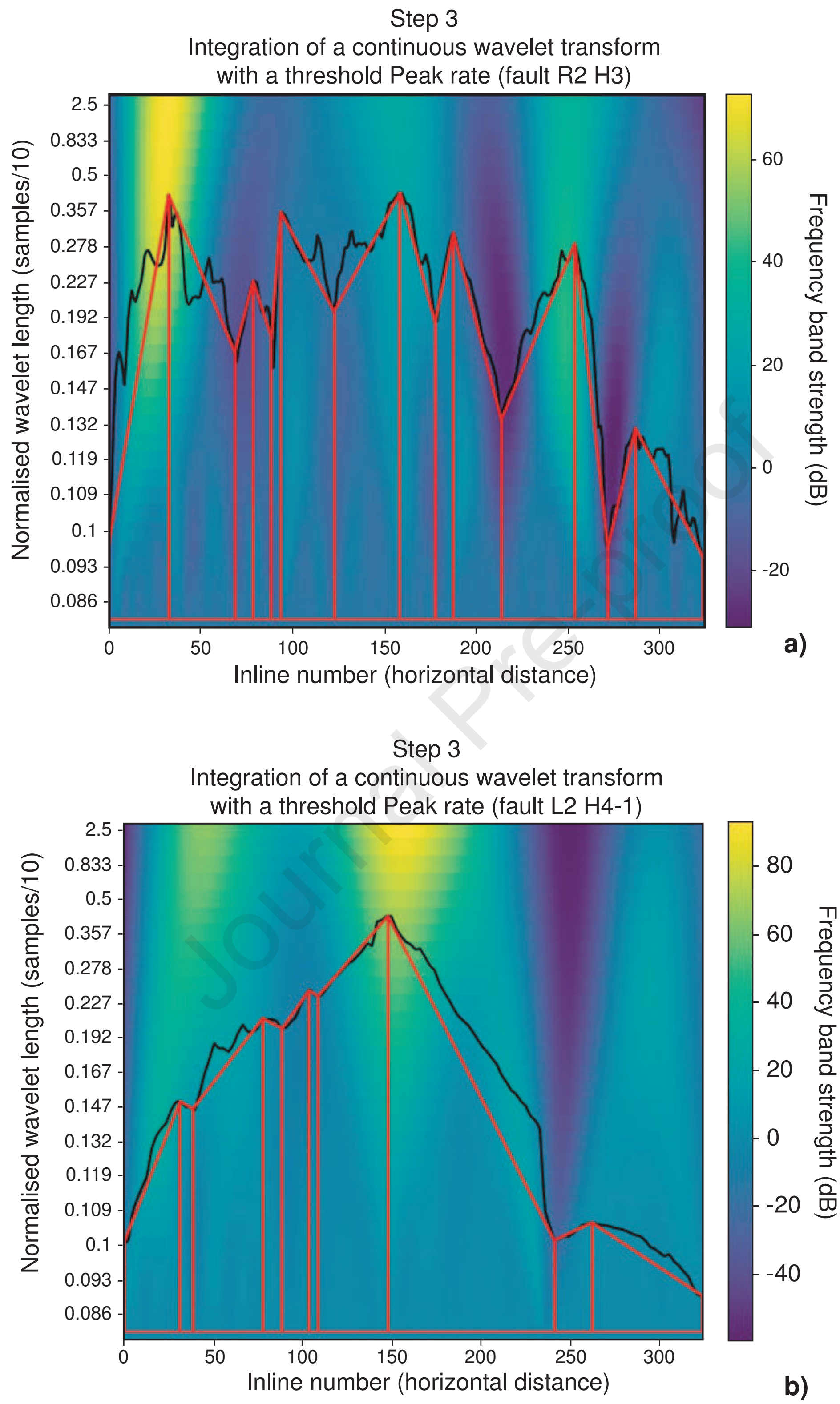


Figure 11

Journal Pre-proof

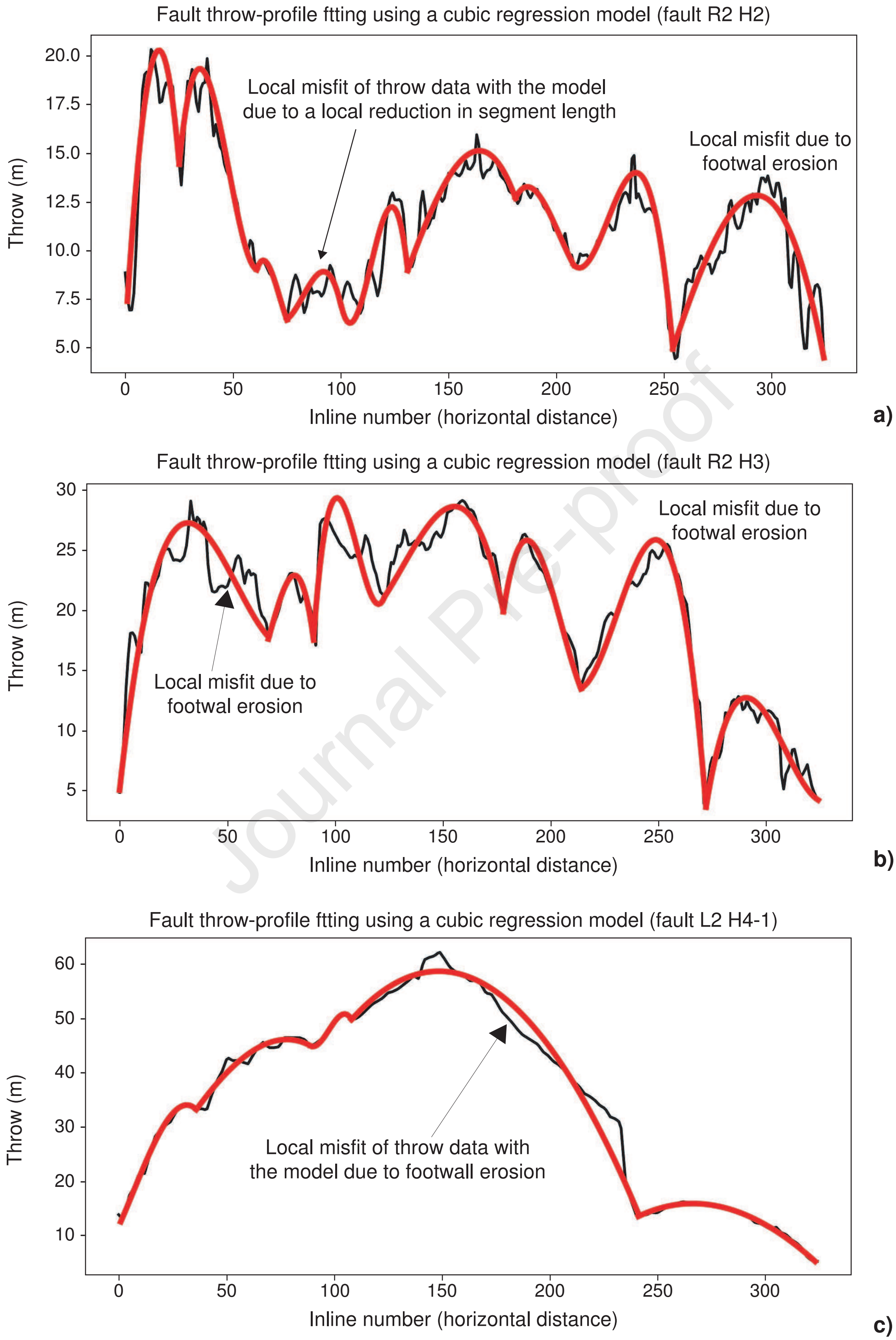
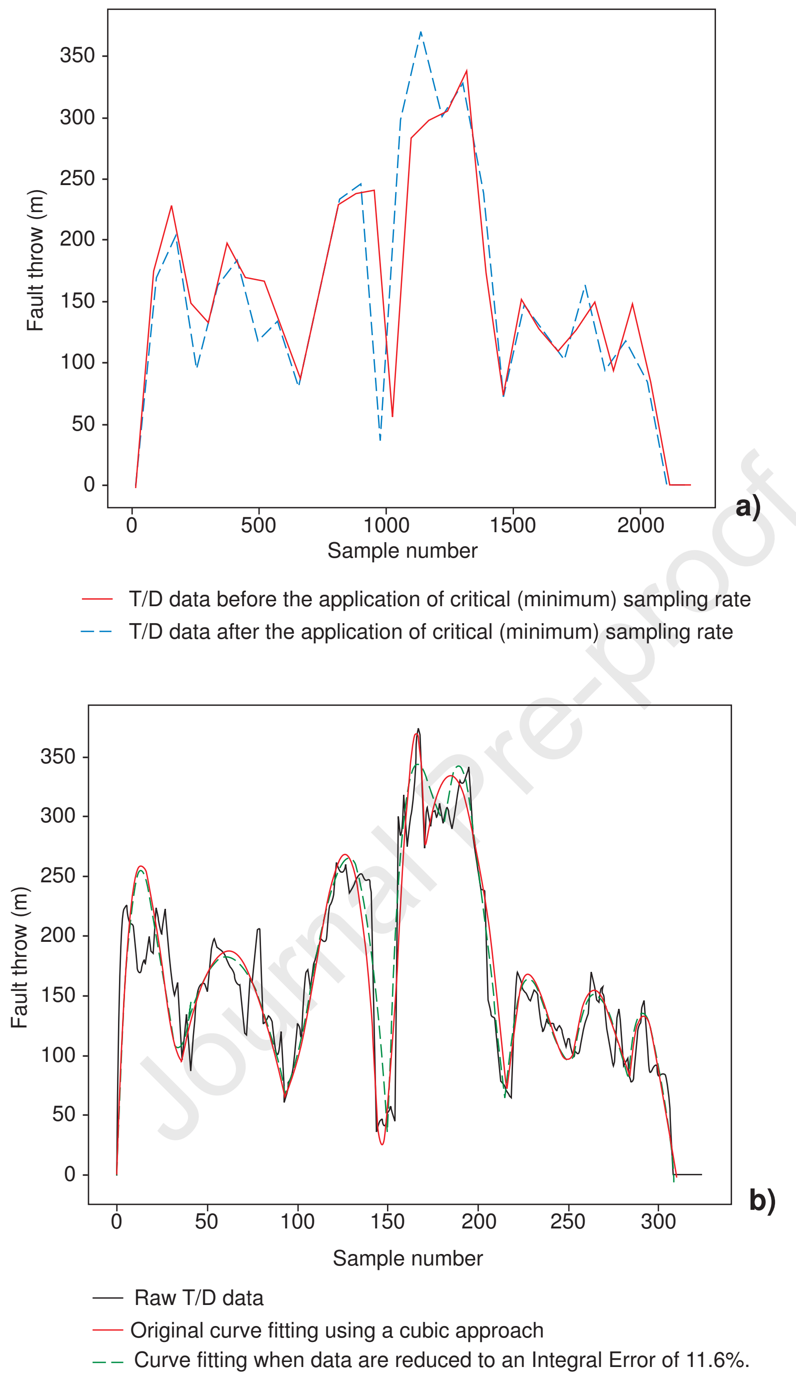


Figure 12

Journal Pre-proof



Journal Pre-proof

Figure 13

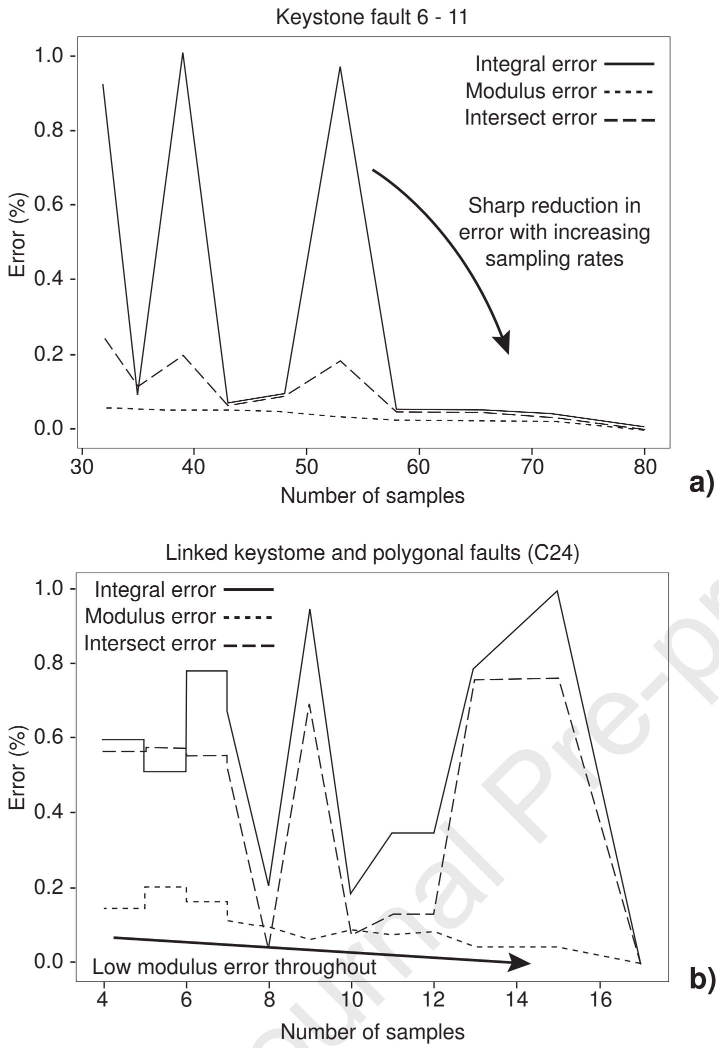


Figure 14

Journal Pre-proof

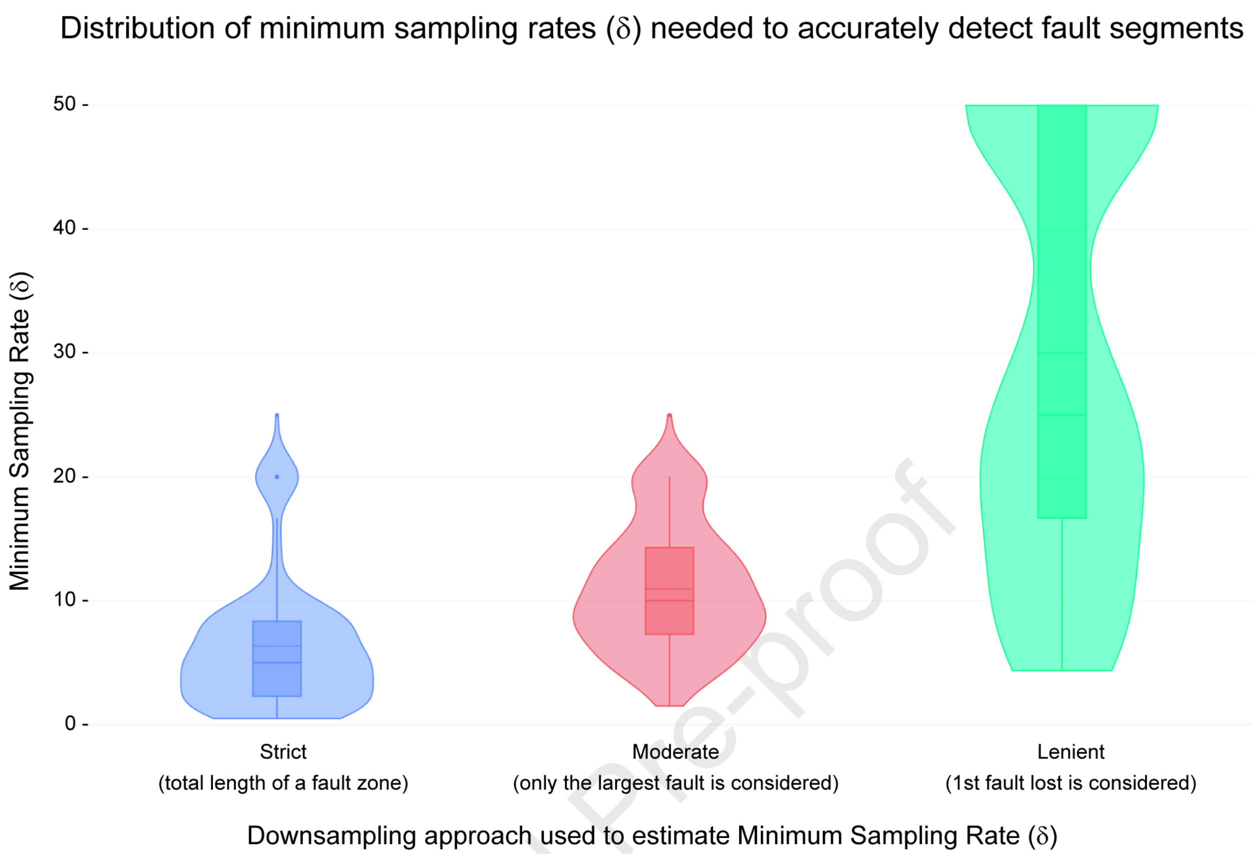


Figure 15

Journal Pre-proof

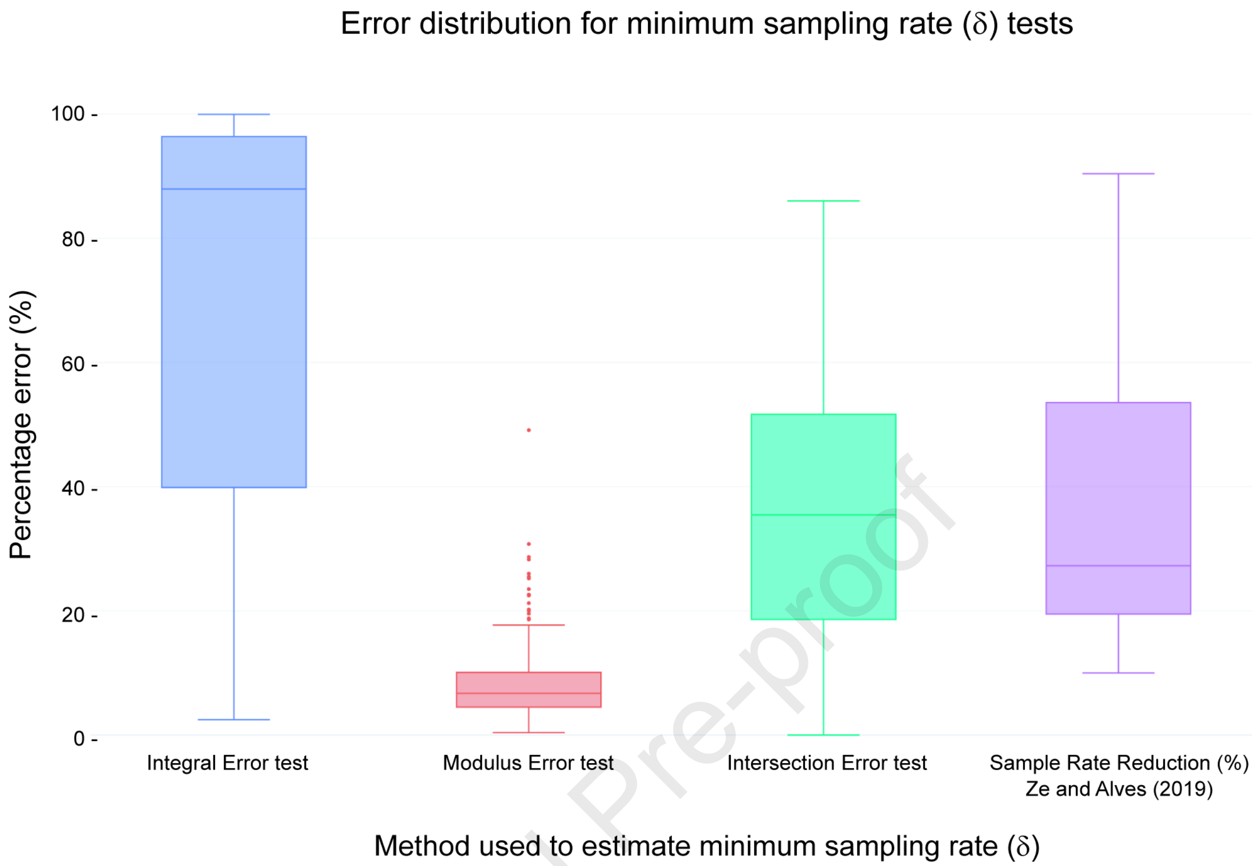


Figure 16

Journal Pre-proof

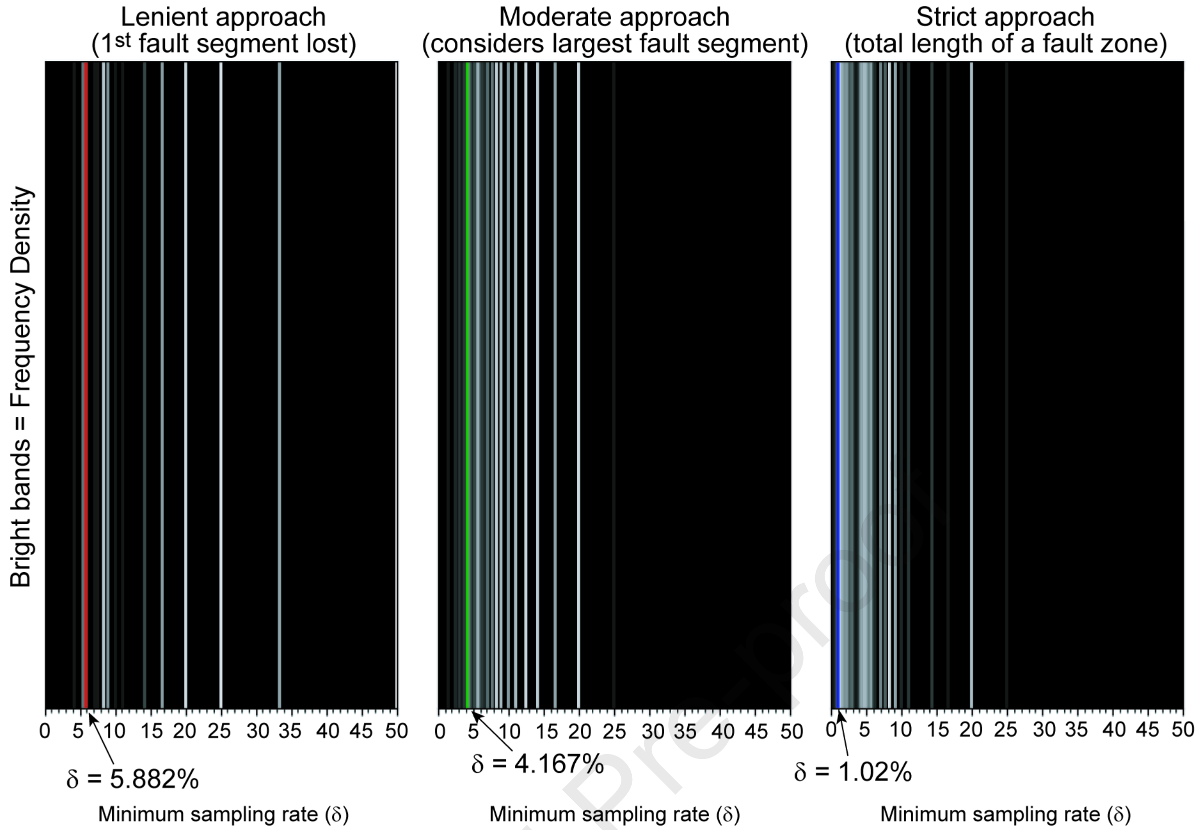


Figure 17

Journal Pre-proof

Journal Pre-proof

Conflicts of interest statement

The authors declare that they have no conflicts of interest regarding this work.

Journal Pre-proof

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof