Acta
Geophysica

RESEARCH ARTICLE - SOLID EARTH SCIENCES

Lithospheric mantle anisotropy from local events beneath the Sunda–Banda arc transition and its geodynamic implications

Syuhada Syuhada[1] · **Nugroho D. Hananto**[2] · **Chalid I. Abdullah**[3] · **Nanang T. Puspito**[4] · **Titi Anggono**[1] ·
Febty Febriani[1] · **Bogie Soedjatmiko**[1]

Abstract

Shear wave splitting analysis to characterise lithospheric mantle anisotropy has been performed to provide better knowledge about lithospheric deformation and mantle flow beneath the Sunda–Banda arc transition, Indonesia. The tectonic setting of the study area is very complex characterised by the transition from subduction along Sunda arc to collision in Banda arc. The splitting measurements show lateral and vertical variation in the fast directions of the S-waves in this region. When the splitting results are analysed through 2D delay-time tomography and spatial averaging, systematic patterns in delay times and fast polarisation become more visible. In the subduction domain, the spatial averages of fast directions are dominated by two distinct fast polarisations: perpendicular and parallel to the plate motion for shallow and deep events, respectively. The results suggest that anisotropy in this area is not only controlled by anisotropic source related to the simple mantle flow model, but also by anisotropic fabric in the mantle deformed under influence of high stresses, high water contents and low temperatures. In addition, there might also be contribution from the anisotropic body in the upper layer. In the collision domain, spatially averaged fast directions show mostly perpendicular to the plate motion for all deep levels. For shallow level in this region, this trend is mainly governed by the lithospheric deformation process due to the continent-arc collision as also shown by delay time tomographic inversion. For deeper part of the region, the result of tomographic inversion and spatial averaging reveals a high anisotropy followed by rotational pattern of fast directions in the north of Timor. We suggest that this pattern might be related to the induced mantle flow due to lateral tearing of the slab.

Keywords Shear wave splitting · Seismic anisotropy · Mantle flow · Sunda–Banda arc transition zone

Introduction

Seismic anisotropy analysis is a seismological tool that can be used to investigate the style of deformation occurring in the crust and mantle, and consequently to reveal the geodynamic evolution of the region. Seismic anisotropy itself is defined as the directional dependence of the seismic wave propagation in a medium within the Earth and is usually observed using shear wave splitting. This occurs when a shear wave propagating through an anisotropic medium splits into two orthogonally polarised shear waves with different velocities (e.g. Vinnik et al. 1984; Collings et al. 2013). Thus, shear wave splitting analysis consists in measuring the propagation direction of the fast wave (ϕ), and estimation of the time lag (δt) between the two arrival waves. The fast orientation provides information about the geometry of anisotropic body, and the split time corresponds to the intensity of anisotropy and a thickness of the anisotropic layer (e.g. Silver 1996).

Shear wave splitting method has been used for many different tectonic environments with various results, including both fast polarisation parallel and perpendicular to the absolute plate motion and variation in delay time values (e.g.

✉ Syuhada Syuhada
syuhada@lipi.go.id; hadda9@gmail.com

[1] Research Centre for Physics, Indonesian Institute of Sciences (LIPI), Tangerang Selatan, Banten, Indonesia

[2] Research Centre for Deep Sea-LIPI, Ambon, Indonesia

[3] Faculty of Earth Sciences and Technology, Bandung Institute of Technology, Kota Bandung, Indonesia

[4] Faculty of Mining and Petroleum Engineering, Bandung Institute of Technology, Kota Bandung, Indonesia

Long and van der Hilst 2005; Greve et al. 2008; Leo et al. 2012; Collings et al. 2013). In a single subduction system, the variation in splitting results might also be observed along the system (e.g. Hammond et al. 2010). This variation reflects the complex anisotropic scenarios of subduction systems with anisotropic contribution originating from various parts of the subduction zone such as the wedge, slab and sub-slab mantle (Long and Silver 2009). At shallow depth, seismic anisotropy is commonly produced by preferred orientation of cracks, fractures and aligned inclusions oriented in the regional direction of stress-field (Crampin 1994). Conversely, seismic anisotropy generated at deeper depth is often ascribed to the lattice preferred orientation (LPO) of anisotropic minerals, which may be fossilised into the lithosphere resulted from past tectonic deformation, or associated with recent asthenospheric flow (Savage 1999; Miller et al. 2013).

Sunda–Banda arc transition is complex tectonic setting involving the interaction of two tectonic regimes: subduction and collision. The study area is situated around Sumba Island and characterised by transition in tectonic setting from subduction in west of Sumba Island to collision in east of Sumba Island. This situation can be considered as

the best modern examples of the initial stages in transition from subduction to arc-continent collision (Fig. 1). In the west of Sumba, the Indo-Australian plate moves and subducts beneath the Eurasian plate in the direction N13E, at a rate of ~70 mm/year (Curray 1989). The relative plate motion then gradually decreases to as low as ~15 mm/year near the Timor Trough (Bock et al. 2003). Seismic reflection and refraction studies (e.g. Shulgin et al. 2009; Luschen et al. 2011) found that this area is highly deformed providing variation in structure and morphology. This area is also marked by the presence of two major structure discontinuities dividing the region into three tectonic blocks (Nugroho et al. 2009). First, major tectonic discontinuity separating between Sunda arc and Banda arc has been hypothesised based on the discontinuity line of volcanism located between Sumbawa and Flores Islands (Audley-Charles 1975). Another hypothesised structure discontinuity based on the uplift different of coral reefs is located along Pantar Strait (Nishimura and Suparka 1986). Furthermore, seismicity studies show that there are spatial gaps in shallow and intermediate depth as observed for example beneath Wetar Island and Timor Island (e.g. McCaffrey et al. 1985; Ely and Sandiford 2010). These
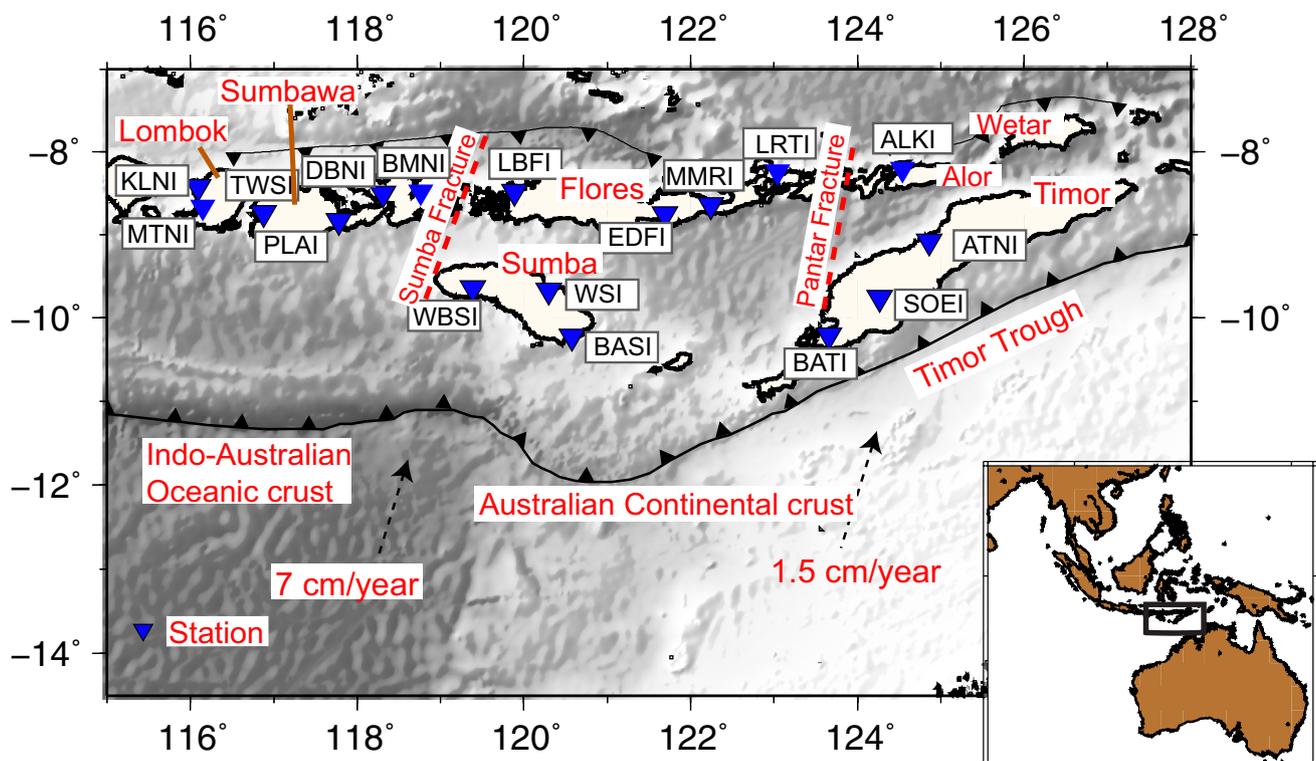

Fig. 1 The plate tectonic setting around Sunda–Banda transition and location of the study area with the features described in the text and bathymetry derived from Smith and Sandwell (1997). Dashed black arrows show displacement vector of the Indo-Australian oceanic lithosphere and the Australian continental lithosphere with respect to

the Eurasian lithosphere. Inverted blue triangles mark the broadband seismic stations used in this research. The red dashed lines reflect the structure discontinuity zones suggested by Nishimura and Suparka (1986)

gaps might correspond to the termination of volcanic belt or to tearing of the subducted oceanic slab separated from the continental part (McCaffrey et al. 1985; Ely and Sandiford 2010). Thus, this complex tectonic setting provides an ideal location to study the geodynamic evolution of the study area during the initial stages of arc–continent collision.

Recent seismological studies also have been carried out in order to substantiate the geodynamic and tectonic evolution of the region as result of the change of tectonic environment from subduction to collision (Syuhada et al. 2016, 2017). A previous crustal study (Syuhada et al. 2016) shows that the Moho depth varies between 26 and 38 km, and the deeper Mohos are found at seismic stations located in the collision domain. This study suggests that the thicker crust around the collision zone is governed by the insertion of buoyant Australian continental crust during the collision process. The study area also has high Vp/Vs that could be correlated to the local geological features such as the existence of mafic and ultramafic rocks, fluid-filled cracks and partial melt produced by magmatic processes in the upper mantle. The crustal anisotropy study shows that the fast polarisations in the subduction domain are in agreement with the general orientations of the compressive principal strain rates (Syuhada et al. 2017), suggesting that anisotropy in this area is mainly generated by stress-induced anisotropy. On the contrary, more scattered fast directions are shown in the collision domain, indicating that anisotropy is either produced by aligned cracks connected to the stress field or structural features related to the aligned macroscopic fabrics and alignment anisotropic minerals. As these studies were focused on the crustal deformation, the information regarding the influence of the lithospheric mantle deformation during the transition processes of tectonic regime changes is still unclear. This information is crucial to complete our understanding of the tectonic evolution of the study area.

In this region, the only information-related mantle anisotropy comes from shear wave splitting measurements using SKS phases and local S from limited seismic station coverage conducted by Hammond et al. (2010). These measurements obtained complex fast orientation around the study area. Hammond et al. (2010) suggested that the complex pattern of fast polarisations arises from the slab bending underneath the seismic station. Leo et al. (2012) interpreted that this pattern may be produced by fossil anisotropy formed from past tectonic events. However, those measurements used limited seismic station coverage causing incomplete information about mantle dynamics around the region. In this current work, we extend our research about lithospheric deformation in this area by studying mantle seismic anisotropy with significantly improved spatial coverage from local seismic stations.

Data and method

We use waveform data from 17 stations of GEOFON-IA network (GEOFON Data Centre 1993) to determine lithospheric anisotropy at Sunda–Banda transition (Fig. 2). We employ local intermediate depth earthquakes (40–300 km) with distance less than 300 km occurring during 2008 and 2015. This criterion is chosen to accommodate the interpretation of the result, and to ensure that the rays have the incidence angle less than 35° to remove effect of S–P converted phase from the surface. Lithosphere thickness in this area is around 120 km (McCaffrey et al. 1985) allowing us to divide the anisotropic analysis based on earthquake sources: shallow (< 100 km) and depth (> 100 km). The events with the above criteria are then searched from the relocated GEOFON-IA network catalogue (Nugraha et al. 2015).

The splitting parameters in this paper are estimated using the automatic shear wave splitting algorithm (MFAST) proposed by Savage et al. (2010). The package is based on the eigenvalue minimisation of Silver and Chan (1991) and the cluster analysis method of Teanby et al. (2004). The eigenvalue minimisation algorithm uses an inverse splitting operator to correct the splitting S-wave by finding the most linear particle motion. To find the best solution, the method applies a grid search over all possible solutions of fast directions (ϕ) between $-90°$ and $90°$ with a 1° increment and delay times (δt) between 0.00 and 0.80 s with a 0.01 increment. The pair of ϕ and δt providing the most linear particle motion with the smallest eigenvalue is chosen the best solution. In this stage, the eigenvector of the larger eigenvalue provides an approximate direction of the initial polarisation. Uncertainties of the measurements are then computed by searching the 95% confidence interval for the optimum splitting parameters using an F-test for the selected time window (Silver and Chan 1991).

We perform a series of bandpass filters ranging from 0.4 to 10 Hz to estimate the optimum filter, which is based on the maximum value of the signal-to-noise ratio (SNR)-bandwidth. In this analysis, we only use records with SNR greater than 3 for the splitting measurements. The SNR is defined as the average root mean square ratio of the signal and noise windows in the horizontal component of seismograms. The dominant frequency of the signal is determined using the frequency that has the highest spectral amplitude calculated from 3 s window after the S-arrival. To reduce subjectivity in choosing too large or too small windows causing ambiguous results, MFAST implements cluster analysis over a number of time windows to search the most stable fitting splitting parameters (Teanby et al. 2004). The quality of the measurements is then evaluated
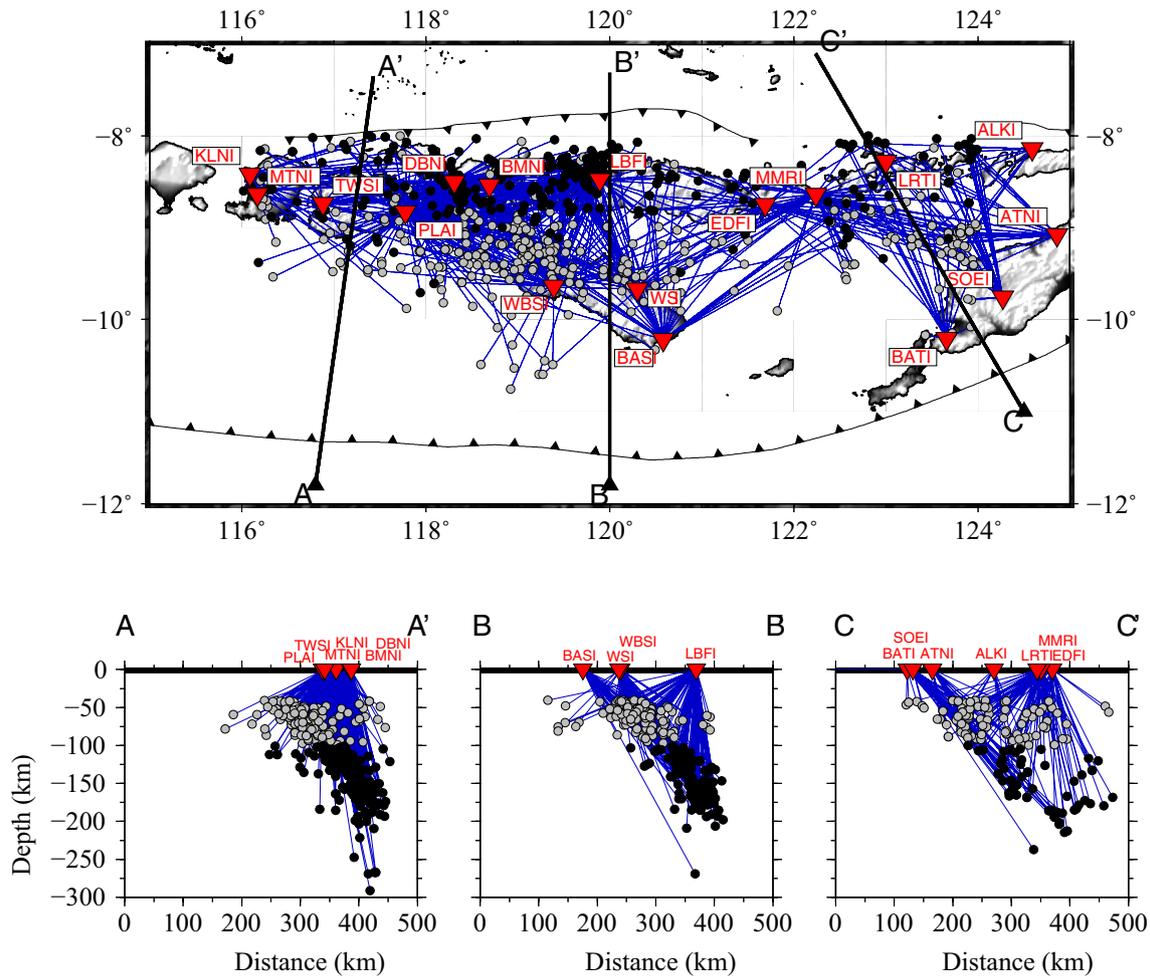
Fig. 2 The station network and the earthquake data used for the splitting analysis. The light and dark circles mark the shallow and deep events, respectively

based on grading system from A to D that represents good and bad results. This grading scheme depends not only on the SNR and uncertainty, but also stability of the clusters (Savage et al. 2010). In this research, only the results with grade of "B" or higher (here after AB grade) are used for further analysis. In this program, the AB grading measurements are achieved if: (1) the SNR > 3, (2) $\delta t < 0.8 * t\text{lag}_{max}$ ($t\text{lag}_{max}$ is defined as the maximum value of δt set in grid search), (3) the maximum error of 25° for the fast polarisation, and if the measurements have a A or B cluster grading in the cluster analysis.

In addition to splitting parameters, we also calculate uncertainties (e.g. standard deviation, standard error) for each single station measurements based on directional statistics (Mardia 1972; Gerst and Savage 2004; Karalliyadda and Savage 2013). Directional statistics is necessary to explain the estimated error and accuracy of mean fast azimuth. Due to 180° ambiguity of angular data, the directional statistics differs from the normal statistics. The obtained statistical

parameters from directional statistics, such as mean fast azimuth ϕ, length of the mean resultant vector of ϕ (R), circular standard deviation of ϕ (CSD) and standard error of (SE) are derived from the fast azimuths. The R is an estimation of circular spread from the calculated mean angular azimuth ϕ. SE is calculated from the obtained R and concentration parameter (κ), which is of a Von Mises distribution VM(ϕ,κ). The obtained SE indicates the accuracy of the mean fast azimuth ϕ. Tables 1, 2 and 3 summarise the obtained statistical parameters that are calculated from the splitting measurement from each station (Fig. 3).

2D delay-time tomography and spatial averaging method

To resolve and infer the spatial heterogeneity of splitting results, we apply 2D delay-time tomography and spatial averaging technique (TESSA) (Johnson et al. 2011;

Table 1 Computed splitting parameters from all stations for all depth

Station	N	Mean ϕ	SE of ϕ	CSD of ϕ	R	Mean δt	SE of δt	SD of δt
ALKI	9	76.43	16.67	39.37	0.39	0.23	0.04	0.11
ATNI	21	59.86	24.25	52.98	0.18	0.31	0.03	0.14
BASI	48	86.25	17.24	54.07	0.17	0.28	0.03	0.19
BATI	28	61.44	14.03	46.50	0.27	0.34	0.04	0.21
BMNI	9	27.95	55.40	58.83	0.12	0.22	0.03	0.10
DBNI	179	− 51.30	14.81	61.22	0.10	0.23	0.01	0.15
EDFI	55	− 55.29	23.80	59.66	0.11	0.22	0.02	0.15
KLNI	30	35.27	20.20	52.92	0.18	0.21	0.03	0.15
LBFI	133	− 8.78	3.77	36.65	0.44	0.25	0.01	0.15
LRTI	5	− 75.94	19.74	36.96	0.43	0.31	0.09	0.20
MMRI	72	73.93	8.65	46.30	0.27	0.27	0.02	0.19
MTNI	6	26.93	10.68	25.95	0.66	0.27	0.09	0.21
PLAI	175	72.77	6.27	48.36	0.24	0.20	0.01	0.11
SOEI	36	− 63.61	6.70	35.07	0.47	0.31	0.03	0.16
TWSI	24	− 37.14	13.78	44.86	0.29	0.27	0.04	0.21
WBSI	80	81.20	8.51	46.93	0.26	0.20	0.02	0.15
WSI	37	51.33	33.50	61.59	0.10	0.23	0.02	0.13

N number of events, *Mean ϕ* mean fast azimuth, *SE* standard error, *CSD* circular standard deviation, *R* mean resultant vector of ϕ, *Mean δt* mean delay time, *SD* standard deviation

Table 2 Computed splitting parameters from all stations for depth less than 100 km

Station	N	Mean ϕ	SE of ϕ	CSD of ϕ	R	Mean δt	SE of δt	SD of δt
ALKI	4	− 79.40	13.24	26.21	0.66	0.25	0.06	0.11
ATNI	13	14.85	70.90	64.53	0.08	0.31	0.04	0.12
BASI	19	− 77.95	9.79	36.29	0.45	0.25	0.04	0.19
BATI	20	70.24	9.10	35.33	0.47	0.26	0.03	0.13
BMNI	7	0.44	23.18	43.16	0.32	0.22	0.04	0.11
DBNI	56	− 29.68	29.34	62.57	0.09	0.18	0.01	0.08
EDFI	30	− 28.81	16.41	49.66	0.22	0.17	0.03	0.14
KLNI	14	66.04	26.68	51.33	0.20	0.14	0.02	0.08
LBFI	49	− 22.30	10.54	46.40	0.27	0.24	0.02	0.14
LRTI	3	− 49.64	53.50	50.15	0.22	0.25	0.08	0.14
MMRI	43	− 89.93	8.23	40.83	0.36	0.25	0.03	0.18
MTNI	4	14.95	14.47	28.15	0.62	0.36	0.11	0.21
PLAI	69	59.47	10.27	48.82	0.23	0.18	0.01	0.11
SOEI	21	− 79.51	10.32	38.32	0.41	0.26	0.03	0.14
TWSI	7	− 69.98	21.01	41.37	0.35	0.15	0.04	0.10
WBSI	44	− 88.13	8.19	40.95	0.36	0.16	0.02	0.12
WSI	14	50.90	28.39	52.30	0.19	0.20	0.02	0.09

N number of events, *Mean ϕ* mean fast azimuth, *SE* standard error, *CSD* circular standard deviation, *R* mean resultant vector of ϕ, *Mean δt* mean delay time, *SD* standard deviation

Karalliyadda and Savage 2013). The methods may allow us to estimate the spatial distribution of anisotropic structure in the first-order approximation. The delay time (δt) from the shear wave splitting measurement is assumed to be the summation of anisotropic structure along the raypath (e.g. Crampin 1991; Johnson et al. 2011), and it is proportional to the path length of the ray passing the anisotropic medium (Johnson et al. 2011; Karalliyadda and Savage 2013). TESSA assumes that the total delay time (δt) is the summation of the delay time from each

Table 3 Computed splitting parameters from all stations for depth larger than 100 km

Station	N	Mean ϕ	SE of ϕ	CSD of ϕ	R	Mean δt	SE of δt	SD of δt
ALKI	5	52.27	15.82	32.44	0.53	0.21	0.05	0.11
ATNI	8	67.44	13.57	34.13	0.49	0.33	0.05	0.15
BASI	29	46.51	23.88	55.17	0.16	0.30	0.04	0.19
BATI	8	3.18	17.39	39.06	0.39	0.54	0.09	0.25
BMNI	2	76.00	7.87	11.14	0.93	0.19	0.00	0.01
DBNI	123	−58.15	15.00	58.84	0.12	0.25	0.01	0.16
EDFI	25	−88.76	17.24	48.99	0.23	0.27	0.03	0.13
KLNI	16	19.75	16.51	44.48	0.30	0.27	0.05	0.18
LBFI	84	−5.13	3.56	30.71	0.56	0.26	0.02	0.15
LRTI	2	−84.00	7.87	11.14	0.93	0.40	0.17	0.24
MMRI	29	47.57	10.08	40.92	0.36	0.31	0.03	0.19
MTNI	2	42.00	1.41	2.00	1.00	0.09	0.03	0.04
PLAI	106	80.10	7.16	46.38	0.27	0.21	0.01	0.11
SOEI	15	−51.12	6.03	23.48	0.71	0.37	0.04	0.16
TWSI	17	−26.91	12.49	39.94	0.38	0.32	0.06	0.23
WBSI	36	59.64	14.22	48.82	0.23	0.25	0.03	0.16
WSI	23	52.45	94.76	71.46	0.04	0.25	0.03	0.15

N number of events, *Mean ϕ* mean fast azimuth, *SE* standard error, *CSD* circular standard deviation, *R* mean resultant vector of ϕ, *Mean δt* mean delay time, *SD* standard deviation

grid block with its strength anisotropy (s_b). To determine the strength anisotropy (s_b) for each block, the solution was estimated using medium-scale optimisation inversion function in MATLAB (lsqlin) (Johnson et al. 2011; Karalliyadda and Savage 2013). The algorithm solves a feasible initial solution and then converges to a final solution iteratively within the boundary constraints. The boundary constrains are applied so that the minimum strength anisotropy (s_b) is larger than 0 s/km and the maximum strength anisotropy (s_b) is less than $\delta t_{r(max)}/L_b$, where $\delta t_{r(max)}$ is the maximum delay time (δt) observed along a raypath and L_b is the grid size. In this study, we set the grid size to be 25 km. Number of rays for each grid is set to be minimum of 6 and maximum of 50 (see Appendix Fig. 17). To estimate average fast azimuth (ϕ) in each grid, we use tomographic weighting scheme applied in TESSA. The tomographic weighting is assumed to account every variation that occurs in the anisotropic structure (Karalliyadda and Savage 2013). The tomographic weighting function takes account on the strength anisotropy from each grid to estimate the average fast azimuth (ϕ) for each grid block as proposed by Johnson et al. (2011). This manner allows us to identify the blocks which may have more than a single mode of ϕ. The weighting parameter for each grid is estimated from the anisotropy strength along the raypath for each grid (s_b) normalised by the observed delay time for that raypath δt_r. In this study, we plot the ϕ mean direction from each grid in yellow and red bars if standard error of the mean is less and greater than 25°, respectively.

Results and discussion

The splitting results are presented in Tables 1, 2 and 3 as well as in Fig. 4. In general, all depth events give fast polarisation directions parallel or sub-parallel to the absolute plate motion (NE–SW) at some seismic stations (KLNI, MTNI, PLAI and MMRI). Other seismic stations, e.g. DBNI, TWSI and EDFI, show the average fast polarisation of the S waves perpendicular to the absolute plate motion. However, those seismic stations show either bimodal distribution of ϕ or high standard deviation, reflecting lateral and vertical variation of anisotropy in this area (Figs. 5, 6, 7). Delay times at some seismic stations are generally within the range of ~0.14–0.3 s for both shallow and deep events.

For seismic stations located in the subduction zone, the shallow events yield scattered distribution of fast directions, except for seismic stations located in Lombok Island, which are consistent from station to station suggesting localised coherent anisotropy. In Sumbawa Island, seismic stations exhibit two distinct fast azimuths: parallel or sub-parallel and perpendicular to the absolute plate motion, indicating lateral anisotropic variation. Deep events processed at stations in this island provide the same fast direction as shallow events do.

For seismic stations situated in the collision domain (Sumba, Flores and Timor), the splitting results display lateral and vertical variation in fast directions as seen in Figs. 4, 6 and 7. In Sumba Island, shallow events at station WBSI and WSI show E–W fast polarisation azimuths or almost perpendicular to the absolute plate motion. ϕ then becomes approximately NE–SW oriented from deeper

Fig. 3 Example of observed splitting result. **a** Manual S pick (solid black line) on the filtered waveforms from three component broadband seismograms with the splitting windows (dashed lines). The grey shaded region indicates the selected splitting analysis window used in the final step of measurement. **b** The rotated waveforms before and after corrections for splitting measurement with grey shaded area indicating the selected S-wave time window. The dashed lines again represent the splitting windows. **c** and **d** The results of the splitting parameters obtained using cluster analysis with the best solution showed by blue crosses. **e** The particle motion of the waveforms before and after correction for the splitting analysis. **f** The contour diagram for the best solution of the splitting parameters with the smallest eigenvalue (blue cross)
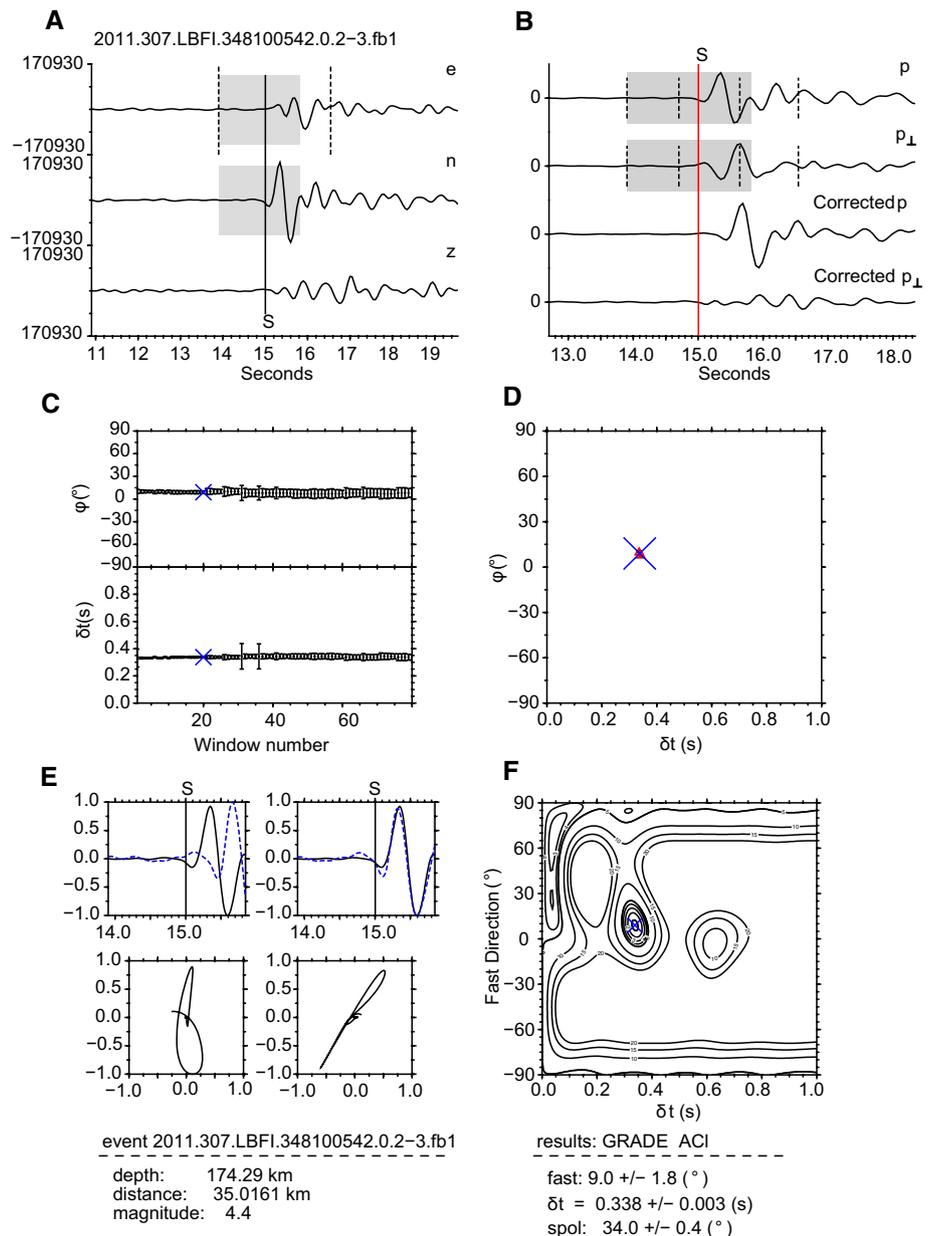
events. Another seismic station in this island displays consistent pattern of ϕ perpendicular to the plate motion for both shallow and deep events. Seismic station in Flores and Alor Islands also shows changes in ϕ with depth (e.g. EDFI, MMRI, ALKI) as seen in Fig. 7. Consistent patterns of ϕ with depth are observed for seismic station located in Timor Island with two distinct fast polarisation azimuths: parallel or sub-parallel (BATI and ATNI) and perpendicular to direction the absolute plate motion (SOEI).

The splitting measurements showing fast directions parallel and perpendicular to the direction of plate motion have been observed in most subduction zone worldwide (e.g. Greve et al. 2008; Long and Silver 2008; Hammond et al. 2010). If mantle anisotropy in this region occurs through the simple lattice preferred orientation (LPO) mechanism, causing the fast axis of olivine to be parallel with the maximum shear direction, then the simple mantle flow model would predict that the S-wave fast directions tend to align with the plate motion (Long 2013). However, the results exhibit that the fast directions at some stations do not always align with the plate motion. These fast directions might originate from olivine fabric deformed under high stress conditions and water content, causing the axis of the predominant mineral olivine to align perpendicular to the direction of mantle flow direction (Jung and Karato 2001). Another plausible mechanism is anisotropy caused by strong contribution of anisotropic body in the upper layer through shape preferred orientation mechanism (e.g. Audoine et al. 2000; Greve

Fig. 4 The spatial distribution of fast polarisations obtained for all events (**a**). **b** and **c** show the spatial distribution of ϕ obtained for event below and above 100 km, respectively. Yellow lines on the rose diagrams represent the average of fast polarisations with their standard deviation shown by the underlying blue area. Black arrow lines represent absolute plate motion direction based on CGPS 2004 model (Prawirodirdjo and Bock 2004)
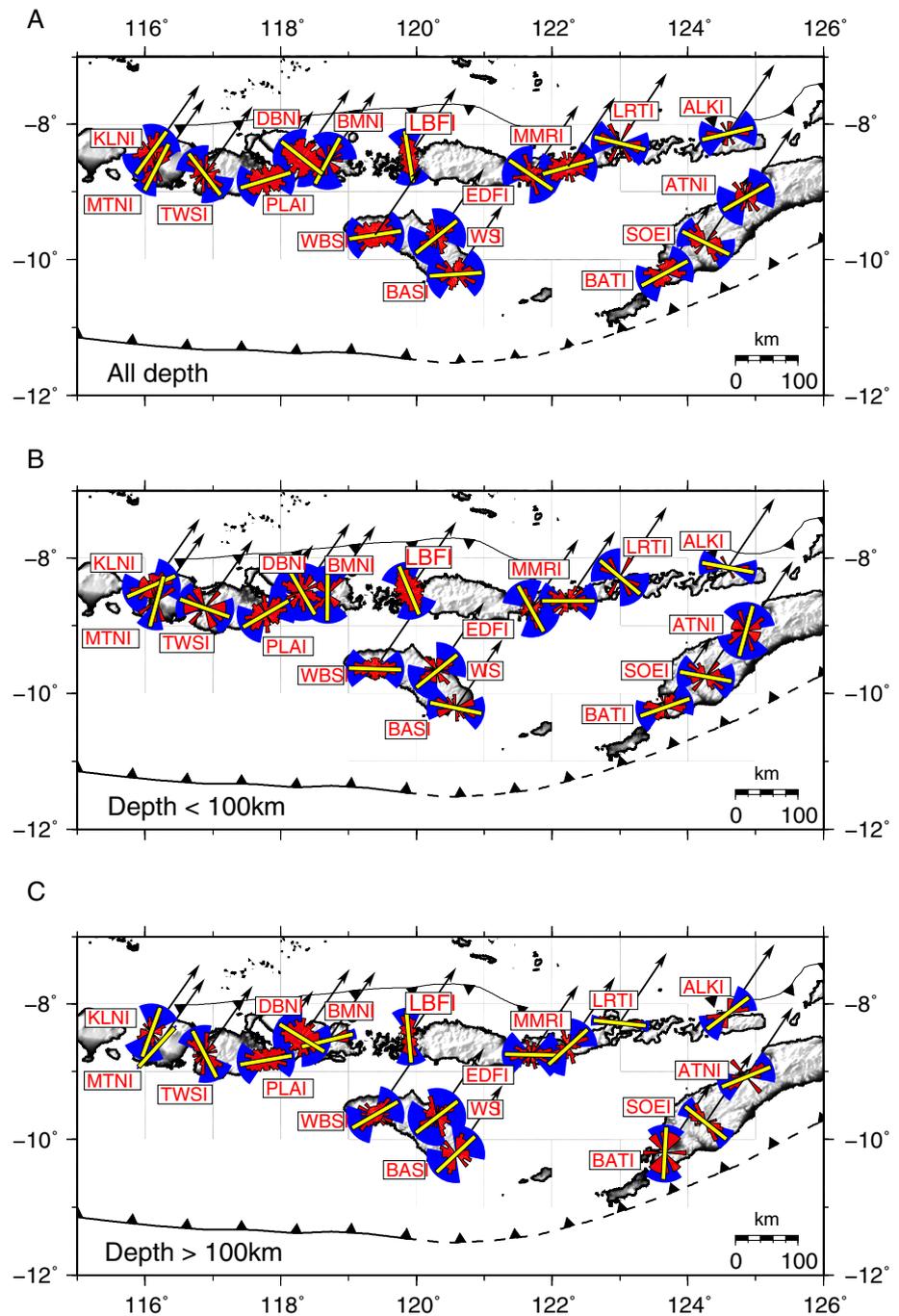
et al. 2008). Furthermore, spatial variations of fast orientations observed in this region might be related to variations in anisotropic geometry fabric (Karalliyadda and Savage 2013). The anisotropic geometry models with two or multiple anisotropic layers might produce unusual effects but with systematic patterns on the splitting parameters (e.g. Silver and Savage 1994; Rumpker and Silver 1998). We then seek any variation of splitting parameters by plotting them with backazimuth and incidence angle at each seismic station. The plots show that there are no any systematic correlation of splitting parameters with either backazimuth or incidence angle (see Appendix Figs. 18 and 19). These suggest that the splitting parameters in this study may be affected by complex anisotropic sources causing lateral variation in anisotropy and it causes difficulty in the interpretation. In this report, we will discuss more detail about the possible causes of lithospheric mantle anisotropy with their implications to geodynamics of this region, including interpretation from 2D delay-time tomography and spatial averaging technique in Sect. 4.
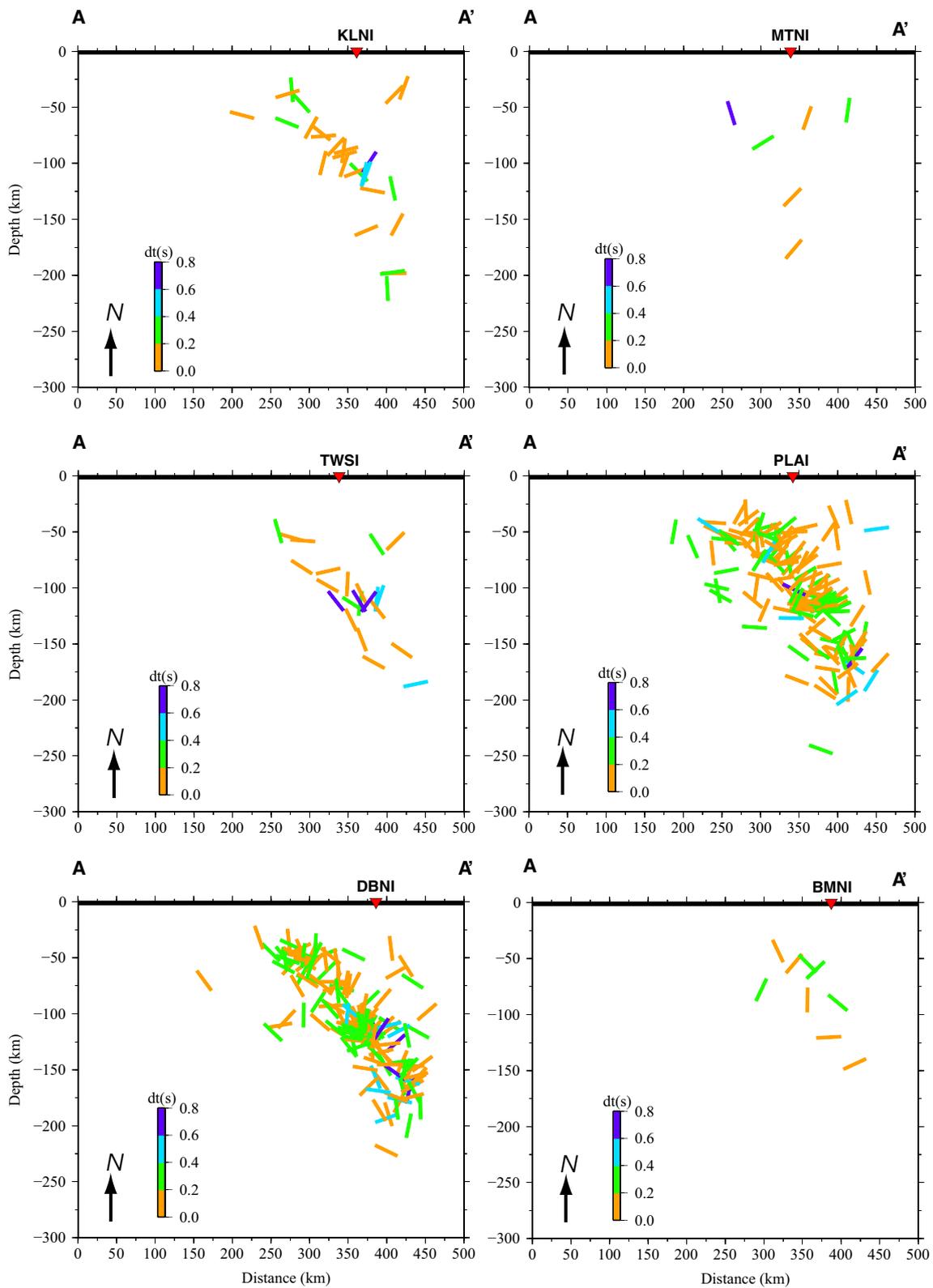
Fig. 5 Change of fast direction (ϕ) with depth along the profile A–A′ (see Fig. 2a). Station measurements are displayed in the upper panel. ϕ is estimated in the horizontal plane and plotted relative to N, which is marked on the figure. Colours of each bar represent the individual δt measurements
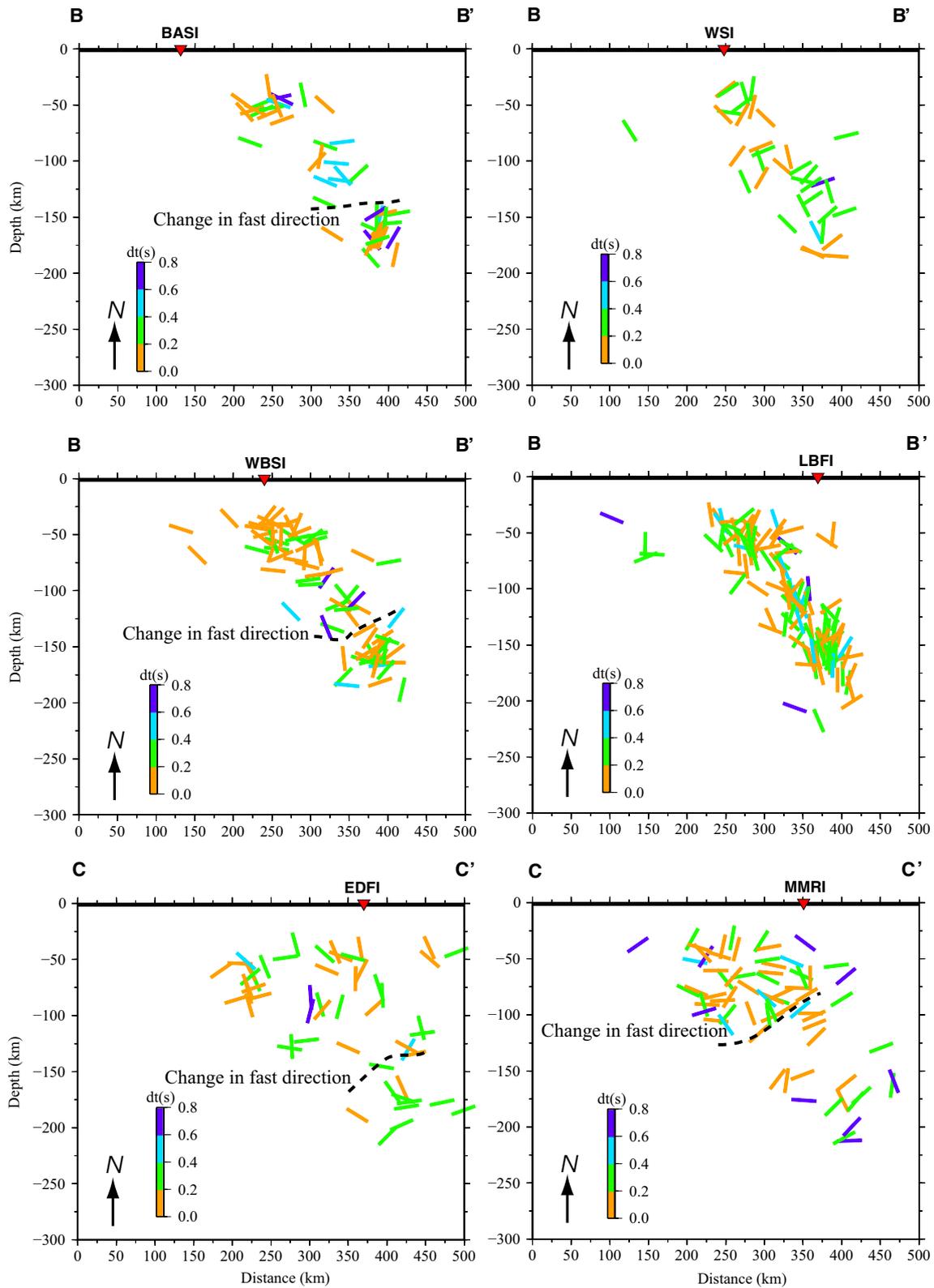
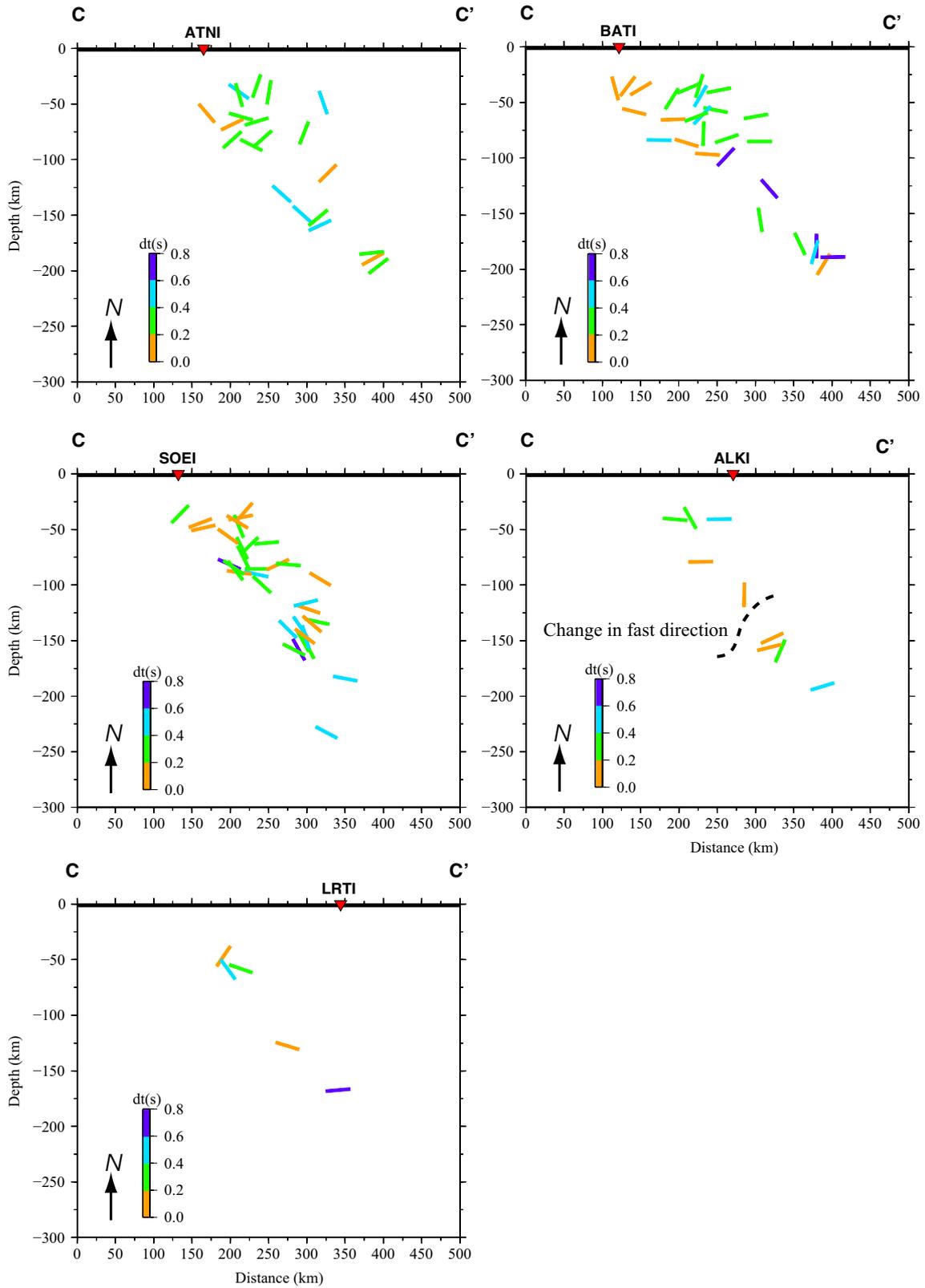**Fig. 6** Same with Fig. 5 but for profiles B–B′ and C–C′

Fig. 7 Same with Fig. 5 but for profile C–C'

Percent anisotropy

To calculate percent anisotropy, we use the relationship between δt (delay time) and L (path length) given by the following linear expression (e.g. Savage 1999):

$$\delta t = L \frac{\delta \hat{\beta}}{\beta}$$

here, β is the average velocity of S waves along the anisotropic path, $\delta \hat{\beta}$ is the fraction of velocity anisotropy computed from the splitting measurements and the path length is proportional to hypocentral distance. For each station, the delay times are plotted versus hypocentral distance and the best slope is calculated using the least squares method. An approximate percent velocity then can be determined by using an average S-wave velocity of about 4.5 km/s for a single anisotropic layer. Our estimations show that velocity anisotropy calculated for some seismic stations located in the collision domain is generally higher than that in the subduction domain (Figs. 8, 9, 10). For example, stations located in Timor Island yield average velocity anisotropy around 0.65–0.85%. In Flores Island, station MMRI shows that the velocity anisotropy is similar to that of seismic stations in Timor Island. In subduction domain, some seismic stations provide velocity anisotropy of about 0.51–0.57%; however, the results of delay times from stations DBNI and BMNI give higher velocity anisotropy of around 0.6%.

The strength anisotropy between 0.65 and 0.85% found in the study area is comparable with the observation conducted in other subduction systems, for instance, anisotropy of 0.5% at depth less than 100 km beneath Peru (Kaneshima and Silver 1995), 1.0% anisotropy in Aleutians (Yang et al. 1995) and 1.5% in the west volcanic front of Japan (Okada et al. 1995). Furthermore, the seismic anisotropy observed from splitting results for the mantle above and below the slabs ranges between 0.5 and 2.0%, (Fouch and Fischer 1996; Okada et al. 1995) and up to 5% within the slabs (e.g., Shih et al. 1991; Kaneshima and Silver 1995; Gledhill and Stuart 1996; Hiramatsu et al. 1997).

Comparison with previous anisotropy studies

The splitting measurements from the previous anisotropy studies were conducted in the small part of the study area with a limited number station (e.g. Hammond et al. 2010), in which the measurements are mainly performed around Java-Sumatra subduction zone. They used station BMI (part of JISNET), which is located at the same location as station BMNI in our study. The fast polarisations obtained in this previous work shows scatter or complicated pattern around the boundary of Sunda subduction zone and Banda collision zone. Their results are consistent with ours for both shallow and deep events. Hammond et al. (2010) also obtained delay time around 0.1–0.9 s (with an average of 0.4 s) and 1.25–1.82 s (with an average of 1.5 s) for local S-wave data and SKS data, respectively. Thus, our estimation of splitting delay time with an average of 0.22 s for BMNI station is within their range estimation for local S-wave measurements. However, our average delay time is about 14% of that obtained by their SKS results. This discrepancy between our estimation and SKS results may suggest that SKS data measure anisotropy sources from deeper part of lithosphere, while our analysis contains higher frequencies, which is more sensitive to re-splitting due to small-scale heterogeneities beneath the station (Karalliyadda and Savage 2013; Wirth and Long 2010).

Interpretation and discussion

Sumbawa zone: two-layer anisotropy

Bimodal distribution or spatial variation of ϕ with high standard deviation suggests the existence of complex anisotropic features beneath the study area. Complex or multi-layers of anisotropy are usually characterised by a $\pi/2$ periodic pattern of splitting parameters as a function of source backazimuth or source initial polarisation (ϕ_p) (Silver and Savage 1994). Initial polarisation is the polarisation azimuth of the unsplit S wave before passing through the anisotropic layer. In the case of simple two anisotropic layers, the S wave entering the lower anisotropic layer will split into the fast and the slow waves while propagating through it. The splitting waves then enter the upper anisotropic layer and resplit with different fast orientation than the lower one. As consequence, the initial polarisation azimuth in the upper layer is equivalent to the fast polarisation of the split waves resulted from the previous anisotropic layer.

In the two-layer anisotropy model, the splitting parameters will produce a characteristic pattern of a $\pi/2$ periodicity as a function of the initial polarisation (Silver and Savage 1994). The patterns can be clearly identified for the wave observed at long periods ($T/\delta t > 5$) (Rumpker and Silver 1998). Therefore, to identify the possibility of multiple anisotropic layers beneath the study area, we select the splitting parameters that satisfy the above requirement. From 17 seismic stations, we only have two seismic stations (DBNI and PLAI) with suitable number of splitting records for multiple anisotropic layer analysis. From the two seismic stations,
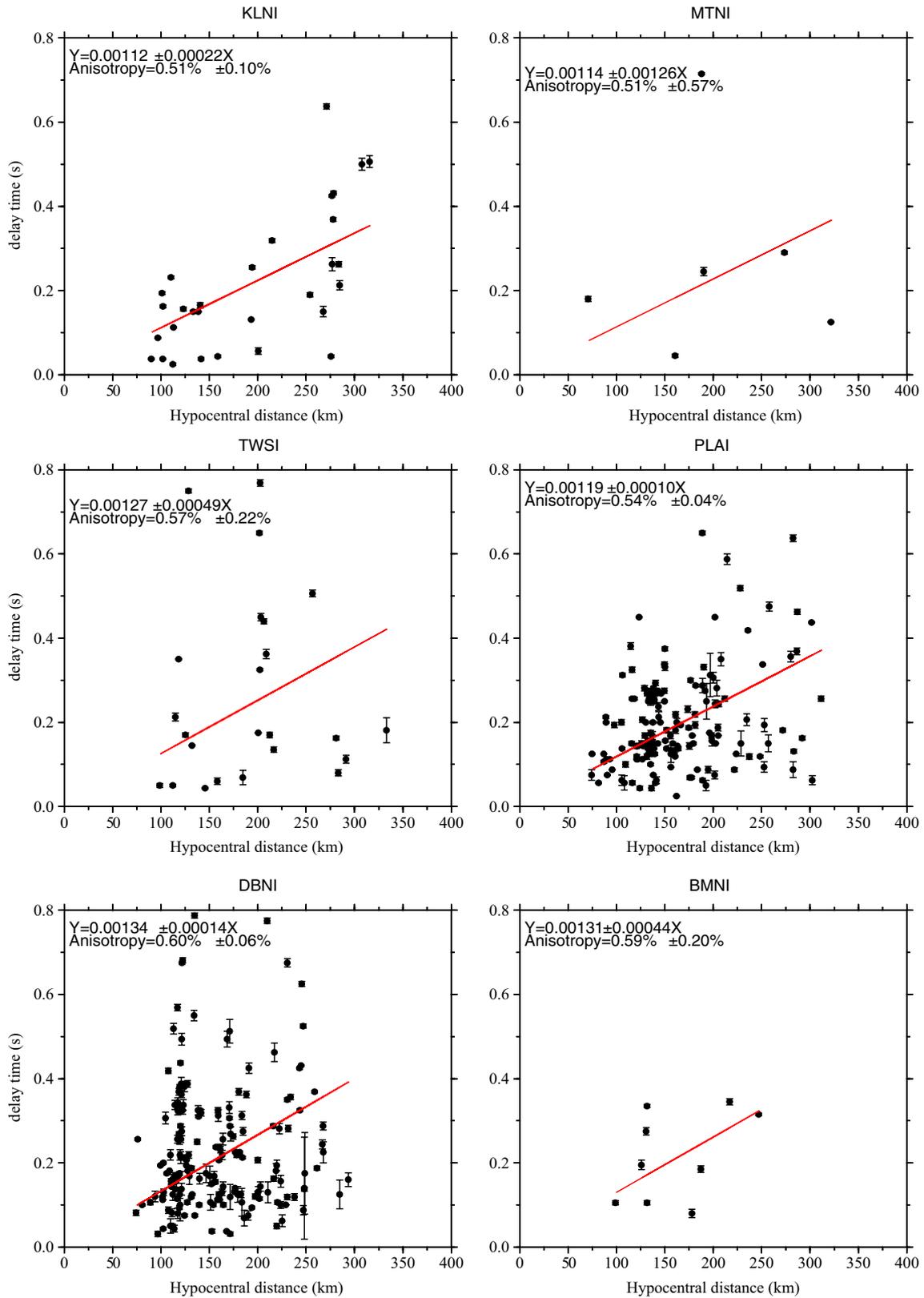
Fig. 8 Delay times versus hypocentral distance for seismic stations located in Sumbawa and Lombok Islands. The solid red lines show the least square fit
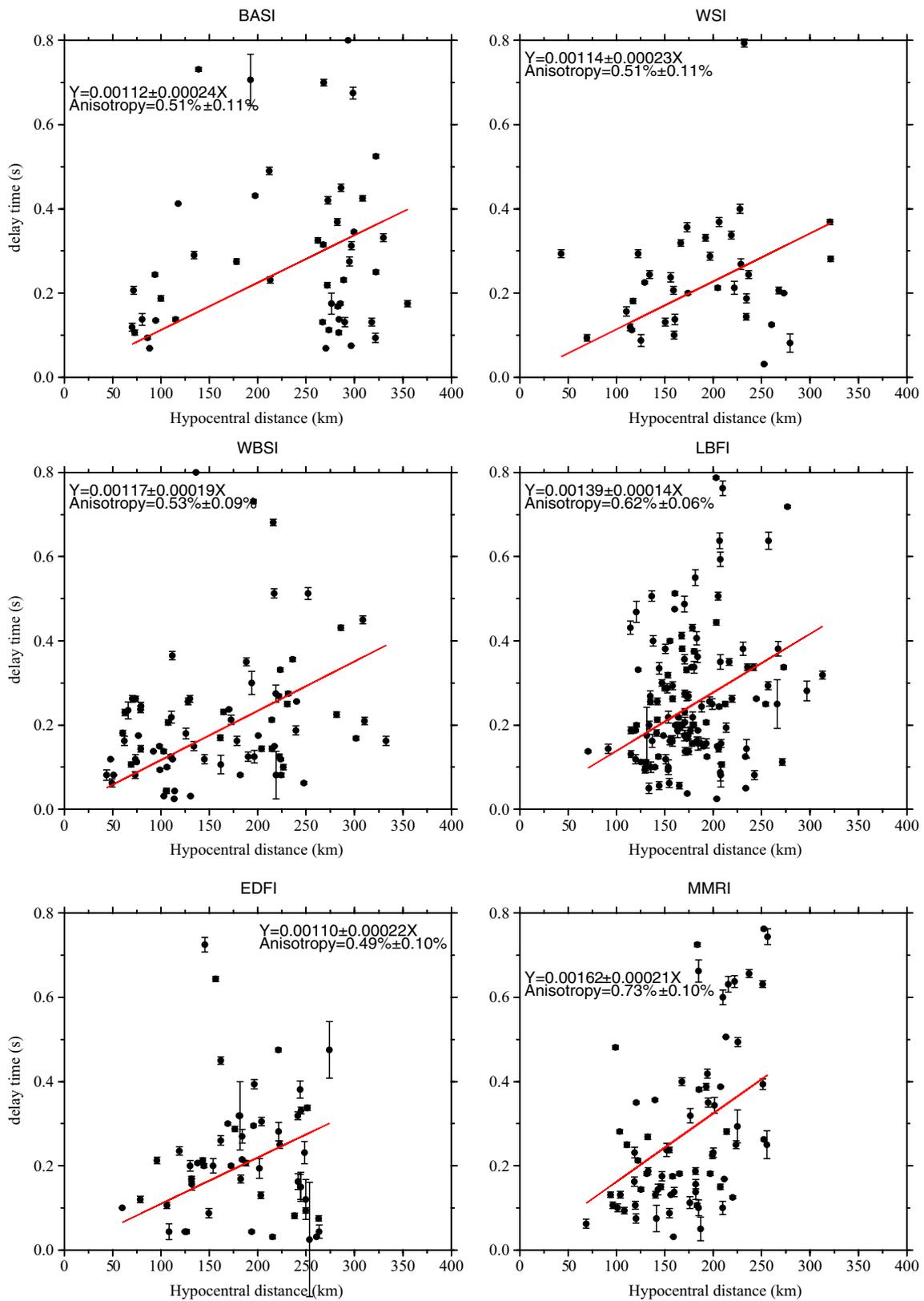
Fig. 9 Delay times versus hypocentral distance for seismic stations located in Flores and Sumba Islands. The solid red lines show the least square fit
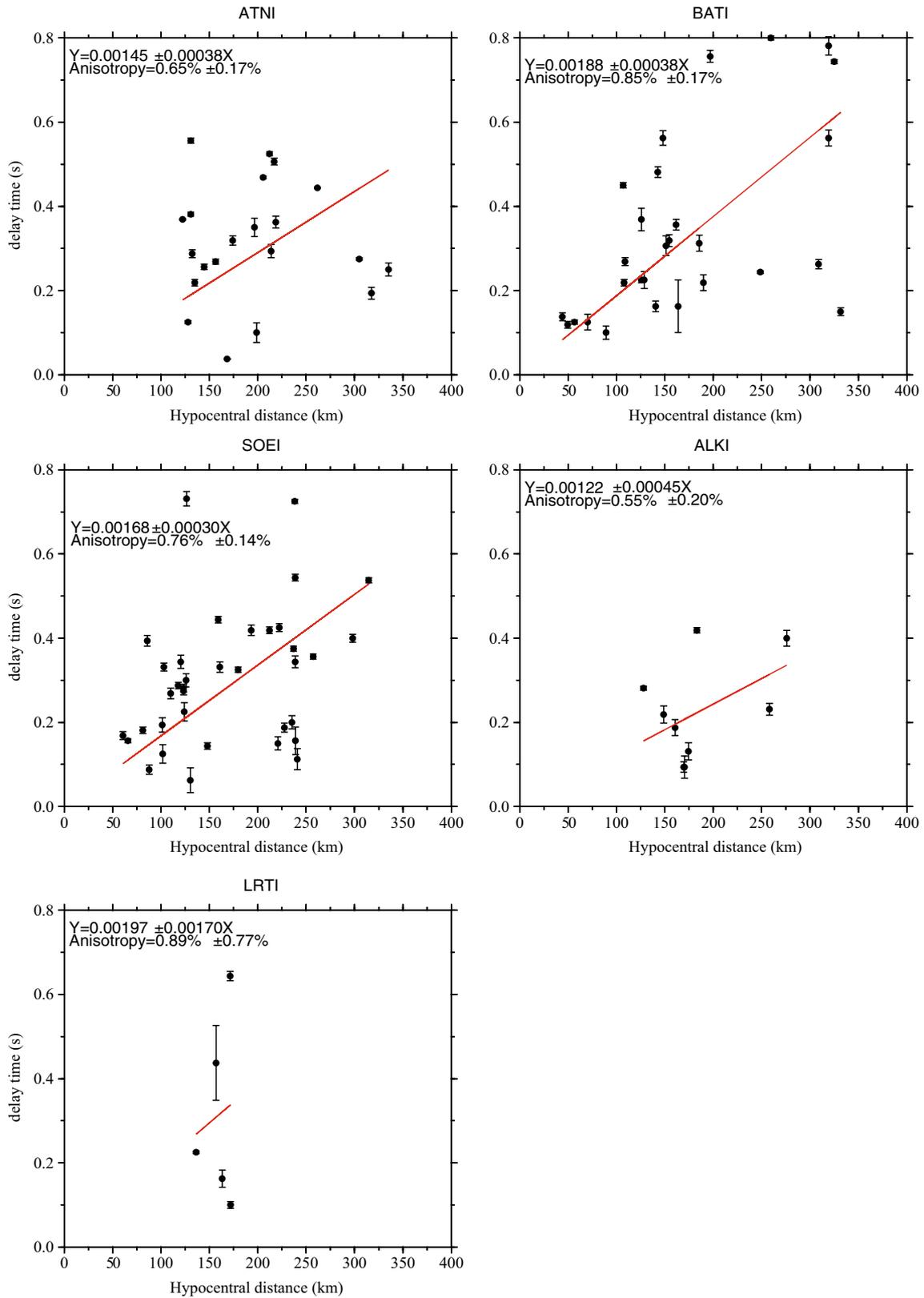
Fig. 10 Delay times versus hypocentral distance for seismic stations located in Flores, Alor and Timor Islands. The solid red lines show the least square fit

only at DBNI station we suggest that a $\pi/2$ periodic pattern in delay time with respect to the initial polarisation ϕ_p can be identified. On the contrary, the fast directions measured at this station do not show any periodic patterns as a function of initial polarisation. We can assume that the system of two anisotropic layers is likely present underneath this station. In this case, we perform forward modelling technique to validate the two-layer anisotropy model with a horizontal symmetry axis (Savage et al. 1990; Silver and Savage 1994). This technique uses the least square method to search the best model of the two-layer anisotropic structure with a horizontal symmetry axis by minimising error between the observed and calculated splitting parameters. The method of finding ϕ and δt (for both the upper and lower layers) is based on grid search techniques from $-90°$ to 90° with intervals of 1° for the fast direction and from 0–dtmax with increment of 0.01 s for the delay time. The optimum model obtained from this method is characterised by $\phi_{lower\ layer} = 33°$, $\delta t_{lower\ layer} = 0.11$ s, $\phi_{upper\ layer} = -52°$ dan $\delta t_{upper\ layer} = 0.04$ s (Fig. 11). However, this model still has limitation due to limited number of data. Therefore, further study may be necessary to model the possible two-layer anisotropic around the study area by involving, for example, larger dataset.

The fast directions on DBNI station obtained from both earthquake sources are consistently NW–SE trending, perpendicular to Sunda–Banda trench. We suggest that anisotropy beneath this station is controlled not only by anisotropic source related to the mantle flow but also anisotropic structure in the upper layer. For further analysis about the existence of two-layer anisotropic layer underneath this station, we plot the rose diagram of the initial polarisations for both earthquake sources (Fig. 12). As stated earlier that

the initial polarisation azimuth of the wave when passing through the upper layer will display the fast polarisation of the lower layer. Here, the initial polarisation from both earthquake sources are generally aligned NE–SW or parallel to the plate motion. Interestingly, the initial polarisations of shallow events are consistent with the forward modelling results showing the NE–SW direction of the fast polarisations for the lower layer, as expected. Thus, this might indicate that anisotropy from above 50 km is so strong causing the S-waves from the event below 50 km completely split when passing through it. Therefore, we suggest that anisotropic source beneath this station is mainly controlled by anisotropic body located in the crust or upper mantle.

As discussed in the previous section that there are two main possible mechanisms related to the fast orientation of splitting wave perpendicular to the plate motion (e.g. Karato et al. 2008; Karato 2004; Fischer et al. 2000). First, if the upper mantle has dominant contribution to the anisotropy, it is more likely due to a change of deformation of olivine aggregates from typical [100] to [010] (B-type olivine) under certain temperature, pressure conditions and high water content (Karato et al. 2008). The subduction of the Indo-Australian plate along the Sunda margin began ~ 45 m year ago (Hall 2002), this long subduction process may provide temperature on the subducted slab to be cold enough to support the high-pressure environment that could contribute to the development of B-type olivine. Furthermore, Karato (2004) proposed that water content in the mantle also may be localised beneath volcanic arc. Thus, the presence of volcanic islands in the north of this research area may suggest the presence of structural column with high water content.
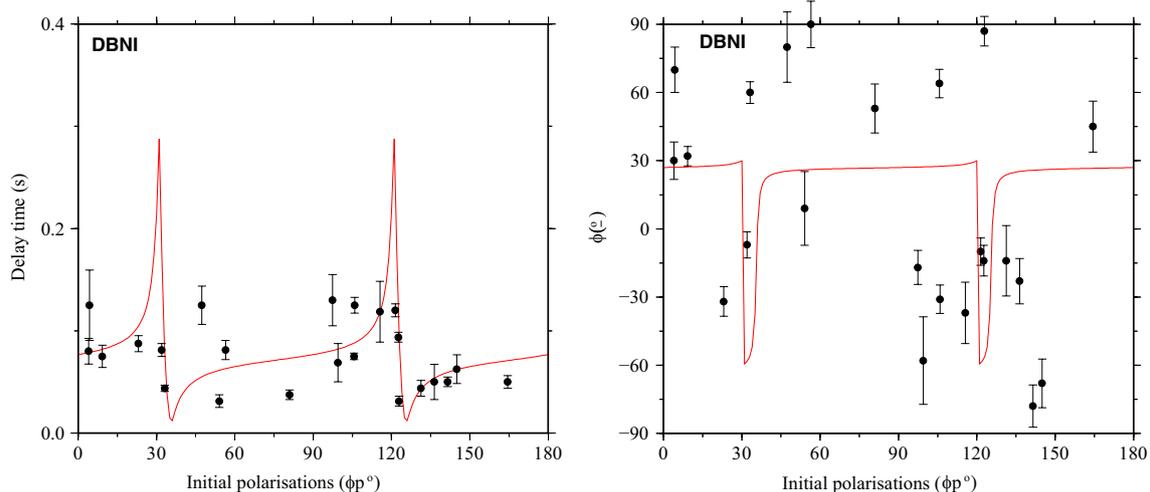

Fig. 11 Delay times versus initial polarisation for splitting parameters with $T/\delta t > 5$ at DBNI stations. Plot suggests a $\pi/2$ periodicity with a favorable model determined using grid search method
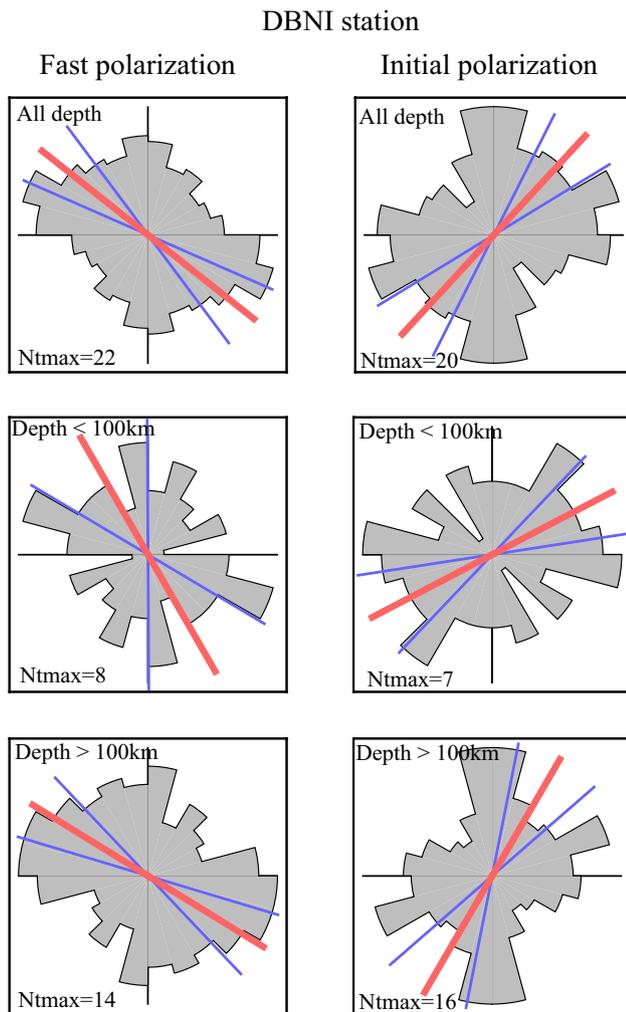
DBNI station

Fast polarization | Initial polarization

Fig. 12 Plot showing the fast polarisation and initial polarisation versus depth. The red and blue lines indicate the mean and error of the measurements, respectively. The splitting result beneath this station shows that the S waves initially have polarisation around NE–SW for both earthquake sources when passing through the upper anisotropic layer. This indicates that the waves leave the lower anisotropic layer with consistent polarisation direction

Another plausible mechanism is anisotropy due to shape preferred orientation (SPO) of the partial melting. Fischer et al. (2000) showed a model of melt filled crack aligned 20°–30° from the maximum deviatoric compressive stress to explain the mechanism anisotropy in the mantle wedge. They demonstrated that melt-filled inclusions could result in anisotropy perpendicular to the direction of subduction. Furthermore, a receiver function study conducted in the

study area observes the presence of partial melt beneath Sumbawa and Flores Islands marked by low-velocity zones with high Vp/Vs (Syuhada et al. 2016). Thus in this case, we suggest that the production of melt is likely restricted to vertical thin channel and concentrated in a small region directly beneath volcanoes (e.g. Long and van der Hilst 2006), as this only occurs at limited stations. This anisotropic mechanism also likely occurs for station TWSI providing plate motion-perpendicular splitting observations.

Flores zone: Anisotropy in the lithosphere or asthenosphere?

In the area around Flores and Alor Islands, the splitting results generally give absolute plate motion-perpendicular and motion-parallel anisotropy for seismic broadband stations located in western part of the area and eastern part of the area, respectively. In the west of the area, at station LBFI, fast polarisations are consistently aligned NW–SE or NNW–SSE for both shallow and deep events. Stations further east of the area indicate that the fast directions vary with depth as shown in Fig. 13. To reveal further about laterally and depth varying anisotropy across this area, we plot spatially the change in fast directions with depth for the stations around this area. Figure 13 shows that at longitude 118°–121° the fast directions are predominantly NW–SE or NNW–SSE at depth less than 100 km, while at longitude 122°–124° the fast polarisations have two dominant directions: around NW–SE and NE–SW. At depth more than 150 km, the spatially change of fast polarisations is more obvious marked by significant rotation from NNW–SSE or N–S in the west of the area to NE–SW in the east of the area.

The complex pattern of anisotropy from events less than 100 km in this area is consistent with the previous anisotropy study using JISNET station network (Hammond et al. 2010; Di Leo et al. 2012). As suggested by Di Leo et al. (2012), the complex pattern of fast directions in this region could be associated to the fossil anisotropy frozen in the boundary of the two tectonic systems (subduction and collision). The joining of Australian continental lithosphere and Indo-Australia oceanic lithosphere may produce high stresses and strains around the boundary zone between the two lithospheric fragments, and thus the fossil anisotropy could remain from this tectonic deformation. Alternatively, Flores Island is situated in the boundary transition zone from subduction to collision, in which the crust of

Fig. 13 Plot showing the fast polarisation versus longitude and depth for seismic stations around Flores and Alor Islands. **a** For all depth. **b** Depth less than 150 km. The pattern of fast polarisation rotates from around NNW–SSE or N–S in the west of Flores, to around NE–SW in the east of Flores
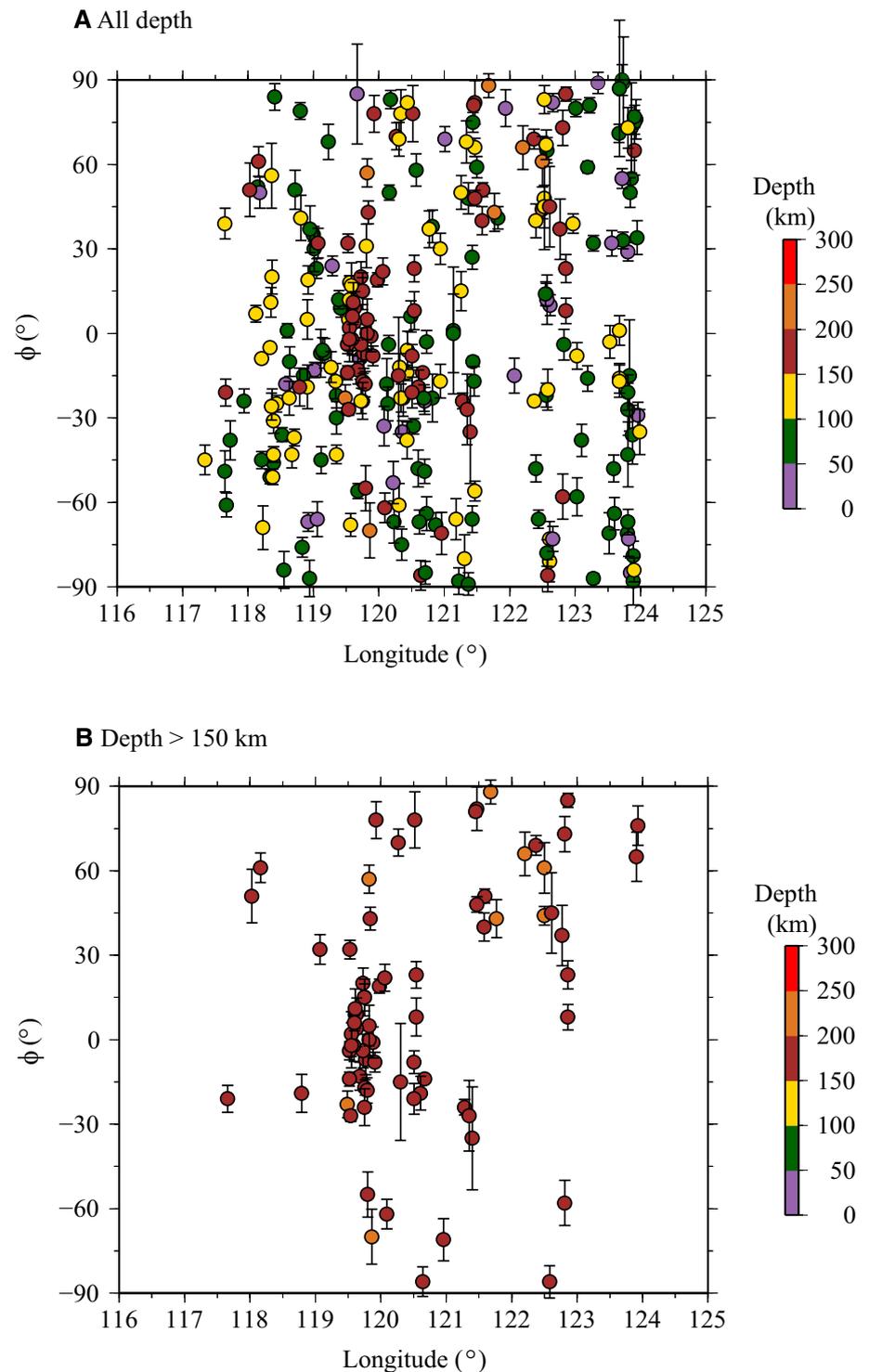
subducted slab consists of material transitional between continental and oceanic crust (McCaffrey et al. 1985; Ely and Sandiford 2010). A seismotectonic study conducted by Ely and Sandiford (2010) showed that the presence of the continental character on the subducted slab, which is more buoyant than oceanic material, could reduce the slab

Fig. 14 Model interpretation for sub-crustal anisotropy beneath Sunda-Banda arc transition. **a** Anisotropy pattern which might be related to the tear slab inducing mantle to flow parallel to strike of the subducting slab. **b** Anisotropy pattern which might be correlated to the fertility change in the mantle wedge
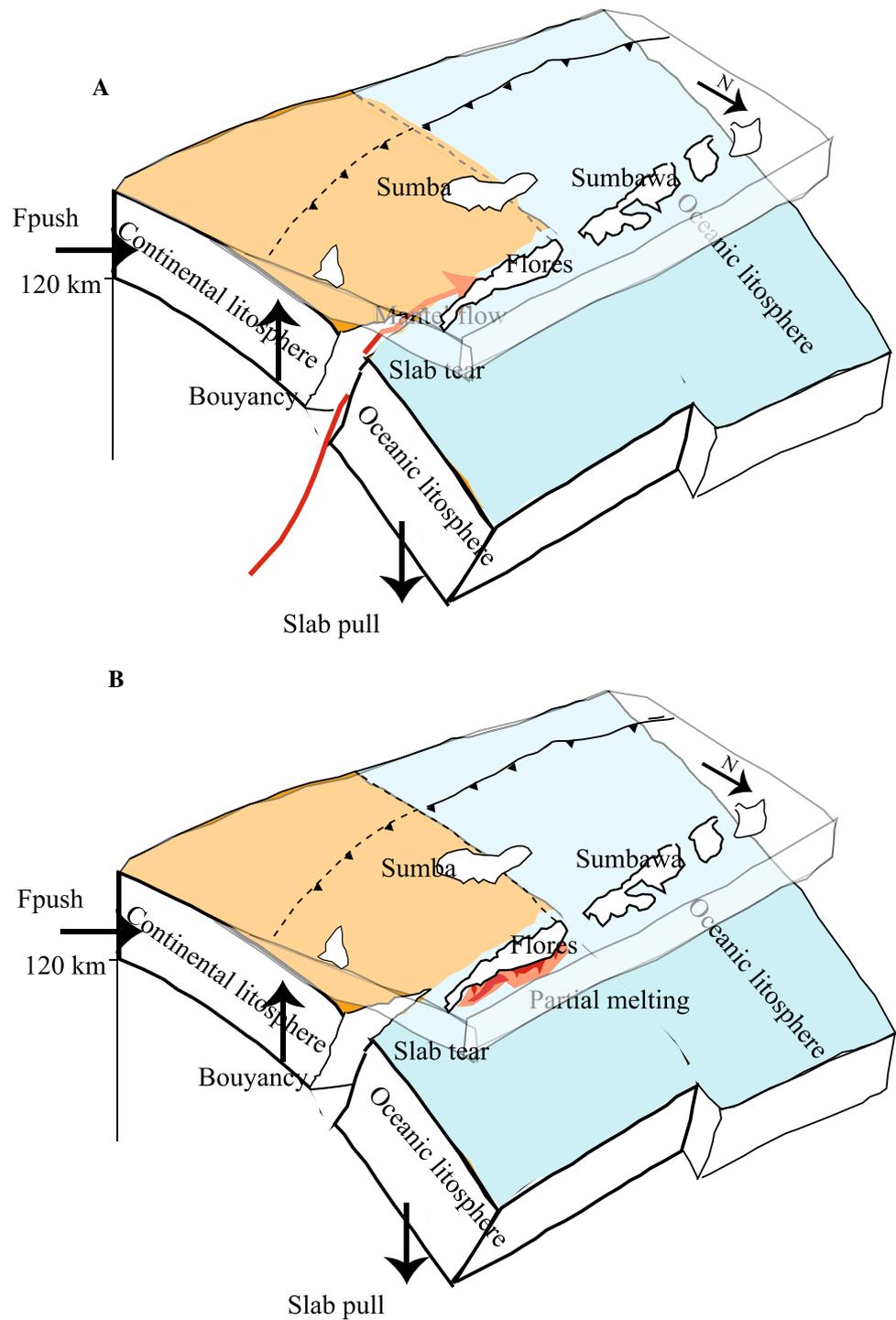
pull force causing the slab to be resistant to subduction. This study also suggested that the material transition has been subducted to around 300-km depth characterised by

the unusual stress state in the subducted slab of down-dip compression. Consequently, the subducted oceanic slab may detach from the continental part as shown by steep

Fig. 15 2-D delay-time tomography from the inversion of splitting data for two earthquake groups with depth ≤ 100 km and depth > 100 km. The region with high or strong anisotropy is indicated by warm colours (red to yellowish green), whereas the region with weak anisotropy is marked by blue colours. The shaded region depicts the areas with poor resolution
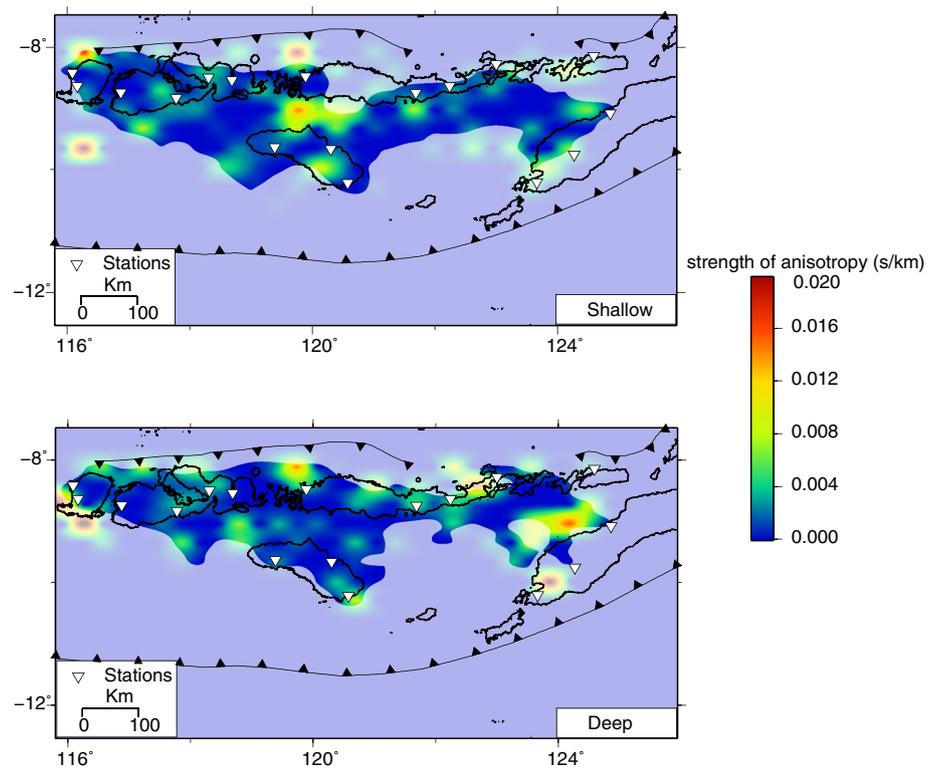

nodal planes observed from earthquake focal mechanisms above 115 km recorded in the north of Timor (McCaffrey et al. 1985; Ely and Sandiford 2010). These studies also suggested that the slab tear is probably propagating eastward at a much faster rate. In this case, we suggest that this lateral tearing of the slab might induce the mantle to flow parallel to the strike of the slab (e.g. Baccheschi et al. 2007) (see the illustration in Fig. 14). Thus, it might correlate to the rotational pattern of the fast directions observed for the events below 150 km. However, further investigations related to the mantle flow model are necessary to reveal the exact pattern and mechanism of the tectonic deformation in this region.

Lastly, Flores zone is unique in which the volcano spacing is quite close compared to the other islands in this arc system. Ely and Sandiford (2010) calculated that Bali and Sumbawa Islands have a large range in volcano spacing of 18–118 km with an average of 68 km and standard deviation of 51 km. The average volcano spacing in the east of Banda Arc is around 72 km with standard deviation of 29 km. In the Flores Zone, the average volcano spacing is about 21 km,

with a range of 6–61 km and standard deviation of 15 km. The variation in volcano spacing may reflect differences in the geometry of magma source and therefore provide information about mantle wedge fertility beneath the study area (de Bremond d'Ars et al. 1995; Baker et al. 2008; Ely and Sandiford 2010). As the mantle wedge production is mainly induced by volatile fluxes from the subducted slab (e.g. Cagnioncle et al. 2007), thus the differences in volcano spacing might reflect variations in the interaction between the mantle wedge and subducting slab (Ely and Sandiford 2010). In this condition, Sunda–Banda arc transition is not only marked by the change in tectonic regime, but also by the change in the mantle wedge fertility corresponding to the boundary between the volcanically active area in Flores and the extinct volcanic area in Wetar (east of Alor). As the mantle flow might affect to the orientation of the fast polarisations (e.g. Greve et al. 2008; Baccheschi et al. 2007), thus we interpret that the rotation of fast directions at depth below 150 km from around NNW–SSE in the west of Flores Island to NE–SW or parallel to the plate motion in the east near Wetar Island is likely to reflect this volcanic boundary zone.

Fig. 16 Maps of spatially averaged fast shear wave directions computed for two earthquake groups: shallow (depth ≤ 100 km) and deep (depth > 100 km) levels. Yellow colour bars mark the spatial average of ϕ with standard error ≤ 25°, and the red bars denote the standard error of spatial averaged ϕ > 25°
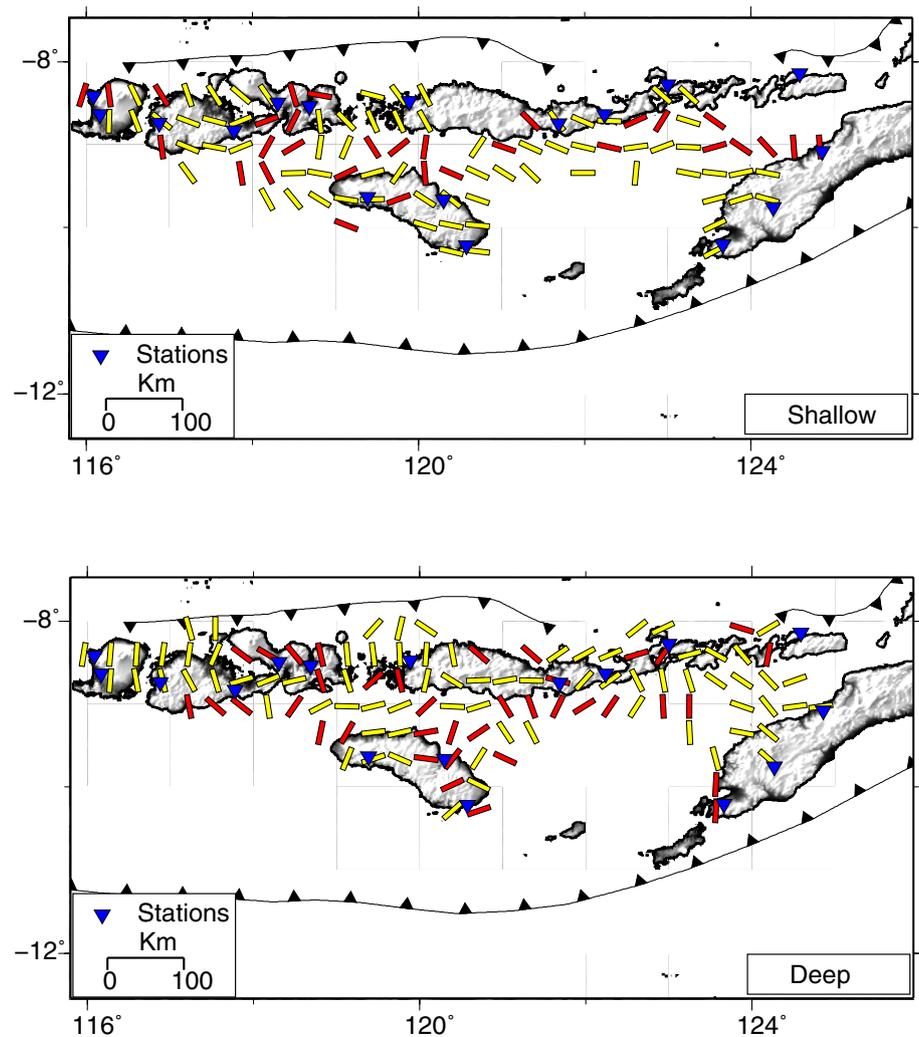
Sumba and timor zone: heterogeneous anisotropic structure

The change in δt and spatially varying ϕ with depth resulted from seismic broadband stations situated in Sumba and Timor Islands imply localised complex sub-crustal anisotropic structure. In Sumba Island, the fast polarisation directions vary from nearly E–W in the east, NE–SW in the middle and back to E–W in the west for earthquake sources above 100 km, while for events below 100 km consistently yield fast directions parallel to the absolute plate motion. We suggest that anisotropic source in this area, first, is likely as a result of the interaction between the lower part of crustal root beneath Sumba with the promontory Australian continental crust (Fleury et al. 2009). This interaction might

generate frictional process at the plate boundary and create heterogeneous anisotropic structure, for instance, the weak or fracture area in the lower crust which might have variation fast polarisation direction in each seismic station. This might imply that the geological condition in Sumba Island is more heterogeneous and complex. Alternatively, if the variation in fast polarisation direction is due to anisotropic sources originating from the upper mantle, this is probably caused by fossil anisotropy frozen into the lithosphere stemming from counter clockwise rotation since the late Cretaceous (Wensink 1994).

In Timor Island, the directions of fast shear waves also spatially vary. At stations BATI and ATNI, the fast orientations are consistently parallel to the plate motion for both earthquake sources. Conversely, at station SOEI, the fast

polarisations are consistently perpendicular to the absolute plate motion for both event groups. These consistent patterns imply that anisotropy is probably controlled by localised shallow anisotropic structure. Furthermore, the high percent anisotropy found in this island indicates the presence of strong anisotropic materials beneath the island. The geological structure of this island is very complex consisted of a wide range of rocks such as a variation of metamorphic and igneous rocks including high pressure types, continental crystalline fragments and ophiolites (Barber et al. 1977; Barber 1978; Charlton 1989), which could be brought to the surface by overthrusting of mantle rocks as a result of continental-arc collision (Kaye 1989). A recent receiver function investigation (Syuhada et al. 2016) found that the deeper part of crust beneath the island is marked by the existence of low-velocity zone and structure with relatively high Vp/Vs ratios. This study suggests that the crust in this island could be dominated by serpentinised rocks deformed during collision between Australian passive margin and volcanic arc. The previous geological study also found the presence of spinel peridotites partially recrystallised to tremolite-talc and serpentine schist, and mafic rocks in West Timor (Helmers et al. 1989). Therefore, we suggest that anisotropy beneath this island is mainly influenced by preferred orientation of high anisotropic materials such as spinel peridotites and serpentine schist.

Maps of spatially averaged data

In the previous section, our result displays inconsistent pattern of splitting parameters at some closely seismic stations suggesting that there may be lateral variation anisotropy around the region. This spatial variation causes difficulty in the interpretation. Thus, using 2D spatial average δt and ϕ method may help in estimating the average anisotropy of heterogeneous anisotropic structure. Figure 15 shows 2D spatial δt distribution maps for two depth earthquake groups (depth $\leq$ 100 km and depth > 100 km). These maps provide the variation of strength anisotropy indicated by warm colours (red to yellowish green) for strong anisotropy and the blue one for weak anisotropy. The shaded area corresponds

to the region with poor resolution or large residual (see Appendix Fig. 20).

The inversion results in the region with high resolution displays several interesting features which are consistent with the splitting interpretation in the previous section. For shallow strength anisotropic profile, there is a region of high anisotropy to the north of Sumba Island, which reasonably agree with the lithospheric boundary between an intra-oceanic lithosphere in the west and an arc-continent lithosphere in the east as proposed by Audley-Charles (1975) and Nishimura and Suparka (1986). There is a region of high anisotropy beneath Sumba Island, which corresponds to the area that has crustal thickening due to the insertion of Australian continental lithosphere underneath southern part of the Banda arc (Fleury et al. 2009). Thus, these deformation processes due to continent-arc collision could be the main mechanisms that control anisotropic sources for the shallow part of this region. For the deep anisotropic profile, the most striking feature is the high anisotropic region located at longitude 124° to the north of Timor, which coincides quite well with the region that has slab detachment as suggested by (McCaffrey et al. 1985; Ely and Sandiford 2010). However, for this profile, it is still unclear how much the contribution of past tectonic deformation and present-day mantle and lithospheric dynamic processes to the anisotropic mechanisms around this area.

The weighted average of ϕ obtained for each grid block of shallow and deep splitting measurements is shown in Fig. 16. For shallow event data, we observe that the direction of ϕ spatial averages with standard error less than 25° is generally perpendicular or sub-perpendicular to the absolute plate motion (NE–SW). As discussed earlier in the previous section that there are some possible causes for the fast direction perpendicular to the absolute plate motion. For Lombok and Sumbawa zone located in the subduction domain, this pattern could be due to the change of deformation of olivine aggregates (Karato et al. 2008), shape preferred orientation (SPO) of the partial melting (Fischer et al. 2000). For other zones located in the collision domain, this anisotropic trend could be mainly controlled by the lithospheric deformation process due to the continent-arc collision (e.g. Audley-Charles 1975; Fleury et al. 2009; Kaye 1989). For deeper

events, spatial averages of ϕ show consistent pattern with the plate motion direction for the zone in the subduction domain as expected. This pattern suggests that anisotropy in this level could be mainly caused by the simple lattice preferred orientation (LPO) mechanism coming from the mantle material. This mechanism produces an olivine a-axis alignment with the maximum shear direction and hence the fast direction would be roughly parallel to the absolute plate motion (Long 2013). Similar pattern has been reported for anisotropic studies in many subduction zones, for example, Tonga subduction zone (Fisher and Wiens 1996), Izu-Bonin (Fouch and Fischer 1996) and Kurils Islands (Fischer and Yang 1994). Meanwhile towards the region under influence of collision domain, the spatial averages of fast directions are more complex. The most interesting feature of this result is the rotation pattern of spatially averaged fast azimuths found at longitude 123°–124°. The pattern is probably associated with the induction of mantle flow due to lateral tearing of the slab as inferred from other geophysical studies (e.g. McCaffrey et al. 1985; Ely and Sandiford 2010). Furthermore, the complex pattern observed around the boundary of transition subduction-collision zone is also consistent with the tomographic delay time result and the previous anisotropy measurements (e.g. Di Leo et al. 2012) as already discussed in the previous section.

Conclusions

Here, we investigate lithospheric anisotropy along the Sunda–Banda arc transition using shear wave splitting method derived from local earthquake data recorded from seventeen permanent seismic stations. This study provides new insight into lithospheric deformation due to the change in tectonic regime from subduction to collision and its implication to the geodynamics in this study area. The results reveal the presence of complex anisotropic patterns that might be related to the change in tectonic regime. When the results are analysed via 2D delay-time tomography and spatial averaging of ϕ method, the systematic

and consistent patterns related to deformation and dynamic process around the region are more obvious. Therefore, we may draw some conclusions showing various geodynamical processes beneath the region due to the change in tectonic setting. First, seismic stations in the subduction domain not only show vertically consistent pattern but also spatially varying anisotropy. The stations in this area also exhibit two distinct fast directions: parallel and perpendicular to the plate motion. The results suggest that anisotropy around this area is not only dominated by simple type-olivine fabric in the mantle flow, but also deformed type-olivine fabric under influence of low temperature, high pressure and high water content. In addition, the anisotropic source in this area might also originate from the strong anisotropic structure in shallow depth through mechanism of shape-preferred orientation related to the upwelling of partial melting. Second, the complex anisotropy characterised by spatial and vertical variations of δt and ϕ is observed at stations located in collision domain. These complexities in splitting parameters observed could be either associated to the fossil fabrics from past tectonic deformation, the slab tear allowing mantle to flow parallel to the strike of the slab, or the distribution of the mantle wedge fertility and heterogeneity of local geological structure as result of collision process.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix

See Figs. 17, 18, 19, 20.

Fig. 17 The maps show the rays and the grid blocks used for the analysis. Shaded boxes represent the boxes traversed by fewer than 6 rays and are not used in the estimation of average fast azimuth
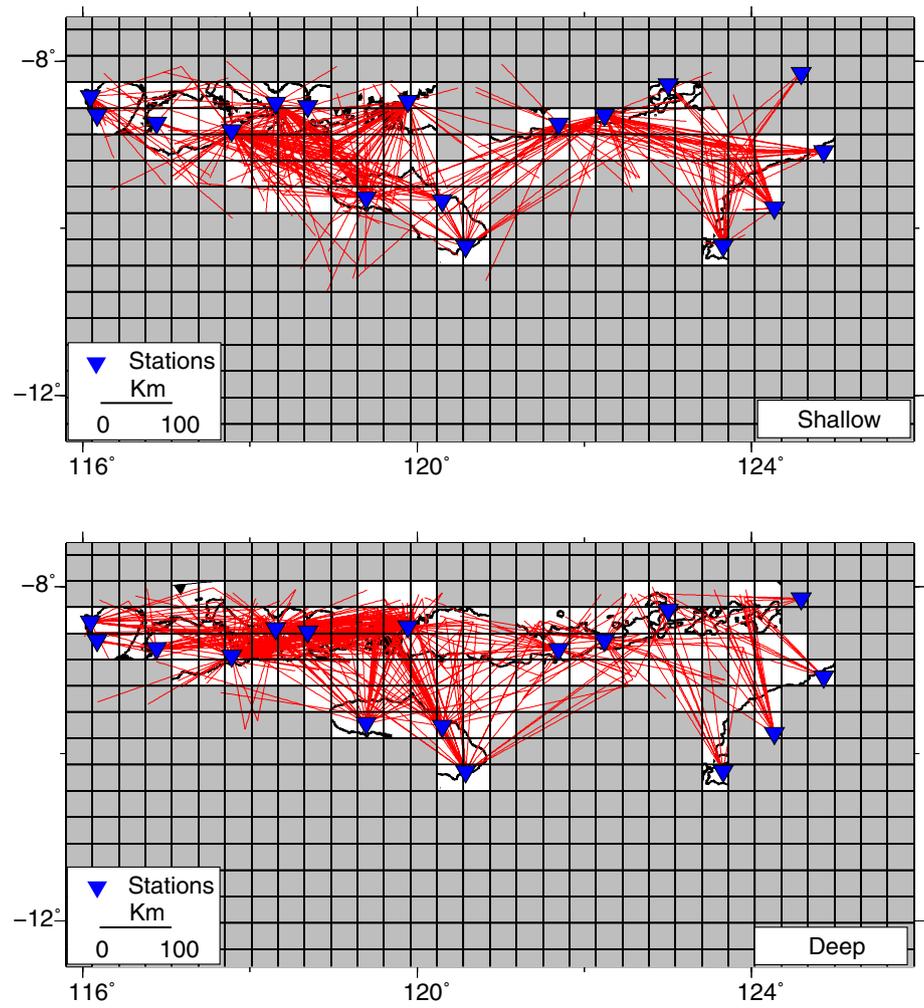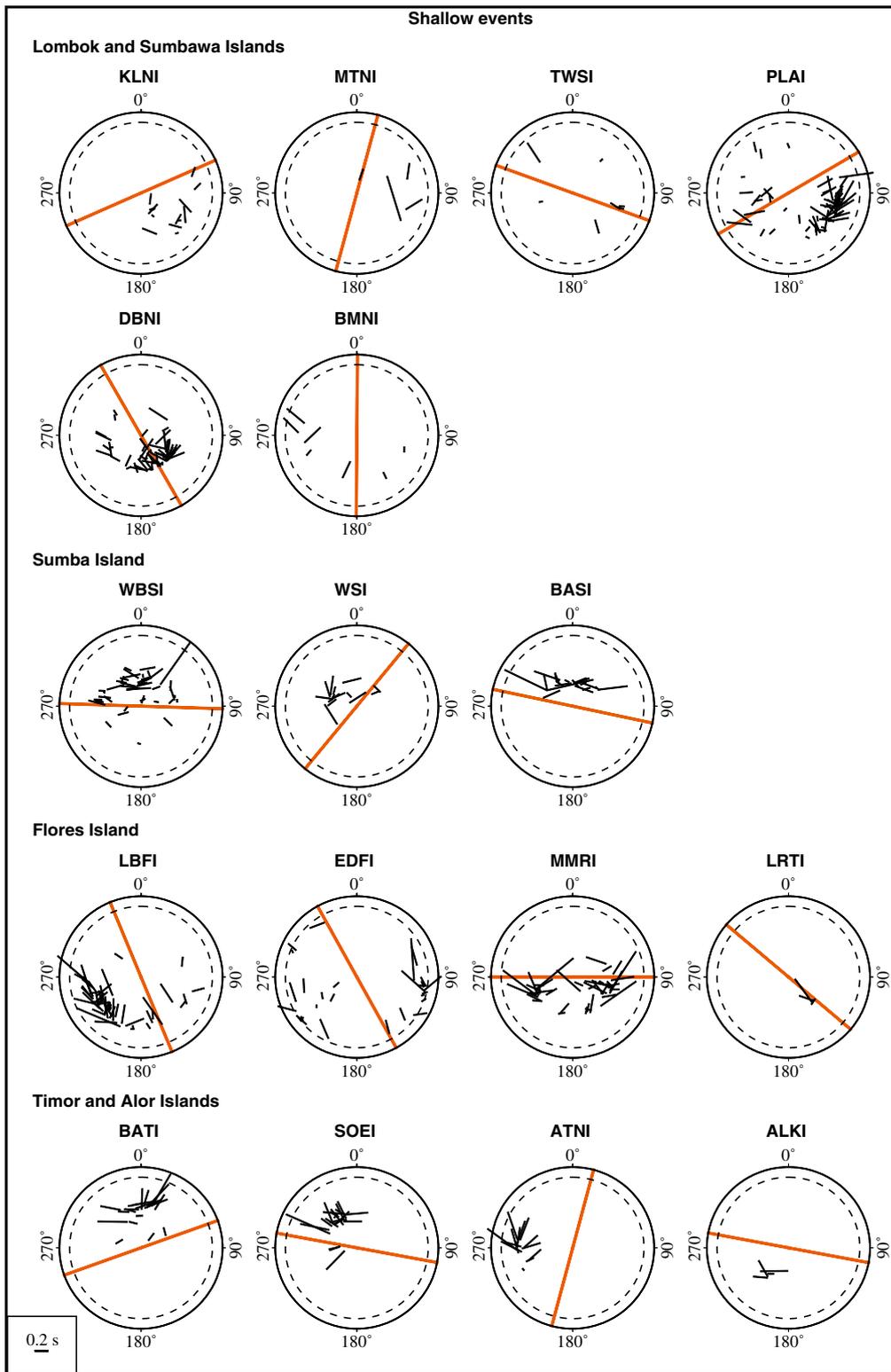
Fig. 18 Equal area plots showing splitting parameters as function of backazimuth and incidence angle for all seismic stations using shallow-event data. The radial position of the plots represents the incidence angle ranging from 0° to 40° and the dashed line represents 35° incidence angle. The red lines denote the mean fast direction. The bar direction represents the fast polarisation and the length of the bar is proportional to the delay time (δt)
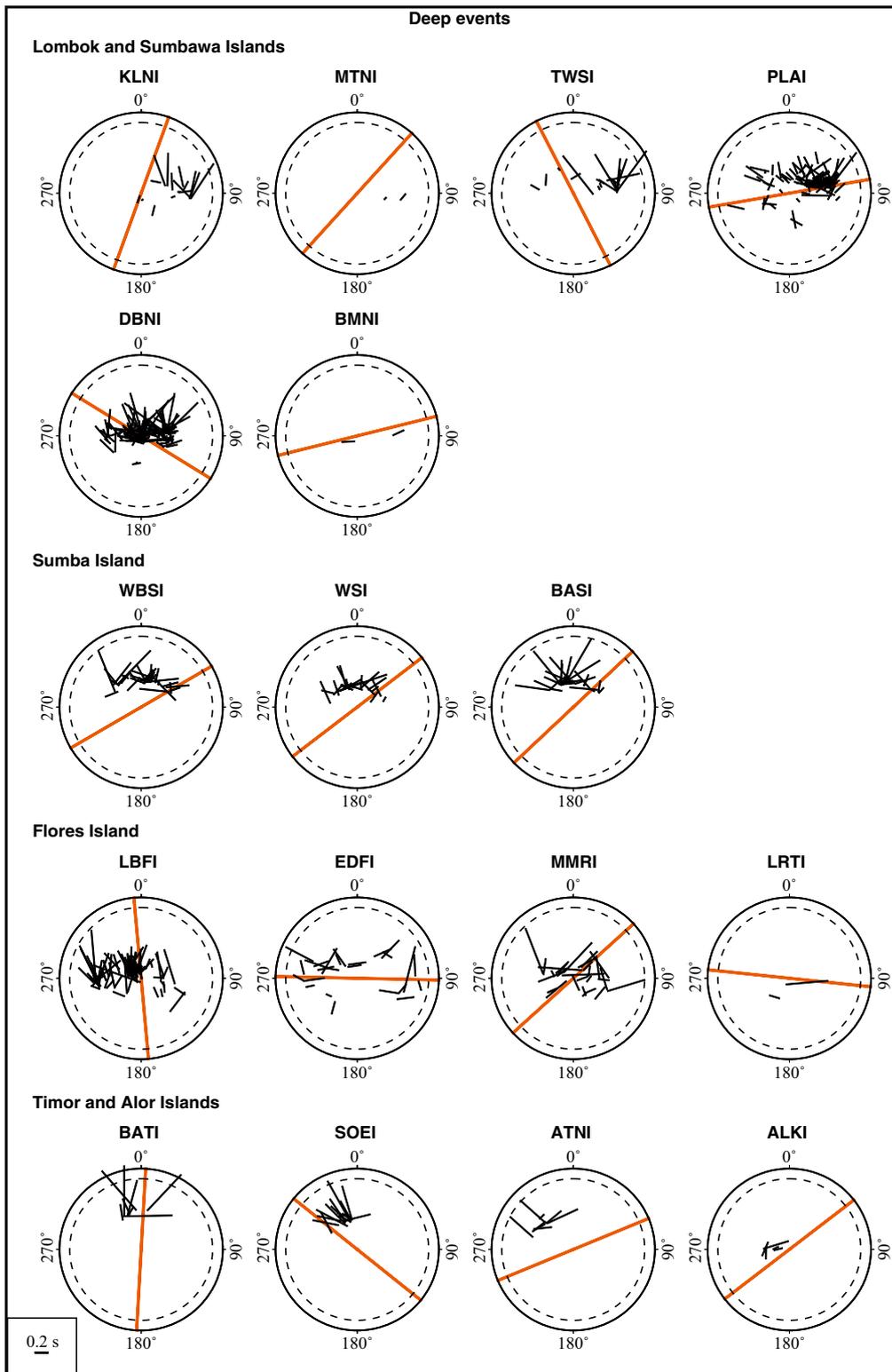
Fig. 19 Equal area plots showing splitting parameters as function of backazimuth and incidence angle for all seismic stations using deep-event data. The radial position of the plots represents the incidence angle ranging from 0° to 40° and the dashed line represents 35° inci-dence angle. The red lines denote the mean fast direction. The bar direction represents the fast polarisation and the length of the bar is proportional to the delay time (δt)

Fig. 20 Resolution maps obtained from the inversion of delay time. The contour line (grey line) corresponds to the estimated limit of resolution chosen based on the minimum number of the raypaths intersecting the blocks
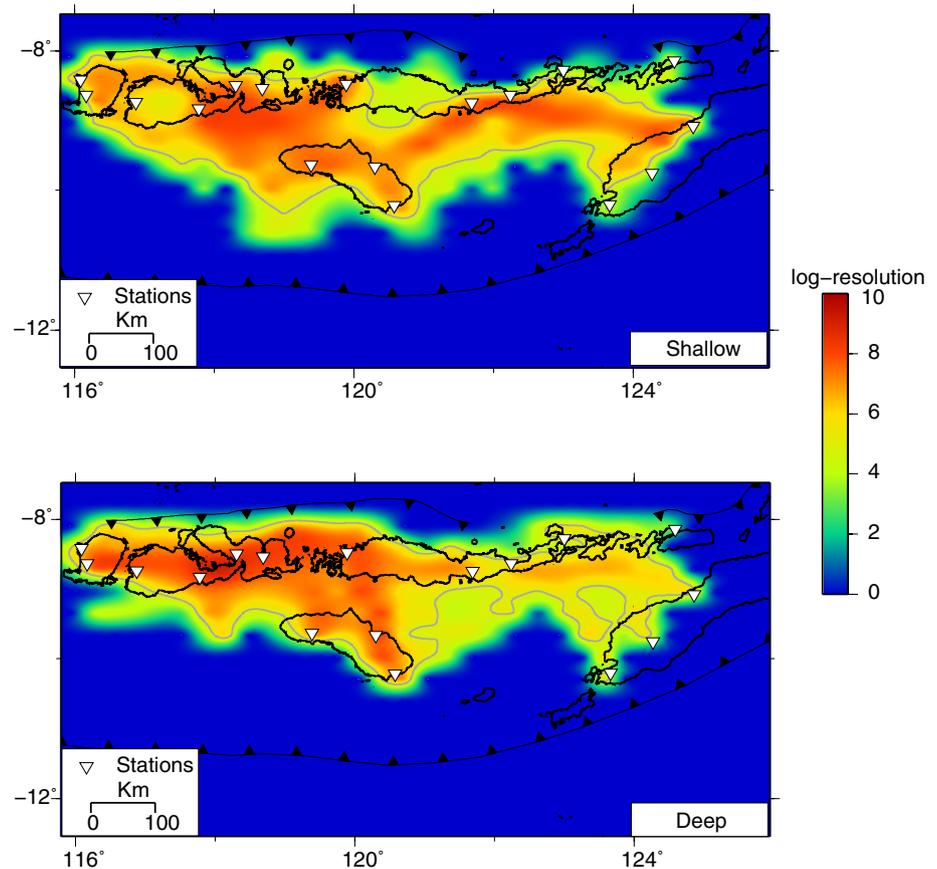
References

Audley-Charles MG (1975) The Sumba fracture: a major discontinuity between Eastern and Western Indonesia. Tectonophysics 26:213–228. https://doi.org/10.1016/0040-1951(75)90091-8

Audoine E, Savage MK, Gledhill K (2000) Seismic anisotropy from local earthquakes in the transition region from a subduction to a strike-slip plate boundary, New Zealand. J Geophys Res Solid Earth 105:8013–8033. https://doi.org/10.1029/1999JB900444

Baccheschi P, Margheriti L, Steckler MS (2007) Seismic anisotropy reveals focused mantle flow around the Calabrian slab (Southern Italy). Geophys Res Lett 34:L05302. https://doi.org/10.1029/2006GL028899

Baker ET, Embley RW, Walker SL, Resing JA, Lupton JE, Nakamura K, de Ronde CEJ, Massoth GJ (2008) Hydrothermal activity and volcano distribution along the Mariana arc. J Geophys Res Solid Earth 113:B08S09. https://doi.org/10.1029/2007jb005423

Barber AJ (1978) Structural interpretations of the island of Timor, eastern Indonesia. Proc Southeast Asia Pet Soc 4:9–21

Barber AJ, Audley-Charles MG, Carter DJ (1977) Thrust tectonics in Timor. J Geol Soc Aust 24:51–62. https://doi.org/10.1080/00167617708728966

Bock Y, Prawirodirdjo L, Genrich JF, Stevens CW, McCaffrey R, Subarya C, Puntodewo SSO, Calais E (2003) Crustal motion in Indonesia from global positioning system measurements. J Geophys Res 108(B8):2367. https://doi.org/10.1029/2001JB000324

Cagnioncle AM, Parmentier EM, Elkins-Tanton LT (2007) Effect of solid flow above a subducting slab on water distribution and melting at convergent plate boundaries. J Geophys Res 112(B9):B09402. https://doi.org/10.1029/2007jb004934

Charlton TR (1989) Stratigraphic correlation across an arc–continent collision zone: Timor and the Australian northwest shelf. Aust J Earth Sci 36:263–274. https://doi.org/10.1080/08120099890 8729485

Collings R, Rietbrock A, Tilmann F, Lange D, Nippres S, Natawidjaja D (2013) Seismic anisotropy in the Sumatra subduction zone. J Geophys Res 118:5372–5390. https://doi.org/10.1002/jgrb.50157

Crampin S (1991) Wave propagation through fluid-filled inclusions of various shapes: interpretation of extensive-dilatancy anisotropy. Geophys J Int 107:611–623. https://doi.org/10.1111/j.1365-246X.1991.tb05705.x

Crampin S (1994) The fracture criticality of crustal rocks. Geophys J Int 118:428–438. https://doi.org/10.1111/j.1365-246X.1994.tb03974.x

Curray JR (1989) The Sunda arc: a for oblique plate convergence. Neth J Sea Res 24:131–140. https://doi.org/10.1016/0077-7579(89)90144-0

de Bremond DJ, Jaupart C, Sparks RSJ (1995) Distribution of volcanoes in active margins. J Geophys Res 100:20421–20432. https://doi.org/10.1029/95JB02153

Di Leo JF, Wookey J, Hammond JOS, Kendall JM, Kaneshima S, Inoue H, Yamashina T, Harjadi P (2012) Mantle flow in regions of complex tectonics: insights from Indonesia. Geochem Geophys Geosyst 13:Q12008. https://doi.org/10.1029/2012GC004417

Ely KS, Sandiford M (2010) Seismic response to slab rupture and variation in lithospheric structure beneath the Savu Sea.

Tectonophysics 483:112–124. https://doi.org/10.1016/j.tecto.2009.08.027

Fischer KM, Yang X (1994) Anisotropy in Kuril-Kamchatka subduction zone structure. Geophys Res Lett 21:5–8. https://doi.org/10.1029/93GL03161

Fischer KM, Wiens DA (1996) The depth distribution of mantle anisotropy beneath the Tonga subduction zone. Earth Planet Sci Lett 142:253–260. https://doi.org/10.1016/0012-821X(96)00084-2

Fischer KM, Parmentier EM, Stine AR, Wolf ER (2000) Modeling anisotropy and plate-driven flow in the Tonga subduction zone back arc. J Geophys Res-Solid Earth 105:16181–16191. https://doi.org/10.1029/1999JB900441

Fleury JM, Pubellier M, Urreiztieta M (2009) Structural expression of forearc crust uplift due to subducting asperity. Lithosphere 113:318–330. https://doi.org/10.1016/j.lithos.2009.07.007

Fouch MJ, Fischer KM (1996) Mantle anisotropy beneath northwest Pacific subduction zones. J Geophys Res 101(15):987–16002. https://doi.org/10.1029/96JB00881

GEOFON data centre (1993) GEOFON seismic network. Deutsches GeoForschungsZentrum GFZ. Other/Seismic Network. https://doi.org/10.14470/TR560404

Gerst A, Savage MK (2004) Seismic anisotropy beneath Ruapehu volcano: a possible eruption forecasting tool. Science 306(5701):1543–1547

Gledhill K, Stuart G (1996) Seismic anisotropy in the fore-arc region of the Hikurangi subduction zone, New Zealand. Phys Earth Planet Inter 95:211–225. https://doi.org/10.1016/0031-9201(95)03117-0

Greve SM, Savage MK, Hofmann SD (2008) Strong variations in seismic anisotropy across the Hikurangi subduction zone, North Island, New Zealand. Tectonophysics 462:7–21. https://doi.org/10.1016/j.tecto.2007.07.011

Hall R (2002) Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer-based reconstructions, model and animations. J Asian Earth Sci 20:353–434. https://doi.org/10.1016/S1367-9120(01)00069-4

Hammond JOS, Wookey J, Kaneshima S, Inoue H, Yamashina T, Harjadi P (2010) Systematic variation in anisotropy beneath the mantle wedge in the Java-Sumatra subduction system from shear-wave splitting. Phys Earth Planet Inter 178:189–201. https://doi.org/10.1016/j.pepi.2009.10.003

Helmers J, Sopaheluwaken J, Tjokrosapoetro S, Nila ES (1989) High grade metamorphism related to peridotite emplacement near Atapupu, Timor, with reference to the Kaibobo peridotite on Seram, Indonesia. Neth J Sea Res 24:357–371. https://doi.org/10.1016/0077-7579(89)90161-0

Hiramatsu Y, Ando M, Ishikawa Y (1997) ScS wave splitting of deep earthquakes around Japan. Geophys J Int 128:409–424. https://doi.org/10.1111/j.1365-246X.1997.tb01564.x

Johnson JH, Savage MK, Townend J (2011) Distinguishing between stress-induced and structural anisotropy at mount ruapehu volcano, New Zealand. J Geophys Res Solid Earth 116(B12):1978–2012. https://doi.org/10.1029/2011JB008308

Jung H, Karato S (2001) Water-induced fabric transitions in olivine. Science 293:1460–1463. https://doi.org/10.1126/science.1062235

Kaneshima S, Silver PG (1995) Anisotropic loci in the mantle beneath central Peru. Phys Earth Planet Inter 88:257–272. https://doi.org/10.1016/0031-9201(94)02981-G

Karalliyadda SC, Savage MK (2013) Seismic anisotropy and lithospheric deformation of the plate-boundary zone in South Island, New Zealand: inferences from local S-wave splitting. Geophys J Int 193:507–530. https://doi.org/10.1093/gji/ggt022

Karato S (2004) Mapping water content in the upper mantle. In: Eiler J (ed) Inside the subduction factory. American Geophysical Union, Washington, D.C

Karato S, Jung H, Katayama I, Skemer P (2008) Geodynamic significance of seismic anisotropy of the upper mantle: new insights from laboratory studies. Annu Rev Earth Plant Sci 36:59–95. https://doi.org/10.1146/Annurev.Earth.36.031207.124120

Kaye SJ (1989) The structure of Eastern Indonesia: an approach via Gravity and other geophysical methods. Ph. D. Thesis, University of London

Long MD (2013) Constraints on subduction geodynamics from seismic anisotropy. Rev Geophys 51:76–112. https://doi.org/10.1002/rog.20008

Long MD, Silver PG (2008) The subduction zone flow field from seismic anisotropy: a global view. Science 319:315–318. https://doi.org/10.1126/science.1150809

Long MD, Silver PG (2009) Shear wave splitting and mantle anisotropy: measurements, interpretations, and new directions. Surv Geophys 30:407–461. https://doi.org/10.1007/s10712-009-9075-1

Long MD, van der Hilst RD (2005) Upper mantle anisotropy beneath Japan from shear wave splitting. Phys Earth Planet Inter 151:206–222. https://doi.org/10.1016/j.pepi.2005.03.003

Long MD, van der Hilst RD (2006) Shear wave splitting from local events beneath the Ryukyu arc: Trench-parallel anisotropy in the mantle wedge. Phys Earth Planet Inter 155:300–312. https://doi.org/10.1016/j.pepi.2006.01.003

Lueschen E, Mueller C, Kopp H, Engels M, Lutz R, Planert L, Shulgin A, Djajadihardja Y (2011) Structure, evolution and tectonic activity at the Eastern Sunda forearc, Indonesia, from marine seismic investigations. Tectonophysics 508:1–4. https://doi.org/10.1016/j.tecto.2010.06.008

Mardia KV (1972) Statistics of directional data. Academic Press, London and New York

McCaffrey R, Molnar P, Roecker SW (1985) Microearthquake seismicity and fault plane solutions related to arc-continent collision in the Eastern Sunda arc, Indonesia. J Geophys Res 90:4511–4528. https://doi.org/10.1029/JB090iB06p04511

Miller MS, Allam A, Becker TW, Di Leo JF, Wookey J (2013) Constraints on the tectonic evolution of the westernmost Mediterranean and northwestern Africa from shear wave splitting analysis. Earth Planet Sci Lett 375:234–243. https://doi.org/10.1016/j.epsl.2013.05.036

Nishimura S, Suparka S (1986) Tectonic development of East Indonesia. J Southeast Asian Earth Sci 1:45–57. https://doi.org/10.1016/0743-9547(86)90006-1

Nugroho H, Harris RA, Amin WL, Bilal M (2009) Plate boundary reorganization in the active Banda Arc–continent collision: insights from new GPS measurements. Tectonophysics 479:52–65. https://doi.org/10.1016/j.tecto.2009.01.026

Nugraha AD, Ash-Shiddiqi H, Widiyantoro S, Ramdhan M, Wandono S, Handayani T, Nugroho H (2015) Preliminary results of teleseismic double-difference relocation of earthquakes around Indonesian archipelago region. AIP Conf Proc. https://doi.org/10.1063/1.4915010

Okada T, Matsuzawa T, Hasegawa A (1995) Shear-wave polarisation anisotropy beneath the north-eastern part of Honshu, Japan. Geophys J Int 123:781–797. https://doi.org/10.1111/j.1365-246X.1995.tb06890.x

Prawirodirdjo L, Bock Y (2004) Instantaneous global plate motion model from 12 years of continuous GPS observations. J Geophys Res 109:B08405. https://doi.org/10.1029/2003JB002944

Rumpker G, Silver PG (1998) Apparent shear–wave splitting parameters in the presence of vertically varying anisotropy. Geophys J Int 135:790–800. https://doi.org/10.1046/j.1365-246X.1998.00660.x

Savage MK (1999) Seismic anisotropy and mantle deformation: What have we learned from shear wave splitting? Rev Geophys 37:65–106. https://doi.org/10.1029/98RG02075

Savage MK, Peppin WA, Vetter UR (1990) Shear wave anisotropy and stress direction in and near long valley Caldera, California,

1979–1988. J Geophys Res 95:11165–11177. https://doi.org/10.1029/JB095iB07p11165

Savage MK, Wessel A, Teanby NA, Hurst AW (2010) Automatic measurement of shear wave splitting and applications to time varying anisotropy at Mount Ruapehu volcano, New Zealand. J Geophys Res 115:B12321. https://doi.org/10.1029/2010JB007722

Shih XR, Meyer RP, Schneider JF (1991) Seismic anisotropy above a subducting plate. Geology 19:807–810. https://doi.org/10.1130/0091-7613(1991)019<0807:SAAASP>2.3.CO;2

Shulgin A, Kopp H, Mueller C, Lueschen E, Planert L, Engels M, Flueh ER, Krabbenhoeft A, Djajadihardja Y (2009) Sunda–Banda arc transition: incipient continent–island arc collision (northwest Australia). Geophys Res Lett 36:L10304. https://doi.org/10.1029/2009GL037533

Silver PG (1996) Seismic anisotropy beneath the continents: probing the depths of geology. Annu Rev Earth Planet Sci 24:385–432. https://doi.org/10.1146/annurev.earth.24.1.385

Silver P, Chan G (1991) Shear wave splitting and subcontinental mantle deformation. J Geophys Res 96:16429–16454. https://doi.org/10.1029/91JB00899

Silver PG, Savage MK (1994) The interpretation of shear-wave splitting parameters in the presence of two anisotropic layers. Geophys J Int 119:949–963. https://doi.org/10.1111/j.1365-246X.1994.tb04027.x

Smith WHF, Sandwell DT (1997) Global seafloor topography from satellite altimetry and ship depth soundings. Science 277:1956–1962. https://doi.org/10.1126/science.277.5334.1956

Syuhada S, Hananto ND, Abdullah CI, Puspito NT, Anggono T, Yudistira T (2016) Crustal structure along Sunda-Banda Arc transition zone from teleseismic receiver functions. Acta Geophys 64:2020–2050. https://doi.org/10.1515/acgeo-2015-0098

Syuhada S, Hananto ND, Abdullah CI, Puspito NT, Anggono T, Yudistira T, Ramdhan M (2017) Crustal Anisotropy along the Sunda–Banda Arc transition zone from shear wave splitting measurements. J Geodyn 103:1–11. https://doi.org/10.1016/j.jog.2016.10.006

Teanby NA, Kendall JM, van der Baan M (2004) Automation of shear-wave splitting measurements using cluster analysis. Bull Seismol Soc Am 94:453–463. https://doi.org/10.1785/0120030123

Vinnik LP, Kosarev GL, Makeyeva LI (1984) Anisotropy of the lithosphere from the observations of SKS and SKKS. Proc Acad Sci USSR 278:1335–1339

Wensink H (1994) Paleomagnetism of rocks from Sumba: tectonic implications since the late Cretaceous. J Southest Asian Earth Sci 9:51–65. https://doi.org/10.1016/0743-9547(94)90065-5

Wirth E, Long MD (2010) Frequency-dependent shear wave splitting beneath the Japan and Izu-Bonin subduction zones. Phys Earth Planet Inter 181:141–154. https://doi.org/10.1016/j.pepi.2010.05.006

Yang X, Fischer KM, Abers GA (1995) Seismic anisotropy beneath the Shumagin Islands segment of the Aleutian-Alaska subduction zone. J Geophys Res 100:18165–18177. https://doi.org/10.1029/95JB01425

Principal Slip Zone determination in the Wenchuan earthquake Fault Scientific Drilling project-hole 1: considering the Bayesian discriminant function

Sinan Fang[1,2] · Zhansong Zhang[1,2] · Zhi Wang[3] · Heping Pan[4] · Ting Du[2]

Abstract
Accurate determination of the Principal Slip Zone (PSZ) of earthquake fault zones is a key task of earthquake Fault Scientific Drilling for future earthquake control. The fault zone structure of Wenchuan earthquake is complex, and there are many strong earthquakes recorded on the fault zone, which make determining the PSZ in the Wenchuan earthquake Fault Scientific Drilling project-hole 1 (WFSD-1) difficult. At present, core analysis of whole coring is the decisive method for determining PSZ depth, and the fresh fault gouge at 589.2 m is the PSZ in WFSD-1. Abundant and comprehensive logging data can only be used as evidence to judge the PSZ. Based on the discrimination function and hyperplane equation in Bayesian discriminant classification, we derive a new algorithm for computing the PSZ possibility using a Bayesian Discrimination function (PSZP-BDF) based on the simplified model, and set up a mode to determine the PSZ directly using machine learning of well logging. For the verification of WFSD-1, the fault gouges are successfully identified and the PSZ depth is accurately located. The algorithm objectively learns the sample data, which is naturally adaptive to the region. The calculation procedure is simple and does not require expensive coring data or heavy core tests in the well. The calculation speed is fast, using multiple physical data types. The PSZP-BDF algorithm is suitable for processing and interpreting earthquake fault scientific drilling data.

Introduction

Weak dynamic seismic stress waves from distant earthquakes are proved to abruptly change the states of many natural geological systems include permeability(Xue et al. 2013; Weaver et al. 2019), liquefaction (Wang 2007), a

✉ Sinan Fang
fangsinan@163.com

1 Key Laboratory of Exploration Technologies for Oil and Gas Resources, Ministry of Education, Yangtze University, Wuhan, China

2 College of Geophysics and Petroleum Resources, Yangtze University, Wuhan, China

3 Electronics and Information School, Yangtze University, Jingzhou, China

4 Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan, China

sudden increase of seismicity and the dynamic triggering of earthquakes (Peng et al. 2011). Therefore, the theory that the transient seismic-wave pressure in a fluid-filled fracture could increase more than 2 orders of magnitude may be a key technology for artificially controlling earthquakes(Zheng 2018), which requires first predicting the Principal Slip Zone (PSZ) location for the next earthquake. In order to make the fracture more controllable, we will give priority to the PSZ of the last earthquake that is easier to slide like a natural earthquake. Here we prove a new method to identify the PSZ of Wenchuan earthquake Fault Scientific Drilling.

The Wenchuan earthquake (Mw 7.9) on May 12, 2008 killed more than 80,000 people, injured over 370,000 people, destroyed millions homes, and caused trillions of yuan in property losses. However, the Longmenshan fault zone where the Wenchuan earthquake occurred recorded no strong earthquake above Mw 7 since earthquake records, and the displacement obtained from geological surveying and GPS monitoring on the surface is small. A detailed

investigation of the Longmenshan fault zone, including extensive drilling of the fault zone, deep in situ observations, laboratory analysis of fault zone samples, and long-term monitoring is necessary to avoid future tragedies like the Wenchuan earthquake (Xu et al. 2018). Most scientific drilling of active fault zones is located in the circum Pacific seismic belt (Brodsky et al. 2009), while Wenchuan earthquake Fault Scientific Drilling (WFSD) is the first scientific drilling project of a seismic fault zone within the interior of the continent, which can provide insight into the long-term evolution of continental seismic faults (Fig. 1).

The Wenchuan earthquake was a complex earthquake that has never occurred in recorded history, in which the fracture generated along two fault zones simultaneously. The two fault zone were the Yingxiu-Beichuan fault, a dextral strike-slip thrust fault and the Guanxian-Anxian fault, a pure thrust fault, which respectively produced 270 km and 80 km of coseismic fractures within 90 s (Li et al. 2018). Six boreholes were drilled and well logged in the two fracture zones in WFSD. WFSD-1 began on November 6, 2008 and finished on July 12, 2009 at a depth of 1201.15 m and coring rate of 85% (main fault intervals are fully cored). The well is located in Hongkou, Dujiangyan City and is located in Pengguan complex south of Yingxiu-Beichuan fracture zone.

The Neoproterozoic Pengguan complex, which extends to a depth of 585 m in the WFSD-1 well, is primarily composed of igneous rock and pushed over the sedimentary rocks of the Upper Triassic Xujiahe formation (Fig. 2). The fault zone is continuously distributed at a depth of 585-759 m where the Yingxiu-Beichuan fault zone is undercutting, that is the target layer of WFSD. There are more than 105 m fault rocks in the multi-core model, indicating that there have been numerous strong earthquakes along the Yingxiu-Beichuan fault zone in the geologic past that caused the rapid uplift of Longmenshan Mountain (Li et al. 2013a, b).

For the study, the first task is to determine the location and depth of the main fault, then the slip mechanism, slip process, physical and chemical properties of the fault, and fluid behavior can be examined (Tanaka et al. 2001). Based on this, it is possible to ultimately control earthquakes. In general, the strain at the center of the main fault is the largest, and the representative products are fault gouge, pseudotachylites, and hypercataclasite (Woodcock and Mort 2008). Strain decreases away from the fault zone center, and the representative products are fault breccia and cataclasite. Fault gouge can form a good caprock for oil and gas reservoirs combined with shale because of its low permeability, while fault breccia and cataclasite contain a large number
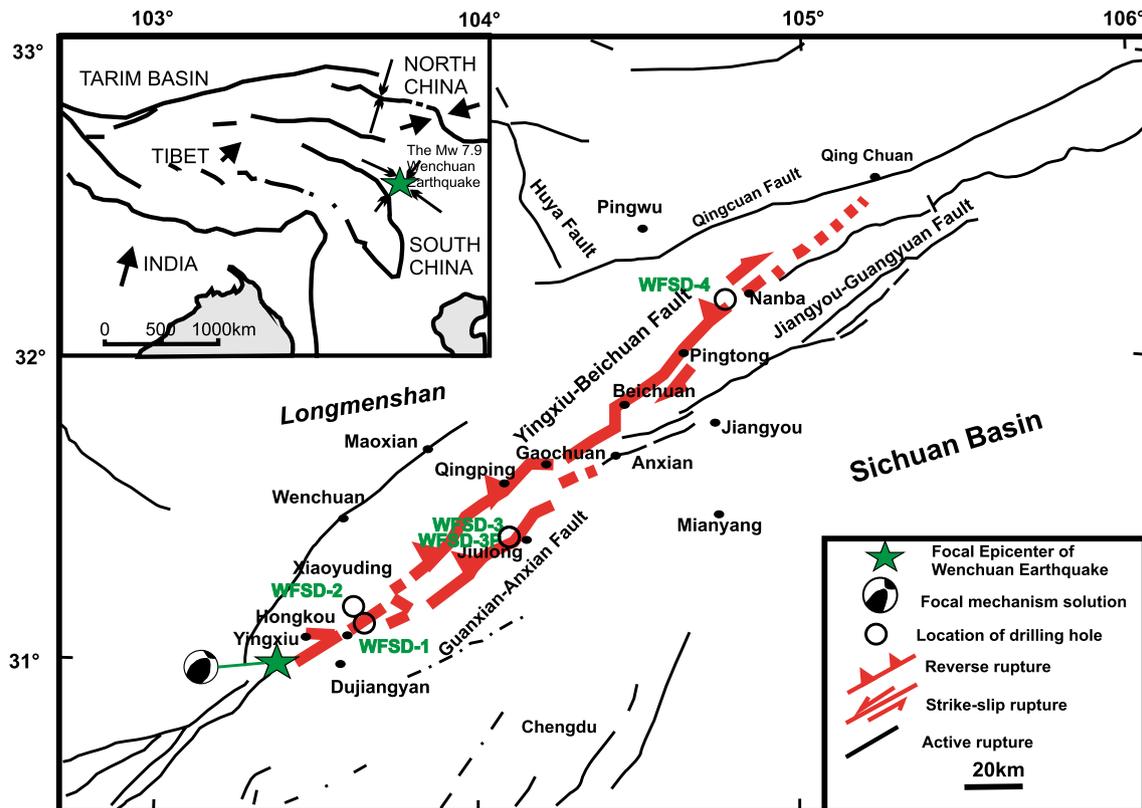

Fig. 1 Map of the Longmen Shan and western Sichuan basin, showing the epicenter of the 2008 Wenchuan earthquake as well as major faults and sites from the WFSD. Modified from Konaté et al. (2017)
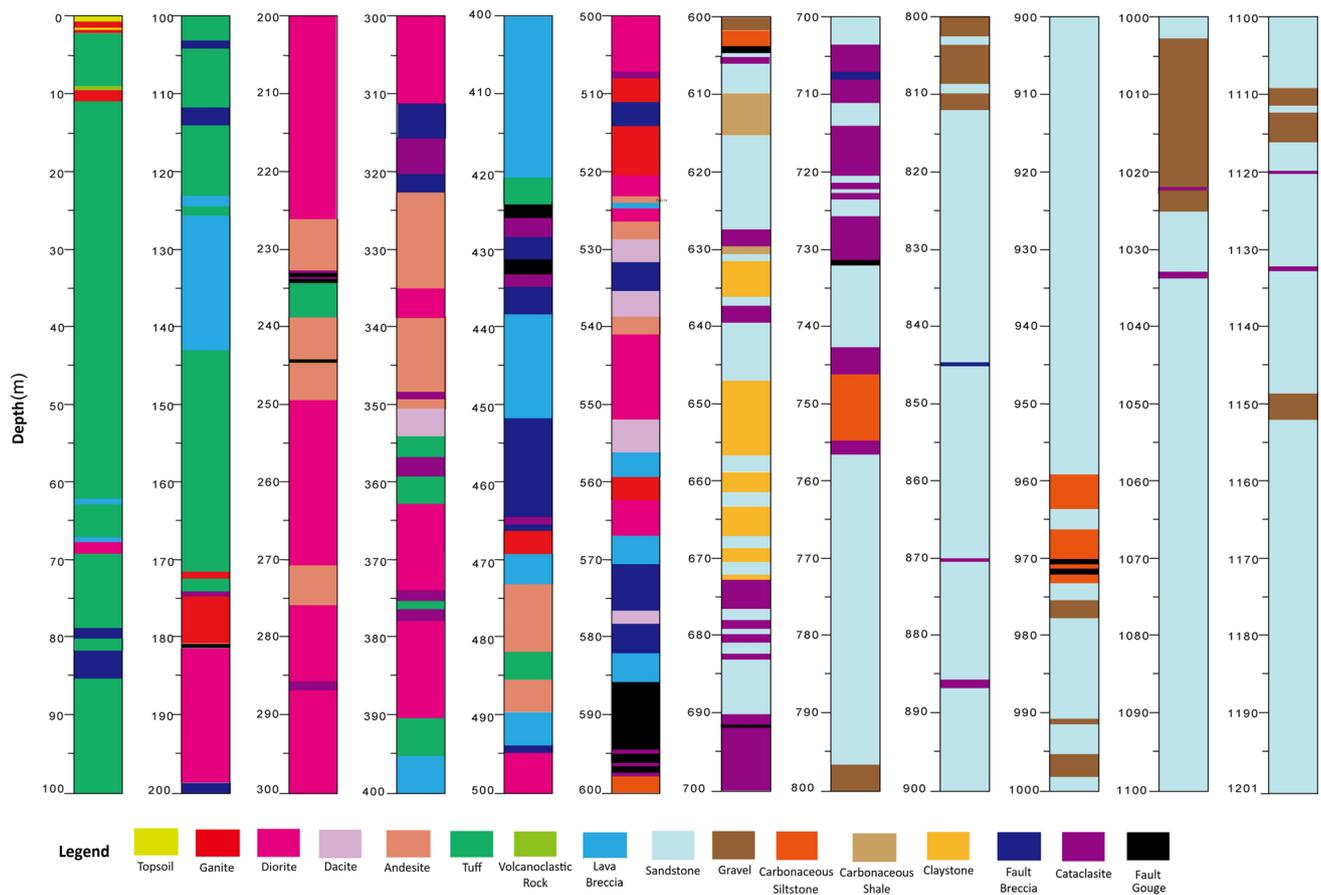
Fig. 2 Lithology of the WFSD-1

of fractures, which can form a range of migration channels. In the actual fault system, the spatial relationship among fault gouge, fault breccia, and cataclasite are very complex; hence, it is necessary to examine the lithologic distribution of the main fault and improve interpretations of the migration mechanism and caprock characteristics in oil and gas traps.

Cataclasite in the seismic zone primarily refers to rocks with a cataclastic structure, which forms after intensive crushing and deformation. Fault breccia is a kind of structural breccia produced by crushing, breaking, or shearing during fault movement. Fault gouge is unconsolidated or weakly consolidated argillaceous rock developed in the fault zone, primarily composed of clay formed from broken powder and broken gravel during shear sliding, breaking, grinding, and geochemical alteration reactions of the fault (Brodsky et al. 2009; Lacroix et al. 2015). Pseudotachylite, a unique product of earthquakes, is a rock formed on the fault surface due to rapid sliding friction, heat generation, and melting-solidification (Spray 1992; Lin et al. 2005). We can determine the primary fault of the earthquake by the degree of alteration or devitrification of pseudotachylite. In

addition, organic biomarkers (Rabinowitz et al. 2020) and absolute gouge dating (Tsakalos et al. 2020) could be used to identify recently active PSZs. However, locating an observable physical response is difficult because of the tiny scale (≤ 1 mm) and small amount of deformation in the cataclasite (Wang et al. 2018.). This contribution focuses on the Principal Slip Zone (PSZ), specifically on the identification and comparison of fault gouge in the center of the fault zone. Generally, fault gouge thickness ranges from several millimeters to tens of centimeters (Sibson 2003), and there are 12 obvious fault gouges in the WFSD-1 fault zone, with thickness ranging from several millimeters to seven meters, and colors ranging from light gray to dark black. The PSZ of the Wenchuan earthquake can be determined from these fault gouges.

General method of determining the PSZ

Determination of the PSZ commonly comes from analyzing core samples.. An obvious calcite vein appears in stable fault gouges formed by ancient earthquakes, but not in the

Wenchuan earthquake fault gouge. Mineral particles in the PSZ are small, dark, soft, unconsolidated, and have certain viscosity when encountering water (Hirose and Shimamoto 2005; Ma et al. 2006). In the 590 m fault zone in WFSD-1, there are at least 12 fault gouge layers with different characteristics, only one of which, the deep black fault gouge, looks the freshest. In addition, the magnetic susceptibility is relatively high, the core scanning photos do not show calcite fragments or veins, and the analysis of major, trace, and rare earth elements shows abnormal changes. Core analysis reveals that the core of the Wenchuan earthquake fault zone has a depth of 590 m, and the fresh fault gouge between 589.21 m and 589.22 m is the PSZ of WFSD-1 (Xu et al. 2018). Core analysis has the advantages of high accuracy, high reliability, and rich geological information, but it is also expensive and time consuming. With the rapid development of drilling and logging technology, we can extract the PSZ information from in situ logging data and apply it to future scientific drilling projects.

Due to frictional heat in the PSZ, a positive geothermal anomaly can be observed up to 6 years after an earthquake (Kano et al. 2006); however, the range of the anomalous geothermal gradient is commonly tens of meters, and the temperature difference is small. Therefore, when locating PSZs, only temperature logging is considered as a proof of identification. The high temperature action caused by friction during fault sliding leads to an increase in magnetic mineral content and magnetic anomalies in fault gouge (Hirono et al. 2008). The high magnetic susceptibility anomaly at the WFSD-1 PSZ has been confirmed in core samples (Pei et al. 2014), but there is no obvious change in the logging curve. Due to the fact that the logging data longitudinal sampling rate is not precise enough, and the abnormal magnetic susceptibility is covered by the noise from the surrounding rock, identify the PSZ from magnetic susceptibility log information is difficult. The abundant formation water in the cataclasite fractures increases the shale content in the fault gouge, gradually producing a higher GR (Konaté et al. 2017). However, the GR of the PSZ is primarily controlled by the shale content of the original rock, so GR is not suitable for locating the PSZ. As a product of clayization during fracturing, unconsolidated or weakly consolidated rock has large porosity and complex anisotropy, which should manifest as low acoustic velocity and low resistivity. Fault gouges in the PSZ have high natural gamma, high magnetic susceptibility, low resistivity, and low acoustic velocity (Zheng et al. 2016), however, there is no theoretical model or experimental law proving the quantitative relationship between logging data and fault gouges in the PSZ. Therefore, logging information can only be used for auxiliary positioning the PSZ.

However, as a high-resolution in situ monitoring method, well logging can provide an abundance information and truly restore the physical characteristics of fault gouge, given the

condition that many logging methods are responsive to fault gouge and the PSZ. The logging tool used at whole depths of WFSD-1 is Micro Logger II, a small diameter digital logging tool from the RG company in the UK. Due to poor borehole conditions, the quality of the logging curve is reduced at some depths. The effective logging curves on the target interval are: Caliper (CAL), Gamma Ray (GR), Spontaneous Potential (SP), Acoustic interval (AC), Compensated Neutron Log (CNL), Resistivity (Rt), and Magnetic Susceptibility (MS). Well logging series meet the requirements of the PSZ related well logging methods described above. Although the accuracy and reliability of the small caliber logging tool used in WFSD-1 is lower than that of the petroleum logging tool, and has not been calibrated strictly, a quantitative interpretation can still be made according to the relative changes in the logging curves.

According to the logging data from the WFSD-1 for different lithologies, we select four logging methods that can be used to identify fault gouges to generate box-whisker plots (Fig. 3). The fluctuation amplitude of fault gouge GR, RT, and AC is very large, showing the uncertainty and instability of the fault gouge. Fault gouge may have a similar mineral composition with cataclasite or shale, so we think the fault gouge GR, RT, and CNL are primarily distributed between cataclasite and shale. Over 75% of fault gouges have larger GR, AC, and CNL than the cataclasite and fault breccia. Over 75% of fault gouges have smaller RT than the cataclasite and fault breccias. These logging responses are consistent with the composition and mineralogy analysis (Konaté et al. 2017). Therefore, GR, AC, RT, and CNL can be used to distinguish fault gouge from cataclasite and fault breccia; however, excluding shale or cataclasite is difficult, which is consistent with the fact that fault gouge has high shale content and a similar mineral composition to shale or cataclasite.

The red frame represents the change range of logging data from the PSZ (Fig. 3). Compared with most fault gouges, the of Wenchuan earthquake PSZ from 2008 are all in the upper quartile, indicating that the probability of GR, AC, and CNL in the newly fractured fault is 75% higher than for other stable fault gouges. Also because they are in the upper quartile, GR, AC, and CNL cannot be used as the basis for the final PSZ determinization, which is why logging information is only used as partial evidence. It is exciting that the PSZ RT is the lowest among all fault gouges. Although this value is easily confused with shale, when fault gouges can be accurately identified, we think RT value can be an effective basis for identifying the PSZ.

To further prove this relationship, we usually achieve it by empirical formula of core experiment and physical simulation of theoretical model. In the absence of sufficient experimental calibration data, determining a complex dynamic relationship between RT and the PSZ is difficult
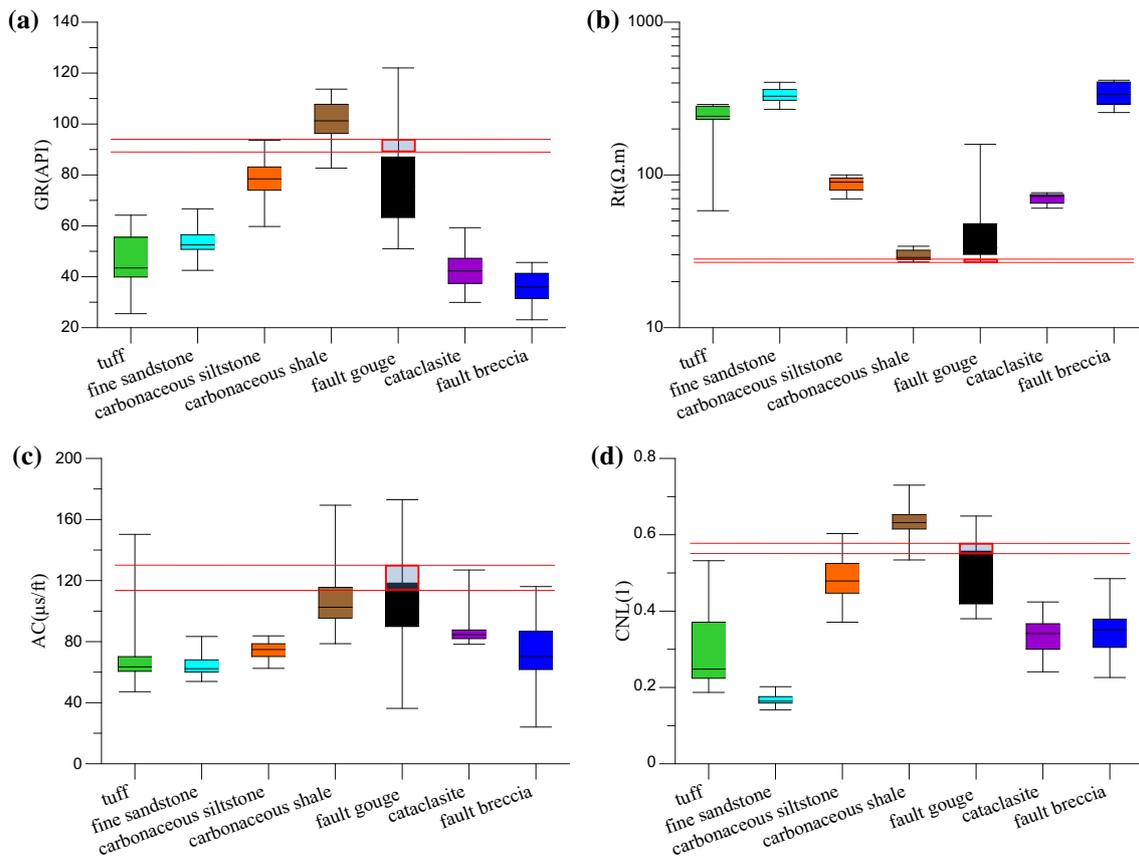
Fig. 3 Box-whisker plots of logging responses from the lithologies in WFSD-1, the red rectangles represent the logging value range of fault gouge

with changes in seismic intensity, fault healing process, matrix composition, formation mechanism of fault gouge, and resistivity of formation water as well as the less obvious relationship between GR, AC, CNL, and PSZ. In fact, many PSZ parameters that remain to be interpreted. The fault gouge core is loose, thin, and rare, so the items that can be tested in the laboratory are very limited, and it is unrealistic to prove the relationship through core analysis.

To establish the relationship between logging parameters and the PSZ, a theoretical simulation is needed; however, the formation and compaction mechanisms of fault gouges are complex and unstable, the source and mineral composition of fault gouges are uncertain, and the physical properties of thin-layer fault gouges are significantly affected by the surrounding rocks, which together lead to significant uncertainty in the theoretical logging response. Therefore, future research on the PSZ should establish a huge complex multi-physical field model, which is very difficult to accomplish. By understanding the correlation between single logging parameters and formation characteristics, we believe that by using machine learning and a comprehensive analysis of multi-physical field information

from logging data, the multi-solution problem of solving the objective function can be effectively reduced, making accurately locating the PSZ possible.

The theory and methods of machine learning have been widely used for data mining in engineering applications and scientific fields. Machine learning methods use multiple types of logging data simultaneously in formation identification, realize the division of high-dimensional logging space better than a histogram (one-dimensional space division) or crossplot (two-dimensional space division) in traditional logging interpretation, making machine learning more suitable for fault gouge with multiple logging response characteristics. Machine learning methods reduce the average relative error based on the sample experience and discriminate formations from a purely mathematical perspective to check for gaps, which is more objective than manual logging interpretation. Based on these advantages, well logging combined with machine learning methods are capable of identifying the lithologies where there was no core recovery and mainly extrapolated through experience, and even finding the possible missing fault gouge.

In geoscience data mining, the primary machine learning tools are artificial neural networks, support vector machines, and Bayesian discrimination. Support vector machines are suitable for strongly nonlinear binary classifications. The discrimination process of artificial neural networks is significantly affected by the final samples; hence, the discrimination quality is dependent on the sample quality and quantity. Bayesian discrimination is more suitable for the small sample classification problem in the WFSD since all samples can be learned simultaneously using a matrix solution. Bayesian discrimination considers the uncertainty of all levels of the model synthetically and regards the parameters and hidden variables in the model as random variables, so the inference of super parameters is not affected by the number and order of samples. Comparing the WFSD discrimination results shows that the generalization of the artificial neural network is weaker than that of Bayesian discrimination, which leads to a prediction error greater than the Bayesian discrimination error when the number of fault gouges samples is small. Bayesian discrimination is an algorithm based on probability theory, with clear statistical significance, complete theoretical support, applicability to any geological situation, and good potential for further information mining. Therefore, this study implements Bayesian discrimination to identify fault gouge and proves the feasibility of using the Bayesian discrimination function to determine the PSZ. Finally, the depth of the PSZ in WFSD-1 is calculated purely from logging data and compared with geological research results to verify the result.

PSZ possibility from Bayesian discrimination function

The naive Bayesian algorithm (Ramoni and Sebastiani 2001) and the Bayesian discriminant algorithm (Liu et al. 2016) are widely used in lithology discrimination. Theoretically, the former is applicable to cases when the relationship between logging data and class values is significant, while the Bayesian discriminant algorithm is more suitable for fault gouge identification and PSZ determination in WFSD. The Bayesian discriminant algorithm comes from the Bayesian theorem and Gauss distribution law. It takes the minimum average loss of misjudgment as the discrimination idea and considers both the prior probability and distribution of each lithology as well as the difference in misjudgment loss.

Lithologic classification using Bayesian discrimination

Considering the prior estimation of the logging data to be predicted, the Bayesian discriminant method learns the conditional distribution of the sample logging data relative to lithologies based on the Bayesian theorem, calculates the posterior probability distribution of the logging data to be predicted, and takes the corresponding lithology with the maximum posterior probability as the predicted result. The algorithm includes both learning and prediction stages. The Bayesian discrimination function is established using classified training samples in the learning stage, then classifies lithologies from training data and unknown data in the prediction stage (Liu et al. 2016).

Before machine learning can be applied, logging data such as depth correction and environment correction must be cleaned and combined with lithology class from intuitive core data and logging interpretation. The logging data at the target depth is set as the sample set. x_{ijk} represents the k-th logging data point of the sample at the j-th depth of the i-th lithology. Let the lithology class number i have a total of M groups (i.e., $i = 1, 2… M$). The sample number j of each lithology has a total of N_i rows (i.e., $j = 1, 2… N_i$). All samples have a total of N rows, and the effective logging parameter number k has a total of L columns (i.e., $j = 1, 2… L$). Let the vector Y_j represent the logging data of the j-th row, where y_{jk} is the k-th logging parameter. The sample data from all N sampling points can be re-divided into M groups, at which point the logging data at each depth is expressed as $X_{ij} = (x_{ij1}, x_{ij2} … x_{ijL})$. The central coordinate of sample aggregation in L-dimensional space can be expressed by the average value of the training samples:

$$\overline{x_{ik}} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ijk} \tag{1}$$

The Bayesian discrimination function based on the minimum error rate and Gaussian kernel function can be written as:

$$F_{ij}(Y_j) = \sum_{k_1=1}^{L} (C_{ik_1} y_{jk}) - \frac{1}{2} \sum_{k_1=1}^{L} (C_{ik_1} \overline{x_{ik_1}}) + \ln\left(\frac{n_i}{N}\right) \tag{2}$$

where parameters in the discrimination function coefficient matrix are

$$C_{ik_1} = \sum_{k_2=1}^{L} p_{k_1 k_2}^{-1} \overline{x_{ik_2}} \tag{3}$$

$$p_{k_1 k_2} = \frac{1}{N-M} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left[\left(x_{ijk_1} - \overline{x_{ik_1}} \right) \left(x_{ijk_2} - \overline{x_{ik_2}} \right) \right] \tag{4}$$

The discrimination function formula (2) is the grouping basis of the Bayesian discriminant algorithm. The i value corresponding to the maximum $F_{ij}(Y_j)$ is the lithology class value corresponding to the logging data.

PSZ determination using an exponential model based on the Bayesian discrimination function

Up to now, there exists no core experiment that gives the formulas of the time-varying logging curves of fault gouges during the earthquake and fault healing processes. In order to extract information from the Bayesian discrimination function, we have to simplify some conditions to locate the PSZ in WFSD. Unlike conventional machine learning algorithms which directly correlate logging and seismic data with known formation parameters, this paper discusses the Principal Slip Zone Possibility (PSZP), which is determined using the Bayesian discrimination function without using experimental test data and regional experience parameters.

The simplified hypotheses are put forward with the help of logging parameter crossplots between different lithologies. The four crossplots comprehensively reflect the relationship between GR, AC, CNL, Rt, and SP in WFSD-1

(Fig. 4). Compared with Fig. 3, two-dimensional crossplots can distinguish various lithologies better than the box-whisker plots. Because fault gouge commonly exists as an interlayer in cataclasite or shale, the logging values at depth will be strongly affected by the complex surrounding rock and varying detection range. After compaction for nearly a year after the Wenchuan earthquake, the fault gouge, and shale have similar physical properties, making distinguishing the fault gouge and shale more difficult. The difficulty in fault gouge identification primarily lies where the logging values coincide with cataclasite and shale (Figs. 3 and 4); hence, the need for a higher-dimensional Bayesian identification algorithm.

Considering the poor consolidation degree of fault gouges in the active earthquake zone, it is inevitable that multiple seismic displacements have occurred in each fault gouge layer. The mineral composition and structural morphology of the PSZ and stable fault gouges that have experienced
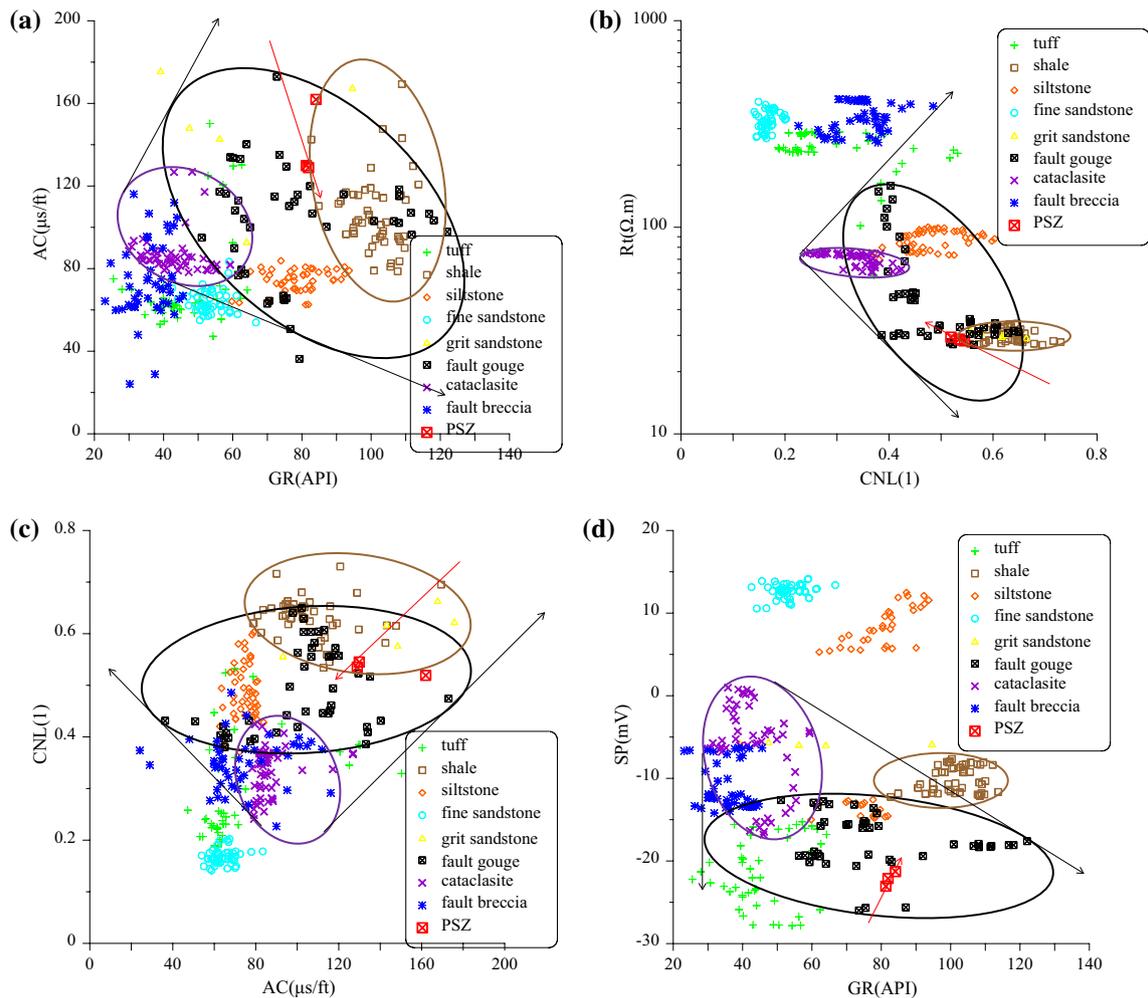

Fig. 4 Crossplots of between the log parameters GR, SP, AC, CNL, and Rt. (Black ellipses are the logging response areas of stable fault gouges, red arrows represent the logging response change direction)

multiple crushing and compaction events are similar. The core of the black ellipses in Fig. 4 are the logging response areas of stable fault gouges. We can make a simplified hypothesis (1): before the earthquake, the logging response of the PSZ can be expressed as that of stable fault gouge $y0_{jk}$.

During the earthquake, there will be some frictional melting (Mckenzie and Brune 1972), porosity creating, thermal decomposition (Han et al. 2007), etc. at the PSZ depth. As a result, PSZ logging responses also change greatly at the moment of the earthquake. Since the fault gouge is likely to be homologous with the surrounding cataclasite, while its logging responses are clearly separated from the cataclasite, we may suppose that the PSZ responses may be distributed over a large range between the black arrows in Fig. 4 after the earthquake. To describe this conjecture, we can give a simplified hypothesis (2): when an earthquake occurs, the physical properties of the stable fault gouge without a fracture remain unchanged at $y0_{jk}$, and only the fault gouge at the PSZ records a huge disturbance in logging response z_k.

During the fault healing period, several years to decades after the earthquake, shale content will increase, and porosity will decrease in the fault gouge at the PSZ depth. The precise change law of the logging response is hard to determine. Since pressure-solution rates are predicted to diminish exponentially with time (Gratier et al. 2003), and some of the healing processes could be consistent with pressure-solution mechanisms, we think the exponential decline is a good approximation. The logging response of the fault gouge will gradually recover to the logging response range of most stable fault gouges (Fig. 4 red arrows). Since there is no recorded seismic event in the Yingxiu-Beichuan fault zone for the last thousand years, the logging response of stable fault gouge can be set as $y0_{jk}$, and the high-dimensional logging response coordinate of the new PSZ is the farthest from the stable fault gouge. A simplified hypothesis (3) is obtained to describe this phenomenon: the logging response of the PSZ after the earthquake will change with time t, and its change law follows an exponential decline $e^{-\beta t}$, finally returning to $y0_{jk}$.

The distribution range of fault gouge and cataclastic rock is within the ellipse (Fig. 4), that is to say, there can be simplified hypothesis (4): the training sample also satisfies Gaussian distribution in the high-dimensional logging space, which conforms to the Gaussian density function model in the Bayesian discrimination. After obtaining accurate experimental data, the above simplified hypothesis can be complicated to match the real fault evolution.

In the conventional Bayesian discriminant algorithm, $F_{ij}(Y_j)$ only compares the maximum value after being calculated using Formula (2), but its meaning has not been further interpreted. When comparing the size of the Bayesian discrimination function $F_{ij}(Y_j)$ for all M type lithologies in row j, the discrimination function coefficient matrix is a constant matrix, so each logging data point is linearly related to $F_{ij}(Y_j)$. According to the above assumptions, if the PSZ is at the j-th depth, the k-th logging data of the j-th depth can be expressed as $y1_{jk}$:

$$y1_{jk} = y0_{jk} + e^{-\beta t} z_k \tag{5}$$

where $y0_{jk}$ is the logging data of the PSZ before the earthquake. $L\ y0_{jk}$ constitutes vector $Y0_j$, and $L\ y1_{jk}$ constitutes vector $Y1_j$. z_k refers to the change in the k-th logging value of fault gouge under compression, friction, and melting of the earthquake. Note that z_k here is not necessarily positive. Since the Bayesian discrimination function coefficient matrix comes from sample data and has no influence on the data to be discriminated, the PSZ discrimination function can also be decomposed into:

$$F_{Pj}(Y1_j) = \sum_{k_1=1}^{L} \left[C_{Pk_1}(y0_{jk} + e^{-\beta t} z_k) \right] - \frac{1}{2} \sum_{k_1=1}^{L} (C_{Pk_1} \overline{x_{Pk_1}}) + \ln\left(\frac{n_P}{N}\right) \tag{6}$$

Here, the angle sign $i = P$ indicates the PSZ lithology number, while $i = F$ indicates the stable fault gouge lithology number. After the rupture of fault zone, the change in PSZ discrimination function with time can be written as:

$$F_{Pj}(Y1_j) - F_{Pj}(Y0_j) = \sum_{k_1=1}^{L} (C_{Pk_1} e^{-\beta t} z_k) \tag{7}$$

where $F_{Pj}(Y1_j)$ is the PSZ discrimination function from measured data, $F_{Pj}(Y0_j)$ is the virtual PSZ discrimination function before the earthquake. The logging data coordinates are still unknown, so $Y0_j$ here cannot be solved directly. Because the PSZ logging response decreases relative to that of the stable fault gouge within several years or longer, a parameter reflecting the PSZ possibility can be introduced as P_j. Even if some petrophysical properties that affect logging data failed to differentiate between PSZs and 'stable fault gouges, the Bayesian discrimination algorithm will automatically highlight effective logging data in the process of sample learning. Let $P_j = e^{-\beta t}$; therefore, this time related item can be expressed as

$$P_j = \frac{F_{Pj}(Y1_j) - F_{Pj}(Y0_j)}{\sum_{k_1=1}^{L} (C_{Pk_1} z_{k_1})} \tag{8}$$

z_k and $F_{pj}(Y0_j)$ in Eq. (8) are both unknown quantities to be solved. An accurate z_k requires that logging data acquisition be completed in a short time after the earthquake, which is impossible. The determination of $F_{pj}(Y0_j)$ requires continuous observation for a long time after the earthquake, which is difficult to achieve. Therefore, Eq. (8) cannot be solved directly. According to hypothesis (3), the PSZ logging response will be restored exponentially to the stable

fault gouge range, and its change path is unidirectional. If the PSZ does exist in the sample, the PSZ probability can be assigned as $P_j=1$, then we can filter out $F_{Pj}(Y1_j)$ by taking back samples. As for the calculation of $F_{Pj}(Y0_j)$, because Bayesian discrimination distinguishes the PSZ from stable fault gouge by decision hyperplane, when the disturbance of PSZ logging response is reduced to 0, the discrimination function $F_{pj}(Y0_j)$ will appear on the decision hyperplane. It should be noted that according to hypothesis (3), PSZ logging data will not necessarily generate the largest $F_{Pj}(Y1_j)$, but the largest $F_{Pj}(Y1_j) - F_{Pj}(Y0_j)$.

By studying the Bayesian discrimination function (2), we can imagine that the Bayesian discrimination algorithm finally shows that different sample groups are divided in L-dimensional logging parameter space through decision hyperplane or line (see Fig. 5 for $L=2$). The PSZ discrimination function is the plane where the orange box is located, the discrimination function of stable fault gouge is the plane where the purple box is located, and the two-dimensional space where the black box is located is the sample data space (Fig. 5). When the L-dimensional sample data are projected into the upper $L+1$-dimensional discrimination function space, and there is a maximum value in the PSZ plane, the stratum can be classified as

affect the value of z_k. Therefore, the general movement path of PSZ logging data is represented by parallel vectors in L-dimensional space, and the direction of the vectors is $\vec{r} = (z_1, z_2 \ldots z_L)$. Let the projection vector from the logging data $Y1_j$ to the decision hyperplane be $\vec{d}_j$, and the angle between vectors $\vec{d}_j$ and $\vec{r}$ is θ. Equation (8) can be rewritten as:

$$P_j = \frac{\left|\vec{d}_j\right|}{\cos\theta \sum_{k_1=1}^{L}(C_{Pk_1}z_{k_1})} \tag{9}$$

$F_{Pj}(Y0_j) = F_{Fj}(Y0_j)$ is on the decision hyperplane. After the PSZ is restored to stable fault gouge, the new PSZ logging data $y0_{jk}$ satisfies the hyperplane equation

$$\sum_{k_1=1}^{L}\left[(C_{Pk_1} - C_{Fk_1})(y0_{jk})\right] - \frac{1}{2}\sum_{k_1=1}^{L}(C_{Pk_1}\overline{x_{Pk_1}})$$
$$+ \frac{1}{2}\sum_{k_1=1}^{L}(C_{Fk_1}\overline{x_{Fk_1}}) + \ln\left(\frac{n_P}{N}\right) - \ln\left(\frac{n_F}{N}\right) = 0 \tag{10}$$

Then, we calculate the projection distance of each training sample $F_{Pj}(Y1_j)$ on the hyperplane Eq. (10)

$$\left|\vec{d}_j\right| = \frac{\left|\sum_{k_1=1}^{L}\left[(C_{Pk_1} - C_{Fk_1})(y_{jk})\right] - \frac{1}{2}\sum_{k_1=1}^{L}(C_{Pk_1}\overline{x_{Pk_1}}) + \frac{1}{2}\sum_{k_1=1}^{L}(C_{Fk_1}\overline{x_{Fk_1}}) + \ln\left(\frac{n_P}{N}\right) - \ln\left(\frac{n_F}{N}\right)\right|}{\sqrt{\sum_{k_1=1}^{L}\left[(C_{Pk_1} - C_{Fk_1})^2\right]}} \tag{11}$$

the PSZ. According to Eq. (5), the healing stages of different fault gouges throughout history theoretically do not
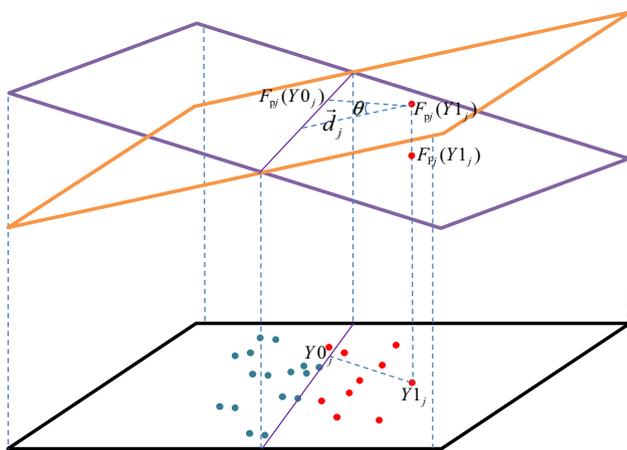
Substituting discrimination function (2), Eq. (11) can be simplified as:

$$\left|\vec{d}_j\right| = \frac{F_{Pj}(Y1_j) - F_{Fj}(Y1_j)}{\sqrt{\sum_{k_1=1}^{L}\left[(C_{Pk_1} - C_{Fk_1})^2\right]}} \tag{12}$$

Finally, the equation that can be used to judge the PSZ possibility is obtained:

$$P_j = \frac{(F_{Pj}(Y1_j) - F_{Fj}(Y1_j))}{\cos\theta \sqrt{\sum_{k_1=1}^{L}\left[(C_{Pk_1} - C_{Fk_1})^2\right]} \sum_{k_1=1}^{L}(C_{Pk_1}z_{k_1})} \tag{13}$$

There are still unknown constants z_k and θ in Eq. (13), which do not affect the judgment of the maximum P_j for the samples. The serial number of the sample is marked with the maximum P_j as J, and $P_J = 1$, so that the PSZ possibility at other depths can be obtained through comparison. Equation (14) is the PSZ Possibility from the Bayesian discrimination function (PSZP-BDF):

Fig. 5 Schematic diagram of the calculation principle of the Bayesian discrimination function

$$P_j = \frac{P_J \left[F_{Pj}(Y1_j) - F_{Fj}(Y1_j) \right]}{F_{Pj}(Y1_J) - F_{Fj}(Y1_J)} \tag{14}$$

It should be emphasized that the selection of $Y1_J$ has a significant influence on the results of Eq. (14). When obtaining time-lapse logging data and core test data is difficult, the selection of $Y1_J$ can only be obtained through the samples themselves. Therefore, during sample selection, we should first increase the number of fault gouge samples as much as possible, and simultaneously remove the abnormal data caused by instrument failure. In addition, even if the hypothesis (2) is not true, that is, the value of S_J is uncertain, then in the same well, the P_j calculated by Eq. (14) has relative comparative significance in order to give a reliable PSZ depth.

From the derivation of P_j, it is found that the advantages of PSZP-BDF algorithm are as follows. (1) The PSZP-BDF model is derived from the statistical segmentation of the sample space in the target layer, which gives an increased statistical significance, better regional pertinence, and wide applicability compared to the empirical formula summarized by the experiment and the theoretical formula derived from the physical model. (2) The parameters in PSZP-BDF are completely from the core and logging data selected from layers with significant logging response; hence, so the calculated P_j is less affected by subjective human factors. (3) Bayesian discrimination is a high-dimensional data analysis method which is more suitable for multi-physical field analysis of fault gouge The PSZP-BDF can make full use of valuable logging data, compared to the limited analysis of one-dimensional histograms and two-dimensional crossplots. (4) After sample selection and logging data pre-processing, the PSZP-BDF algorithm can quickly calculate the results and even provide timely PSZ depth suggestions simultaneously with data acquisition. The rapid positioning of PSZ is of great significance for understanding the fault zone structure and deploying follow-up scientific drilling. The PSZP-BDF algorithm can screen out core data that requires further testing, which can greatly reduce the cost of whole well coring and the workload of experimental testing in earthquake scientific drilling.

Application of PSZP-BDF algorithm

After theoretical derivation of the PSZP-BDF algorithm, we further apply it to the calculation of WFSD-1. The lithology number in the WFSD-1 sample is $M = 9$ (Fig. 4). GR, AC, CNL, Rt, SP, and MS are used as dependent variables, which are enough to calculate the shale content, porosity, permeability, and other formation information. Because the model, series, and calibration standard of logging tools used in other WFSD wells are different from those of WFSD-1, and the lithologies of different wells are quite different, standardize among the 5 wells is difficult, so only the logging data from WFSD-1 is analyzed here. For the increase of logging parameters in other scientific drilling expeditions, more logging series can be introduced, and the PSZP-BDF algorithm is still Eq. (14). Considering the thick fault gouge in WFSD-1, point by point discrimination method is adopted for all data in this paper. The thin layer within 0.5 m is ignored and included in the upper or lower formation.

It can be seen from Eq. (14) that P_j is independent of the prior probability in the discrimination function, so the influence caused by the inaccurate prior probability is avoided in the PSZP-BDF algorithm. However, the fault gouge may be thin and its logging response can easily be masked the surrounding rock. In order to ensure that all fault gouges are identified as much as possible, the prior probability of the fault gouge is appropriately increased in the discrimination algorithm, which significantly improves the accuracy of the fault gouge identification, while the identification accuracy of other lithologies is slightly reduced.

Several core tests show that the PSZ depth of WFSD-1 is between 589.21 m and 590 m (Xu et al. 2018) highlighted as a blue box in Fig. 6. To identify the fault gouge or even PZS position from logging curves, shale content or porosity is difficult. But according to the comparison between tested core logging lithology and predicted Bayesian lithology in the interpreted result in Fig. 6, Bayesian discrimination can successfully identify various lithologies and fault gouges. Due to the influence of point by point discrimination, discriminated stratigraphic boundaries are still difficult to match with core logging data completely. Because the core logging data has certain depth error, the depth provided by logging data is more likely to be accurate when the depth of the fault gouge identified by Bayesian discrimination is not far away. Therefore, the fault gouge at 590 m and 607 m positions can be regarded as the correct depth. From the results of PSZP-BDF, the depth of the maximum PSZ possibility P_j is exactly the same as the PSZ depth determined by geology, with relative high AC and porosity. The peak value of P_j has obvious advantages over others at stable fault gouges, which verifies the effectiveness of the PSZP-BDF algorithm. In addition, the PSZP-BDF algorithm also evaluates the P_j of other stable fault gouges, which may be the PSZ of past earthquakes, or the accompanying structural response generated by Wenchuan earthquake. If multiple across-strike gouge zones slip simultaneously in the same earthquake, theoretically it will produce several obvious peak value of P_j. In this case, researches on the cutoff value of P_j will become meaningful.
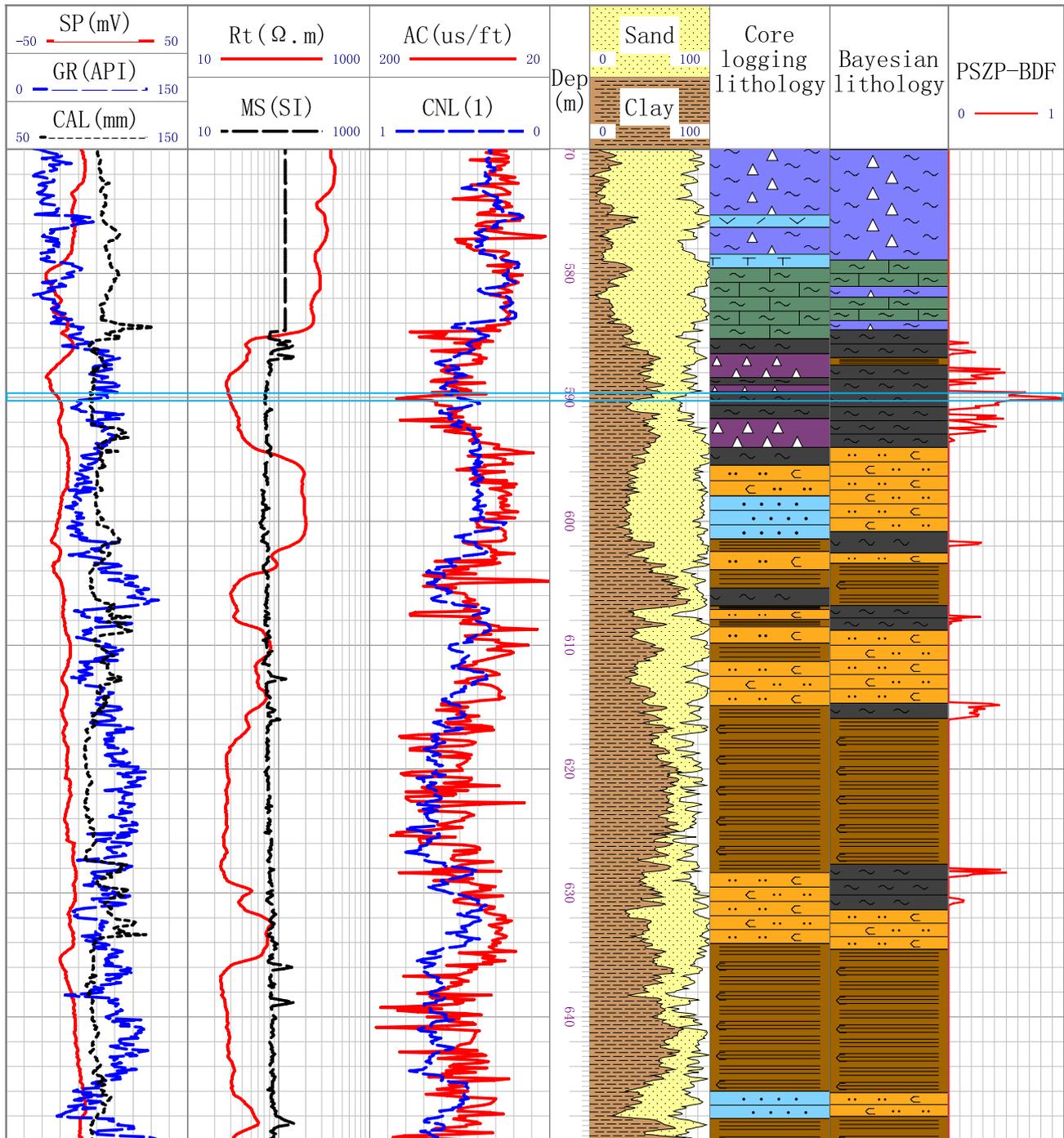
Fig. 6 Logging and core data of the main fault zone at 570–650 m depth. The maximum anomalies of P_J appear at 590 m-depth

Conclusion

Using information mining via Bayesian discrimination, this paper validates the PSZP-BDF algorithm based on four simplified hypotheses and puts forward a simple and practical

Eq. (14) to evaluate the PSZ possibility just using logging data.

Through theoretical derivation of a Bayesian discrimination function and field data test from WFSD-1, we find that the PSZP-BDF algorithm has the following advantages: (1) The accuracy of the PSZP-BDF algorithm is high enough

to determine the PSZ. (2) The algorithm requires little geological data and is widely applicable to PSZ evaluation in varying geological conditions. (3) The PSZ possibility is objective, because the algorithm parameters come completely from the samples. (4) The algorithm can be used to comprehensively interpret the PSZ multi-physical field response and reduce the multiple solutions of PSZ problem determined by a single parameter. (5) The algorithm is simple and fast, which can provide quick PSZ suggestions and save on coring and test costs.

The PSZP-BDF algorithm also needs to be improved, especially for cases with significant data noise, in which the fault gouge discrimination accuracy is not high enough, resulting in the loss of possible PSZs. It's only an expedient way to increase the judgment rate of fault gouge through the improvement of Bayesian prior probability. A machine learning method with strong generalization ability and high judgment accuracy was used to carry out the Bayesian discrimination. In addition, the hypothesis has certain limitations and is different from the real situation; therefore, monitoring the time-lapse logging for the dynamic logging response of the PSZ and making physical property analysis for fault gouge core data is necessary to obtain more accurate model parameters.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Brodsky EE, Kuo-Fong M, Jim M et al (2009) Rapid response fault drilling past, present, and future. Sci Drill 8(8):66–74

Gratier JP, Favreau P, Renard F (2003) Modeling fluid transfer along California faults when integrating pressure solution crack sealing and compaction processes. J Geophys Res 108(B2):1–25

Han R, Shimamoto T, Hirose T et al (2007) Ultralow friction of carbonate faults caused by thermal decomposition. Science 316(5826):878–881

Hirono T, Fujimoto K, Yokoyama T et al (2008) Clay mineral reactions caused by frictional heating during an earthquake: an example from the Taiwan Chelungpu fault. Geophys Res Lett 35(16):L16303

Hirose T, Shimamoto T (2005) Growth of molten zone as a mechanism of slip weakening of simulated faults in gabbro during frictional melting. J Geophys Res (Solid Earth) 110:B05202

Kano Y, Mori J, Fujio R et al (2006) Heat signature on the chelungpu fault associated with the 1999 chi-chi, taiwan earthquake. Geophys Res Lett 33(14):L14306

Konaté AA, Pan HP, Ma HL et al (2017) Use of spectral gamma ray as a lithology guide for fault rocks: a case study from the Wenchuan Earthquake Fault Scientific Drilling project Borehole 4 (WFSD-4). Appl Radiat Isotopes 128:75–85

Lacroix B, Tesei T, Oliot E et al (2015) Early weakening processes inside thrust fault. Tectonics 34(7–8):1396–1411

Li HB, Wang H, Xu ZQ et al (2013a) Characteristics of the fault-related rocks, fault zone structures and the principal slip zone of the wenchuan earthquake in wfsd drilling cores. Tectonophysics 584:23–42

Li HB, Xu ZQ, Wang H et al (2013b) The Principle Slip Zone of the 2008 Wenchuan earthquake: a thrust fault oblique cutting the Yingxiu-Beichuan fault zone. Geol China 40(1):121–139

Li HB, Xu ZQ, Wang H et al (2018) Fault behavior, physical properties and seismic activity of the Wenchuan earthquake fault zone: evidences from the Wenchuan earthquake Fault Scientific Drilling project (WFSD). Chin J Geophys (in Chin) 61(5):1680–1697

Lin A, Maruyama T, Aaron S et al (2005) Propagation of seismic slip from brittle to ductile crust: evidence from pseudotachylyte of the Woodroffe thrust, central Australia. Tectonophysics 402(4):1–35

Liu XY, Chen XH, Li JY et al (2016) Reservoir physical property prediction based on kernel-Bayes discriminant method. Acta Petrol Sin 37:878–886

Ma KF, Tanaka H, Song SR et al (2006) Slip zone and energetics of a large earthquake from the taiwan chelungpu-fault drilling project. Nature 444(7118):473–476

Mckenzie D, Brune JN (1972) Melting on fault planes during large earthquakes. Geophys J Roy Astron Soc 29(1):65–78

Pei J, Li H, Wang H et al (2014) Magnetic properties of the Wenchuan Earthquake Fault Scientific Drilling Project Hole-1 (WFSD-1), Sichuan Province, China. Earth Planets Space 66:23

Peng ZG, Wu CQ, Aiken C (2011) Delayed triggering of microearthquakes by multiple surface waves circling the Earth. Geophys Res Lett 38(4):L04306

Rabinowitz HS, Savage HM, Polissar PJ et al (2020) Earthquake slip surfaces identified by biomarker thermal maturity within the 2011 Tohoku-Oki earthquake fault zone. Nat Commun 11(1):533

Ramoni M, Sebastiani P (2001) Robust Bayes classifiers. Artif Intell 125:207–224

Sibson RH (2003) Thickness of the seismic slip zone. Bull Seismol Soc Am 93(3):1169–1178

Spray JG (1992) A physical basis for the frictional melting of some rock-forming minerals. Tectonophysics 204(3–4):205–221

Tanaka H, Fujimoto K, Ohtani T et al (2001) Structural and chemical characterization of shear zones in the freshly activated Nojima fault, Awaji island, southwest Japan. J Geophys Res 106(B5):8789–8810

Tsakalos E, Lin A, Kazantzaki M et al (2020) Absolute Dating of Past Seismic Events Using the OSL Technique on Fault Gouge Material—A Case Study of the Nojima Fault Zone, SW Japan. J Geophys Res: Solid Earth 125(8):1

Wang CY (2007) Liquefaction beyond the near field. Seismol Res Lett 78(5):512–517

Wang H, Li HB, Zhang L et al (2018) Pseudotachylytes in the Longmen Shan fault zone and fault weakening mechanisms. Chin J Geophys 61(5):1698–1714 **(in Chinese)**

Weaver KC, Doan ML, Cox SC et al (2019) Tidal behavior and water-level changes in gravel aquifers in response to multiple earthquakes: a case study from New Zealand. Water Resour Res 55(2):1263–1278

Woodcock NH, Mort K (2008) Classification of fault breccias and related fault rocks. Geol Mag 145(3):435–440

Xu ZQ, Wu ZL, Li HB et al (2018) The most rapid respond to a large earthquake—the Wenchuan earthquake Fault Scientific Drilling Project. Chin J Geophys 61(5):1666–1679 **(in Chinese)**

Xue L, Li HB, Brodsky EE et al (2013) Continuous permeability measurements record healing inside the Wenchuan earthquake fault zone. Science 340(6140):1555–1559

Zheng YC (2018) Transient pressure surge in a fluid-filled fracture. Bull Seismol Soc Am 108:1481–1488

Zheng Y, Li H, Gong Z (2016) Geothermal study at the Wenchuan earthquake Fault Scientific Drilling project-hole 1 (WFSD-1): borehole temperature, thermal conductivity, and well log data. J Asian Earth Sci 117:23–32

A migration-based location method using improved waveform stacking for microseismic events in a borehole system

Qinghui Mao[1,2] · Tahir Azeem[3] · Xuliang Zhang[4] · Yu Zhong[5] · Yuan Fang[6] · Yunxiao Zhang[7]

Abstract

The migration-based microseismic event location methods using waveform stacking algorithms are widely used for hydro-fracturing monitoring. These methods have the advantage of not requiring the accurate first arrival time around a detected event, which is more suitable for noisy data than classical travel time-based methods. However, accuracy of these methods can be affected under the condition of relatively low signal-to-noise ratio (SNR). Therefore, in order to enhance the location accuracy of microseismic events in a borehole system, we have proposed a migration-based location method using improved waveform stacking with polarity correction based on a master-event technique, which optimizes the combination way of *P*- and *S*-wave waveform stacking. This method can enhance the convergence of the objective function and the location accuracy for microseismic events as compared to the conventional waveform stacking. The proposed method has been successfully tested by using synthetic data example and field data recorded from one downhole monitoring well. Our study clearly indicates that the presented method is more viable and stable under low SNR.

Introduction

It is vital to accurately and automatically locate the micro-seismic events for estimating the stimulated reservoir volume. Migration-based source location methods are more applicable to relatively low signal-to-noise ratio (SNR) microseismic data than traditional travel time-based source location methods, because the former methods are far less sensitive to the picking precision than the latter ones (Gharti et al. 2010). During the last two decades, a number of studies have been carried out about the migration-based location of microseismic events based on nonnegative waveform stacking in order to eliminate the influence of waveform polarity changes due to shear source mechanisms (Kao and Shan 2004; Gajewski et al. 2007; Gharti et al. 2010; Drew et al. 2013; Grigoli et al. 2014; Li et al. 2016; Shi et al. 2019). These approaches can make the location results more reliable at the cost of decreasing the location resolution in the case of high SNR. According to Trojanowski and Eisner (2017), stacking of both the positive and negative values of amplitude can provide more reliable results after the polarization correction as compared to any of absolute value-based methods. In the past, the polarity was reliably corrected by using source mechanism inversion (Anikiev et al. 2014). However, it dramatically increases the computational cost. Afterward, Kim et al. (2017) proposed the automatic method for the determination of first-motion polarity which estimates the relative polarities of waveforms at other receivers based on cross-correlation analysis. Later on, Xu

✉ Qinghui Mao
maoqinghuijz@126.com

1 Key Laboratory of Exploration Technologies for Oil and Gas Resources (Yangtze University), Ministry of Education, Wuhan 430100, China

2 Hubei Cooperative Innovation Center of Unconventional Oil and Gas, Yangtze University, No. 111 Daxue Rd. Caidian District, Wuhan 430100, Hubei Province, China

3 Department of Earth Sciences, Quaid-I-Azam University, Islamabad 45320, Pakistan

4 Bureau of Geophysical Prospecting, Zhuozhou 072750, China

5 Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China

6 Development and Research Center, China Geological Survey, Beijing 100037, China

7 Oil and Gas Survey, China Geological Survey, Beijing 100083, China

et al. (2019) corrected the changing polarities by using an amplitude trend least-squares fitting method. These polarization correction methods are effective and comprehensively reduce the computational cost. Therefore, polarization correction methods can be implemented before waveform stacking, which can provide the basic guarantee for obtaining the constructive interference from stacked amplitudes. Moreover, a relative location was also introduced into the migration-based location methods to mitigate the negative effect of an inaccurate velocity model on the location results, which provides the foundation for accurate waveform stacking (Grigoli et al. 2016; Li et al. 2016). In addition, characteristic function can be calculated within *P*- and *S*-wave corridors by selecting an approximate P or S arrival time instead of scanning the whole time period of seismic records, which can significantly accelerate the scanning efficiency (Eaton et al. 2011).

On the basis of a predefined time window centered on the arrival time of signals (namely *P*- and/or *S*-wave corridors) (Eaton et al. 2011), polarization correction algorithm (Kim et al. 2017) and master-event technique (Grigoli et al. 2016), we present a new migration-based location method via improving the combination way of *P*- and *S*-wave waveform stacking in order to increase the resolution of the objective function in the study. Thereafter, a two-dimensional (2D) model is used to prove the feasibility of the presented method. Finally, we assess the performance of this method on real data.

Method

This section briefly introduces the master-event technique and mainly describes the objective functions formulated by using the conventional and improved waveform stacking, as well as the location flow for the migration-based method to locate a selected microseismic event.

Master-event technique

The master-event technique, proposed by Grigoli et al. (2016), can reduce the dependency of the location results on the precision of the adopted velocity model, which inherits the main features of relative location algorithms. Here a master event should have the characteristics of high SNR waveform and reliable source location. The perforation shot

event or high SNR events may be selected as master events. Through this master-event technique, we can calculate the master-event time corrections of all receivers for *P* and *S* waves, respectively. Then, these corrections can be applied to Eq. (1) as discussed in "Conventional waveform stacking function" section in order to reduce the influence of an inaccurate velocity model on the location results.

Formulation of the objective function

The objective function for the migration-based location method is generally constructed by the way of waveform stacking. The waveform stacking function can be broadly classified into two main categories: the first, which only stacks the positive values of seismograms (such as absolute values, STA/LTA ratio, envelopes and phase square) and the second, which stacks both the positive and negative values of seismograms. Although the former methods, namely absolute-value-based approaches, can avoid the influence of changing polarities due to source radiation pattern, the imaging resolution of their objective functions may be reduced for microseismic data with relatively low SNR. However, a simple stacking can yield better results than the former methods when the signal polarization correction is applied to the latter methods (Trojanowski and Eisner 2017). So, in this study, we only discuss that method which can stack both the positive and negative values of seismograms with polarization correction. In addition, the waveform stacking function based on the combination of *P* and *S* waves is applied to migration algorithms in order to enhance the accuracy of event location as recommended by Gharti et al. (2010).

Conventional waveform stacking function

The waveform stacking function can be calculated within *P*- and *S*-wave corridors by selecting an approximate arrival time instead of scanning the whole time period of seismic records (Eaton et al. 2011). We take the selection of an approximate *S*-wave arrival time because the weaker *P*-wave energy is more easily submerged in noise.

By assuming the *P*- or *S*-wave calculated travel times at the *i*th receiver (t_i^k) and the *S*-wave calculated travel time at the *m*th receiver (t_m^S), we can obtain their corresponding corrected travel times (tc_i^k and tc_m^S) after applying the master-event time corrections calculated through the master-event technique (Grigoli et al. 2016). Then, the calculated travel time difference between the *i*th receiver for *P/S*-wave and the *m*th receiver for *S*-wave is given by Eq. (1):

$$\Delta t_i^k = \begin{cases} t_i^k - t_m^S & \text{if not applying the master} - \text{event time corrections} \\ tc_i^k - tc_m^S & \text{if applying the master} - \text{event time corrections} \end{cases} \qquad k \in \{P, S\} \qquad (1)$$

For the purpose of discussions, the basic unit of waveform stacking function ($L_k(j)$) can be defined by Eq. (2):

$$L_k(j) = \sum_{i=1}^{n} \left(\text{sgn}_i^k u_i(\Delta t_i^k + j) \right) \quad k \in \{P, S\} \tag{2}$$

where n and j denote the number of receivers and an index of the discrete-time signal, respectively. u_i represents the signal amplitude at the ith receiver. sgn_i^k is the corresponding polarity correction sign of the raw amplitude u_i at the ith receiver for P- or S-wave, which is automatically determined using the method proposed by Kim et al. (2017). $L_k(j)$ represents the waveform stacking with polarity correction at time index j from all receivers for P- or S-wave after we eliminate the moveout (Δt_i^k) of the corresponding trace. Subsequently, the conventional waveform stacking function with a predefined time window centered on the S-wave arrival of this event is constructed based on P- and S-wave stacking basic unit information as expressed by Eq. (3):

$$\text{CWS}(t) = \sum_{j=t-w_1}^{t+w_1} L_P(j) \times \sum_{j=t-w_1}^{t+w_1} L_S(j) \ \left(t = (t_{\text{mFA}}^S - w_2), \ldots (t_{\text{mFA}}^S + w_2) \right) \tag{3}$$

where w_1 represents the size of the inner window and it controls the SNR of the waveform stacking function. j is centered on the index of time t and its range is from $t - w_1$ to $t + w_1$. t_{mFA}^S is an approximate estimate for the S-wave first arrival at the mth receiver for this event. w_2 represents the size of the outer time window. And it is a sliding window range centered on the t_{mFA}^S mentioned above, which is used to scan through this event in order to avoid searching through the entire time.

Improved waveform stacking function

For enhancing the convergence and resolution of the objective function, we build the objective function by improving the combination way of P- and S-wave waveform stacking, which calculates the multiplication of the basic units of P- and S-wave waveform stacking functions before summing within the inner time window $[t - w_1, t + w_1]$. Thus, the improved waveform stacking function with the time window range from $t_{mFA}^S - w_2$ to $t_{mFA}^S + w_2$ can be expressed by the following Eq. (4):

$$\text{IWS}(t) = \sum_{j=t-w_1}^{t+w_1} L_P(j) \times L_S(j) \ \left(t = (t_{\text{mFA}}^S - w_2), \ldots (t_{\text{mFA}}^S + w_2) \right) \tag{4}$$

where L_P, L_S, t_{mFA}^S, w_1 and w_2 are the same as those described previously in "Conventional waveform stacking function" Section.

Location flow

Below-mentioned procedure is adopted to calculate the location of a selected microseismic event.

1. Input an adjusted velocity model and create a lookup table by calculating the travel times from every potential event location to all of the receivers based on the ray-tracing technique (Moser 1991).
2. Calculate the master-event time corrections by using the master-event technique (Grigoli et al. 2016) and apply these corrections to Eq. (1).
3. Only input a single approximate S-wave first arrival time for this event.
4. Calculate the objective function by using the method discussed in "Conventional waveform stacking function" or "Improved waveform stacking function" section.
5. Output the corresponding location when we determine the global maximum value of the corresponding objective function by utilizing the grid search method (Mao et al. 2019).

Synthetic data example

In this study, we focus on the accuracy of the conventional and improved waveform stacking functions. For simplicity, we used a 2D homogeneous medium model with P-wave velocity 2000 m/s, S-wave velocity 1200 m/s and density 2.0 g/cm^3 as shown in Fig. 1.

To cover the depth range from 950 to 1050 m, eleven receivers were placed at 10-m spacing intervals along a vertical observation well. The wellhead was located at $x = 200$ m. A Ricker wavelet having dominant frequency of 60 Hz was used as the microseismic source time function. Time interval was 0.5 ms. The master event and target
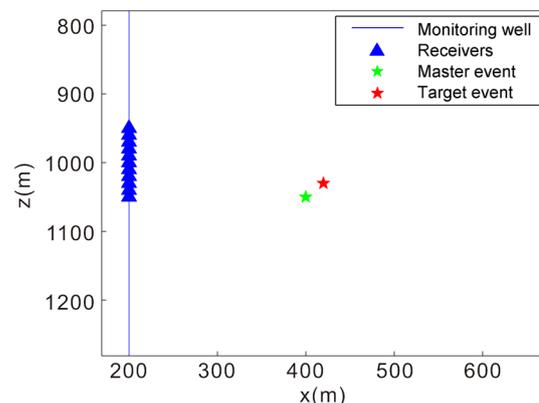

Fig. 1 Acquisition geometry for the 2D model with the target event (red pentagram) to be located, the master event (green pentagram) to be known and receivers (blue triangles). The x- and z-axes denote the offset and depth directions, respectively
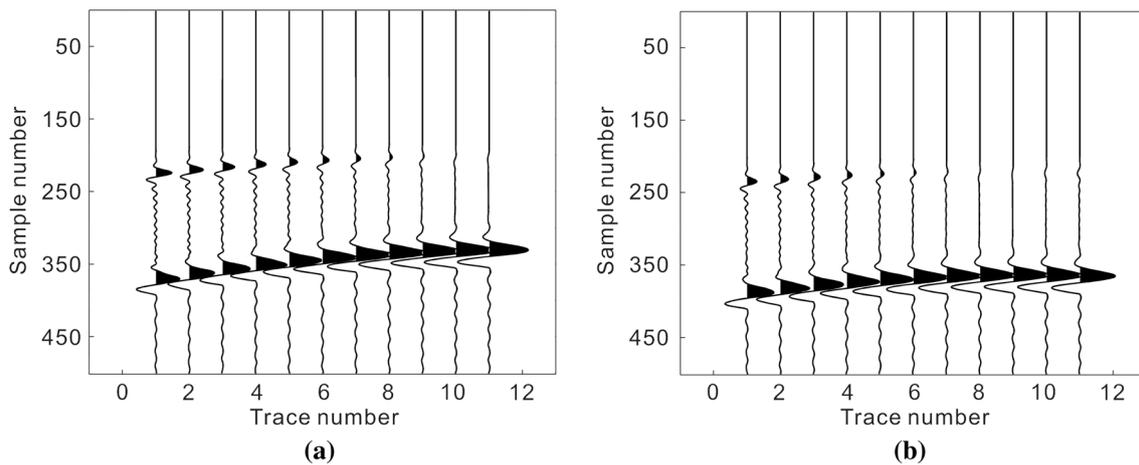
Fig. 2 Record section of the synthetic noise-free microseismic data (vertical component) for the **a** master and **b** target events

event were set at $(x, z) = (400, 1050)$ m and $(x, z) = (420, 1030)$ m, respectively. Two-dimensional elastic microseismic records were simulated using staggered-grid finite differences (Graves 1996). The vertical component records of the synthetic noise-free data for the master event and target event are shown in Fig. 2a, b, respectively. Here, we defined the grid search area which was 200 m both in x and z directions centered on the location of the true event. Meanwhile, the search step length was 1 m both in x and z directions. Then, we calculated the values of different objective functions for all discrete points in the search area using the grid search method.

To analyze the global convergence and resolution of the objective functions based on the conventional and improved waveform stacking constructions mentioned by Eqs. (3) and (4), we specially discussed the distribution of

the waveform stacking functions from the conventional and improved methods without noise and velocity model errors. Figure 3 clearly shows the distribution of the conventional and improved waveform stacking functions calculated by using Eqs. (3) and (4) based on the noise-free target event record in Fig. 2b, respectively. As we can see from Fig. 3, the deep red zone denotes the potential microseismic event location area most possibly. From Fig. 3a, b, it can be easily observed that the value in red area, calculated through the improved waveform stacking method, is obviously compressed as compared to that of the conventional waveform stacking method. As a result, the convergence and resolution of the improved waveform stacking function are far better than those of the conventional waveform stacking function.

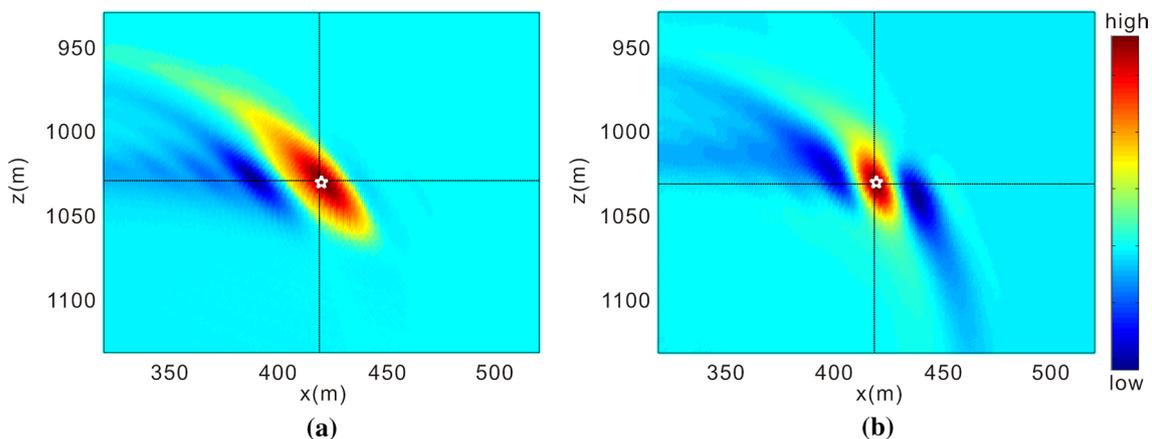To test the noise immunity of the conventional and improved methods, we applied Monte Carlo relocation

Fig. 3 Waveform stacking function distribution based on the **a** conventional and **b** improved methods. The color scale indicates the waveform stacking function value (deep red for the maximum values and deep blue for the minimum values). In each subplot, the white

pentagrams denote the true location of synthetic event (target event) and the intersections of black dashed lines, namely the positions of the maximum waveform stacking function values, represent the calculated locations
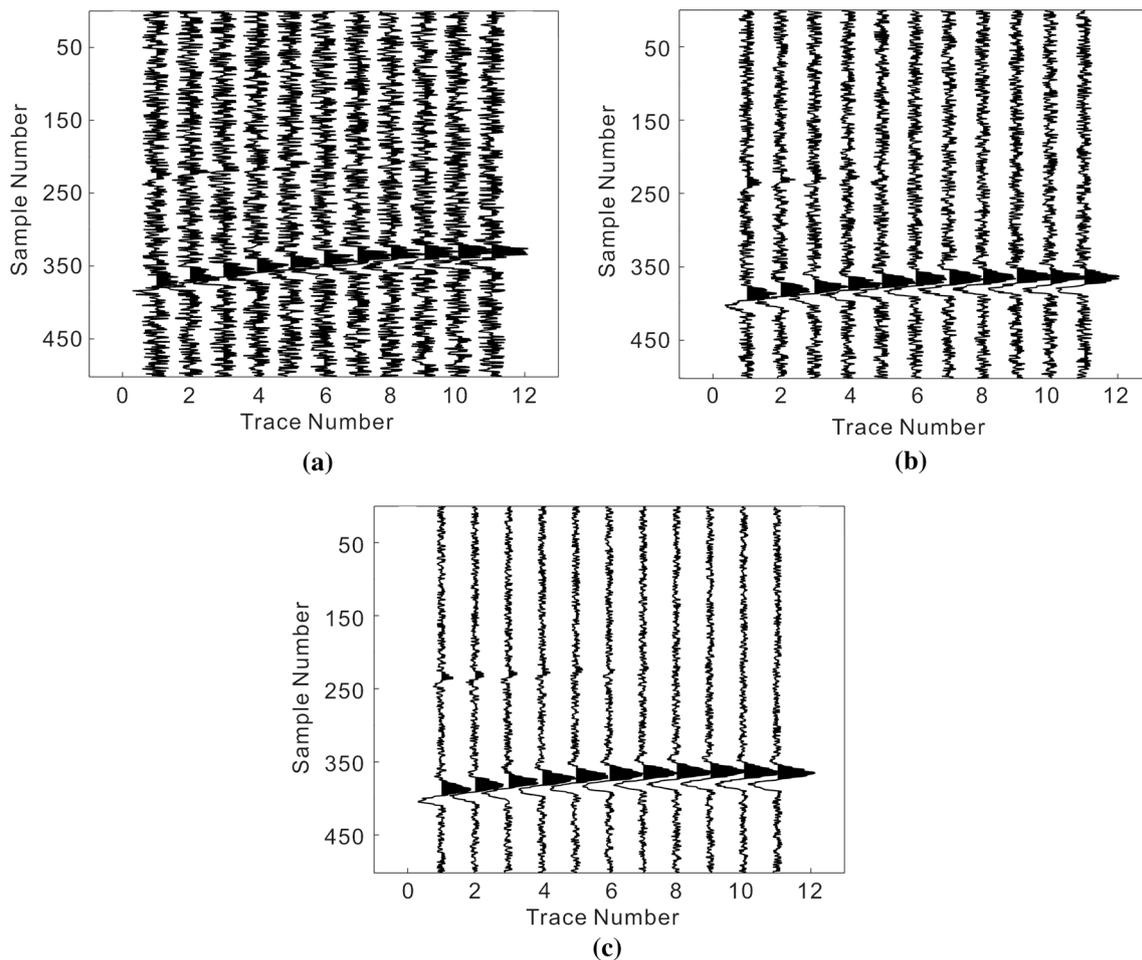
Fig. 4 Record section of one separate noisy data of the target event with the SNR = **a** 0.5, **b** 1 and **c** 2 as one realization mentioned in Fig. 5. The record section of the different SNR data is generated by adding random noise with different amplitudes to noise-free data of the target event as in Fig. 2b

with 100 realizations of random noise for different SNR. Figure 4 shows one separate noisy data of the target event for SNR = 0.5, 1 and 2, which is only an input data for one realization mentioned in Fig. 5.

It is clear from Fig. 5 that the relocation results based on the conventional and improved methods gradually converge to the true event position as the SNR increases. What is more, the relocation results based on the improved method are more convergent to the true event position than those of the conventional method for all SNR (0.5, 1 and 2). In addition, Fig. 6 shows the comparison with absolute location errors through the use of the same corresponding 100 realizations as in Fig. 5. From Fig. 6, it is obvious that the accuracy of the improved method is much higher than that of the conventional method because it estimates less absolute location error compared to the conventional method for the same corresponding SNR as in Fig. 5.

As we know, the accuracy of many traditional location methods greatly depends upon the precision of the velocity model. But relative location methods can generally mitigate the negative effect of an inaccurate velocity model. Hence, the conventional and improved methods in this paper were further extended by using the master-event technique (Grigoli et al. 2016).

To test the sensitivity of the conventional and improved methods with the master-event technique to the velocity model errors, we conducted a Monte Carlo relocation with 100 realizations of random velocity perturbation. Both the P- and S-wave velocity values were perturbed by (-3%, 3%), respectively. In order to isolate the effect of velocity method errors, we used the noise-free data.

The upper row of Fig. 7a, b shows the relocation results based on the conventional and improved methods without applying the master-event time corrections calculated through the master-event technique, which are greatly divergent. However, after applying the master-event time corrections, the location results become more convergent (Fig. 7c, d). Thus, with the addition of master-event time
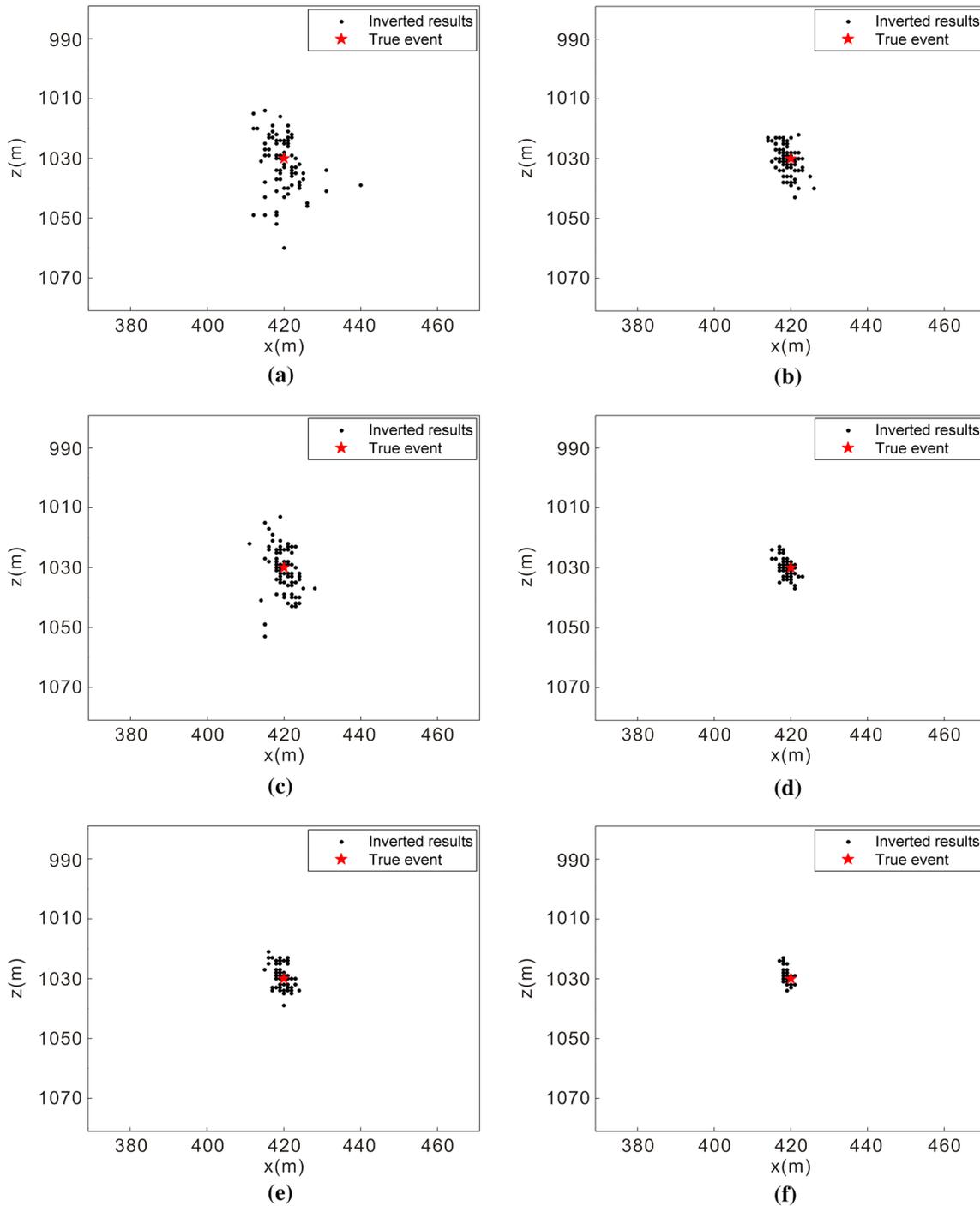
Fig. 5 One hundred realizations from Monte Carlo relocation using the conventional [left column (**a**, **c**, and **e**)] and improved [right column (**b**, **d**, and **f**)] methods, under random noise for different SNR. The SNR = (**a**, **b**) 0.5, (**c**, **d**) 1 and (**e**, **f**) 2. The red pentagram and black dots represent the true event (target event) and inverted results (calculated locations), respectively

corrections, the location accuracy has been dramatically improved under the velocity model errors. Consequently, the master-event technique can make our waveform stacking location method more reliable under velocity model inaccuracies.

Field data application

For verifying the feasibility of the presented method in field data processing, we performed one real data test with microseismic data obtained during the hydraulic
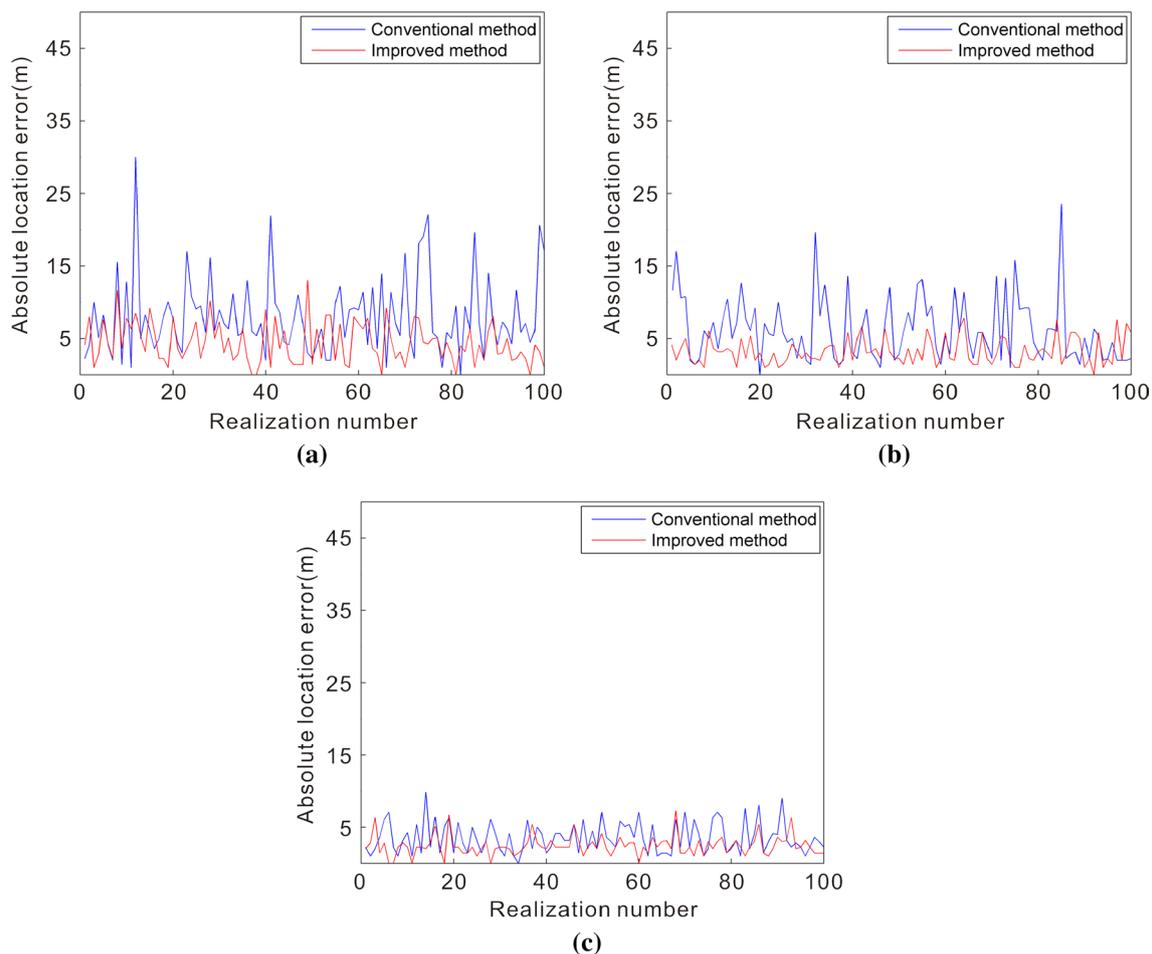
Fig. 6 Comparison with absolute location errors based on the same corresponding 100 realizations as in Fig. 5. The SNR is **a** 0.5, **b** 1 and **c** 2

fracturing of tight sandstone from a Chinese oil field. An array of ten three-component (3C) receivers was placed in an inclined monitoring well whose wellhead was located at $x = 415.9$ m and $y = 608.1$ m. The depth of the placed array ranged from 2455.1 to 2545.1 m with 10-m spacing intervals. Afterward, the midpoint location of the perforation interval was (360.6, 281.9, 2538.0) m. To calibrate the initial velocity model and the orientation of 3C receivers, we used the known perforation information. For determining the global maximum value of the waveform stacking function, we implemented the global grid search algorithm based on back-azimuth constraint. The volume set for the three-dimensional (3D) grid search was 400 m in the x, y and z directions centered on the midpoint of the perforation interval. Then, we chose the search step size 1 m in all three directions (x, y and z directions) and set the acceptable back-azimuth error 4° for the conventional and improved methods, respectively. The perforation event and high SNR events were selected as master events for this study.

According to the location flow described in "Location flow" section, more than 170 microseismic events were located by utilizing the conventional and improved methods, respectively. The location results of the conventional and improved methods are shown in Fig. 8a–f in the form of 3D space, top view of xy-plane and side view of xz-plane.

As we can observe from 3D space (Fig. 8a, b), although the location results based on the methods mentioned above are near to the perforation interval, the inverted results determined by the improved method are more spatially clustering compared with those found through the conventional method. We can clearly find that the location results based on the improved method (Fig. 8d) are significantly more clustering than those of the conventional method (Fig. 8c) from top view in xy-plane. Moreover, as we can notice from side view in xz-plane (Fig. 8e, f), the inverted results obtained through the use of the improved method are clustered very well than those determined by the conventional method. The real data test suggests that the location results obtained through the improved method are more acceptable.
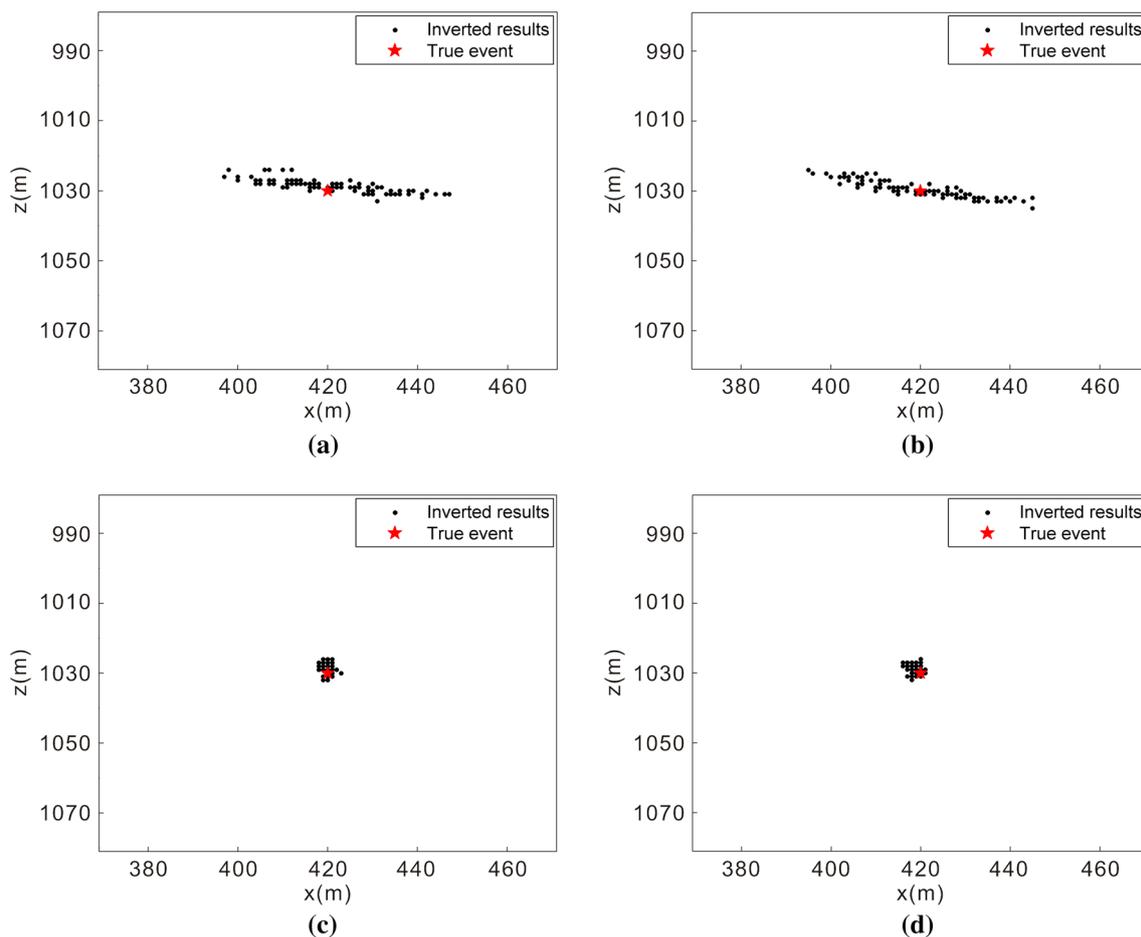
Fig. 7 One hundred realizations from Monte Carlo relocation under random velocity perturbation. **a** and **b** show the relocation results using the conventional and improved methods without applying the master-event time corrections. **c** and **d** show the relocation results using the conventional and improved methods applying the master-event time corrections

Discussion

Our results indicate that the proposed migration-based location method through improving the combination way of *P*- and *S*-wave waveform stacking can significantly increase the resolution and convergence of the objective function on the basis of a predefined time window centered on the arrival time of signals (Eaton et al. 2011), polarization correction algorithm (Kim et al. 2017) and master-event technique (Grigoli et al. 2016).

In the synthetic data example, the distributions of the conventional and improved waveform stacking functions without noise and velocity model errors clearly indicate that the global convergence and resolution of the improved waveform stacking function (Fig. 3b) are much better than those of the conventional waveform stacking function (Fig. 3a). Subsequently, under different SNR (0.5, 1 and 2) tests, the improved method has achieved more convergent relocation results compared with the conventional method (Fig. 5) and the former has also obtained higher precision results than

the latter (Fig. 6). Therefore, the improved method can be more resistant to noise than the conventional method. Furthermore, the former method can lead to better stability than the latter under varying noise levels. Moreover, even under relatively low SNR, the synthetic data results confirm the accuracy of this proposed method. Hence, it is clear that the above two methods can mitigate the dependency upon the precision of the velocity model after the application of master-event time corrections, which is in agreement with the conclusion gained by Grigoli et al. (2016). For simplicity and reducing the computational cost, we chose 2D model in the synthetic test. Once the azimuth information is estimated, these conclusions may easily be confirmed in 3D synthetic model test.

In the field data application, the location results obtained by the improved method are more clustered than those based on the conventional method both from top view in *xy*-plane (Fig. 8c, d) and side view in *xz*-plane (Fig. 8e, f). Thus, our inverted results also indicate that the improved method can more accurately locate the microseismic events than the
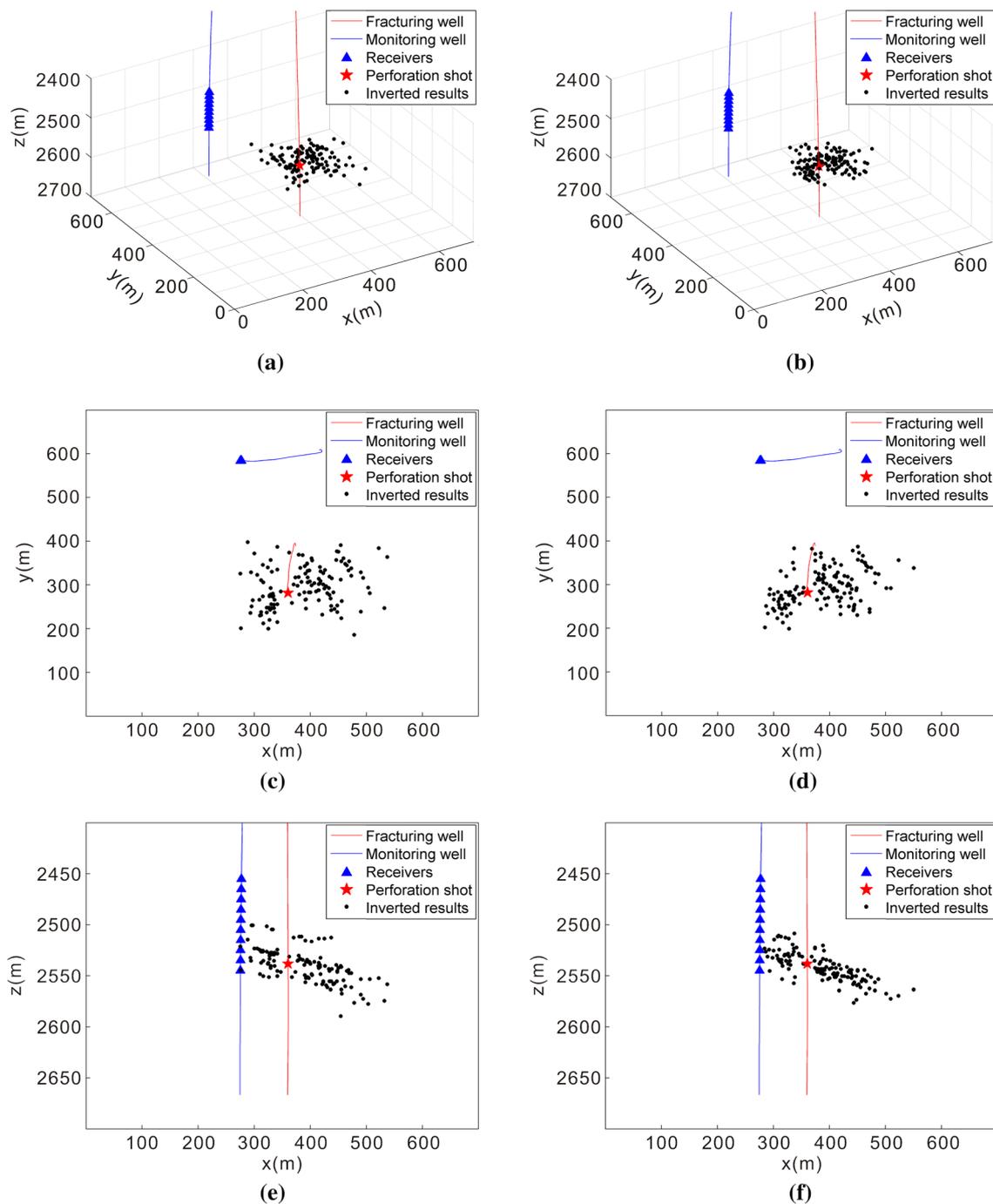
Fig. 8 Location results based on the conventional [left column (**a**, **c**, and **e**)] and improved [right column (**b**, **d** and **f**)] methods. Three-dimensional space view for using the **a** conventional and **b** improved methods, respectively. Top view in *xy*-plane for the corresponding methods mentioned above is observed in **c** and **d**, respectively. Side view in *xz*-plane for the corresponding methods mentioned above is observed in **e** and **f**, respectively. The red pentagram and blue triangles indicate the perforation shot and receivers, respectively. The blue and red curves denote the monitoring and fracturing wells, respectively. The black dots represent inverted locations determined by the conventional and improved methods. The *x*-, *y*- and *z*-axes denote the east, north and depth directions, respectively

conventional method, which is almost consistent with the above conclusion from the synthetic test.

To sum up, the proposed method can obviously enhance the location precision compared with the

conventional method under varying noise levels. What is more, the former method seems to be more applicable to process low SNR microseismic data than the latter method. This may provide the reliable foundation for

precisely implementing the location procedure even in the presence of low SNR.

Conclusion

This study presents a migration-based location method using improved waveform stacking with polarity correction based on a master-event technique. Compared with the conventional waveform stacking location method, the proposed method can improve the convergence of the objective function. Furthermore, this new method may effectively provide more reliable and accurate location results even in low SNR circumstances. Thus, this method may be completely robust and practicable for relatively low SNR microseismic data in the hydro-fracturing monitoring. Applications to synthetic and field data have validated the feasibility and reliability of the new proposed method.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Anikiev D, Valenta J, Stanek F et al (2014) Joint location and source mechanism inversion of microseismic events: benchmarking on seismicity induced by hydraulic fracturing. Geophys J Int 198(1):249–258

Drew J, White RS, Tilmann F et al (2013) Coalescence microseismic mapping. Geophys J Int 195(3):1773–1785

Eaton DW, Akram J, St-Onge A et al (2011) Determining microseismic event locations by semblance-weighted stacking. In: Presented at the proceedings of the CSPG CSEG CWLS convention

Gajewski D, Anikiev D, Kashtan B et al (2007) Localization of seismic events by diffraction stacking. SEG Tech Progr Expand Abstr 26:1287–1291

Gharti HN, Oye V, Roth M et al (2010) Automated microearthquake location using envelope stacking and robust global optimization. Geophysics 75(4):MA27–MA46

Graves RW (1996) Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences. Bull Seismol Soc Am 86(4):1091–1106

Grigoli F, Cesca S, Amoroso O et al (2014) Automated seismic event location by waveform coherence analysis. Geophys J Int 196(3):1742–1753

Grigoli F, Cesca S, Krieger L et al (2016) Automated microseismic event location using master-event waveform stacking. Sci Rep 6:25744

Kao H, Shan SJ (2004) The source-scanning algorithm: mapping the distribution of seismic sources in time and space. Geophys J Int 157(2):589–594

Kim J, Woo JU, Rhie J et al (2017) Automatic determination of first-motion polarity and its application to focal mechanism analysis of microseismic events. Geosci J 21(5):695–702

Li L, Chen H, Wang XM (2016) Relative elastic interferometric imaging for microseismic source location. J Geophys Eng 13(5):733–744

Moser TJ (1991) Shortest path calculation of seismic rays. Geophysics 56(1):59–67

Mao QH, Wang P, Azeem T (2019) Microseismic event location using an improved global grid search and its extended method in a downhole monitoring system. J Geophys Eng 16(1):159–174

Shi PD, Angus D, Rost S et al (2019) Automated seismic waveform location using multichannel coherency migration (MCM)–I: theory. Geophys J Int 216(3):1842–1866

Trojanowski J, Eisner L (2017) Comparison of migration-based location and detection methods for microseismic events. Geophys Prospect 65(1):47–63

Xu JC, Zhang W, Chen XF et al (2019) Minimum semblance weighted stacking with polarity correction for surface microseismic data processing. Lead Edge 38(8):630–636

Rock-physics-based estimation of quartz content in the Shenhu area, South China Sea

Wei Deng[1,2] · Jinqiang Liang[2,3] · Yulin He[1] · Zenggui Kuang[2] · Miaomiao Meng[3]

Abstract

Estimation of quartz content (QC) is important for gas hydrate production. However, the existing methods pay more attention to estimate saturations of hydrate or free gas instead of QC. QC is difficult to be estimated because of its limited and unclear influences on velocities in the hydrate-bearing sediments. A workflow was proposed to estimate QC from core to logging to seismic inversion whose core technologies were an unconsolidated anisotropic model (UAM) and an inverse modeling approach. We used the UAM to construct the quantitative relationships between physical properties including QC and velocity. Then, the velocities of the reservoirs were obtained by logging-constrained seismic inversion. Finally, QC can be scaled by the inverse modeling of the UAM. To build the UAM, we analyzed the physical properties of hydrate reservoirs based on the cores and logging while drilling (LWD) data in the Shenhu area, South China Sea, and characterized unconsolidated sediments with horizontally layered hydrates and gas occurrences. The calculated P-velocities and S-velocities from the UAM agreed with the LWD data when the input variables were QC, porosity, hydrate saturation, and gas saturation. Conversely, for a group of P-velocity and S-velocity from seismic inversion, the corresponding QC can be scaled out as well as the other parameters based on the UAM, which was defined as an inverse modeling. Because the significant parameters such as hydrate saturation have been considered as independent variables in the model, we can effectively avoid the correlation between QC and the others. Prediction multiplicity can be reduced. The estimated QC was consistent with the drilling and geological understanding in the field application, indicating that the method proposed is effective and practical.

Keywords Gas hydrate · Quartz content · Gas saturation · Rock physics · Seismic inversion · Shenhu

Introduction

Natural gas hydrate is an ice-like solid formed by the adsorption of small gas molecules in a cage structure. Among them, methane hydrate is the most widely distributed hydrate in the world (Dickens 2001; Sloan and Koh 2007). Marine gas hydrates generally occur within the top few hundred meters of sediments in continental margins and continental permafrost regions worldwide (Xu and Rupple 1999). As a clean alternative energy source with huge resource potential,

methane hydrate has attracted widespread attention. Shenhu area, with the first successful gas hydrate production test in the South China Sea (SCS) (Li et al. 2018), has been considered as one of the most promising areas with commercial potential (Zhang et al. 2017).

Quartz content (QC) estimation is important for the exploration of hydrate reservoirs with commercial potential. Boswell and Collett (2006) suggested four types of reservoir: (1) sand reservoir, (2) clay-rich fractured reservoir, (3) submarine sediments with mid-high saturation gas hydrates, and (4) low saturation hydrates dispersed in impermeable clays. The first two types of reservoirs are most potential to achieve prospective exploration and commercial utilization because they can provide the permeability required for a high concentration of gas hydrate accumulation (Moridis et al. 2009). Rock physics modeling is an effective way to estimated QC. Hydrate occurrence as the main character in a hydrate reservoir must be considered when estimating QC. Two types of gas hydrate morphology exist in a variety of

✉ Wei Deng
hahens@163.com

1 Guangzhou Marine Geological Survey, China Geological Survey, Guangzhou, China

2 Natural Gas Hydrate Engineering Technology Center, China Geological Survey, Guangzhou, China

3 Southern Marine Science and Engineering Guangdong Laboratory, Guangzhou, China

sediments ranging from fine-grained clays to coarse-grained sands. Pore-filling gas hydrates in fine-grained clays are wildly distributing in Shenhu area (Zhang et al. 2017), which partially replace the pore fluid and act as a part of the pore fluid or a part of the solid frame, but little change to the sediment structure (Holland et al. 2008). Several pore-filling gas hydrate reservoirs in fine-grained clay–silt and coarse-grained sands have also been successfully identified in the Gulf of Mexico (Boswell et al. 2012), offshore India (Riedel et al. 2010), offshore Korea (Ryu et al. 2013), and the SCS (Zhang et al. 2017). The velocities of pure hydrate have been measured in the laboratory (Waite et al. 2000), whose P wave velocity is 3650 m/s, S wave velocity is 1890 m/s, and density is 900 kg/m^3. The velocities are higher than the velocities of hydrate-bearing sediments, and the accumulation of a large amount of hydrates can decrease the Poisson's ratio (Hamilton 1980). The velocities also depend on the hydrate formation pattern. If the hydrates cement the matrix, the velocities increase significantly (Helgerud et al. 1999; Sava and Hardage 2006). However, free gas in the pore affects the velocity more complicatedly. Even if there is less gas (< 1%) in the pores, the P wave velocity changes significantly (Domenico 1976), and S wave velocity keeps unchanged. The coexisting of free gas and hydrate has been found in the Shenhu area, SCS (Wang et al. 2013); therefore, the variations of hydrates and free gas make the elastic characteristics more complicated. Since gas hydrates exhibit relatively high P wave velocity compared to pore-filling fluids such as water (Waite et al. 2000), and free gas owns much lower P wave velocities, so P wave velocity and S wave velocity are the most commonly used parameters to predict gas hydrate saturation and gas saturation both in fine-grained clays and in coarse-grained sands (Lee and Collett 2011), using a time average equation (Miller et al. 1991; Wood et al. 1994), a cement model (Dvorkin et al. 1994), effective medium theory (Helgerud et al. 1999; Jakobsen et al. 2000), a three-phase Biot-type equation (Carcione et al. 2005; Carcione and Tinivella 2000), weighted average equations (Lee and Collett 1999; Ojha and Sain 2007), or a simplified three-phase equation (Lee 2008; Lee and Waite 2008).

Due to the significant influences on the elastic characteristics caused by gas hydrate and free gas occurrences, the existing methods are effectively used to estimate saturation of hydrate or free gas instead of QC. With the exploration demand of hydrate reservoirs with commercial potential, the prediction of QC has become more and more important. But the relationship between QC and velocity is unclear; nearly no studies have been done to estimate QC in the hydrate-bearing sediments with rock physics and seismic data. Sediments with various combinations of hydrate saturation, gas saturation, porosity, and QC may own similar velocity (Kumar et al. 2007). The above models often assume that

QC is a constant or simplified by empirical or semiempirical formulas. Besides, the above models do not consider sediment grain, unconsolidation, and hydrate or gas occurrence simultaneously, which may reduce accuracy.

In this paper, we first analyzed the hydrate occurrence through Geo-vision resistivity (GVR) imaging and X-ray, and element capture logging and X-ray diffraction (XRD) were used to obtain QC. The average particle size and sorting coefficient were obtained through particle size analysis. Then, we proposed an unconsolidated anisotropy model (UAM) in which layered hydrates manifested as vertical transversely isotropic (VTI) media were added to an unconsolidated matrix. The Hertz–Mindlin (H–M) particle contact model (Mindlin 1949) was used to construct the unconsolidated matrix with high Poisson's ratio (Domenico 1977) where grain size and sorting were considered. Then, based on the periodic layered medium and the Backus average (Backus 1962), the modulus of the dry matrix with horizontal layered hydrates was calculated. Finally, the Brown–Korringa equation (Brown and Korringa 1975) was used to add fluid to dry micropores. To reduce errors caused by the inappropriate correlation between multiple parameters, we made QC as well as hydrate saturation (HS), gas saturation (GS), porosity-independent variables in the proposed model. P wave velocities, S wave velocities, and impedances of hydrate-bearing sediments with any physical combinations can be obtained using UAM. Finally, based on an inverse modeling method, we can estimate QC with high precision.

Geological setting

Shenhu area is located in the northern slope of the South China Sea (Fig. 1a), which is structurally affiliated with Zhu-II Depression of the Pearl River Mouth Basin (Wu et al. 2009). The water depth ranges from 600 to 1800 m. The topography of the seafloor is quite undulating and generally appears to be sloping from northwest to southeast (Fig. 1b). Due to gravity flow and underflow, large-scale submarine landslides happened, causing a pattern of submarine ridges and erosion trenches arranged in a north–south direction. The bottom simulating reflectors (BSRs) have been identified mainly in the ridges, knolls, and edges (Fig. 1c). The temperature of seafloor is 2–4 °C, and the geothermal gradient is 45–67 °C/km (Zhang et al. 2017). Hydrates have been found at 150–300 m below seafloor (mbsf) by LWD and coring.

The tectonic evolution of the Pearl River Mouth Basin can be divided into two stages with 23.8 Ma as the boundary: the Eocene–Oligocene rifting stage and the Neogene–Quaternary post-rifting stage (Ru and Pogott 1986). There were three peak periods of fault activity in the study area: Paleozoic (or earlier)–early Oligocene, early Miocene–mid
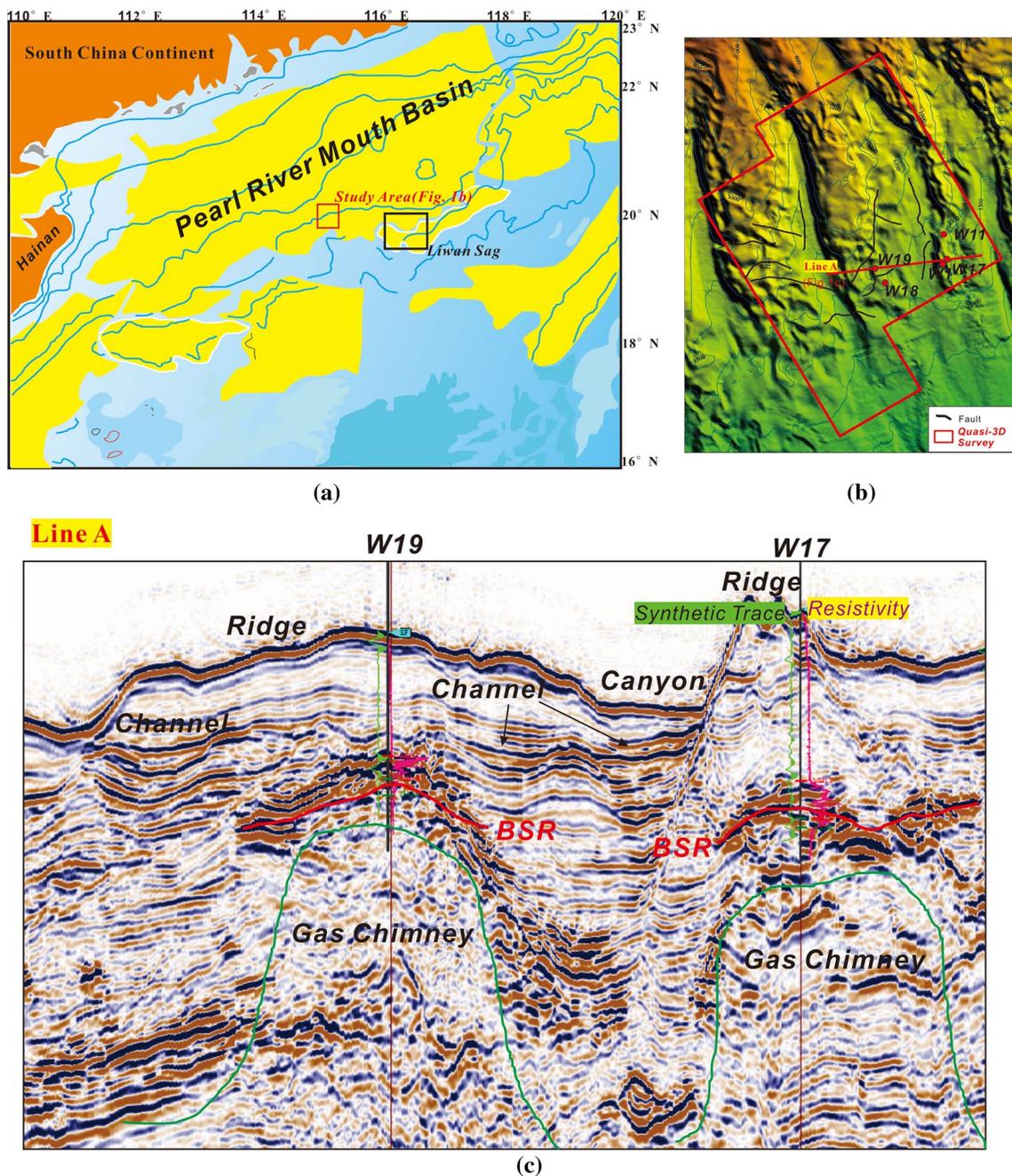
(a)

(b)

(c)

Fig. 1 Location of the study area. **a** The location of the study area. The red rectangle indicates the location. **b** Enlarged view of the red rectangle in Fig. 1a. The base map shows the submarine landform. Location of the wells used in the paper is displayed with red circles and **c** seismic section of Line A. W17 and W19 located on the line. Red curves refer to the BSRs. Resistivity logs are displayed after time–depth conversion. The synthetics agree with the seismic data, indicating that time–depth conversion is rational

Miocene and late Miocene, and faults were particularly active since late Miocene (Wu et al. 2005). When these faults were active, a large number of cracks were generated, making surrounding area high porosity and high permeability. The fault system connected the deep gas source with the shallow gas, providing a channel for fluid migration (Kuang et al. 2018). The neotectonic movement also led

to the extensive development of gas chimney. Gas chimney used to be a pressure relief channel for hot fluids, which was closely related to gas seepage, migration, and accumulation.

The gas composition of the hydrate samples was characterized by the absolute dominance of methane (Zhang et al. 2017). The methane content is higher than 98.5%, and C1/C2 increases with depth overall. At the same time, there is a

sudden change near the top of the hydrate reservoir. C1/C2 of the gas samples in the hydrate reservoir is less than 300, but the value in the shallow sediments is greater than 1000. It is speculated that the gas source of hydrate is mainly the mixture of pyrolysis gas and microbial gas, and the variation of ethane and propane is similar, indicating that the proportion of thermogenic gas gradually increases as the depth increases (Wei et al. 2018). According to previous study on source rock comparison in the Liwan Sag (He et al. 2008), combined with the thermal evolution and hydrocarbon generation history, the thermogenic gas in the natural gas hydrate in the Shenhu Sea area is likely to be mainly derived from the lacustrine source rocks of the Enping and Wenchang Formations.

Since the Miocene, the tectonic subsidence rate was abnormally high due to the Dongsha Movement (Yu et al. 2012). The sedimentary filling sequence in the same period was dominated by marine sediments, including shelf-margin delta, deepwater turbidite channel, and

mass transport deposit (Ludmann et al. 2001). During this period of geological history, the structure-sedimentary background of the Shenhu area was consistent with the Pearl River Mouth Basin. According to drilling data, the average deposition rate of the Pleistocene in this area was 6.46 cm/k.y., and the average deposition rate of the Pliocene was 3.93 cm/k.y (Yu et al. 2012; Zhang et al. 2017). We studied stratigraphic sequences and sedimentary facies for the Pliocene–Quaternary and Quaternary in the study area (Fig. 2a). According to previous studies (Yu et al. 2012; Wu et al. 2005), the Upper Miocene–Quaternary in the study area was divided into three sequences, of which the sequence interfaces of T3 and T2 represent the two episodes of the Dongsha Movement in late Middle Miocene and Pliocene (10.4 Ma and 5.2 Ma), respectively, and T1 refers to the end of the Pliocene (1.64 Ma). Gas hydrates were mainly accumulated in the unconsolidated sediments of the Late Miocene and Pliocene (Su et al. 2015), or T1–T2. Sedimentary facies were delta
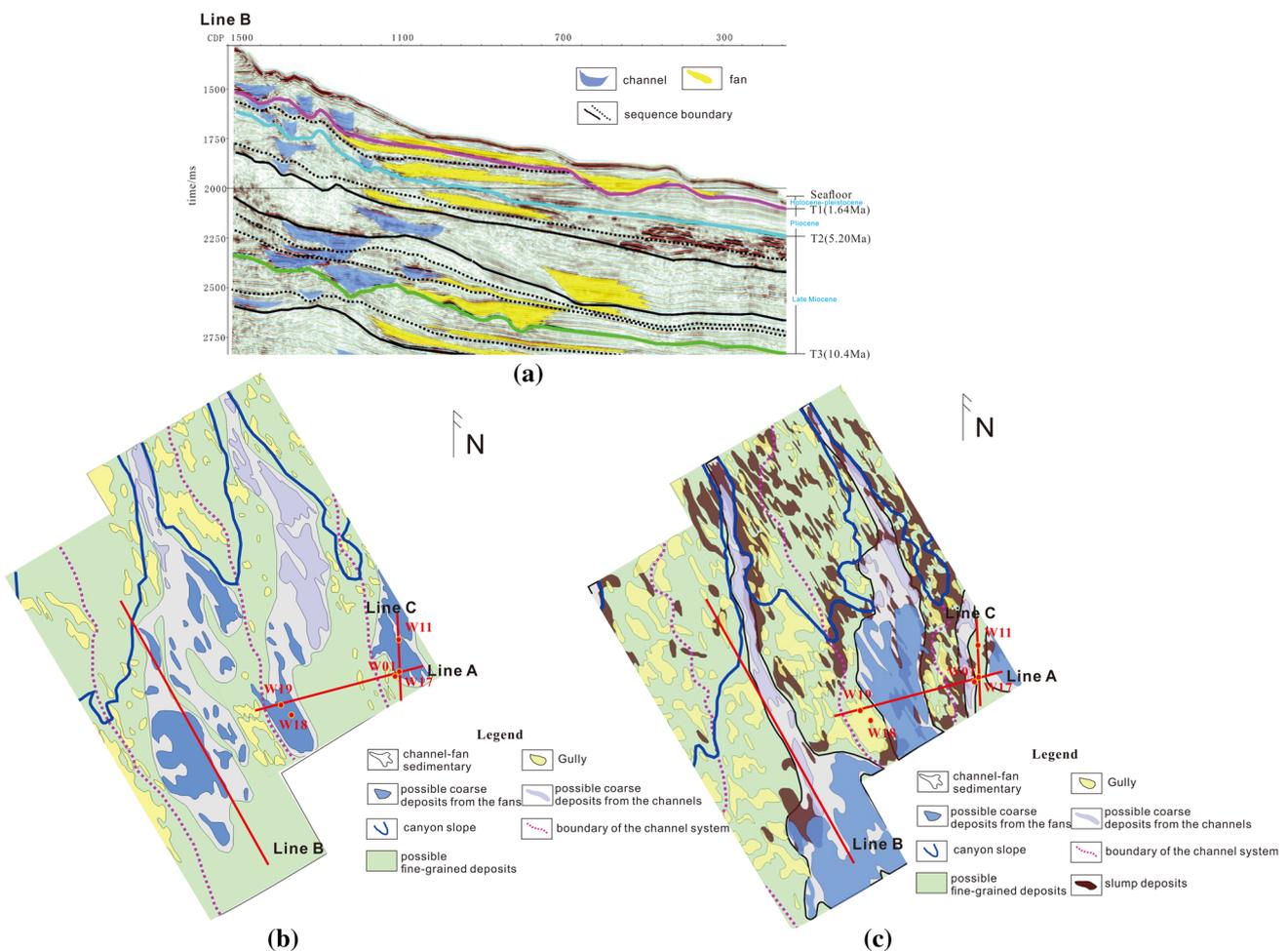

Fig. 2 Sedimentary facies for the Pliocene–Quaternary and Quaternary in the study area. a Sedimentary facies and stratigraphic sequences in a seismic profile. b Sedimentary facies in the Pliocene–Quaternary; and c sedimentary facies in the Quaternary

front with high porosity, underwater fan, channel, and slump deposition. The strong hydrodynamic environment mainly occurs at the initial stage of a sequence, causing the coarse-grained sedimentation by gravity flow separation to be above the sequence interface. Therefore, the relatively coarse-grained sediments in the deep-sea slope environment are mainly deposited at the bottoms of the channel or the roots of the fan (Kuang and Guo 2011), as shown in Fig. 2b. Besides, erosion gullies of different sizes were distributed on the levees on both sides of the channel. The strong underflow in the early stage of a sequence could bring coarse sediments to these lowlands close to the gullies. On the plane, there were three channels in the study area (Fig. 2b), and three large-scale underwater fans were developed at the fronts of the channel. The Quaternary inherited sedimentary pattern in the Pliocene and mainly developed channels, fans, erosion gullies, and slump deposits (Fig. 2c). Since the hydrates ever found occurred before T2 (Zhang et al. 2017), possible coarse-grained deposits in the study area were identified (Fig. 2b, c). Wells W11, W17, W18, and W19 are all located in the possible coarse-grained sediments of the fan developed during T1–T2. By analyzing the relationship between the particle size and hydrate saturation in Shenhu sea area, Chen et al. (2011) pointed out that there is a good correspondence between the reservoirs with high saturation of hydrate and sand, coarse silt layers. Coarse-grained sediments are more conducive to hydrate formation.

Method

The unconsolidated anisotropic model

Characteristics of hydrate-bearing sediments

1. Density of unconsolidated sediments

A total of 150 samples from wells W11, W17, W18, and W19 were tested for acquiring natural water content, density, and porosity. As shown in Table 1, the porosity of hydrate-bearing sediments is 39–71%, the maximum water content is about 88%, and the density is 1.4–2.1 g/cm^3. These data provided references for the calculation of dry rock modulus. When calculating the modulus, the total density of different mineral components and porosity should be within the range of the densities and porosities in Table 1.

2. Lithology and particle size

We used the H–M particle contact model to characterize the equivalent modulus of unconsolidated sediments (Winkler 1983; Walton 1987). Particle size and number of particle contacts are important parameters in this model.

Figure 3 shows the grain size analysis of well W11. The mean diameter was 6.17–7.88Φ, and the kurtosis was 0.87–1.18. The sorting coefficient was between 1.43 and

Table 1 Water content, density, and porosity of shallow samples in the study area

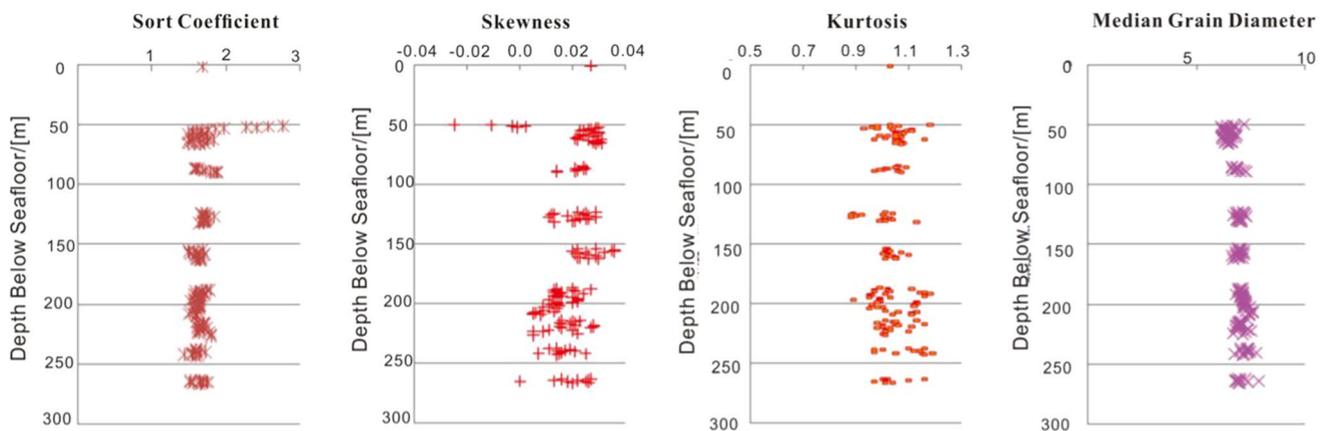| Well name (%)	Natural water content	Wet density (g/cm^3)	Dry density (g/cm^3)	Porosity (%)
W19	20.0–73.5	1.42–1.92	1.18–1.44	47.3–69.9
W18	19.6–86.1	1.4–1.86	1.17–1.39	48.3–69.7
W11	21.1–83.6	1.54–2.09	1.27–1.65	39.1–68.8
W17	23.0–87.8	1.48–1.98	1.20–1.57	42.3–71.0

Fig. 3 The sort coefficient, skewness, kurtosis, and the median grain diameters in well W11

2.78Φ, indicating that the sorting was poor, so the number of particle contacts should be larger than that of conventional sandstone. As a result, we assume that the mean diameter was 7.5Φ, and the number of contacts was 6.

Figure 4a shows the contents of sand, silt, and clay of the wells W17 and W19. At about 140–160 mbsf, the sand content increases significantly, and the silt content is greater than 50% (Fig. 4a), making QC close to 50% (Fig. 4b). It indicates that there are reservoirs with high

QC in this area, which should be due to the underwater fan deposits (Fig. 2b).

3 Hydrate occurrence

Different types of hydrate occurrence were identified in GVR images. The horizontal thick hydrates in the study area were common (Fig. 5a), which were generally distributed on the top of the hydrate reservoir and behaved as continuous
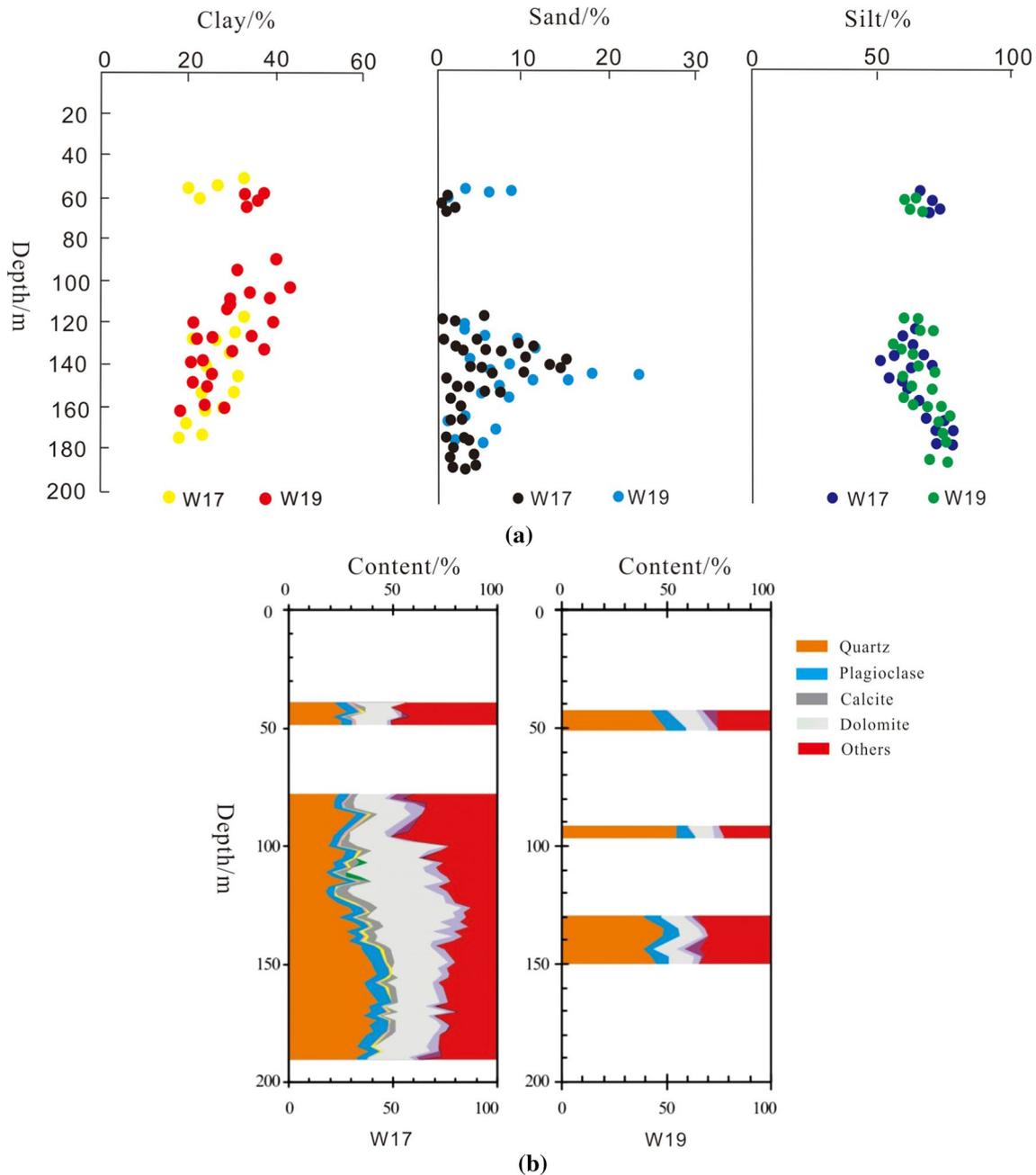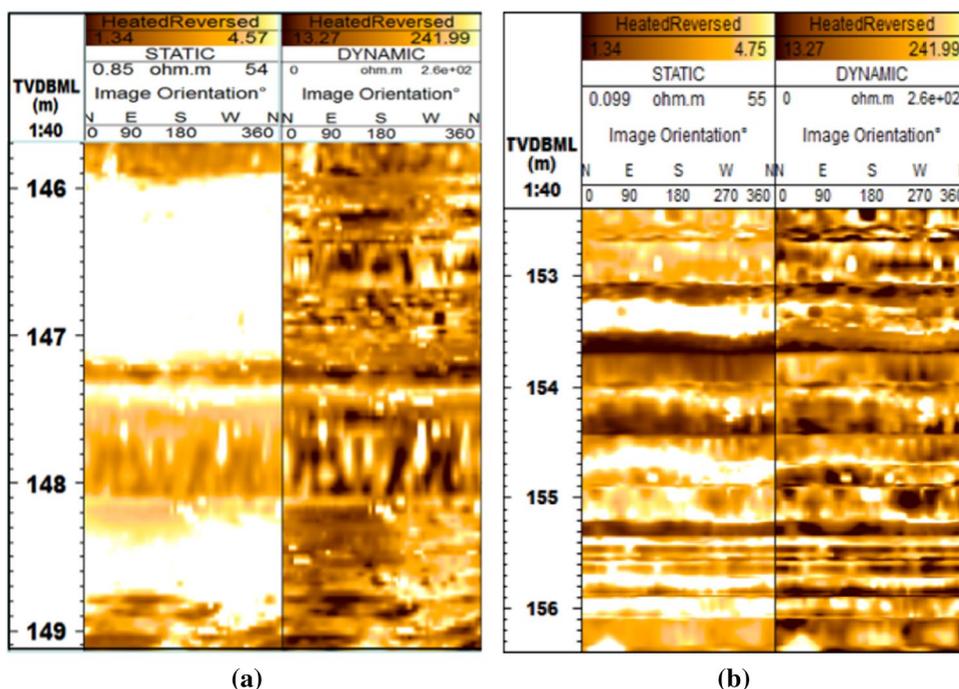

Fig. 4 Lithological components and mineral content in wells W17 and W19; **a** sand content, silt content, and clay content obtained by particle size analysis; **b** QC and other mineral components obtained by XRD

Fig. 5 GVR image of hydrates with different thicknesses; **a** GVR image of thick hydrates in well W19; **b** GVR image of thin hydrates in well W18

(a)   (b)

highlights on the static images. Thick layered hydrates with different saturations were found in W17, W11, and W19 (Wang et al. 2011, 2013). Well W11 was confirmed to contain a very thick (~ 90 m) interval of gas-hydrate-bearing sediment (average concentration 35% of pore volume) by resistivity. Gas hydrate concentrations were also calculated from porewater freshening, based on chloride as a conservative species. Significant fresh departures from the chlorinity baseline were observed between 131–202 mbsf. The maximum gas hydrate concentration from porewater freshening was 47% of pore volume in well W11 (Fig. 6). Thinner hydrate reservoirs were found in wells W17, W18, and W19. Maximum gas hydrate concentration from porewater freshening was 66% of pore volume in well W18. Pressure core depressurization has also been done to calculate the hydrate saturation, which allowed quantification of the total amount of methane that was present in a given pressure core and estimation of the quantity of methane hydrate or free methane gas. The calculation results based on porewater freshening were consistent with the results of the depressurization of the pressure cores.

Thin layer hydrates were found in almost all mentioned wells, distributing hundreds of meters below the seafloor. Such thin layered hydrates may not be identified through conventional logging curves (neutron, density, P-sonic), but they appear as bright highlights on the GVR image (Fig. 5b). From the X-ray images of the pressure cores in wells W17, W18, and W19 (Fig. 7), the hydrates were mainly horizontally distributed in a more macroscopic view.

Rock physics modeling

Based on the physical property of hydrate reservoirs, we built an unconsolidated matrix with multiple minerals including quartz, hydrate, and clay. The modulus and densities of different components are shown in Table 2. Then, we constructed a periodic layered matrix with connected pores (Fig. 8). Specifically, we divided the modeling into three steps.

Modulus of hydrate-bearing sediments with different lithologic grains According to the H–M model, the bulk modulus and shear modulus of randomly stacked equal-sized spherical particle aggregates can be expressed as Eq. 1:

$$K_{HM} = \frac{n(1-\phi)}{12\pi R}S_n; G_{HM} = \frac{n(1-\phi)}{20\pi R}(S_n + 1.5S_t) \quad (1)$$

where n is the number of contacts. ϕ is the porosity. R is the radius of the particle. S_n and S_t are the normal and tangential stiffness of the contact, respectively. Then, we calculated the equivalent modulus of sand, clay, and hydrate, respectively. The unconsolidated tensor $C_{unconsolid}$ of each component can be obtained by Eq. 2:

$$C_{unconsolid} = \begin{bmatrix} C'_{11} & C'_{11} - 2C'_{44} & C'_{11} - 2C'_{44} & 0 & 0 & 0 \\ C'_{11} - 2C'_{44} & C'_{11} & C'_{11} - 2C'_{44} & 0 & 0 & 0 \\ C'_{11} - 2C'_{44} & C'_{11} - 2C'_{44} & C'_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & C'_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C'_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & C'_{44} \end{bmatrix}$$
$$(2)$$

Fig. 6 Hydrate saturations calculated by porewater freshening and pressure core in well W11
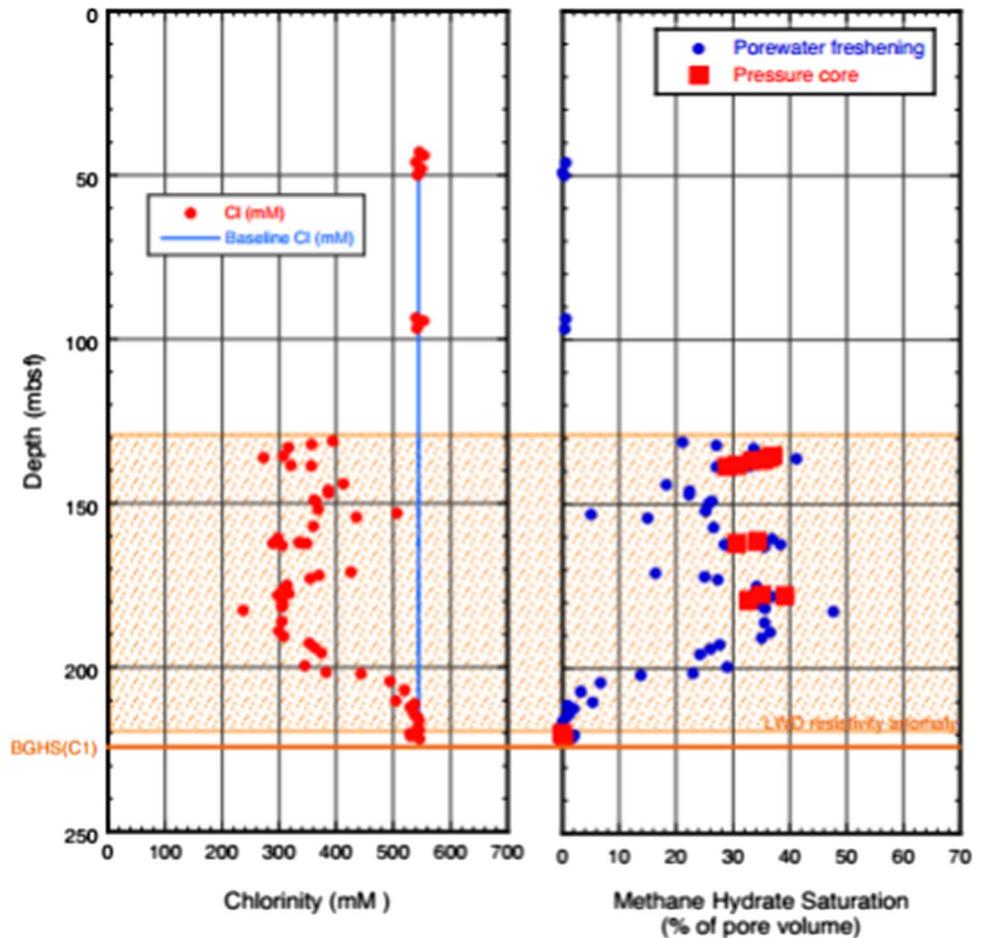
where C'_{ij} is unconsolidated (isotropic) elastic stiffness tensor of each component, and $C'_{11} = K_{HM} + \frac{4}{3}G_{HM}, C'_{44} = G_{HM}$.

VTI matrix and modulus calculation Vertical transversely isotropic (VTI) media is the most common anisotropic model. VTI media can be used to describe the characteristics of layered hydrated and thin layers (Figs. 5, 7). Postma (1955) proposed the periodic thin layer (PTL) anisotropic model, and White et al. (1983) studied the stiffness tensor of the model (Eq. 3):

$$C_{VTI} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & 0 & 0 \\ C_{12} & C_{11} & C_{13} & 0 & 0 & 0 \\ C_{13} & C_{13} & C_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5(C_{11} - C_{12}) \end{bmatrix} \quad (3)$$

According to the formula proposed by Backus (1962), each elastic coefficient in the tensor is shown separately.

$$C_{11} = \left\langle \frac{4\mu(\lambda + \mu)}{\lambda + 2\mu} \right\rangle + \left\langle \frac{1}{\lambda + 2\mu} \right\rangle^{-1} \left\langle \frac{\lambda}{\lambda + 2\mu} \right\rangle^2,$$

$$C_{12} = \left\langle \frac{2\mu\lambda}{\lambda + 2\mu} \right\rangle + \left\langle \frac{1}{\lambda + 2\mu} \right\rangle^{-1} \left\langle \frac{\lambda}{\lambda + 2\mu} \right\rangle^2,$$

$$C_{13} = \left\langle \frac{1}{\lambda + 2\mu} \right\rangle^{-1} \left\langle \frac{\lambda}{\lambda + 2\mu} \right\rangle,$$

$$C_{33} = \left\langle \frac{1}{\lambda + 2\mu} \right\rangle^{-1}, C_{44} = \left\langle \frac{1}{\mu} \right\rangle^{-1} \quad (4)$$

where λ and μ refer to the Lame parameters of each isotropic thin layer (or each component) obtained by Eqs. 1 and 2. Operator $< >$ represents the weighted average of the variables in Eq. 4. Then, elastic tensor of layered hydrate-bearing matrix can be calculated by Eq. 3.

Fluid substitution The Wood formula (Wood 1955) was used to calculate the modulus and density of mixing fluids.
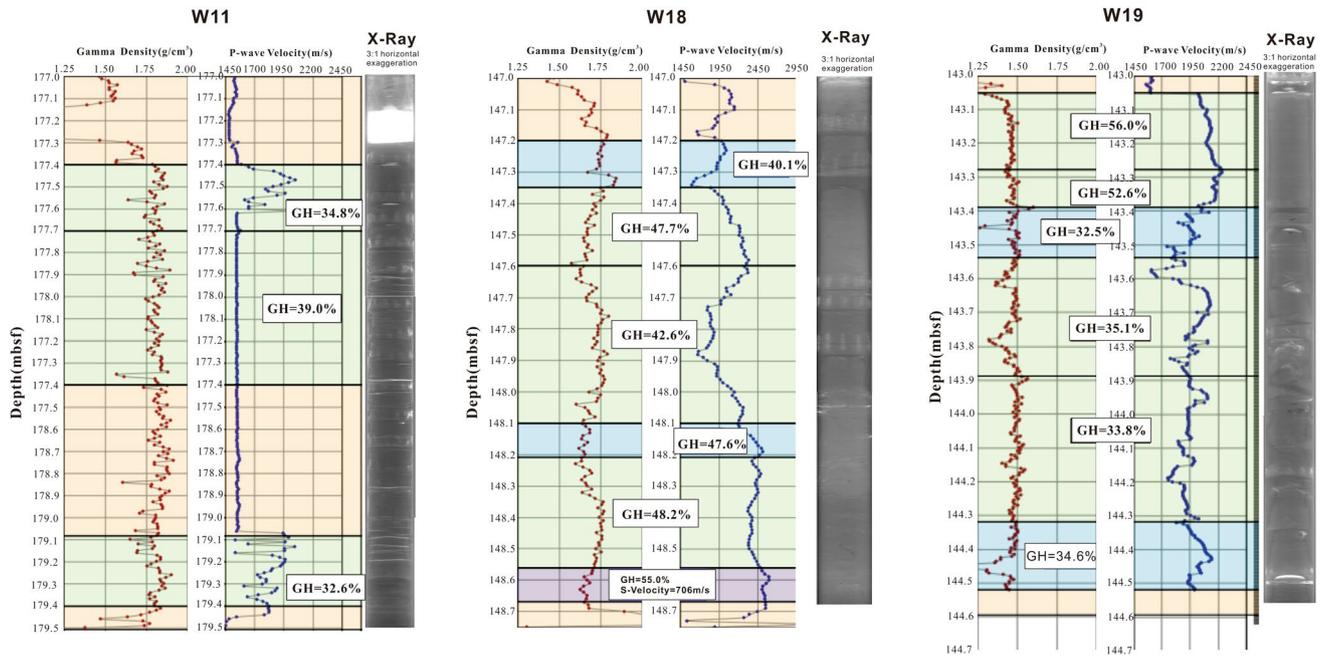
Fig. 7 X-ray images, P wave velocities, and densities of pressure cores in wells W17, W18, and W19. Saturations of the gas hydrate (GH) were calculated by the measured velocities and densities. Gamma densities of the pressure cores were collected by Geotek Pressure Core Analysis and Transfer System (PCATS), which were measured through the attenuation of gamma rays from a ^{137}Cs source and calibrated with water and aluminum. Gamma densities is an inferred density, which is close to the bulk density

Table 2 Constants used in the model

Mineral components	Bulk modulus (Gpa)	Shear modulus (Gpa)	Density (g/cm³)
Matrix			
Quartz	36.6	45	2.65
Shale	20.9	6.85	2.58
Dolomite	61.5	41.1	2.79
Calcite	76.8	32	2.71
Fluid			
Water	2.55	–	1.05
Gas	0.01	–	0.1
Hydrate			
–	7.7	3.2	0.92

$$\frac{1}{K_R} = \sum_{i=1}^{N} \frac{f_i}{K_i} \tag{5}$$

$$\rho = \sum_{i=1}^{N} f_i \rho_i \tag{6}$$

where K_R is the bulk modulus of the mixture and ρ is the average density f_i, K_i and ρ_i are the volume content, bulk modulus, and density of each component, respectively. The Brown–Korringa equation (Brown and Korringa 1975) was used to add fluid to the connected dry micropores. Fluid substitution for anisotropic rocks can be done with Eq. 7:
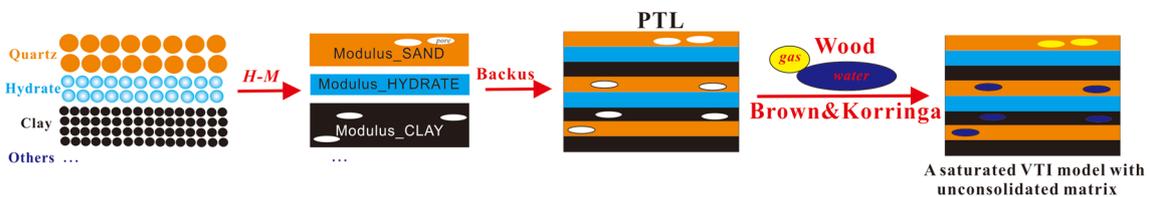

Fig. 8 Schematic diagram of the UAM modeling

$$s_{ijkl}^{\text{sat}} = s_{ijkl}^{\text{dry}} - \frac{(s_{ijaa}^{\text{dry}} - s_{ijaa}^{0})(s_{bbkl}^{\text{dry}} - s_{bbkl}^{0})}{\phi'(\beta_{f1} - \beta_0) + (s_{ccdd}^{\text{dry}} - s_{ccdd}^{0})} \quad (7)$$

where s_{ijkl}^{sat}, s_{ijkl}^{dry} are the compliance tensors for saturated rock and dry rock, respectively. s_{ijkl}^{0} is the compliance tensor for a mineral β_{f1}, β_0 are compressibility for the fluid and mineral, respectively, which are related to the content and elastic characteristics of each component of the fluid. ϕ' is the nuclear magnetic porosity that will reduce with an increase in hydrate saturation. The total porosity ϕ, nuclear magnetic porosity ϕ', and hydrate saturation (HS) should satisfy Eq. 8:

$$\phi' = \phi - \text{HS} \cdot \phi \quad (8)$$

Porosity inversion

Porosity variation affects velocity significantly (Wang 2012). The porosity of hydrate reservoirs can be obtained by the relationship between porosity reflection coefficient (R_{por}) and impedance reflection coefficient (R_{ip}) (Chatterjee et al. 2016; Kumar et al. 2016). Figure 9a shows the linear relation between R_{por} and R_{ip} in the non-hydrated and non-gas layers in wells W11 and W17, which can be written as Eq. 9. The relationships between the hydrate layer, gas layer, and other layers are piecewise linear. Figure 7b shows the linear relation between R_{por} and R_{ip} in the hydrate layers.

$$R_{\text{por}} = -1.14 R_{\text{ip}} \quad (9)$$

where $R_{\text{por}} = 2(\phi_2 - \phi_1)/(\phi_2 + \phi_1)$; $R_{\text{ip}} = 2(\text{ip}_2 - \text{ip}_1)/(\text{ip}_2 + \text{ip}_1)$

Because the reflection integral obtains the relative porosity, the low-frequency part of the integrated result is almost absent. Therefore, we merged the integral result without the low-frequency part and measured low-frequency porosity model to obtain the final porosity.

$$\phi = \phi_{\text{INT}} - \phi_{\text{INT}}^{\text{LFM}} + \phi_{\text{well}}^{\text{LFM}} \quad (10)$$

where ϕ_{INT}, $\phi_{\text{INT}}^{\text{LFM}}$, $\phi_{\text{well}}^{\text{LFM}}$ are integral porosity, low-frequency integrated porosity, and low-frequency measured porosity, respectively.

Seismic inversion

AVO seismic inversion was used to get P wave velocities and S wave velocities. The Aki–Richards approximation is the basis of AVO inversion, which describes the energy distribution on both sides of the seismic reflection interface (Aki and Richards 2002). The expression is as Eq. 11:

$$R_{\text{pp}}(\overline{\theta}) \approx \frac{\sec^2 \overline{\theta}}{2} R_{\text{vp}} - 4 \frac{\overline{v_s}^2}{\overline{v_p}^2} \sin^2 \overline{\theta} R_{\text{vs}} + \frac{1}{2}\left(1 - 4\frac{\overline{v_s}^2}{\overline{v_p}^2} \sin^2 \overline{\theta}\right) R_\rho \quad (11)$$

where $R_{\text{vp}} = \frac{\Delta v_p}{v_p}$; $R_{\text{vs}} = \frac{\Delta v_s}{v_s}$; $R_\rho = \frac{\Delta \rho}{\rho}$. The relationship between the reflection coefficient and seismic data at different incidences can be expressed as Eq. 12:

$$D_{\theta_i} = R_{\text{pp}}(\theta_i) \cdot W_{\theta_i} (i = 1, 2, 3 \ldots) \quad (12)$$

where D is the seismic data. W is the wavelet matrix. The inversion objective function can be written as Eq. 13:

$$J = ||D_{\theta_i} - R_{\text{pp}}(\theta_i) \cdot W_{\theta_i}||_2^2 + \lambda_1 ||R_{\text{pp}}(\theta_i)||_1 + \lambda_2 ||R_{\text{pp}}(\theta_i) - R_0||_1 (i = 1, 2, 3 \ldots) \quad (13)$$

R_0 is a low-frequency model calculated by well interpolation, and λ_1, λ_2 are regularization coefficients. By minimizing J, P wave velocity and S wave velocity can be obtained.
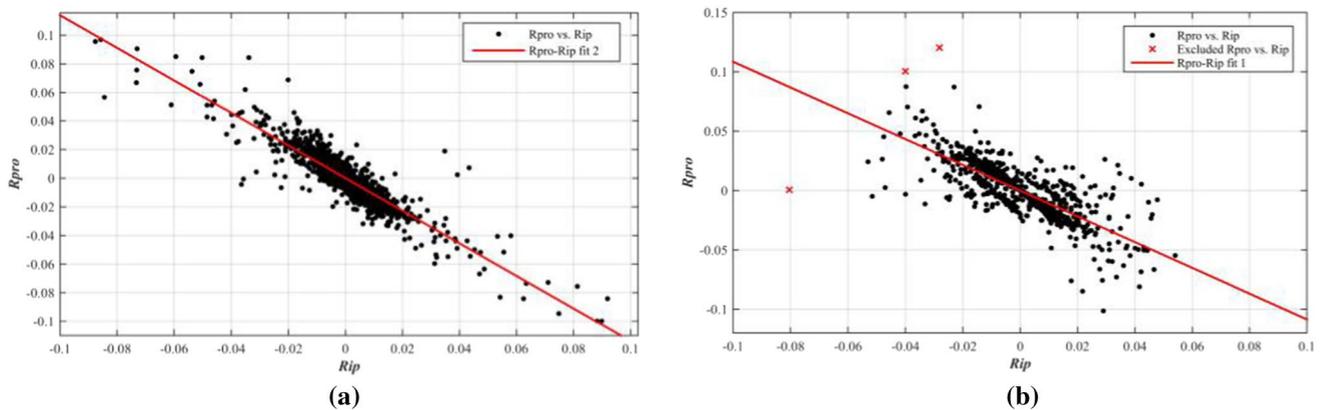

Fig. 9 The relations between R_{por} and R_{ip}; **a** The relation in the non-hydrated and non-gas layer; **b** The relation in the hydrate layer
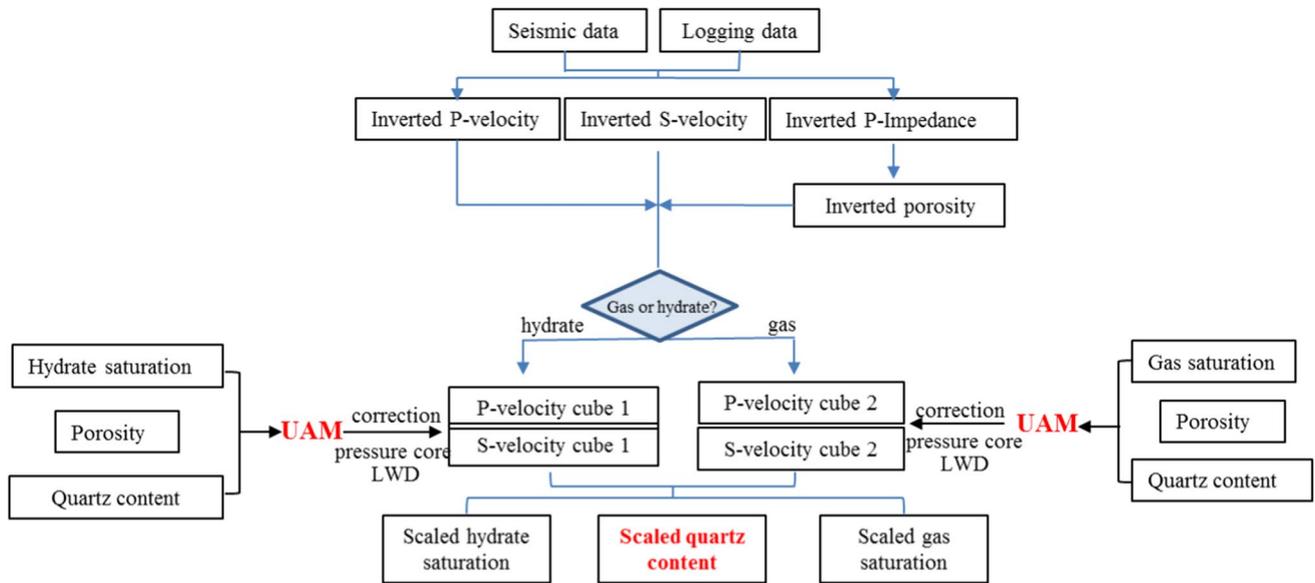
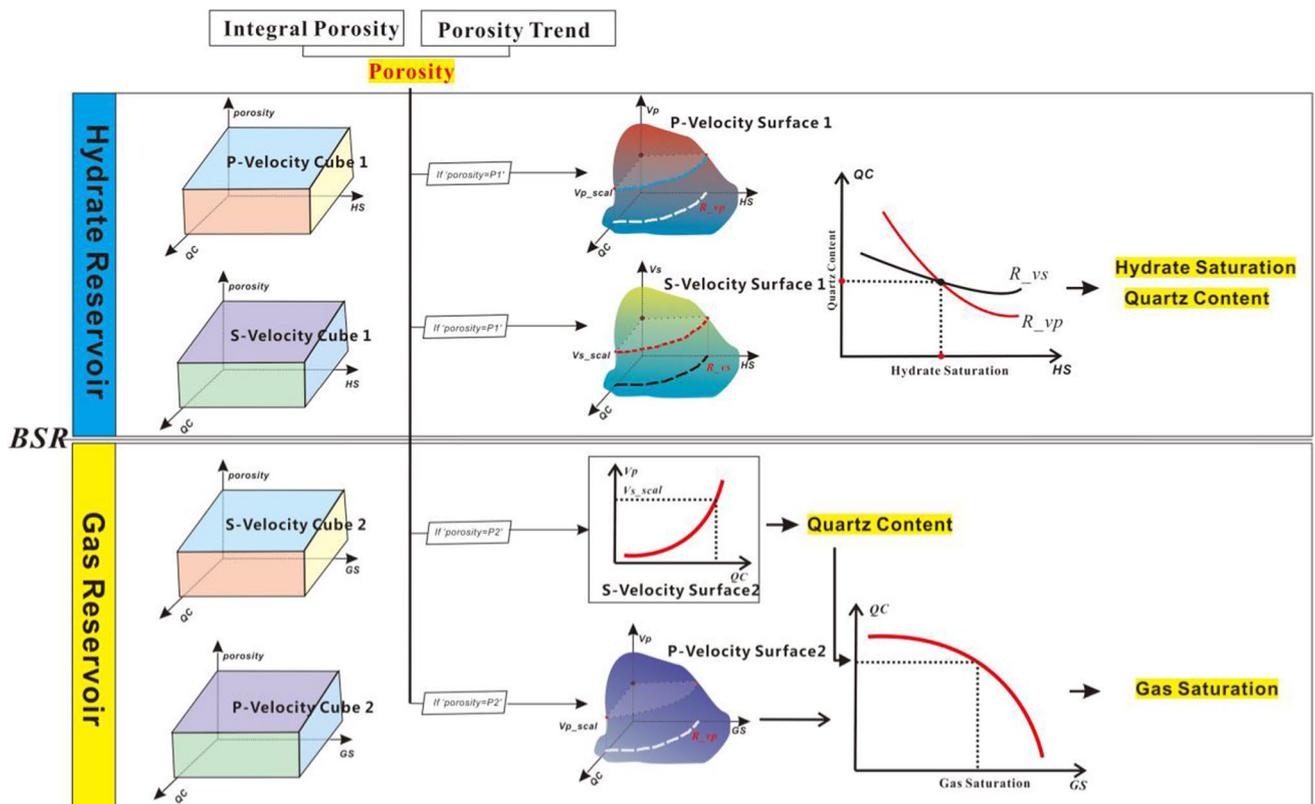**Fig. 10** The workflow of the proposed method

Fig. 11 Schematic diagram of HS, GS and QC estimation by inverse modeling

Estimation of QC by inverse modeling

Figure 10 is the workflow and data flow of QC estimation in this paper. First, seismic data and well logging data were utilized to get the P wave velocity, S wave velocity, and impedance by seismic inversion. We can obtain the porosity using the inverted impedance and Eq. 10. With the priori seismic interpretation, we distinguished the hydrates from the free gas and then used the velocities and inverse modeling to obtain HS, GS, and QC. Figure 11 shows the schematic diagram of the inverse modeling. When estimating QC, the other parameters (HS, GS, porosity) must be considered simultaneously. In the gas hydrate reservoir, making QC, HS, and porosity variables, the model outputs the corresponding P-Velocity Cube 1 and S-Velocity Cube 1 (Fig. 11). In the gas reservoir, making QC, GS, and porosity variables, the model outputs the corresponding P-Velocity Cube 2 and S-Velocity Cube 2 (Fig. 11). When estimating QC, P-velocity and S-velocity are the inputs conversely, which can be obtained by seismic inversion. (vp_scale, vs_scale) is called a scaling sample in this paper. The main factors affecting the velocity in the hydrate reservoir are porosity, HS, and QC. The S wave velocity in the free gas layer is mainly affected by porosity and QC; therefore, the outputs in the hydrate layer are QC and HS, and the outputs in the gas layer are GS and QC.

HS estimation

In the hydrate layer, for a given porosity P1, P-Velocity Surface 1 and S-Velocity Surface 1 corresponding to P1 can be selected from P-Velocity Cube 1 and S-Velocity Cube 1 (Fig. 11). For a scaling sample (vp_scale, vs_scale), vp_scale and vs_scale are used to obtain R_vp and R_vs, respectively. R_vp and R_vs refer to the relations between HS and QC that can satisfy the scaling sample. Only one intersection is produced by R_vp and R_vs, whose abscissa is HS.

QC estimation

When HS and porosity are known, the only variable affecting the S wave velocity in a hydrate reservoir is QC. We can get the relations between HS and QC. QC can be obtained making vs_scale input for S-Velocity surface 1. In the gas reservoir, because gas occurrence nearly has no impact on S wave velocity, QC can be easily scaled out by S-velocity Surface 2 for a known porosity.

GS estimation

GS estimation can verify the correctness of QC. Through seismic interpretation, we can identify the BSR, hydrate layer, and free gas layer. For (vp_scale, vs_scale) in the gas

reservoir, P-Velocity Surface 2 and S-Velocity Surface 2 corresponding to P2 can be selected from P-Velocity Cube 2 and S-Velocity Cube 2 (Fig. 11). Variables in Surface 2 are QC and GS. Another R_vp that refers to the relation between GS and QC can satisfy the scaling sample. QC is known, so we can estimate GS. If the estimated GS is consistent with results obtained by XRD or other methods, the estimated QC is considered reasonable.

Results

Model test

All necessary data are available in well W1, which is close to well W17 (Fig. 2b). The input parameters are porosity, QC, HS, and GS. We assumed that the grain size and number of contacts keep constants for a certain mineral based on core analysis. The shale and quartz contents were calculated by element capture logging. Element logging can provide the main element contents of the sediment. With the peroxy closure and lithology model, the corresponding mineral contents were obtained. Aluminum and iron were converted into clay, silicon was converted into quartz sand, and calcium was converted into calcite. Finally, XRD was used to calibrate this result.

We calculated the velocities with these inputs. The calculated P wave and S wave velocities agreed well with the measured results within tolerance (Fig. 12). There was no other empirical formula involved in this process, indicating that these four inputs are the significant physical parameters for the hydrate-bearing sediments.

Further, we changed the input parameters within a certain range to obtain the possible P wave and S wave velocities with any parameter combination, which are P-Velocity Cube 1, S-Velocity Cube 1, P-Velocity Cube 2, and S-Velocity Cube 2 in Fig. 11. Figures 13 and 14 show the velocities in the hydrate reservoir when the QC is 0.5 and 0.3, respectively. Figures 15 and 16 show the velocities in the gas reservoir when the QC is 0.5 and 0.3, respectively. The measured velocities of well W1 were projected into the figures, and calculated results were more in line with the measured data in the hydrate reservoir when QC was 0.3. It is more reasonable in the free gas reservoir when QC was 0.5, which is due to the lower QC in the hydrate layer, while QC in the gas layer increased in well W1. Therefore, we can find a proper QC so that the calculated results are completely consistent with the measured ones.

QC estimation test

We first estimated HS in well W1 (Fig. 17a). We obtained all the relations that can satisfy S wave velocity at different
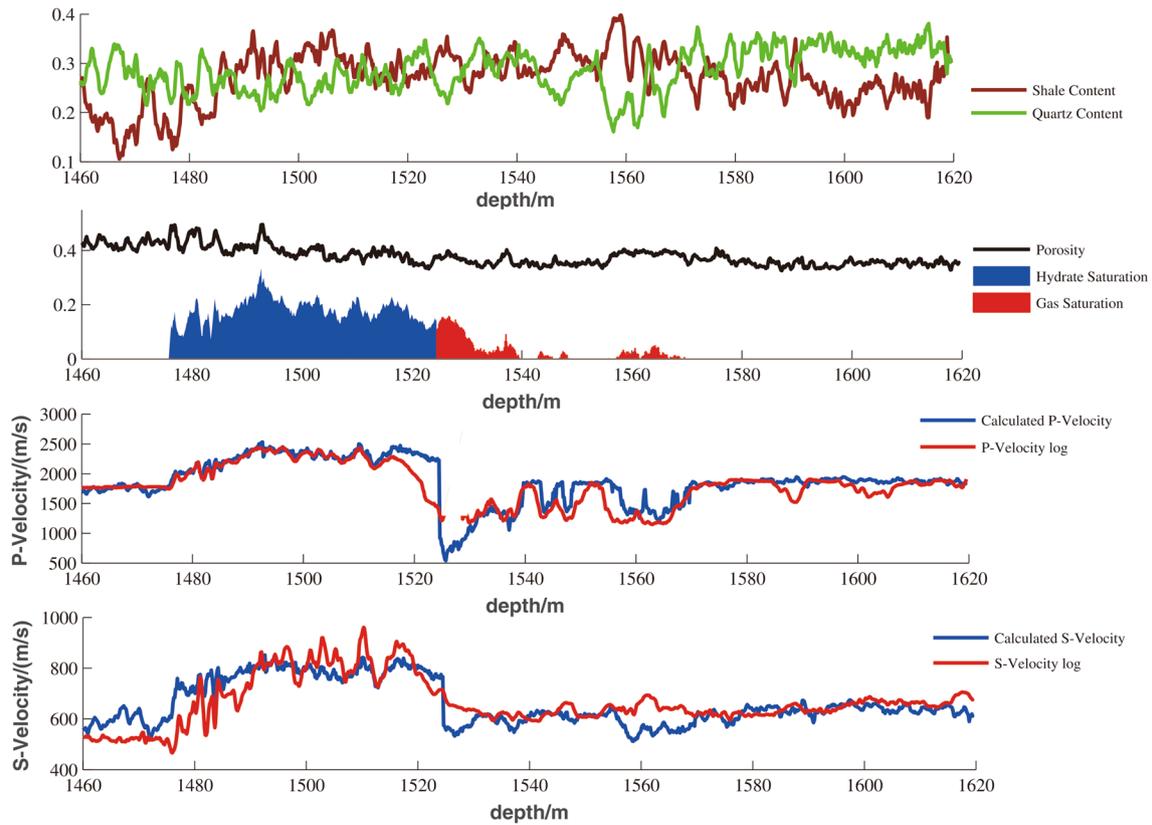
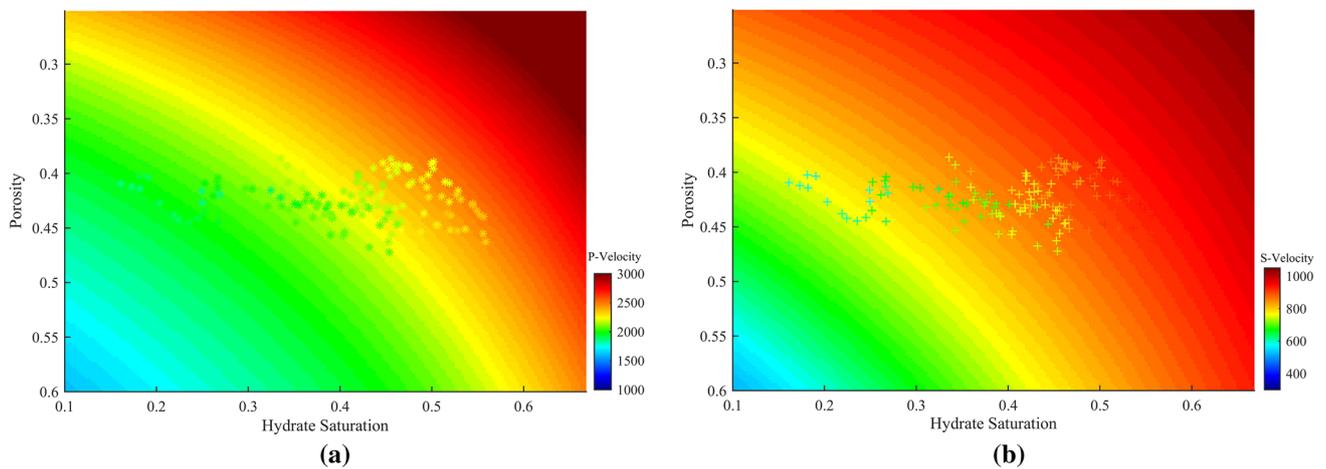**Fig. 12** Input logging curve and modeling results of well W1

Fig. 13 P wave velocity and S wave velocity versus different porosity and HS when QC is 0.5; **a** P wave velocity when porosity and HS are changing; **b** S wave velocity when porosity and HS are changing

depths (Fig. 17b). Then, the corresponding QC can be obtained using the scaled HS (Fig. 17a).

If HS and porosity are known, the S wave velocities corresponding to different QC can be obtained. Figure 17 shows the S wave velocities when QC is 0.4, 0.43, 0.48, and 0.53. Combined with the measured S wave velocity (red bold curve in Fig. 18), QC at different depths was obtained. Significant deviations occurred in the blue rectangular. This was

Fig. 14 P wave velocity and S wave velocity versus different porosity and HS when QC is 0.3; **a** P wave velocity when porosity and HS are changing; **b** S wave velocity when porosity and HS are changing

Fig. 15 P wave velocity and S wave velocity versus different porosity and GS when QC is 0.5; **a** P wave velocity when porosity and GS are changing; **b** S wave velocity when porosity and GS are changing
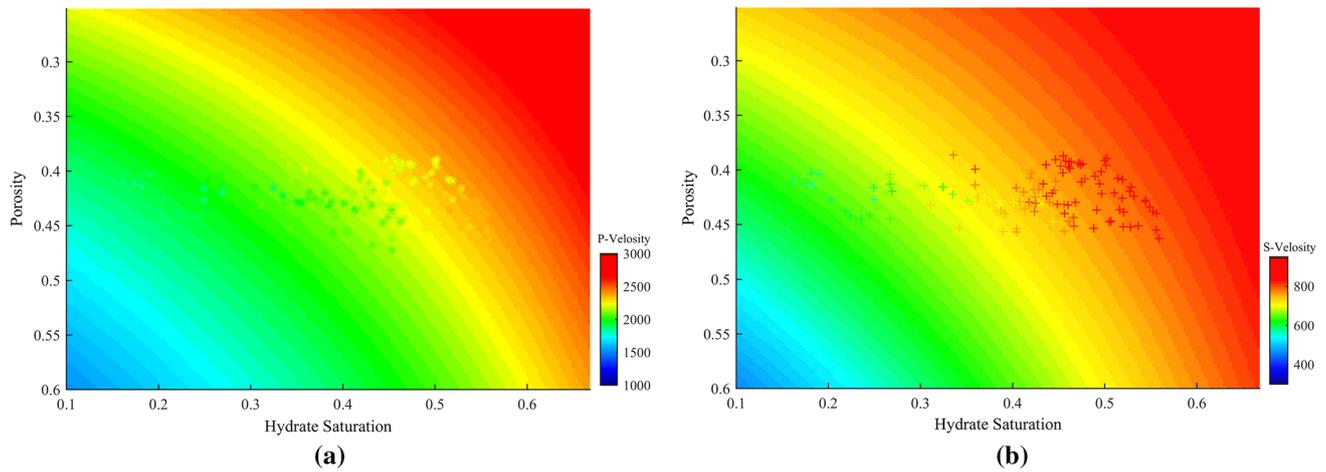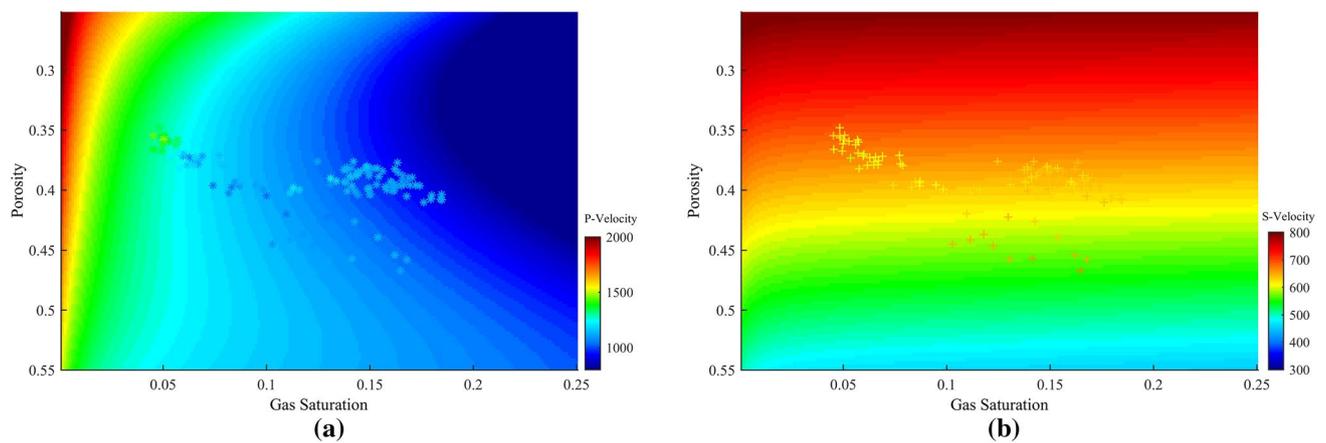

Fig. 16 P wave velocity and S wave velocity with different porosity and GS when QC is 0.3; **a** P wave velocity when porosity and GS are changing; **b** S wave velocity when porosity and GS are changing
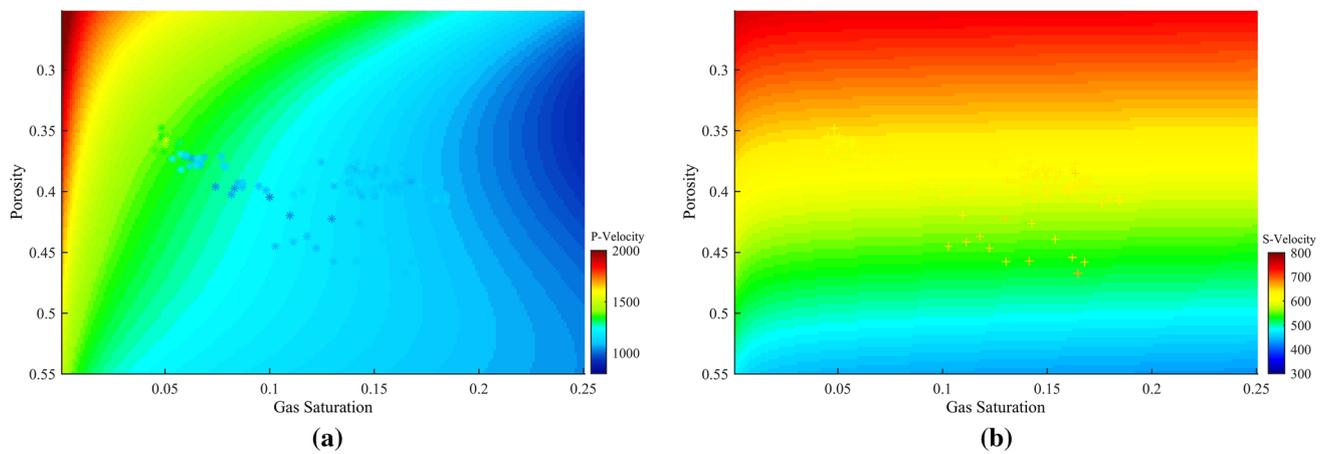
Fig. 17 Relation between scaled HS and QC that can satisfy the measured S wave velocity in well W1; **a** scaled HS in well W1; **b** HS corresponding to QC that can satisfy the measured S wave velocity

(a)

(b)

Fig. 18 Calculated S wave velocity curves when QC is 0.4, 0.43, 0.48, and 0.53 and scaled QC; **a** Calculated S wave velocity curves when QC is 0.4, 0.43, 0.48, and 0.53, red bold curve refers to S wave velocity, and blue bold curve refers to 1000 * QC; **b** Scaled QC by the model and QC curve.

because when the model was constructed, it was assumed that the modulus of sand grain was a constant and bigger than that of clay, which came from the overall understanding of the area (Fig. 19), and only clay and sand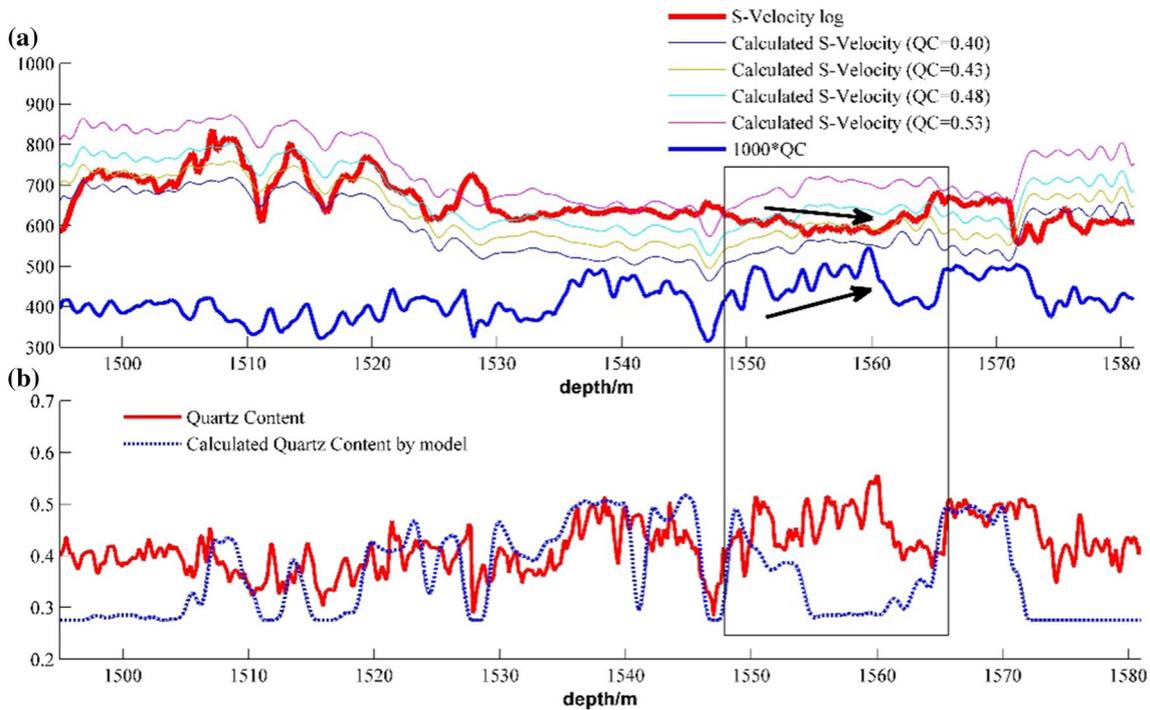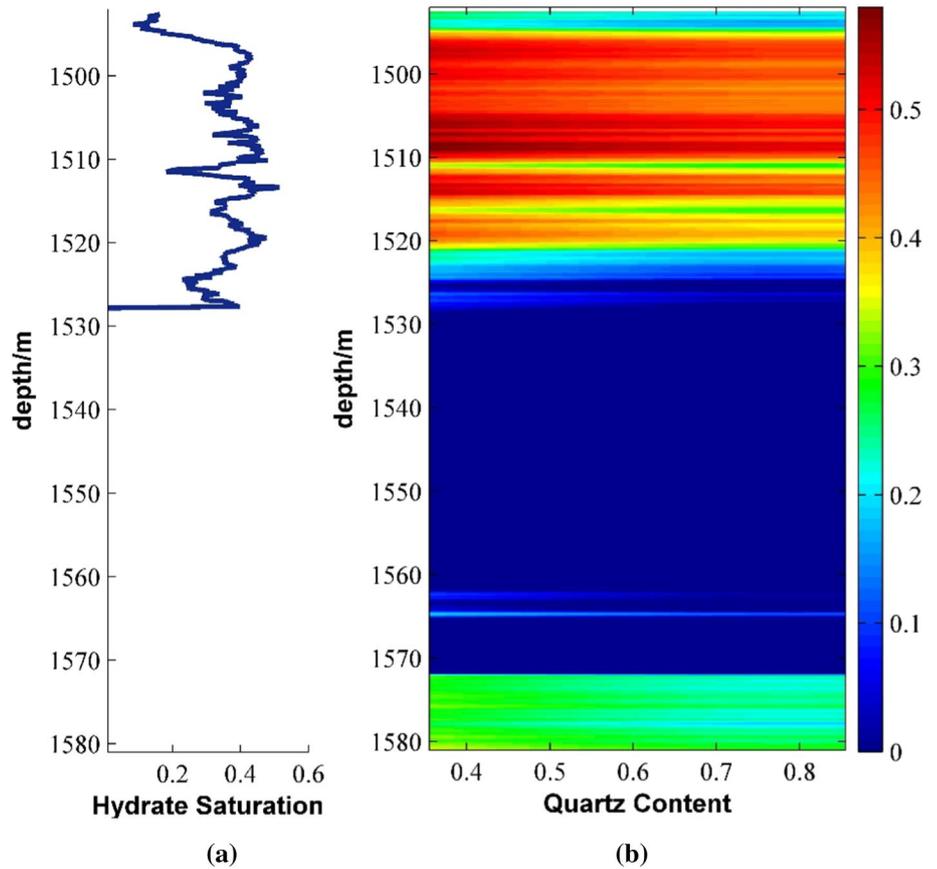 were considered in the model according to lithological analysis. But S wave velocity decreased with an increase in QC, and we could

hardly find a change in grain size. It is speculated that other minerals here have a greater impact on S wave velocity. This problem cannot be solved by this proposed model, because one more input for the model, one more output in the inversion process is required. P wave velocity and S wave velocity

Fig. 19 Relation between clay content and velocity, porosity * velocity can eliminate the effect of porosity to some extent, so velocity decreases with an increase in clay content
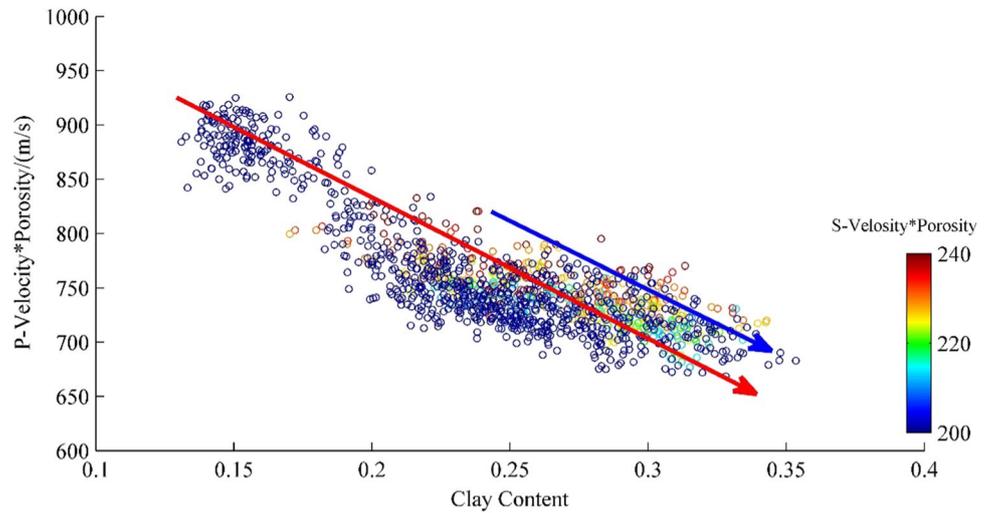

Fig. 20 Comparison of estimated HS, GS and HS, GS calculated by resistivity
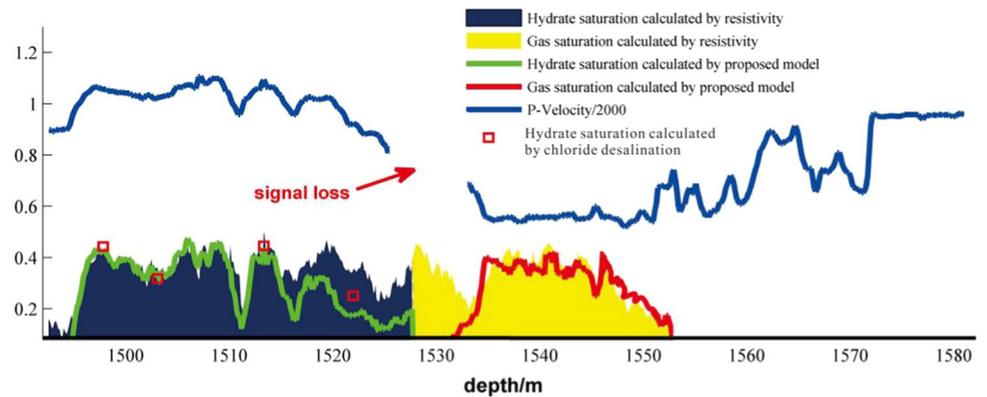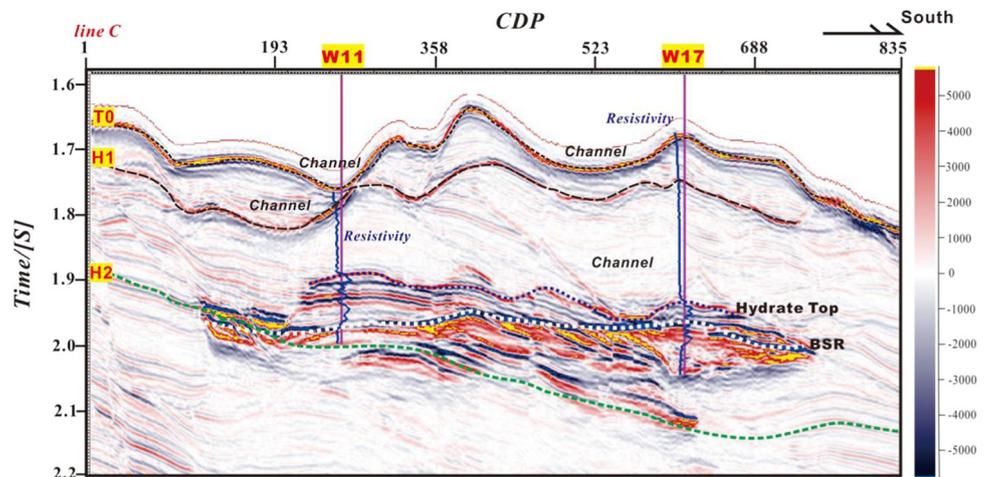

Fig. 21 Seismic section with some interpretation of Line C

are two reliable parameters that can be inverted stably as far as we concerned.

Then, we estimated GS (Fig. 20). The estimated GS and HS agreed with the saturation calculated by resistivity and porewater freshening, indicating that the estimated QC is reasonable.

Field application

Line C crosses the hydrate reservoir from north to south (Fig. 21). A positive reflection indicated the hydrate top, and the BSR exhibited a strong negative reflection which was almost parallel to the seafloor. Gas hydrates and free gas were in different saturations and thicknesses in wells W11 and W17. The lithological interface H2 was unclear on the seismic section due to the occurrence of hydrates and free gas, which should be a sedimentary interface. The deposition may roughly follow the direction of H2, but it may also be washed away by the channels, resulting in sudden changes in lithology laterally.

Based on the partial incidence stack data, the simultaneous AVO inversion of the three parameters (P-velocity, S-velocities, and density) was utilized to obtain the P wave and S wave velocities of survey line C, as shown in Fig. 22. The hydrate-bearing sediments exhibited high P wave velocities. The inversion results were in good agreement with the log curves. The thickness of hydrates in well W11 was larger than well W17, and they were distributed in multiple layers vertically, but almost no free gas has been found in well W11, while the thickness of free gas of well W17 was large. High-velocity anomalies beneath BSR, which were easily noticed at the bottom of W17 (Fig. 22a), may be due to the high QC in these areas. The lithology was almost the same along the H2 horizon, where velocities almost unchanged along with H2.

We first obtained the inverted neutron porosity profile, as shown in Fig. 23. As a whole, the neutron porosity gradually decreases with increasing depths. The porosities of shallow sediments were close to 1, and the porosities below H1 were approximately between 0.3 and 0.7. The porosities above

Fig. 22 The profiles of inverted velocity. **a** The inverted P wave velocity. **b** The inverted S wave velocity

(a)

(b)
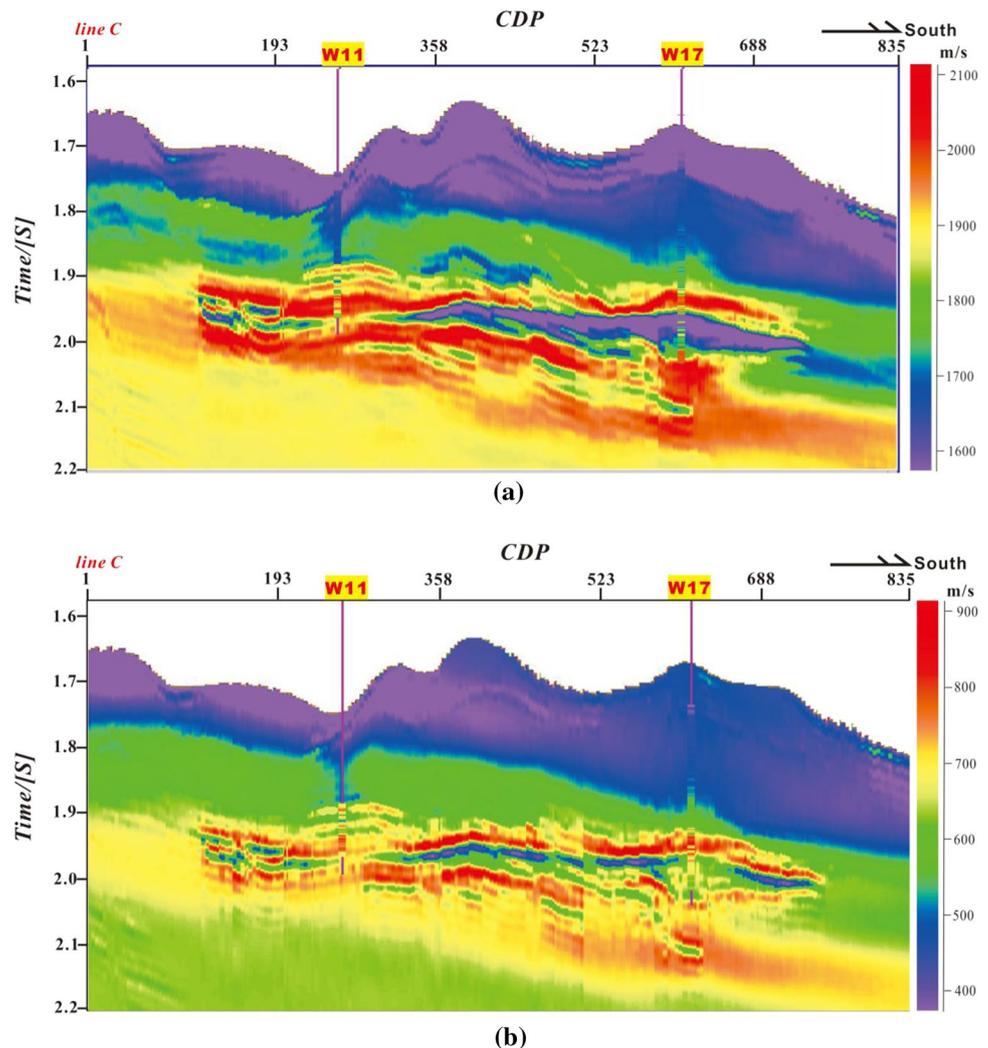
the hydrate reservoir in well W17 were relatively larger than well W11, which may be due to the development of channels near well W17 between T1 and the hydrate top. There was a thin layer with high porosity below the BSR, which provided enough pore space for gas accumulation. The inversion results were in good agreement with the log data, which can provide support for the next step in obtaining the physical parameter.

According to the proposed method, we input the inverted P wave velocities and S wave velocities into the P-Velocity Cubes and S-Velocity Cubes to obtain hydrate saturation. As shown in Fig. 24, the estimated saturation agreed well with the log interpretation. There were mainly three layers of hydrates with different thicknesses, with a maximum saturation of 43%. There was one layer of hydrate in W17.

Fig. 23 The profile of inverted porosity
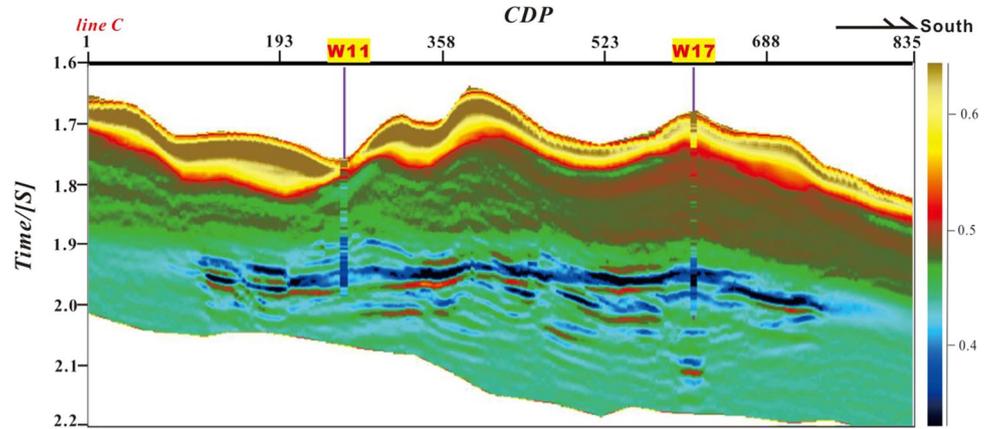
Fig. 24 The profile of estimated HS
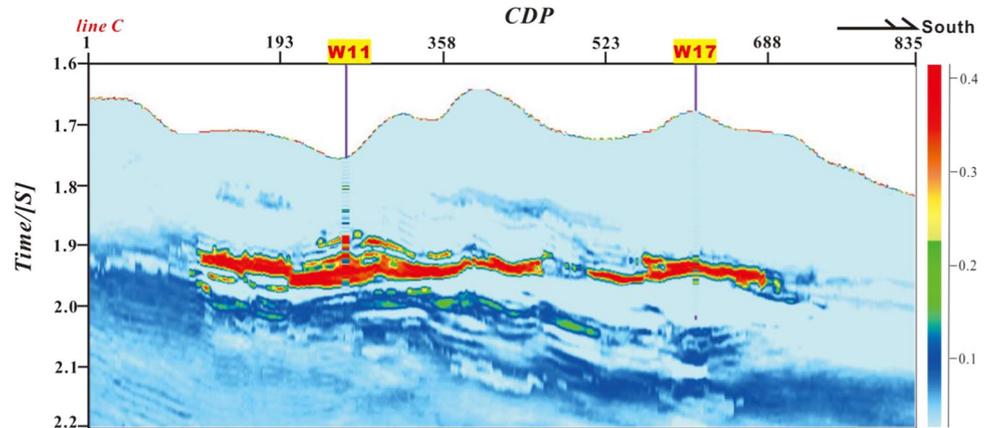
Fig. 25 The profile of inverted QC
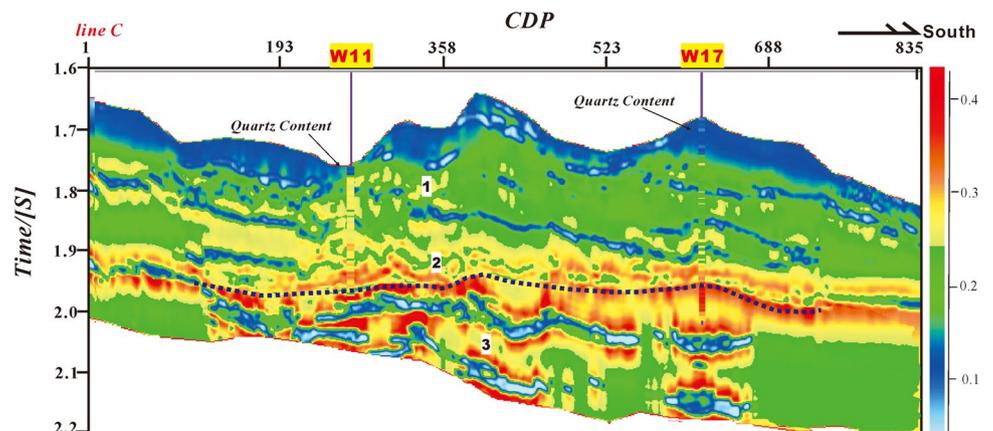
The occurrence of hydrate and free gas should not cause lithological changes, but produce obvious related seismic anomalies. If the conventional multi-parameter fitting or empirical formulas are used to obtain QC, the prediction results may be greatly affected by the occurrence of hydrates and free gas. Figure 25 shows the QC section obtained by the method in this paper. The white dashed line in Fig. 25 refers to the BSR. Mainly three sedimentary strata with relatively high QC can be identified, which were section-1, section-2, and section-3 in Fig. 25. Among them, section-2 owns the highest QC and good lateral continuity, and BSR located in

this section. The extension direction of section-3 was similar to H2. QC of the hydrate layer was higher than that of the upper formation, and the QC of the free gas layer was the largest. The QC of the hydrate and free gas reservoirs was higher than the surrounding sediments, which was consistent with the previous research results in the study area (Chen et al. 2011). The inversion results agreed well with the log data and geological understanding. Figure 26 shows the comparison of the inverted QC and measured QC from element logging and XRD in wells W11 and W17. The inverted
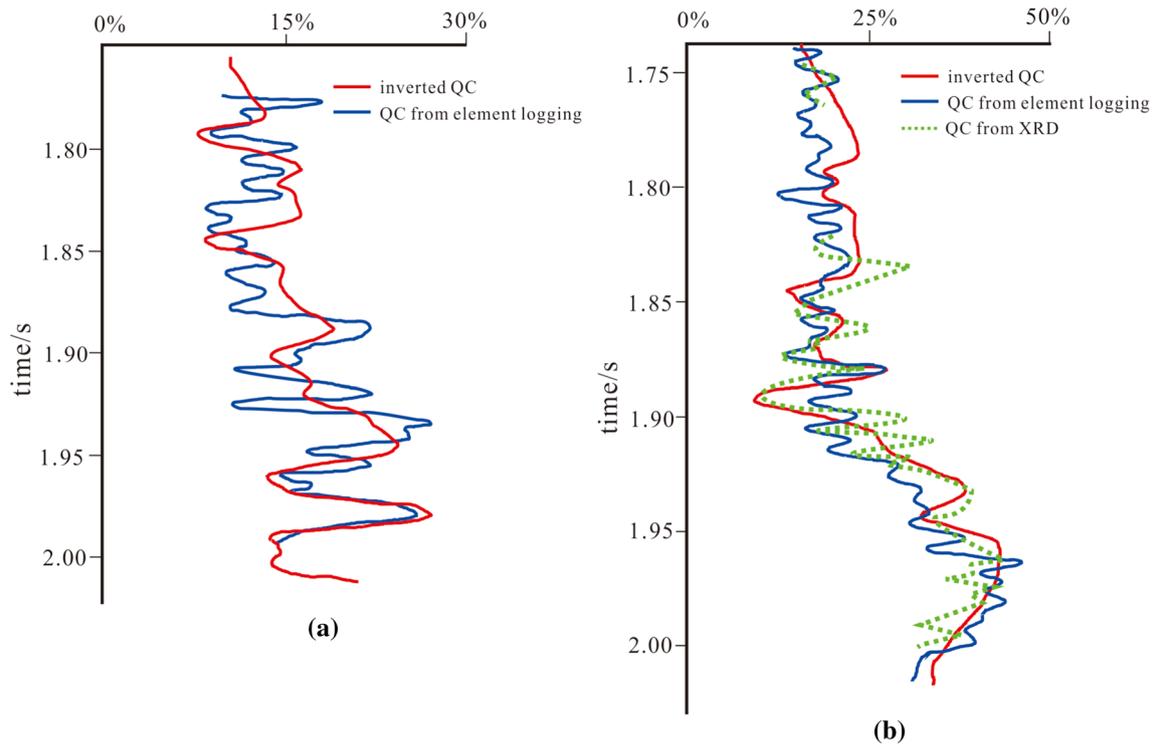
Fig. 26 The inverted QC and measured QC. **a** The inverted QC and QC from element logging in well W11; **b** The inverted QC and QC from element logging and XRD in well W17

Fig. 27 The profile of inverted GS
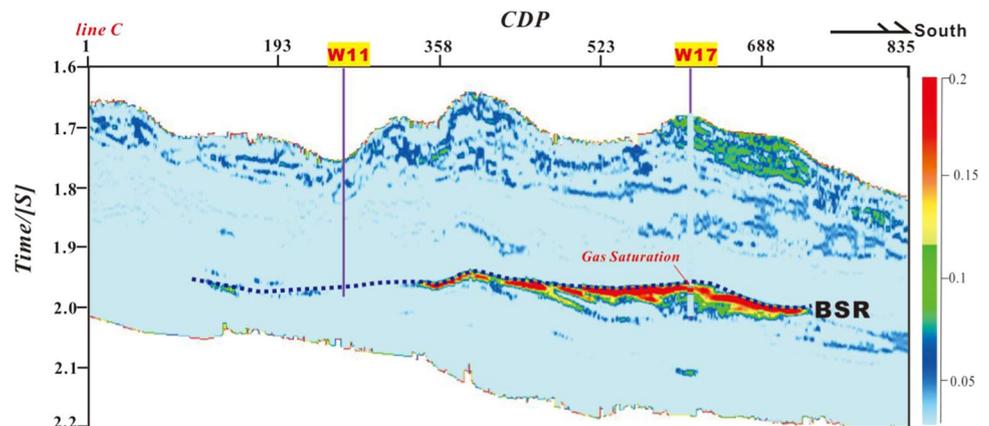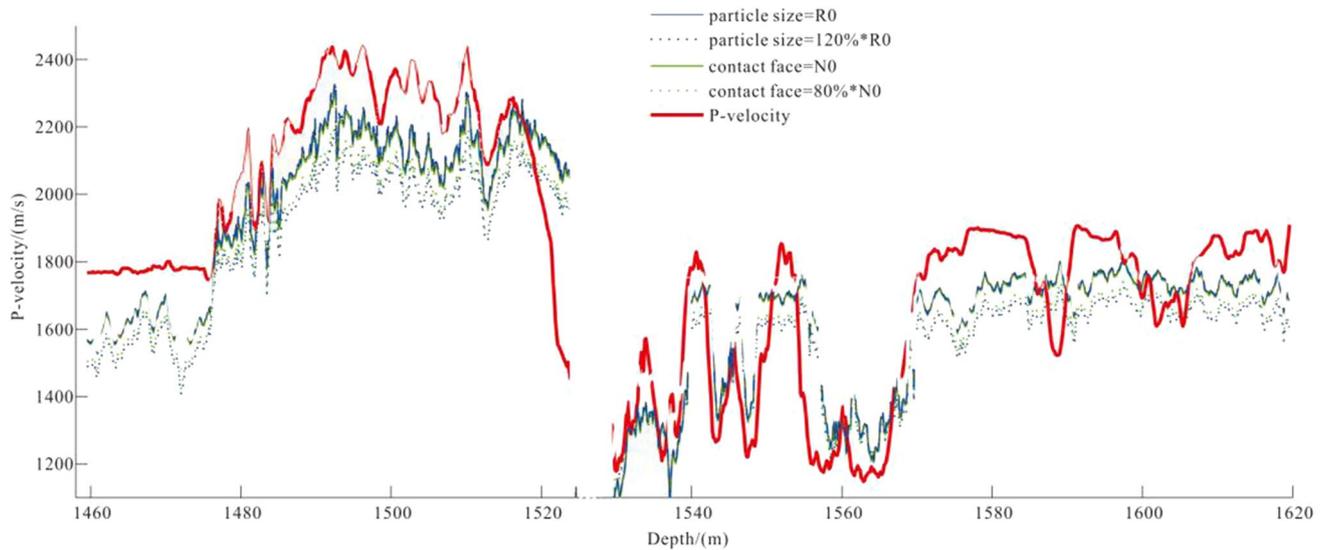
Fig. 28 Calculated velocities in well W1 when modeling constants change within 20%. The velocities caused by the same constant change are shown in the same color. R0, N0 refer to the particle size, contact face set in Fig. 12
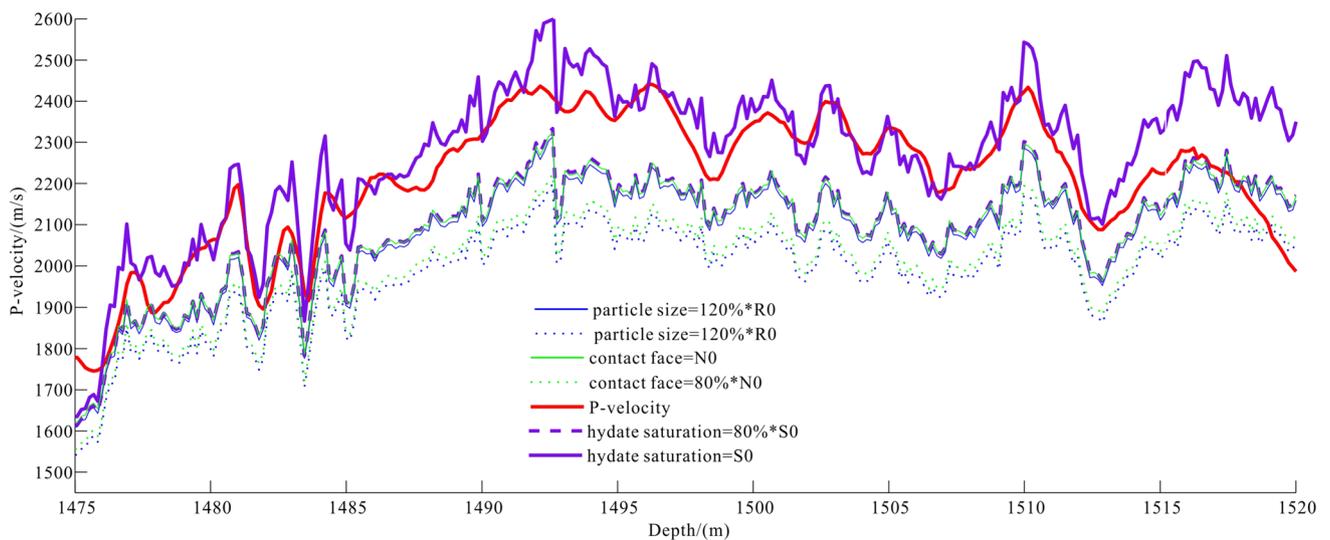
Fig. 29 Calculated velocities in W1 when modeling constants change within 20% in the hydrate reservoir. The purple curves refer to the velocities corresponding to hydrate saturation of S0 and 80% * S0 (S0 is the saturation used in Fig. 12)

QC agreed with the measured QC, but some errors are inevitable due to the low seismic resolution.

Also, we estimated GS based on estimated QC (Fig. 27). The results indicated that almost no free gas exists in W11, and the free gas saturation is up to about 20% in W17. The free gas mainly accumulated near the bottom of the BSR where porosity was relatively high as mentioned, and the vertical saturation of free gas decayed rapidly. These characters have been confirmed by drilling.

Discussion

Factors that cause errors

In this paper, the particle size, number of contacts, and modulus of different lithologic components were obtained from the analysis of samples in the study area. From the previous analysis, it is known that these parameters can be set to constants because the lithology changes little. It can also be seen from Fig. 12 that the calculated velocities deviate
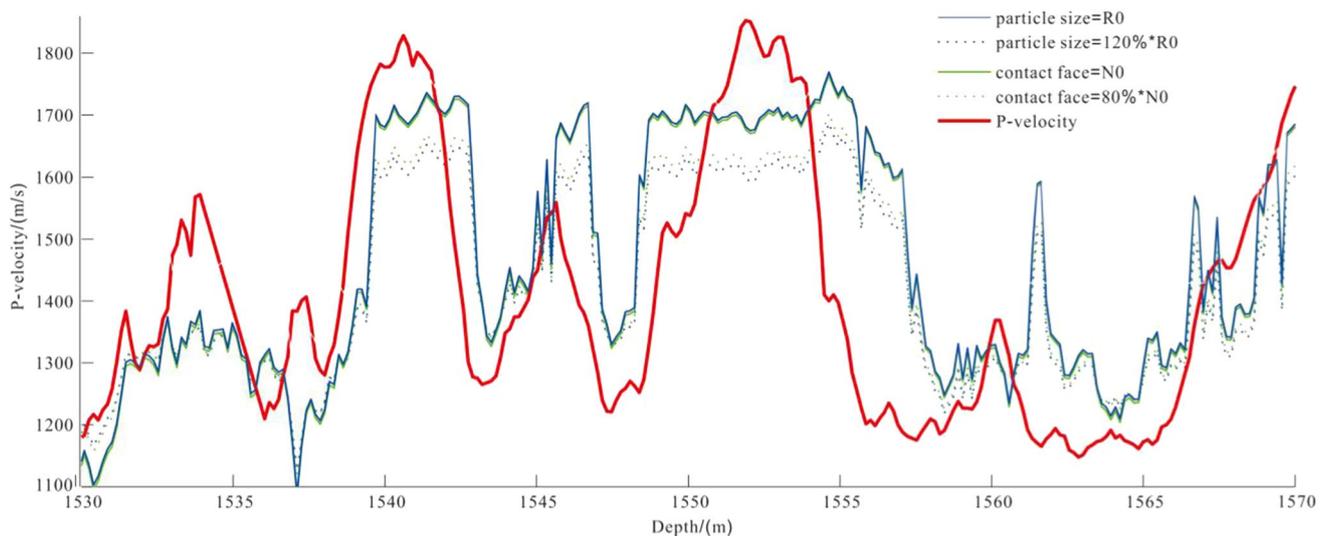
Fig. 30 Calculated velocities in W1 when modeling constants change within 20% in the free gas reservoir

from the actual velocities at some depths, which is unavoidable. However, if these parameters are set as variables, these parameters cannot be obtained by inverse modeling, because the variables in the model in this paper have reached the upper limit. It is, therefore, necessary to ensure that these constants are as reasonable as possible. We discussed which constant has the greatest effect on velocity. We can get calculated velocities when modeling constants change (Fig. 28). Particle size (blue curves) has almost the same effect on velocity as the number of contacts (green curve). We enlarge the hydrate and the free gas reservoirs separately for further analysis.

In the hydrate reservoir, the influence of hydrate saturation on velocity is particularly obvious. The particle size and number of contacts own a slight effect on velocities (Fig. 29). The estimated hydrate saturation in Fig. 20 fits the measured saturation well, because almost no change in particle size. However, in the free gas reservoir where the velocities are lower than 1600 m/s, the influence of particle size and number of contacts is significantly reduced (Fig. 30). As a result, we can obtain a good estimation result (Fig. 20) in the free gas reservoir too.

Therefore, particle size and number of contacts slightly affect the estimation in the hydrate reservoir.

Implication to hydrate exploration

The first hydrate production was conducted in the Shenhu area of the South China Sea in 2017. High-precision prediction of hydrate reservoir physical properties is the basis of the production test. The method proposed in this paper can effectively obtain hydrate saturation and free gas saturation, to provide a basis for hydrate resource estimation;

meanwhile, it can obtain QC and provide a certain reference for the sand control in the production test. The method proposed in this paper is a supplement to existing rock physics models of hydrates and physical property prediction methods, and it is expected to be extended to the estimation of other types of hydrates.

Conclusion

In this paper, the possible distributions of the deposit with high quartz content in the Shenhu area, South China Sea, were discussed, and the characteristics of the hydrate-bearing sediments were studied. Then, a rock physics approach was used to characterize unconsolidated sediments with layered hydrates and gas occurrence. This model has considered sediment grain, unconsolidation, and hydrate or gas occurrence, and well described the relationship between physical properties (quartz content, hydrate saturation, gas saturation, and porosity) and elastic characteristics of hydrate-bearing sediments in the study area.

Then, an inverse modeling method was proposed to obtain quartz content as well as other physical parameters simultaneously combining with seismic inversion. The tests on a well and a seismic section have been done, and the inversion results were consistent with the logging data and geological understandings. It indicates that the method has certain practicability and is helpful for the quantitative interpretation of the hydrate reservoir.

Compliance with ethical standards

References

Aki K, Richards PG (2002) Quantitative seismology. University Science Books, Mill Valley

Backus GE (1962) Long-wave elastic anisotropy produced by horizontal layering. J Geophys Res 67:4427–4440

Brown R, Korringa J (1975) On the dependence of the elastic properties of a porous rock on the compressibility of the pore fluid. Geophysics 40:608–616

Boswell R, Collett TS (2006) The gas hydrates resource pyramid: fire in the ice, methane hydrate newsletter, vol 6, no 3. US Department of Energy, Office of Fossil Energy, National Energy Technology Laboratory, Fall Issue, pp 5–7

Boswell R, Collett T, Frye M, Shedd W, McConnel D, Shelander D (2012) Subsurface gas hydrates in the northern Gulf of Mexico. J Mar Pet Geol 34(1):4–30

Batzle M, Han DH, Hofmann R (2006) Fluid mobility and frequency-dependent seismic velocity—direct measurements. Geophysics 71:N1–N9

Carcione JM, Tinivella U (2000) Bottom-simulating reflectors: seismic velocities and AVO effects. Geophysics 65:54–67

Carcione JM, Gei D, Rossi G, Madrussani G (2005) Estimation of gas-hydrate concentration and free-gas saturation at the Norwegian–Svalbard continental margin. Geophys Prospect 53:803–810

Chapman M (2009) Modeling the effect of multiple sets of mesoscale fractures in porous rock on frequency-dependent anisotropy. Geophysics 74:D97–D103

Chen F, Zhou Y, Su X et al (2011) Gas hydrate saturation and its relation with grain size of the hydrate-bearing sediments in the Shenhu area of northern South China Sea. Mar Geol Quat Geol 31(5):95–100

Chatterjee R, Singha DK, Ojha M, Sen MK, Sain K (2016) Porosity estimation from pre-stack seismic data in gas-hydrate bearing sediments, Krishna–Godavari basin, India. J Nat Gas Sci Eng 33:562–572

Domenico SN (1977) Elastic properties of unconsolidated porous sand reservoirs. Geophysics 42(7):1339–1368

Domenico SN (1976) Effect of brine–gas mixture on velocity in an unconsolidated sand reservoir. J Geophys 41:882–894

Dvokin J, Nur A, Yin H (1994) Effective properties of cemented granular material. Mech Mater 18(4):351–356

Dickens GR (2001) The potential volume of oceanic methane hydrates with variable external conditions. Org Geochem 32(10):1179–1193

Hamilton EL (1980) Geoacoustic modeling of the sea floor. J Acoust Soc Am 68:1313–1340

Helgerud MB, Dvorkin J, Nur A, Sakai A, Collett T (1999) Elastic-wave velocity in marine sediments with gas hydrates: effective medium modeling. Geophys Res Lett 26(13):2021–2024

He JX, Yao YJ, Liu HL, Wan ZF (2008) Genetic types of natural gas and characteristic of the gas source composition in marginal basins of the northern South China Sea. J Geol China 35(5):1007–1016 **(in Chinese with English abstract)**

Holland M, Schultheiss P, Roberts J, Druce M (2008) Observed gas hydrate morphologies in marine sediments. In: 6th international conference on gas hydrates, Vancouver, BC, Canada, 6–10 July 2008

Jakobsen M, Hudson JA, Minshull TA, Singh SC (2000) Elastic properties of hydrate-bearing sediments using effective medium theory. J Geophys Res Solid Earth 105(B1):561–577

Kumar D, Sen MK, Bangs NL (2007) Gas hydrate concentration and characteristics within hydrate ridge inferred from multicomponent seismic reflection data. J Geophys Res Solid Earth. https://doi.org/10.1029/2007JB004993

Kuang ZG, Guo YQ (2011) The sedimentary facies and gas hydrate accumulation models since Neogene of Shenhu Sea Area, northern South China Sea. Earth Sci J Chin Univ Geosci 36(5):914–920 **(in Chinese)**

Kumar R, Das B, Chatterjee R, Sain K (2016) A methodology of porosity estimation from inversion of post-stack seismic data. J Nat Gas Sci Eng 28:356–364

Kuang ZG, Fang YX, Liang JQ et al (2018) Geomorphological–geological–geophysical signatures of high-flux fluid flows in the eastern Pearl River Mouth Basin and effects on gas hydrate accumulation. Sci China Earth Sci 61(7):914–924

Lee MW (2008) Models for gas hydrate-bearing sediments inferred from hydraulic permeability and elastic velocities. Scientific investigations report 2008–5219. J US Geological Survey

Lee MW, Collett TS (1999) Amount of gas hydrate estimated from compressional-and shear-wave velocities at the JAPEX/JNOC/GSC Mallik 2L–38 gas hydrate research well. Bull Geol Surv Can 544:313–322

Lee MW, Collett TS (2011) In-situ gas hydrate saturations estimated from various well logs at Mount Elbert well, North Slope of Alaska. Mar Pet Geol 28:439–449

Lee MW, Waite WF (2008) Estimating pore-space gas hydrate saturations from well-log acoustic data. Geochem Geophys Geosyst 9(7):1–8

Li JF, Ye JL, Qin XW, Qiu HJ, Wu NY, Lu HL, Lu C (2018) The first offshore natural gas hydrate production test in South China Sea. China Geol 1(1):5–16

Ludmann T, Wong HK, Wang PX (2001) Plio-quaternary sedimentation processes and neotectonics of the northern continental margin of the South China Sea. Mar Geol 172(34):331–358

Mindlin RD (1949) Compliance of elastic bodies in contact. J Appl Mech ASME 16:259–268

Miller JJ, Lee MW, von Huene R (1991) An analysis of a seismic reflection from the base of a gas hydrate zone, offshore peru (1). AAPG Bull 75(5):910–924

Moridis GJ, Collett TS, Boswell R, Kurihara M, Reagan MT, Koh C, Sloan ED (2009) Toward production from gas hydrates: current

status, assessment of resources, and simulation-based evaluation of technology and potential. SPE Reserv Eval Eng 12(5):745–771

Ojha M, Sain K (2007) Seismic velocities and quantification of gas-hydrates from AVA modeling in the western continental margin of India. Mar Geophys Res 28(2):101–107

Postma GW (1955) Wave propagation in a stratified medium. Geophysics 20:780–806

Ru K, Pigott JD (1986) Episodic rifting and subsidence in the South China Sea. AAPG Bull 70(9):1136–1155

Riedel M, Collett TS, Kumar P, Sathe AV, Cook A (2010) Seismic imaging of a fractured gas hydrate system in the Krishna–Godavari Basin offshore India. Mar Pet Geol 27:1476–1493

Ryu B, Collett TS, Riedel M, Kim GY, Chun J, Bahk J, Lee J, Kim J, Yoo D (2013) Scientific results of the second gas hydrate drilling expedition in the Ulleung basin (UBGH2). Mar Pet Geol 47:21–29

Sava D, Hardage B (2006) Rock-physics characterization of hydrate-bearing deepwater sediments. Lead Edge 25:616–619

Sloan ED, Koh CA (2007) Clathrate hydrates of natural gases. CRC Press, New York

Su M, Sha ZB, Qiao SH et al (2015) Sedimentary evolution since Quaternary in the Shenhu hydrate drilling area, northern South China Sea. Chin J Geophys 58(8):2975–2985

Winkler KW (1983) Contact stiffness in granular and porous meterials: comparison between theory and experiment. Geophys Res Lett 10:1073–1076

White JE, Martineau-Nicholetis L, Monash C (1983) Measured anisotropy in Pierre shale. Geophys Prospect 31:709–729

Walton K (1987) The effective elastic moduli of a random packing of spheres. J Mech Phys Solids 35(2):213–226

Waite W, Helgerud MB, Nur A, Pinksto JC, Stern L, Kirby S (2000) Laboratory measurements of compressional and shear wave speeds through methane hydrate. Ann N Y Acad Sci 912:1003–1010

Wang Z (2012) Fundamentals of seismic rock physics. Geophysics 66(2):398–412

Wang XJ, Hutchinson DR, Wu SG et al (2011) Elevated gas hydrate saturation within silt and silty clay sediments in the Shenhu area South China Sea. J Geophys Res 116:B05102

Wang XJ, Wu SG, Wang JL et al (2013) Anomalous wireline logging data caused by gas hydrate dissociation in the Shenhu area, northern slope of South China Sea. Chin J Geophys 56(8):2799–2807

Wei JG, Fang YX, Lu HL et al (2018) Distribution and characteristics of natural gas hydrates in the Shenhu Sea Area, South China Sea. Mar Pet Geol 98:622–628

Wood AW (1955) A Textbook of sound[M]. McMillan Co, New York

Wood WT, Stoffa PL, Shipley TH (1994) Quantitative detection of methane hydrate through high-resolution seismic velocity analysis. J Geophys Res Solid Earth 99(B5):9681–9695

Wu SG, Zhang GX, Huang YY et al (2005) Gas hydrate occurrence on the northern slope of the northern South China Sea. Mar Pet Geol 22(3):403–412

Wu NY, Yang SX, Wang HB (2009) Fluid transport system for gas hydrate accumulation in the Shenhu area of the northern slope of the South China Sea. Chin J Geophys 52(6):1641–1650

Xu WY, Rupple C (1999) Predicting the occurrence, distribution, and evolution of methane gas hydrate in porous marine sediments. J Geophys Res 104(B3):5081–5095

Yu XH, Liang JQ, Fang JN et al (2012) Tectonic subsidence characteristics and its relationship to BSR distribution in deep water area of Pearl River Mouth Basin since the Late Miocene. J Palaeogeogr 14(6):787–800 **(in Chinese with English abstract)**

Zhang W, Liang JQ, Lu JA, Wei JG, Su PB, Fang YX, Zhang G (2017) Accumulation features and mechanisms of high saturation natural gas hydrate in Shenhu Area, northern South China Sea. Pet Explor Dev 44(5):708–719

Enhancing the resolution of time–frequency spectrum using directional multichannel matching pursuit

Lu Xu[1,2] · Xingyao Yin[1,2] · Kun Li[1,2]

Abstract

Matching pursuit is able to decompose signals adaptively into a series of wavelets and has been widely applied in signal processing of the geophysical fields. Single-channel matching pursuit could not take into account the lateral continuity of seismic traces, and the recent multichannel matching pursuit exploits the lateral coherence as a constraint, which helps to improve the stability of decomposition results. However, atoms searched by multichannel matching pursuit currently are just shared by lateral seismic traces at the same time slicers. The lack of directionality in multichannel search strategies leads to irrationality in dealing with large dip angle seismic traces. Considering that the waveforms of reflection events are relatively continuous and similar, an improved multichannel matching pursuit is proposed to realize the directional decomposition of adjacent signals. Based on the principle of seismic reflection events tracking and identification, directional multichannel decomposition of seismic traces is realized. The seismic channel to be decomposed is correlated with the time shift of the optimal atom determined by the previous seismic channel. The time position of the maximum correlation indicates the center time of the optimal atom. Optimal atoms identified by one iteration of multichannel decomposition have the same frequency and phase parameters, different center time and amplitude parameters. The center time of the optimal atoms is consistent with seismic reflection events. Tests illustrate that the algorithm can successfully reconstruct 2D seismic data without reducing accuracy. Besides, the application of field data is of great significance for reservoir exploration and hydrocarbon interpretation.

Keywords Matching pursuit (MP) · Reflection event · Time–frequency analysis · Directional multichannel decomposition · Optimal atom

Introduction

The time–frequency analysis technique can convert the signal from the time domain to the time–frequency domain and can be used to extract more attribute characteristics of the signal, which plays an important role in field of geosciences

✉ Xingyao Yin
  xyyin@upc.edu.cn

  Lu Xu
  948687403@qq.com

  Kun Li
  781216086@qq.com

[1] School of Geosciences, China University of Petroleum (East China), Qingdao, Shandong, China

[2] Laboratory for Marine Mineral Resources, Qingdao National Laboratory for Marine Science and Technology, Qingdao, Shandong, China

(Hess-Nielsen and Wickerhauser 1996; Qian and Chen 1999). To estimate the time–frequency representations of the signal, a variety of methods including short-time Fourier transform (STFT), continuous wavelet transform (CWT), S transform (ST), matching pursuit (MP) are usually used (Duijndam and Schonewille 1999; Sinha et al. 2005; Stockwell et al. 2002; Mallat and Zhang 1993). Matching pursuit is widely applied in various subject areas for its superiority of resolution (Gribonval et al. 1996; Salomon and Ur 2004) and has been well discussed in signal processing of different fields (Durka et al. 2005).

The conventional matching pursuit is a greedy strategy to decompose a signal into a linear combination of time–frequency atoms chosen among a Gabor dictionary, which can adaptively sparsely represent signals and has high time–frequency resolution. However, the greedy global search strategy leads to low decomposition efficiency, which limits its widespread application for huge data processing (Lin et al.

2006). Therefore, various improved algorithms to deal with this problem are proposed. In order to reduce the computational cost of the algorithm, genetic algorithm is introduced to matching pursuit achieve this goal successfully (Stefanoiu and Lonescu 2003). Ricker or Morlet atomic libraries which are controlled by three parameters are usually constructed for seismic signal decomposition. The global search of all atoms in the atomic library results in extremely low computational efficiency. The instantaneous properties of the signal estimated by the Hilbert transform can be used to narrow the scanning range of the time–frequency atomic library, which improves the decomposition efficiency of the conventional MP algorithm (Liu et al. 2004; Liu and Marfurt 2005). Since the atoms in the dictionary are not orthogonal, the vertical projection of the signal residual at the selected atom is non-orthogonal, which makes each iteration not optimal but suboptimal. The proposed orthogonal matching pursuit helps to overcome this drawback, which can avoid the repeated selection of the optimal atoms by orthogonalizing the atoms in the dictionary (Tropp and Gilbert 2008; Miandji et al. 2017).

The methods above are usually single-channel decomposition strategies. The operation order of the single-channel matching pursuit has a great negative influence on the decomposition result. Slight changes in the seismic trace will result in changes in the order of matching pursuit, and the atoms obtained by the decomposition will be unstable (Jun et al. 2012). This leads to the lateral instability and non-uniqueness of the time–frequency analysis results, and it is difficult to ensure the spatial correlation of the decomposition result. Also, the decomposition is easily contaminated by the data noise.

To make full use of the spatial continuity of seismic data, the multichannel matching pursuit (MCMP) exploits lateral coherence as a constraint to improve the uniqueness of the solution (Wang 2010). The genetic algorithm is introduced to the multichannel matching pursuit to reduce the complexity of the algorithm (Jun et al. 2012). Multichannel matching pursuit keeps better stability, but there are some limitations. Some improved algorithms are proposed to optimize the decomposition strategy of the algorithm (Ning and Wang 2014).

We know from sedimentology that the layered lithologic body is mainly distributed underground, and the seismic waves of adjacent seismic traces show better horizontal continuity and spatial coherence in the same seismic horizon (Yuan et al. 2015). The optimal atom searched in the multichannel matching pursuit best matches the average of all seismic traces, rather than the best match for each channel. The algorithm does not take the information of seismic reflection events into account, and the matching optimization direction does not match the direction of the seismic event. Therefore, in this study,

we propose an improved directional multichannel matching pursuit algorithm, which can make full use of the lateral continuity of the seismic data and can directionally search for multichannel atoms. The proposed multichannel matching pursuit has been proved to enhance the resolution of the time–frequency spectrum profile, which help to guide reservoir prediction and hydrocarbon detection.

Methods

Single-channel matching pursuit theory

In matching pursuit, the signal can be expressed in a linear combination of multiple atoms obtained by decomposition. Given a seismic trace $f(t)$, we select a parent function according to the characteristics of the signal to create a redundant dictionary $D = \{g_\gamma(t)\}$, where $g_\gamma(t)$ is an atom defined by parameter group $\gamma = (u, \omega, \varphi)$ and $\left\| g_\gamma(t) \right\| = 1$, where u, ω and φ represent time, frequency and phase, respectively. The signal is decomposed as follows.

Firstly, the atom which matches the signal best is selected from the over-complete atomic library and satisfies the following equation (Mallat and Zhang 1993):

$$\left| \left\langle f(t), g_{\gamma_1}(t) \right\rangle \right| = \sup_{i \in (1,2,\dots k)} \left| \left\langle f(t), g_{\gamma_i}(t) \right\rangle \right| \tag{1}$$

where $g_{\gamma_1}(t)$ is the atom that best matches the signal at the first iteration. $g_{\gamma_i}(t)$ presents each atom in the dictionary. sup is an upper bound operator. k is the number of atoms in the redundant dictionary.

The signal can be expressed as follows:

$$f(t) = \left\langle f(t), g_{\gamma_1}(t) \right\rangle g_{\gamma_1}(t) + R_1 f(t) \tag{2}$$

where R is the residual operator. $R_1 f(t)$ is the residual component of the signal after the first iteration. Continuously performing such an iterative process, the nth decomposition process is:

$$R_{n-1} f(t) = \left\langle R_{n-1} f(t), g_{\gamma_n}(t) \right\rangle g_{\gamma_n}(t) + R_n f(t) \tag{3}$$

where $R_{n-1} f(t)$ and $R_n f(t)$ are the residual signal at $n-1$th and nth step, respectively. $g_{\gamma_n}(t)$ is the atom that fits to $R_{n-1} f(t)$ best at the nth iteration, and $g_{\gamma_n}(t)$ satisfies the following equation:

$$\left| \left\langle R_{n-1} f(t), g_{\gamma_n}(t) \right\rangle \right| = \sup_{i \in (1,2,\dots k)} \left| R_{n-1} f(t), g_{\gamma_i}(i) \right|. \tag{4}$$

After nth iterations, the signal becomes:

$$f(t) = \sum_{i=1}^{n} \left\langle R_{i-1} f(t), g_{\gamma_i}(t) \right\rangle g_{\gamma_i}(t) + R_i f(t) \tag{5}$$

here $R_0 f(t) = f(t)$. As discussed above, the signal can be expressed as the sum of the linear combination of all atoms and the residual signal.

At each step of decomposition, the atom extracted from the redundant dictionary satisfies Eq. (4). When the number of iterations reaches the preset value or the signal residual energy is less than a certain threshold, the decomposition process is completed. Although the matching pursuit can be used to yield a sparse representation model of the signal, this algorithm has the main disadvantage of a large amount of calculation (Masood and Al-Naffouri 2013).

Directional multichannel matching pursuit theory

Given M seismic traces $\{f_1(t), f_2(t), \ldots, f_M(t)\}$, we start the first iterative decomposition from the first seismic trace $f_1(t)$. Similar to the strategy of searching for the optimal atom in the single-channel matching pursuit algorithm, we decompose the signal from the time position of the maximum envelope amplitude. Utilize the transient characteristics of the time position to build the dynamic atomic library, and search for the atom $g_{\gamma_1}(t)$ that best matches the signal at the first iteration. The instantaneous phase φ_1, instantaneous frequency ω_1 and central time u_1 parameters of the optimal atom are recorded. Then estimate the amplitude of the optimal atom corresponding to the first trace according to the following equation:

$$a_1 = \frac{\left| \left\langle f_1(t), g_{\gamma_1}(t) \right\rangle \right|}{\left\| g_{\gamma_1}(t) \right\|^2}. \tag{6}$$

The signal residual can be expressed as follows:

$$R f_1(t) = f_1(t) - a_1 g_{\gamma_1}(t) \tag{7}$$

where R represents residual signal. Considering that the same interface's event of reflected waves often has similar waveform characteristics, we use the similarity of seismic wavelet waveform to identify the seismic event and realize the fast decomposition of adjacent seismic traces. The search strategy is shown in Fig. 1.

The red wavelet is the optimal atom determined by the first search. The green dot is the central time position u_1 of the optimal atom. For the search of the next seismic trace, we only need to select the appropriate delay time. After the delay of the optimal wavelet determined by the search of the previous channel, we can do correlation analysis with the cableway to be inspected to obtain the maximum correlation coefficient, which shows that the optimal wavelet is the most
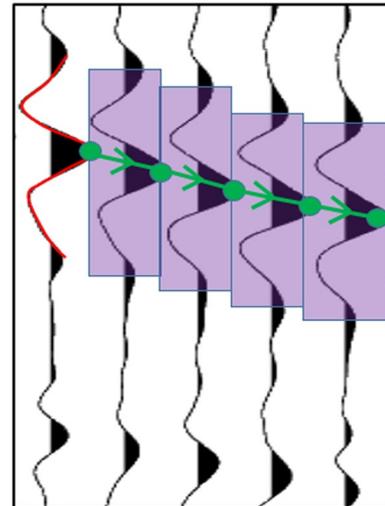
Fig. 1 Directional multichannel search schematic

similar to the seismic channel to be retrieved after the time delay. The correlation coefficient is calculated as follows:

$$g_{\gamma_2}(t) = \arg\max \left| \left\langle f_2(t), g_{\gamma_1}(t - \tau) \right\rangle \right| \tag{8}$$

where τ is the time length of sliding up and down of the optimal wavelet in correlation calculation. As shown in Fig. 1, the purple shadow area represents the size of the sliding window, that is, first determine the time position of the maximum instantaneous amplitude closest to the center time of the optimal atom obtained from the previous seismic channel search, and the time position of the maximum instantaneous amplitude closest to the top and bottom of this position is the upper and lower limit of the sliding window.

The optimal atoms obtained in one iteration of multichannel search have the same frequency and phase parameters, different center time and amplitude parameters. Therefore, it is only necessary to record the center time and amplitude parameters of the optimal atom. For the next seismic channel to be decomposed, we need to update the center time of the optimal atom and determine the sliding window and then calculate the correlation coefficient between the atom and the seismic channel to be decomposed. The amplitude of the adjacent wavelets on the same reflection event changes gradually. When the amplitudes of the adjacent seismic channels change greatly, it is considered that there are discontinuous layers, faults or pinch-out points, etc., and then, the current multichannel decomposition is stopped. Stop the first iteration of the multichannel search until all seismic traces have completed or the amplitude ratio does not satisfy the following formula:

$$\alpha_1 \geq \frac{a_{k+1}}{a_k} \geq \alpha_2 \tag{9}$$

Fig. 2 Estimated results of theoretical data using the proposed multichannel matching pursuit. **a** A synthetic 2D seismic profile. The matching atoms (**b**) and the corresponding residual (**c**) identified by the third iteration, respectively. The matching atoms (**d**) and the corresponding residual (**e**) identified by the sixth iteration, respectively. **f** The reconstructed seismic profile. **g** The final residual

where α_1 and α_2 are the multichannel search threshold to ensure the existence of the fault and the discontinuous layer in the reconstructed seismic traces. After testing, we think

that amplitude ratio $\alpha_1 = 1.5$, $\alpha_2 = 0.5$ can get better decomposition results. It is not perfect to use only the amplitude information as the parameter indicating the optimal atom,

Fig. 3 Estimated results of noisy theoretical data using the proposed multichannel matching pursuit. **a** A noisy synthetic 2D seismic profile with a signal-to-noise ratio of 2:1. The matching atoms (**b**) and the corresponding residual (**c**) identified by the sixth iteration, respectively. **d** The reconstructed seismic profile. **e** The final residual. **f** The added random noise

so it is possible to constrain the lateral search direction by combining coherence and phase parameters in the future.

Results

Synthetic examples

Considering the wide application of Ricker wavelet in modeling, processing, inversion and interpretation of seismic data (Yuan et al. 2019), we use Ricker wavelet dictionary to decompose and reconstruct model data and field data. To verify the feasibility of the directional multichannel

matching pursuit method, we build a geological model of multilayer sand bodies to get the synthetic seismic data. As shown in Fig. 2a, the synthetic seismic data consist of 700 traces with the vertical length of 800 ms. The dominant frequency of the zero-phase Ricker wavelet is 30 Hz, and the vertical sampling interval is 2 ms. We apply the proposed multichannel matching pursuit to decompose and reconstruct the theoretical data.

In order to better understand the decomposition process of the algorithm, we demonstrate the iterative decomposition process in detail. Figure 2b, d shows the matching atoms identified by the third and sixth iteration of the first seismic trace using the proposed multichannel matching

pursuit, respectively. Figure 2c, e displays the corresponding residual seismic profiles, respectively. For the third iteration of the first trace, we first scan the dynamic atomic library to determine parameters of the optimal atom, and then, we just need to calculate the time shift correlation between this atom and the next seismic channel. Continuously update the center time and amplitude parameters of the optimal atom. Figure 2b shows the optimal atoms determined by the above steps. The optimal atoms shared by lateral seismic traces have different time parameters, and the directions are consistent with the seismic events. The atoms estimated by the multichannel search are best matched to each seismic trace, not the best match to the average of all the seismic traces.

Therefore, the proposed multichannel decomposition strategy makes the signal residuals get the fastest convergence and decline. We set the amplitude ratio as the criterion to cease the multichannel search, which can guarantee the accuracy of the reconstructed seismic traces.

We can see from Fig. 2d that the horizon continuity of estimated atoms is preserved well, and the atoms determined by multichannel search keep high consistency with seismic reflection events. The matching atoms shown in Fig. 2d only need to search the atomic library six times, while the single-channel matching pursuit algorithm needs to search the atomic library once to determine each optimal atom. The proposed algorithm greatly reduces the time cost of
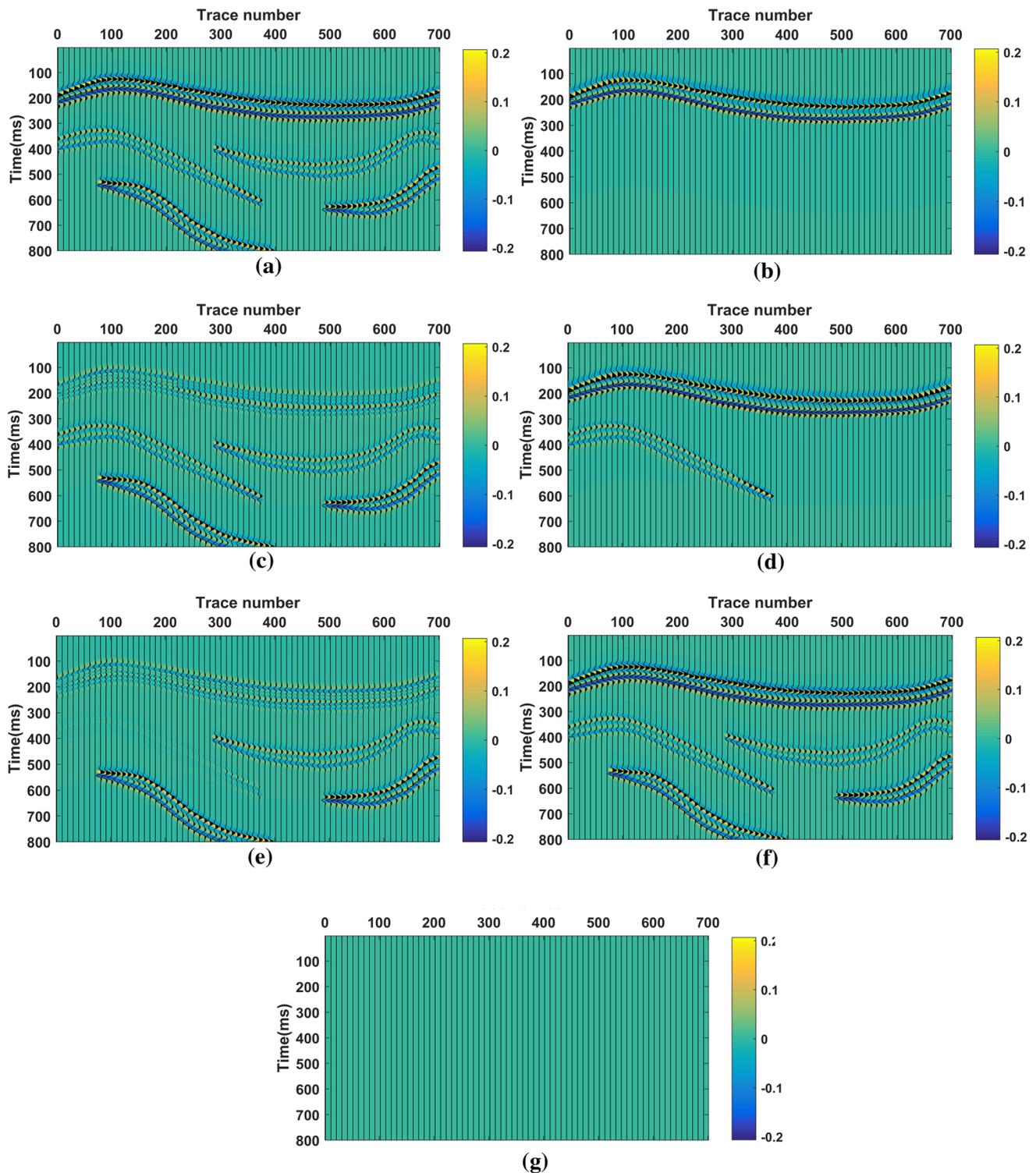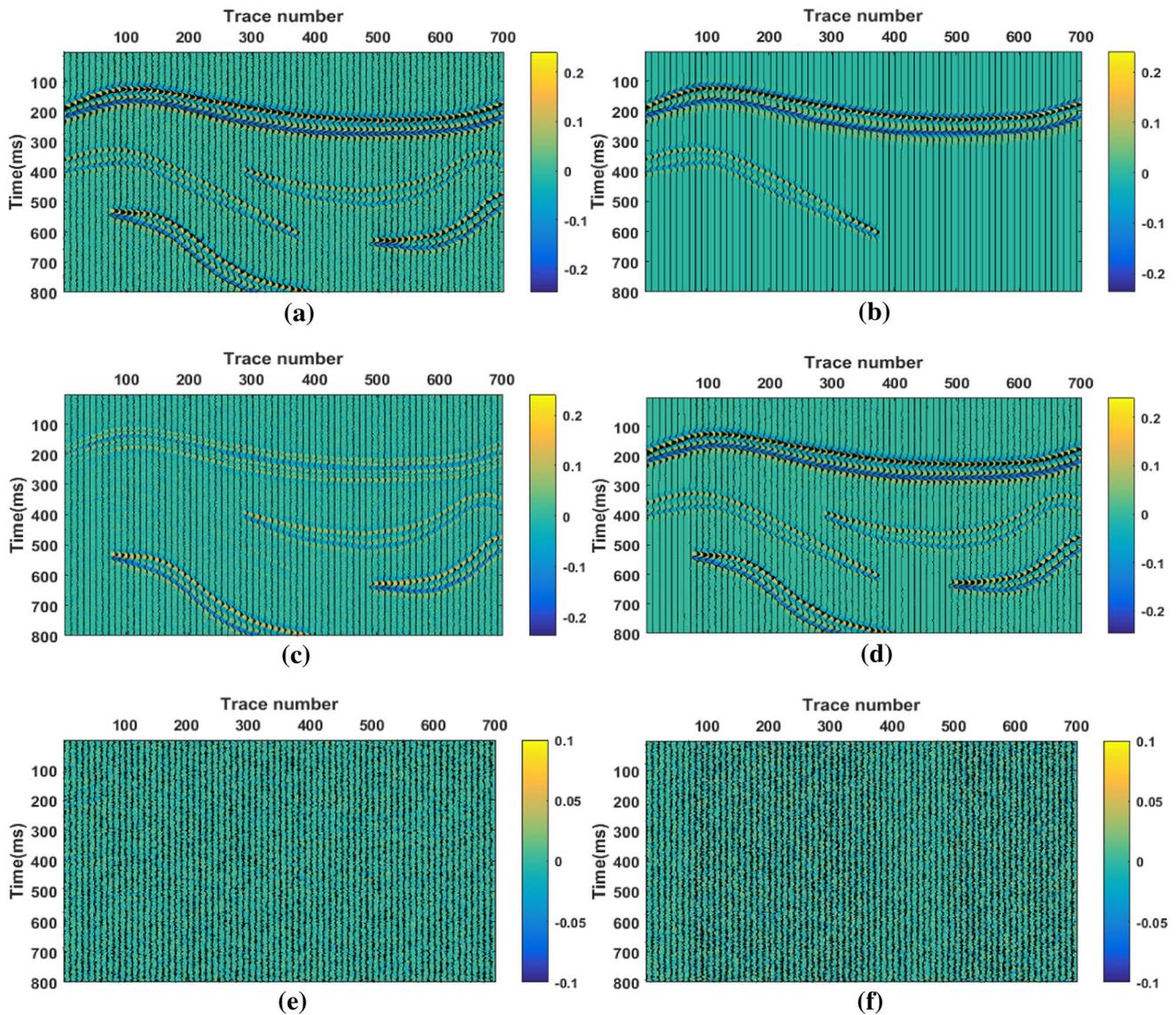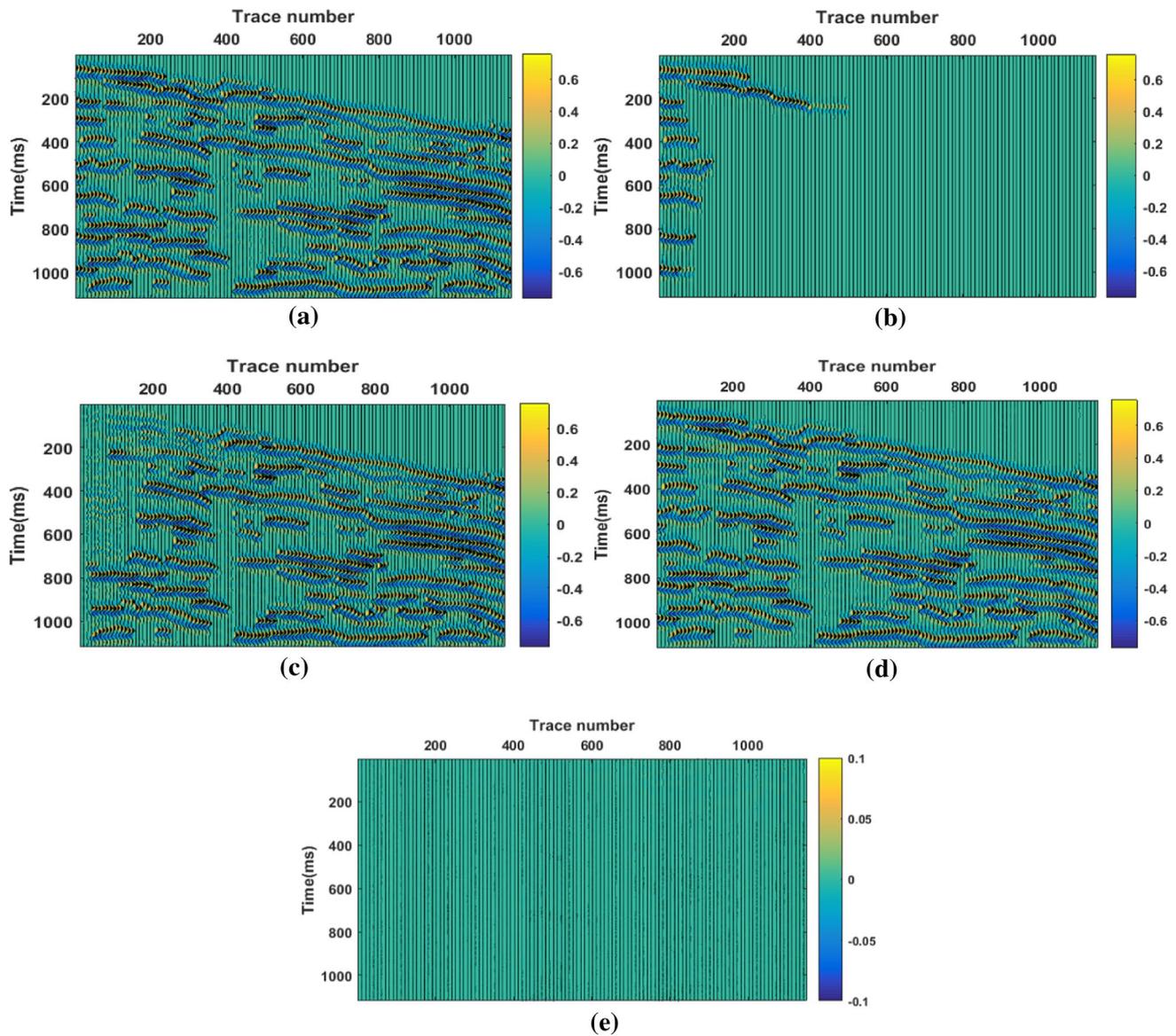

Fig. 4 Estimated results of complex theoretical data using the proposed multichannel matching pursuit. **a** A synthetic 2D seismic profile. The matching atoms (**b**) and the corresponding residual (**c**) identified by the 12th iteration, respectively. **d** The reconstructed seismic profile. **e** The final residual

searching the optimal atom. Figure 2f shows the final result of the reconstructed profile after 120 iterations. The value of the seismic residual shown in Fig. 2g is close to zero, and the reconstructed seismic profile shares the same characteristics with the theoretical data as shown in Fig. 2a.
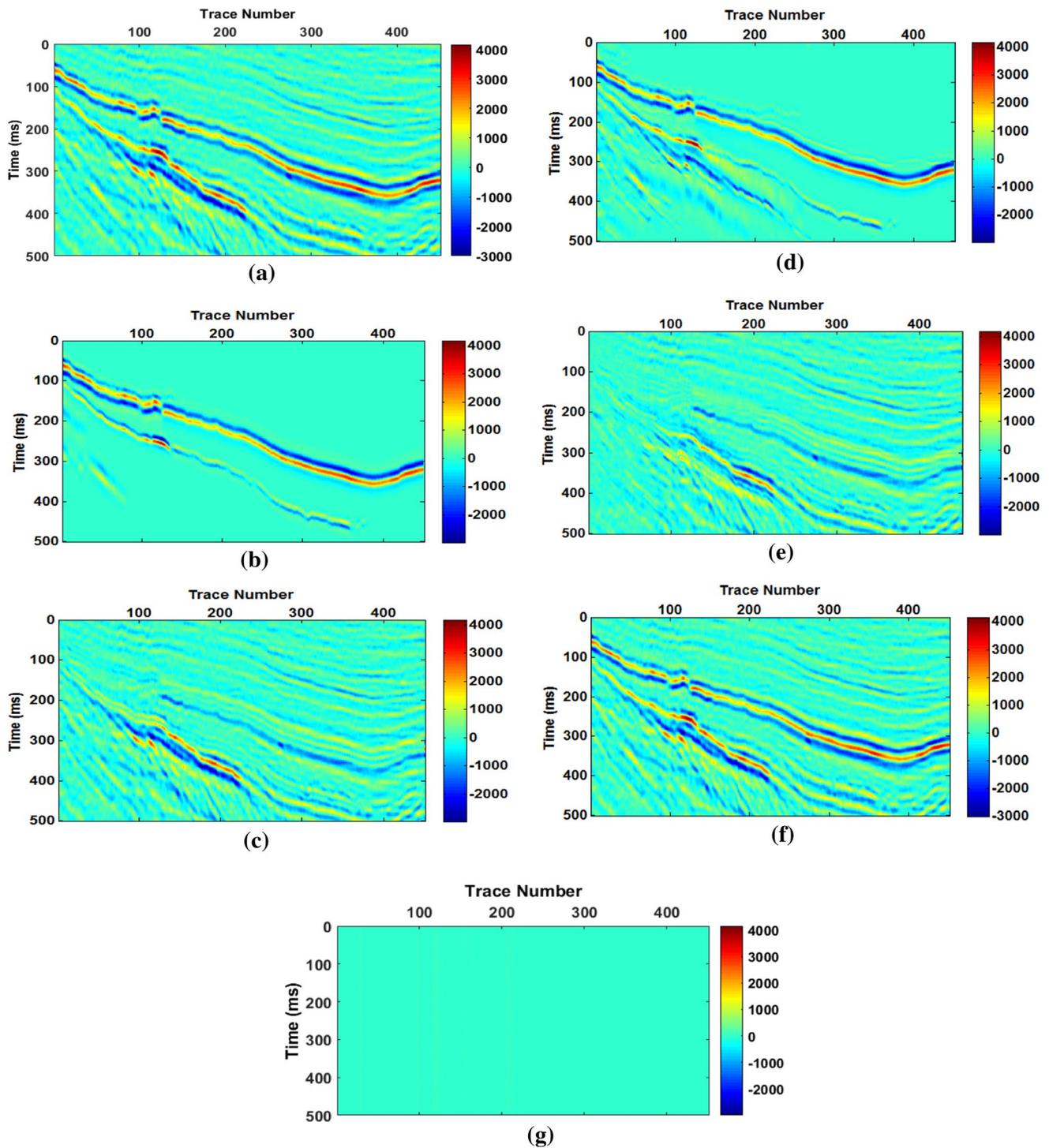
Fig. 5 The iterative decomposition process of field data using the proposed multichannel matching pursuit. **a** A seismic profile extracted from a 3D seismic cube. The matching atoms (**b**) and the corresponding residual (**c**) identified by the fifth iteration, respec-tively. The matching atoms (**d**) and the corresponding residual (**e**) identified by the 30th iteration, respectively. **f** The reconstructed seismic profile. **g** The final residual
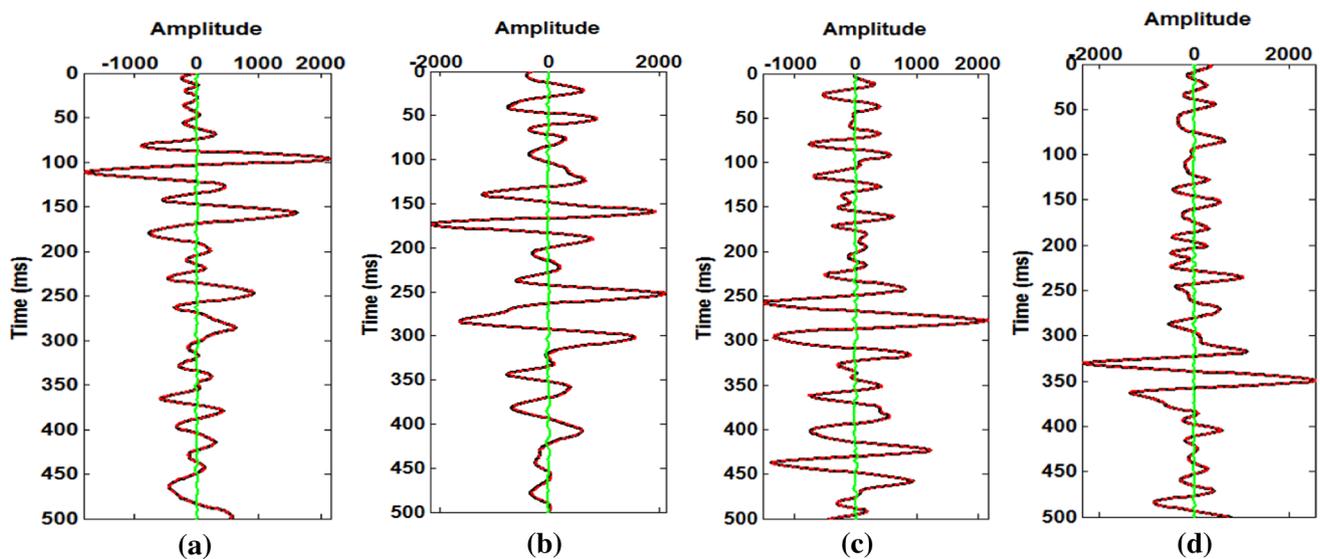
Fig. 6 Reconstruction precision of different trace **a** the 30th trace; **b** the 110th trace; **c** the 260th trace; **d** the 410th trace (red lines represent reconstructed signals with the proposed matching pursuit, black lines represent original signals, green lines represent residual signals)

To test the stability of the proposed method, we add random noise to get the model data shown in Fig. 3a with signal-to-noise ratio of 2:1. Figure 3b, c displays the matching wavelets and the corresponding residual profile identified by the sixth iteration, respectively. We can see that the decomposition iteration process of the proposed method is relatively stable although in the presence of random noise. The matching wavelets shown in Fig. 3b keep good lateral continuity. Figure 3d, e is the final results of the reconstructed profile and the corresponding residual profile. We can see that the reconstructed profile contains less noise than the noisy synthetic data in Fig. 3a. The residual profile shown in Fig. 3e does not contain valid reflection information and keeps a high similarity with the random noise profile shown in Fig. 3f, which proves that the proposed directional multichannel matching pursuit has a certain anti-noise ability.

In addition, we built a complex geological model to verify the applicability of the proposed method. The model data shown in Fig. 4a is composed of 28 Hz Ricker wavelet, and Fig. 4b, c displays the matching wavelets and the corresponding residual identified by the 12th iteration, respectively. We can see that even for complex geological model data, the proposed method can still identify multichannel matching wavelets along the direction of the seismic reflection events. The matching wavelets shown in Fig. 4b are obtained by only iteratively decomposing the first seismic trace for 12 times. The reconstructed profile in Fig. 4d obtained by 430 iterations keeps high consistency with the original profile in Fig. 4a. The residual profile in Fig. 4e does not contain effective reflection information, which proves that the proposed matching pursuit can realize the directional decomposition of the synthetic data.

Field examples

Furthermore, the proposed directional multichannel matching pursuit is applied to the field data as displayed in Fig. 5a to test the stability and reliability. The data consist of 450 traces with a longitudinal time of 500 ms. The vertical sampling interval is 1 ms. Figure 5b, d shows the matching wavelets identified by the 5th and 30th iteration of the first seismic trace using the proposed multichannel matching pursuit, respectively. Figure 5c, e displays the corresponding residual seismic profiles. Although we only decompose the first seismic trace, we can accurately estimate the optimal atoms of adjacent seismic traces shown in Fig. 5b, d. The matching wavelets between adjacent seismic traces keep better horizon continuity. We can see that the center time of the extracted optimal atoms is different, which coincides with the trend of stratigraphic structure. We can see from Fig. 5d that the proposed algorithm can adaptively break a multichannel search at the location of the pinch out, ensuring the accuracy of reconstructing seismic traces. The algorithm can estimate many matching wavelets by iterating a few times, which can make full use of the lateral continuity of seismic traces.

Figures 5f shows the reconstructed field seismic profile obtained by 870 iterations, which accurately resembles the original seismic section shown in Fig. 5a. We can see that the weak seismic signals can also be better identified. The corresponding decomposition residual is displayed in Figs. 5g. The mismatch value is around zero, which demonstrates the high decomposition and reconstruction precision of the proposed multichannel matching pursuit in the field data.

To verify the accuracy of the algorithm for data reconstruction, we randomly extract several seismic traces from the reconstructed section and the original section shown in Fig. 5a, f, respectively. Figure 6a–d displays the decomposition results of the 30th, 110th, 260th and 410th trace (red means reconstruction signal, black means the original signal, green means residual signal) with the proposed multichannel matching pursuit, respectively. We can see from Fig. 6 that the green curve is close to zero and the red curve is highly consistent with the black line, which proves that the reconstruction accuracy of the algorithm is preserved well and the algorithm can be used for signal processing.

And then, the proposed multichannel matching pursuit is applied to generate time–frequency spectrum for the oil and gas reservoir. Spectral decomposition has been widely used in the field of geophysics, which can convert the seismic traces from the time domain to the time–frequency domain. It can display more characteristics of the target reservoir and play an important role in oil and gas exploration. Figure 7a shows the seismic profile across three wells. The black seismic event at about 40 ms is the gas reservoir, and the black axis at 80 ms is the oil reservoir.

Figure 7b, c shows the 30-Hz spectral profiles estimated by the proposed multichannel matching pursuit, respectively. We can see that the spectral components exhibit strong amplitudes at the gas and oil layers, and the effective time–frequency information gradually decreases with the increase in frequency. Figure 7d, e is the frequency profiles of 45-Hz obtained by single-channel matching pursuit. Through comparison, we find that the formation continuity of the proposed matching pursuit is better than the single-channel matching pursuit, which may be due to the fact that the proposed method searches for multichannel optimal atoms along the seismic reflection events. The frequency profiles shown in Fig. 7d, e are difficult for us to distinguish the boundary between layers. The proposed algorithm has significantly improved time and frequency resolution relative to the single-channel matching pursuit. Therefore, the directional multichannel matching pursuit method can achieve better application results in time–frequency analysis.

Conclusion

A novel multichannel matching pursuit is proposed in this study, which could realize the directional multichannel decomposition of seismic traces. The optimal atoms estimated by the proposed multichannel search are not at the
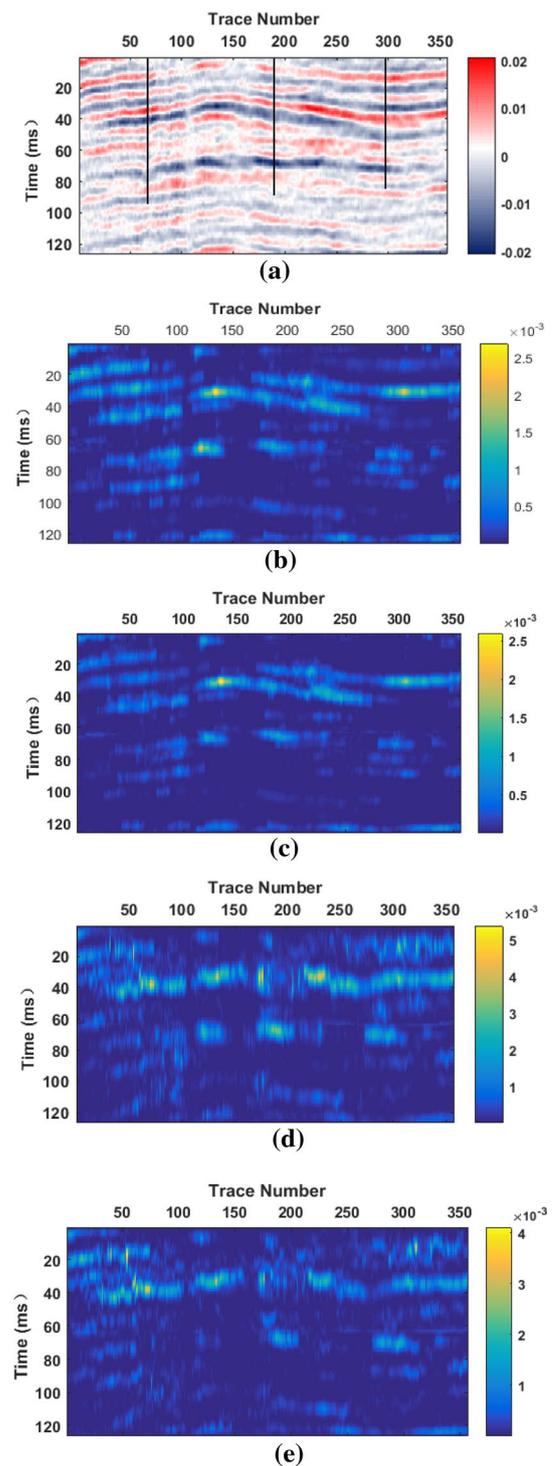

Fig. 7 Time–frequency decomposition to detect reservoir. **a** A seismic line across three wells. **b** 30 Hz and **c** 45 Hz spectral components estimated by proposed multichannel matching pursuit. **d** 30 Hz and **e** 45 Hz spectral components estimated by single-channel matching pursuit

same time slicers, which is consistent with the direction of seismic reflection events. The synthetic test and field data case illustrate the feasibility and stability of the directional multichannel matching pursuit. The application of time–frequency analysis contributes to the identification of oil and gas reservoir. The formation continuity and time–frequency resolution of spectrum profile are inferior to the constant spectrum profile generated by single-channel matching pursuit. The comparisons in Fig. 7 illustrate the superiority of our methodology in time–frequency analysis. The application example shows that the proposed multichannel matching pursuit can be better used for geophysical exploration.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Duijndam AJW, Schonewille MA (1999) Nonuniform fast fourier transform. Geophysics 64(2):539–551

Durka PJ, Matysiak A, Montes EM, Sosa PV, Blinowska KJ (2005) Multichannel matching pursuit and EEG inverse solutions. J Neurosci Methods 148(1):49–59

Gribonval R, Bacry E, Mallat S, Depalle P, Rodet X (1996) Analysis of sound signals with high resolution matching pursuit. In: IEEE-SP International Symposium on Time-frequency & Time-scale Analysis, IEEE. https://doi.org/10.1109/TFSA.1996.546702

Hess-Nielsen N, Wickerhauser MV (1996) Wavelets and time-frequency analysis. Proc IEEE 84(4):523–540

Jun W, Lijun W, Yanjiang W (2012) Seismic signal fast decomposition by multichannel matching pursuit with genetic algorithm. In: Presented at the 11th international conference on signal processing, IEEE

Lin JL, Hwang WL, Pei SC (2006) Multiple blocks matching pursuit update algorithm for low bit rate video coding. IEEE Trans Circuits Syst Video Technol 16(3):331–337

Liu J, Marfurt KJ (2005) Matching pursuit decomposition using Morlet wavelets. In: SEG technical program expanded abstracts, pp 786–789

Liu J, Wu Y, Han D, Li X (2004) Time-frequency decomposition based on Ricker wavelet. In: SEG technical program expanded abstracts, pp 1937–1940

Mallat S, Zhang Z (1993) Matching pursuit with time-frequency dictionaries. IEEE Trans Signal Process 41(12):3377–3415

Masood M, Al-Naffouri TY (2013) Sparse reconstruction using distribution agnostic bayesian matching pursuit. IEEE Trans Signal Process 61(21):5298–5309

Miandji E, Emadi M, Unger J, Afshari E (2017) On probability of support recovery for orthogonal matching pursuit using mutual coherence. IEEE Signal Process Lett 24(11):1646–1650

Ning C, Wang Y (2014) An improved method for multichannel matching pursuit decomposition. In: 76th EAGE Conference and Exhibition 2014, p 1–5. https://doi.org/10.3997/2214-4609.20141566

Qian S, Chen D (1999) Joint time-frequency analysis. IEEE Signal Process Mag 16(2):52–67

Salomon BG, Ur H (2004) Sparse approximations with a high resolution greedy algorithm. In: IEEE International Conference on Electronics IEEE. https://doi.org/10.1109/ICECS.2004.1399685

Sinha S, Routh PS, Anno PD, Castagna JP (2005) Spectral decomposition of seismic data with continuous wavelet transform. Geophysics 70(6):P19–P25

Stefanoiu D, Lonescu F (2003) A genetic matching pursuit algorithm. In: International symposium on signal processing & its applications. IEEE

Stockwell RG, Mansinha L, Lowe RP (2002) Localization of the complex spectrum: the s transform. IEEE Trans Signal Process 44(4):998–1001

Tropp JA, Gilbert AC (2008) Signal recovery from random measurements via orthogonal matching pursuit: the gaussian case. IEEE Trans Inf Theory 53(12):4655–4666

Wang Y (2010) Multichannel matching pursuit for seismic trace decomposition. Geophysics 75:V61–V66

Yuan S, Wang S, Luo C, He Y (2015) Simultaneous multitrace impedance inversion with transform-domain sparsity promotion. Geophysics 80(2):R71–R80

Yuan S, Ji Y, Shi P, Zeng J, Gao J, Wang S (2019) Sparse bayesian learning-based seismic high-resolution time-frequency analysis. IEEE Geosci Remote Sens Lett 16(4):623–627

RESEARCH ARTICLE - APPLIED GEOPHYSICS

Distortion effects caused by target abnormal bodies in CSAMT exploration

Xian-Xiang Wang[1] · Ju-Zhi Deng[1] · Jing-Li Ren[1]

Abstract

In CSAMT exploration, the using of the artificial sources not only improves the signal-to-noise ratio of the data, but also brings a series of distortion effects, such as shadow and source overprint effects. This paper attempts to introduce a distortion effect caused by the target in the survey area. Although it is often ignored, it always plagues the data interpretation. In CSAMT method, the primary current has determined direction due to the source. When the primary current encounters electrical interfaces, the induced charge will accumulate on it and generate local current, causing local distortion. The anomaly body stretches in the direction of the vertical primary current, and a false anomaly with opposite polarity appears on both sides of the target. If the direction of the primary current is different, the accumulation position of the induced charge is also different, which will result in different shapes of the anomalies in observed data. This paper confirms the existence of the distortion by taking four simple models as examples and explains it from the physical mechanism. On this basis, the paper summarizes the relationship between inversion and distortion. If our code can simulate the distortion effect in the forward, we do not need to remove it before the inversion. Otherwise, it must be removed.

Keywords Controlled source audio-frequency magnetotellurics · Distortion effects · Galvanic effects · Target abnormal body

Introduction

Controlled source audio-frequency magnetotellurics (CSAMT) is an artificial source electromagnetic method derived from magnetotellurics (MT). The method was first proposed by Professor Strangway and Goldstein of the University of Toronto (1975). The frequency range of the CSAMT is usually $0.1–10^5$ Hz, and the exploration depth is about 1–2 km. Due to artificial sources, this method has stronger anti-interference ability. In recent years, it has been widely applied in the exploration of metal ore, geothermal, groundwater, and hydrocarbons and has been developed into an effective method in geophysical exploration (Di et al 2002, 2004, 2018; An et al. 2013a, b, 2016; Fu et al. 2013; Wang et al. 2015; Lei et al. 2017a, b; Wynn et al. 2016).

In electromagnetic exploration, distortions due to local anomalies at the surface are ubiquitous. Berdichevsky and Dmitriev (1976) conducted a study on it and divided it into inductive and galvanic distortion. Many techniques have been proposed to remove galvanic effects, such as impedance tensor decomposition (Swift 1967; Groom and Bailey 1989; Bibby et al. 2005; Jones 2012; Neukirch et al. 2019, 2020), equivalent source technique (MacLennan and Li 2013; Tang et al. 2018), 3D inversion considering galvanic distortion (Avdeeva et al. 2015; Li et al. 2016). In artificial-source electromagnetic method, apart from the galvanic effects in MT, the man-made source also brings a series of distortions. The abnormal body beneath the source may influence observed data, which is called source overprint effects (Nabighian 1991). When the abnormal body is between the source and receiver, shadow effects will happen (Nabighian 1991). These distortions affect the reliability of the interpretation to some extent. In recent years, many efforts have been focusing on these distortions and a series of articles have been published (Zonge and Hughes 1991; Kuznetzov 1982; Sternberg and Washburne 1988; Yan and Junmei 2004; Wang et al. 2009; Lei et al. 2017a, b; Zhou et al.

✉ Xian-Xiang Wang
 wangxianxiang09@163.com

1 Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, East China University of Technology, Nanchang 330013, Jiangxi, China

2018). However, it seems that the distortion effect caused by the target abnormal bodies in the exploration area has not been given high priority. Since the distortion is caused by the target, if not taken seriously, it will bring greater difficulty to the interpretation of the CSAMT data. In this work, we addressed this problem modeling such distortion effect, exploring how to deal with it.

Physical mechanism of Galvanic effects

When the current encounters electrical interfaces, induced charges will accumulate on it and start to move under the external electric field to produce galvanic effects. The observed field is composed of two parts: One is the primary electric field and the other is the secondary electric field generated by the induced charges. According to Ohm's law,

$$\mathbf{J} = \sigma \mathbf{E} \tag{1}$$

Here, σ and $\mathbf{E}$ represent the conductivity and electric field, respectively. We can get the following formula based on boundary conditions:

$$D_{1n} - D_{2n} = \rho_s \tag{2}$$

where $\mathbf{D}$ is the electric displacement vector, the subscript n represents the perpendicular to the interface, ρ_s is the surface charge density, 1 represents medium 1, 2 represents medium 2. And the electric displacement vector is:

$$\mathbf{D} = \varepsilon \mathbf{E} \tag{3}$$

Here, ε is the dielectric constant in F/m. Considering Eqs. 1, 2 and 3, we can derive the following expressions:

$$\rho_s = \varepsilon_1 E_{1n} - \varepsilon_2 E_{2n} = J_n \left(\frac{\varepsilon_1}{\sigma_1} - \frac{\varepsilon_2}{\sigma_2} \right)$$
$$\approx J_n \varepsilon_0 \frac{\sigma_2 - \sigma_1}{\sigma_1 \sigma_2} \tag{4}$$

In the low frequency range, according to Coulomb's law, the secondary electric field is (Jiracek 1990):

$$\mathbf{E}_s = -\nabla \varphi = -\frac{1}{4\pi\varepsilon} \int_S \frac{\rho_s}{|r|^3} r \, \mathrm{d}S \tag{5}$$

where φ is the electrostatic potential, $\mathbf{r}$ is the vector from the differential surface element $\mathrm{d}S$ to the observation point.

When the current encounters electrical interfaces, it is known from Eq. 4 that the polarity of surface charge ρ_s is determined by the relative relationship between the conductivity of the two sides, and the direction of the primary current. The same quantity, but opposite polarity charges will accumulate on both sides of the abnormal body. When encountering a conductive body, it is negative in the front

end of the abnormal body and is positive in the rear end. The field distribution, from accumulative charges, is determined by Eq. 5. The electric field distribution should be as shown in Fig. 1. When the abnormal body is conductive, the secondary current, generated due to the accumulation of induced charges, will be as shown in Fig. 1a. This figure shows that the direction of the secondary current is opposite to the primary current in region 1, which will reduce the total current intensity; the direction of the secondary current is the same as the primary current in region 2, which will increase the total current. When the abnormal is resistive, the secondary current is illustrated in Fig. 1b. The direction of the secondary current in region 1 is the same as the direction of the primary current, which will increase the total current intensity; the direction of the secondary current is opposite to the primary current in region 2, which will weaken the total current intensity.

Distortion effects

To better understand the distortion we proposed, we built four models which are model A, model B, model C, model D, respectively. The background of four models is uniform half space with a resistivity of 1000 Ω.m. In each model, there is a block whose parameters are shown in Table 1. We define the direction of the source as the x axis, the z-axis downward, and the coordinate origin at the center of anomalies on the ground. Our forward is based on PIE3D code with the integral equation (IE) method (Zhdanov et al. 2006). The anomalies are discretized into many cells with the size of 10 m × 10 m × 10 m.
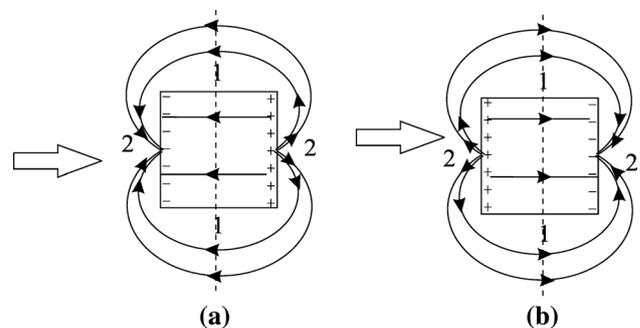
Fig. 1 The current distribution caused by galvanic effects. **a** The current distribution caused by galvanic effects for the conductive body. **b** The current distribution caused by galvanic effects for the resistive body. The arrows indicate the current flow and the dotted lines indicate the cross section perpendicular to the current flow

Table 1 The parameters of four models

Model name	The resistivity (hom.m)	The size	Top plate depth (m)	The source position	The angle between the anomaly and the source (°)
Model A	10	100 m × 100 m × 100 m	50	0, − 10000.0	0
Model B	10	100 m × 400 m × 100 m	50	0, − 10000.0	90
Model C	10	100 m × 400 m × 100 m	50	0, − 10000.0	45
Model D	10	100 m × 100 m × 100 m	50	− 342.02, − 9396.920	0
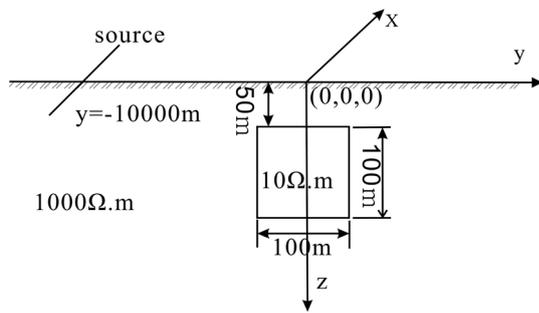
Fig. 2 The sketch of model A with sole conductive anomalous body. The figure is out of scale

Model A

The model A is shown in Fig. 2. The background is uniform half space with a resistivity of 1000 Ω.m. The abnormal body has a size of 100 m × 100 m × 100 m, a top buried depth of 50 m and a resistivity of 10 Ω.m. The length of source is 1 km and the offset is 10 km. A total of 19 survey lines are arranged parallel to the source above the anomalous

body. The length of each survey line is 800 m, the line spacing is 50 m, and the station spacing is 20 m.

Figure 3 shows the plane view of CSAMT apparent resistivity and phase for 100 Hz. Figure 4 shows the MT apparent resistivity and phase. The anomalous body designed in this example is a cube, the projection on the surface should be a square, and the MT synthetic results are basically consistent with this situation. In the plane view of CSAMT apparent resistivity, the abnormal body is stretched along the y direction, and there are two resistive anomalies on both sides of the target. Apart from the above changes, there is also a change in the phase along the y direction, which is mainly caused by the propagation of electromagnetic field (Fig. 3b). The observation results are the superposition of the two changes. From the above analysis, we can see that the CSAMT method can well reflect the single abnormal body, but there are distortions in some areas compared with the MT method. If these distortion effects were not processed properly, they will inevitably cause the difficulty of interpretation.

To explain the distortion effects in physical mechanism, the distribution of electromagnetic field for 100 Hz

Fig. 3 The plane view of CSAMT apparent resistivity and phase for 100 Hz when $z = 0$ m (model A). **a** The plane view of apparent resistivity, **b** the plane view of phase

Fig. 4 The plane view of MT apparent resistivity and phase for 100 Hz when $z = 0$ m (model A). **a** The plane view of apparent resistivity, **b** the plane view of phase

is shown in Fig. 5. Figure 5a shows the total electric field distribution. It can be seen from the figure that the total electric field is along the x direction except the place which is above the anomaly, and the electric field intensity is almost the same. Above the abnormal body, the electric field is obviously weakened, and the flow direction also changed. Figure 5b shows the total magnetic field distribution. The total magnetic field is stronger near the source and weaker further from the source, and there is no obvious anomaly above the target, indicating that the abnormal body has little effect on the magnetic field. Figure 5c shows the distribution of secondary electric field caused by the anomalous body. It is possible to be seen that the secondary electric field intensity is of the same order as the total electric field. As indicated in Fig. 1, in region 1, the direction of secondary electric field is opposite to the primary electric field. So the total electric field is reduced, and a conductive abnormality appears in this region. In region 2, the direction of the secondary electric field follows the main trend of primary field, and the electric field is strengthened, showing a resistive anomaly. These explain the two distortions from the physical mechanism. Figure 5d illustrates that the secondary magnetic field is very weak and is one-tenth of the total magnetic field.

According to Jiracek (1990), the electric field dominates in the galvanic effects, mainly causing the change of the electric field. The magnetic field dominates in the inductive effects, causing changes of the electric field and the magnetic field. Based on the above analysis, the intensity of the electric field caused by the abnormal body is almost the same as the total electric field. The intensity of the magnetic field is much weaker than the total magnetic field. And, the direction of secondary electric field is also very consistent with Fig. 1a. The above two points indicate that the galvanic

effects dominate in this model, and the inductive effect plays a secondary role.

Model B

In CSAMT exploration, survey lines are generally required to cross the target. In model B, the width of the anomalous body is 100 m, the length is 400 m, and the thickness is 100 m. The source and survey lines are perpendicular to the strike of the conductive body. Other parameters are the same as the model A (Fig. 6).

Figure 7 is the plane view of CSAMT apparent resistivity and phase for model B. Similar to model A, the anomaly body still extends in the y direction, and false anomalies appear on both sides of the target (Fig. 7), which shows that the accumulation position of induced charges and the direction of the secondary current in model B are the same as model A (Fig. 8b). As the conductive block stretches in the y direction compared with model A, the anomalies in plane view of CSAMT apparent resistivity and phase, also have the same extension. With the offset increasing, the ground wave gradually attenuates, and the proportion of the surface wave gradually increases. Figure 7b and d are the plane views of apparent resistivity and phase when the offset is 50 km. There are certain changes in apparent resistivity and phase with the offset changing from 10 km to 50 km. The apparent resistivity over the conductive body rises, while the apparent resistivity of false anomalies on both sides of the target decreases, which proves that the distortion effects is reduced. In addition, the CSAMT phase becomes larger and is closer to the MT phase. When the offset is 10 km, the maximum value of total electric field is 8.18×10^{-7} V/m, the maximum value of the secondary electric field is 6.15×10^{-7}
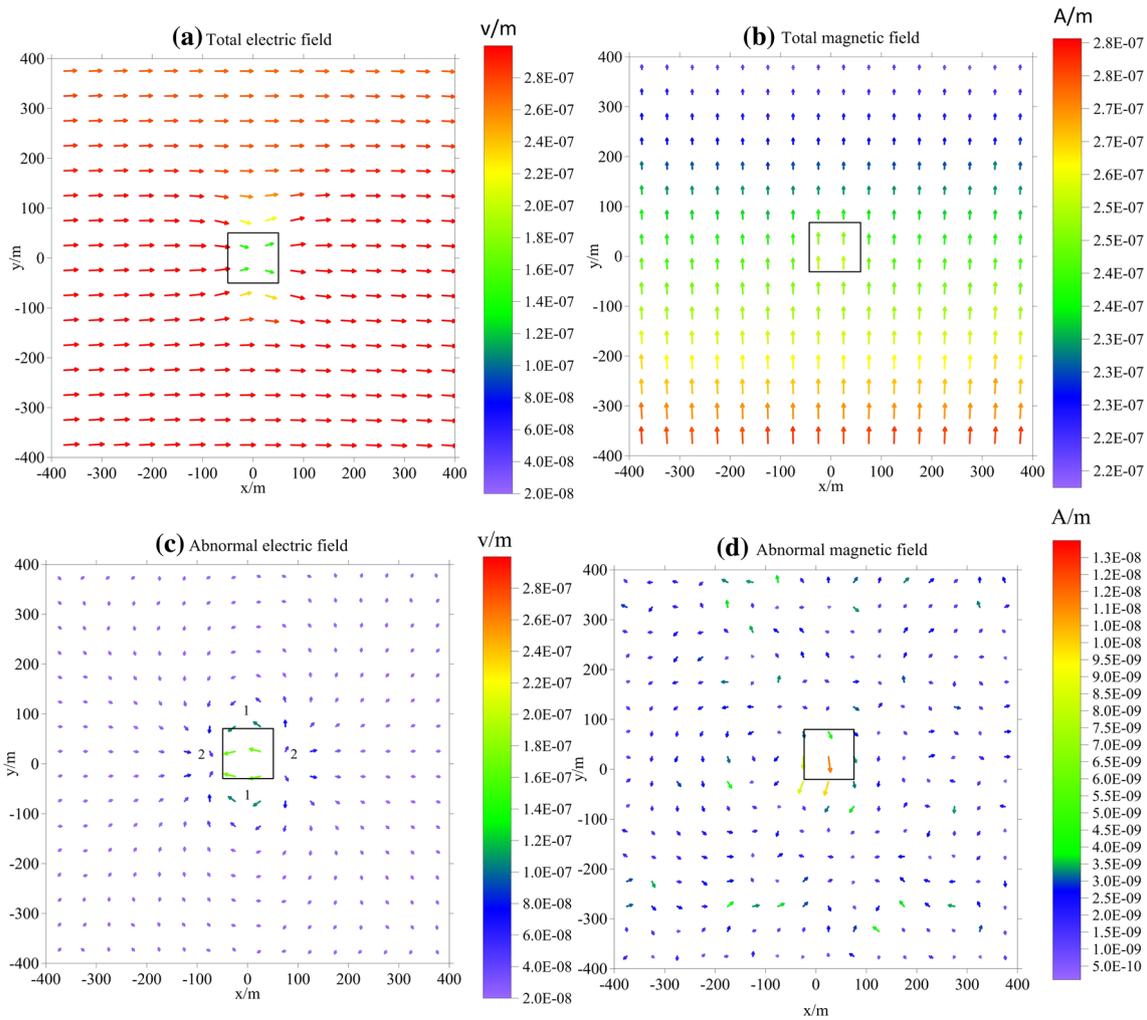
Fig. 5 The distribution of the electromagnetic field when $z = 0$ m (model A). **a** The total electric field, **b** the secondary electric field, **c** the magnetic field when offset = 50 km, **d** the secondary magnetic field
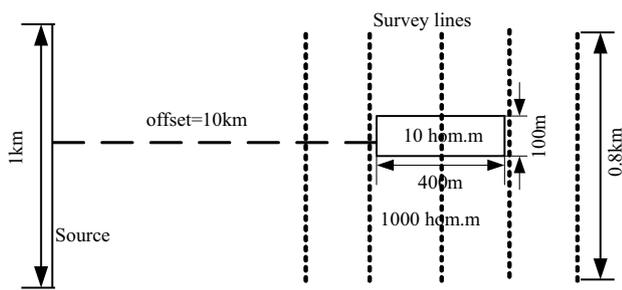
Fig. 6 Plane view of the observation device for model B. The figure is out of scale

V/m which accounts for 75.2% of the total electric field (Fig. 8a, b). When the offset becomes 50 km, the maximum value of the secondary electric field is 2.70×10^{-9} V/m which accounts for 47.3% of the total electric field

(Fig. 9a, b). As the offset increases, the proportion of the secondary electric field decreases significantly, while the proportion of the secondary magnetic field in the total magnetic field increases. The above research shows that the galvanic effects reduce, while the inductive effects increase when the offset increases.

In the CSAMT method, the source determines the direction of the primary current that determines the position of the induced charge and the direction of the secondary current. Natural sources, the direction of which is time-varying, come from thunderstorms and solar activity. In MT method, the accumulation position of induced charge is mainly related to geological structure. There are false anomalies on both sides of the conductive body in the y direction (Fig. 9a, b). The abnormal body becomes wider in x-direction (Fig. 9c, d).

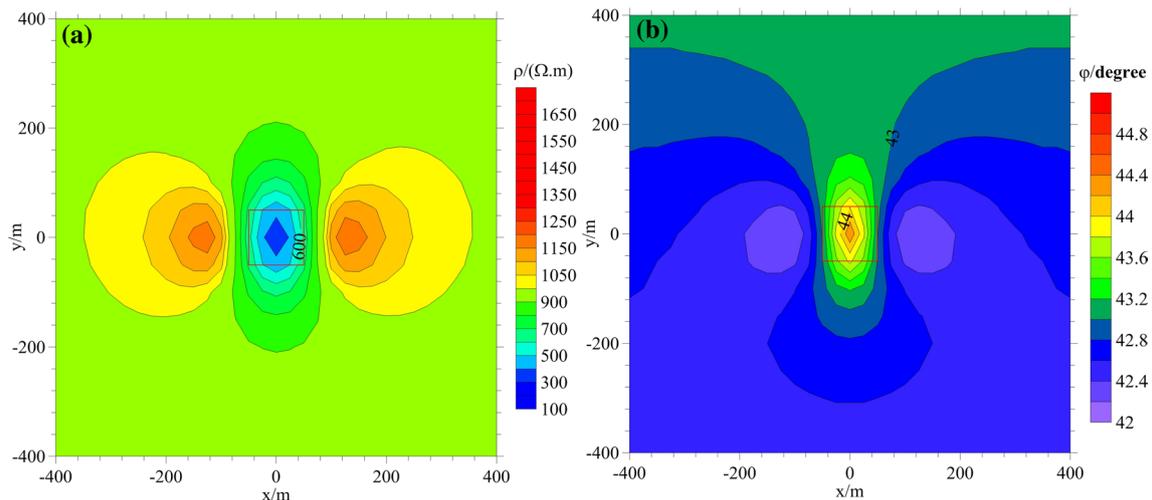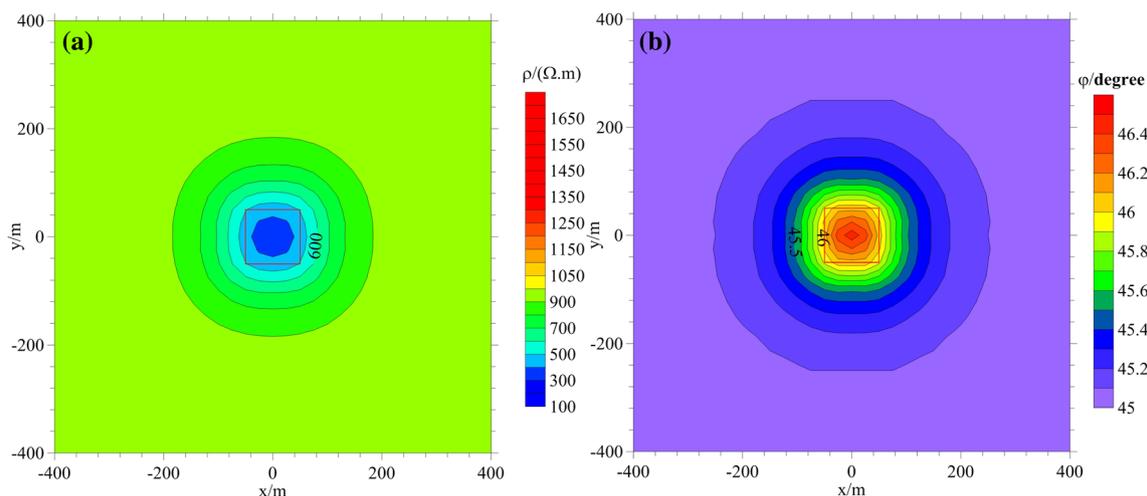**Fig. 7** The plane view of CSAMT apparent resistivity and phase for 100 Hz when $z = 0$ m (model B). **a** The plane view of apparent resistivity when offset = 10 km, **b** the plane view of apparent resistivity when offset = 50 km, **c** the plane view of phase when offset = 10 km, **d** the plane view of phase when offset = 50 km

Model C

Due to the complexity of the geological structure, it is almost impossible that survey lines in actual exploration are s strictly perpendicular to geological strike. When the source direction is not perpendicular to the abnormal body strike, observation data will become complicated due to the distortion effects. In model C (Fig. 10), the source is skewed with the anomalous body, of which the width is 100 m, the length is 400 m, and the thickness is 100 m. Other parameters are the same as the model A.

As shown in Fig. 11 b, the apparent resistivity strike is inconsistent with the body and is biased toward the y-axis direction. Figure 11a, c and d shows the apparent resistivity plane when the resistivity of the abnormal body is 1

Ω.m, 100 Ω.m, and 500 Ω.m, respectively. The deviation becomes smaller when the resistivity difference between the abnormal body and the background becomes smaller. Equation 5 shows that the secondary electric field and galvanic effects become weaker with weaker resistivity difference. Figure 12 shows that the degree of deviation decreases with the increase in frequency. This is because the higher the frequency at the same offset, the faster the ground wave decay, which leads to the weakening of the distortion effect. According to the above analysis, the degree of deviation is related to the resistivity difference and frequency.

In order to explain the distortion, we drew the distribution of the total electric field and the secondary electric field in Fig. 13 when the abnormal body resistivity is 100 Ω.m and the frequency is 100 Hz. As can be seen from this figure,

Fig. 8 The distribution of electric field for 100 Hz when $z=0$ m (model B). **a** The total electric field when offset = 10 km, **b** the secondary electric field when offset = 10 km, **c** the total electric field when offset = 50 km, **d** the secondary electric field when offset = 50 km

there are some differences between the direction of the primary current and the strike of the abnormal body when the source is not parallel to the abnormal body. Due to this, the position of accumulation of the charges also alters. The position is consistent with the resistive region. On the whole, the main reason of the distortion is that the source is not perpendicular to the abnormal body, resulting in the position change of the charges accumulation.

Model D

According to the radiation pattern of electromagnetic fields for the horizontal electric dipole, the CSAMT survey profiles are generally required to be within 60°. In model D, the abnormal body is not on the central axis. The angle with the central axis is 20°. Other parameters are the same as model A (Fig. 14).

Figure 15 illustrates that the strike of apparent resistivity and phase have rotated to one side although the model parameters are the same as model A. One also has to pay

attention to this in data interpretation. Figure 16 shows the total electric and secondary electric field. It shows that the primary current direction is no longer parallel to the x-axis when the abnormal body is not on the central axis. As the direction of the primary electric field changes, the position of accumulation of the charges also changes, causing a change of the secondary current direction as well. In a word, this distortion is also related to the positional relationship between the anomalies and the source.

Relationship between inversion and distortion

The distortion effects mentioned in this paper are widespread. They will appear when there is inhomogeneous structure. Whether the distortion effects need to be removed before the inversion depends on specific situation. If the distortion can be simulated in the forward part of the inversion, it does not need to be removed before
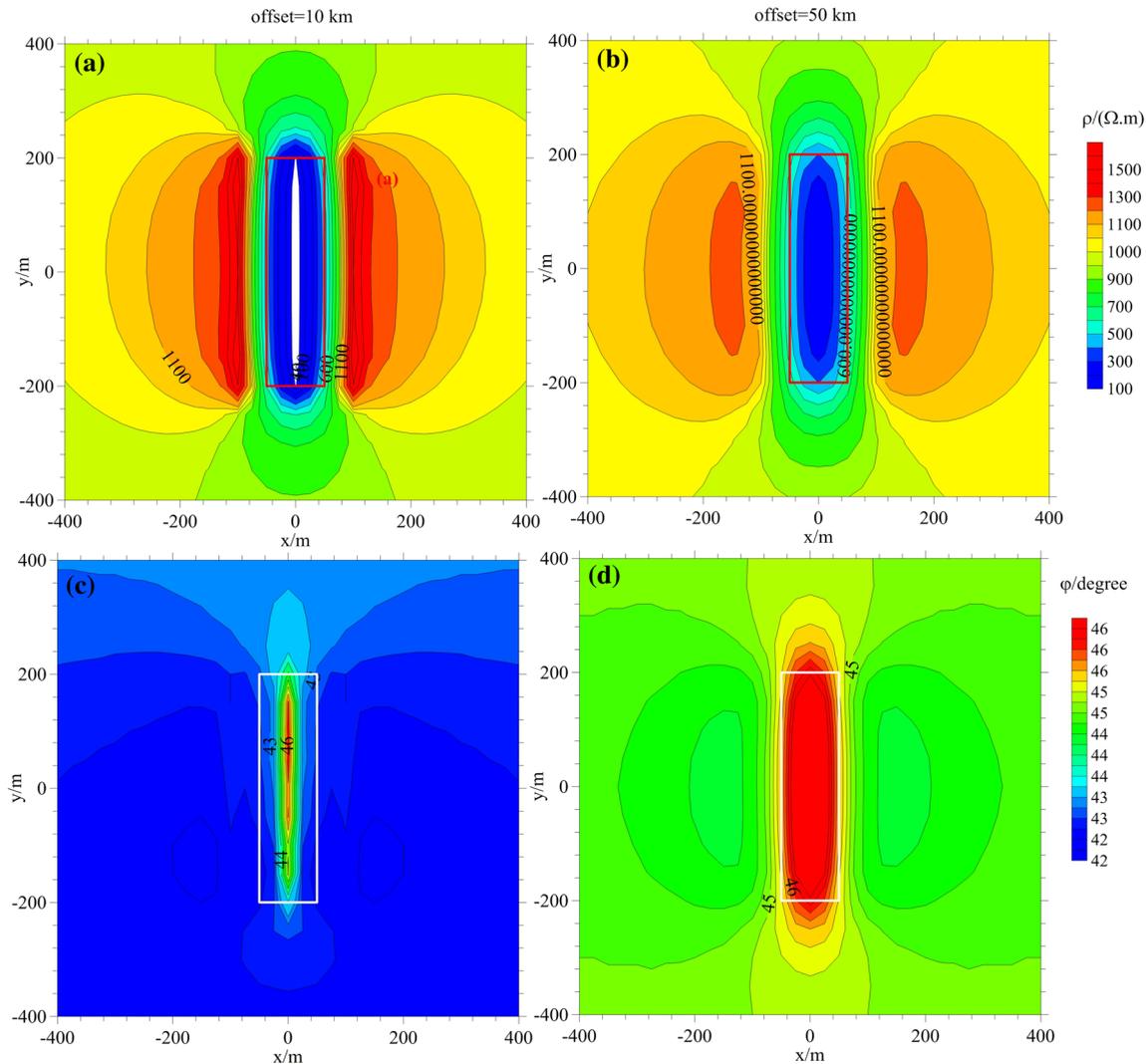
Fig. 9 The plane view of MT apparent resistivity and phase for 100 Hz when $z = 0$ m (model B). **a** The plane view of ρ_{xy}^a. **b** The plane view of ρ_{yx}^a. **c** The plane view of φ_{xy}^a. **d** The plane view of φ_{yx}^a
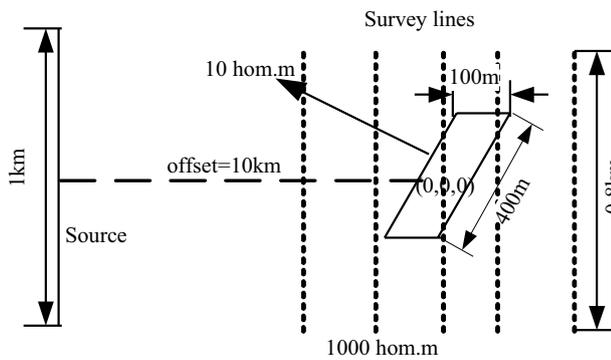
Fig. 10 Plane view of the observation device for model C. The figure is out of scale

the inversion. On the other hand, the distortion effects caused by 2D or 3D structure must be removed in the 1D inversion; otherwise, it may severely affect the data interpretation.

Here, we choose model A in Chapter two to get the CSAMT synthetic data with a frequency range of $1-10^4$ Hz. The data of the profile just crossing the center of conductive body are inverted with 1D and 2D inversion methods described in Di et al. (2004). Before inversion, 5% Gaussian noise was added to the synthetic data. The results are shown in Fig. 17. It is clear that the 1D inversion results are basically the same as those in the apparent resistivity pseudosection (in Fig. 17a). That is because CSAMT 1D forward can only simulate layered structures and cannot simulate the distortion effects caused by 2D or 3D geological structure, so it is difficult to remove these distortions during the 1D inversion process.
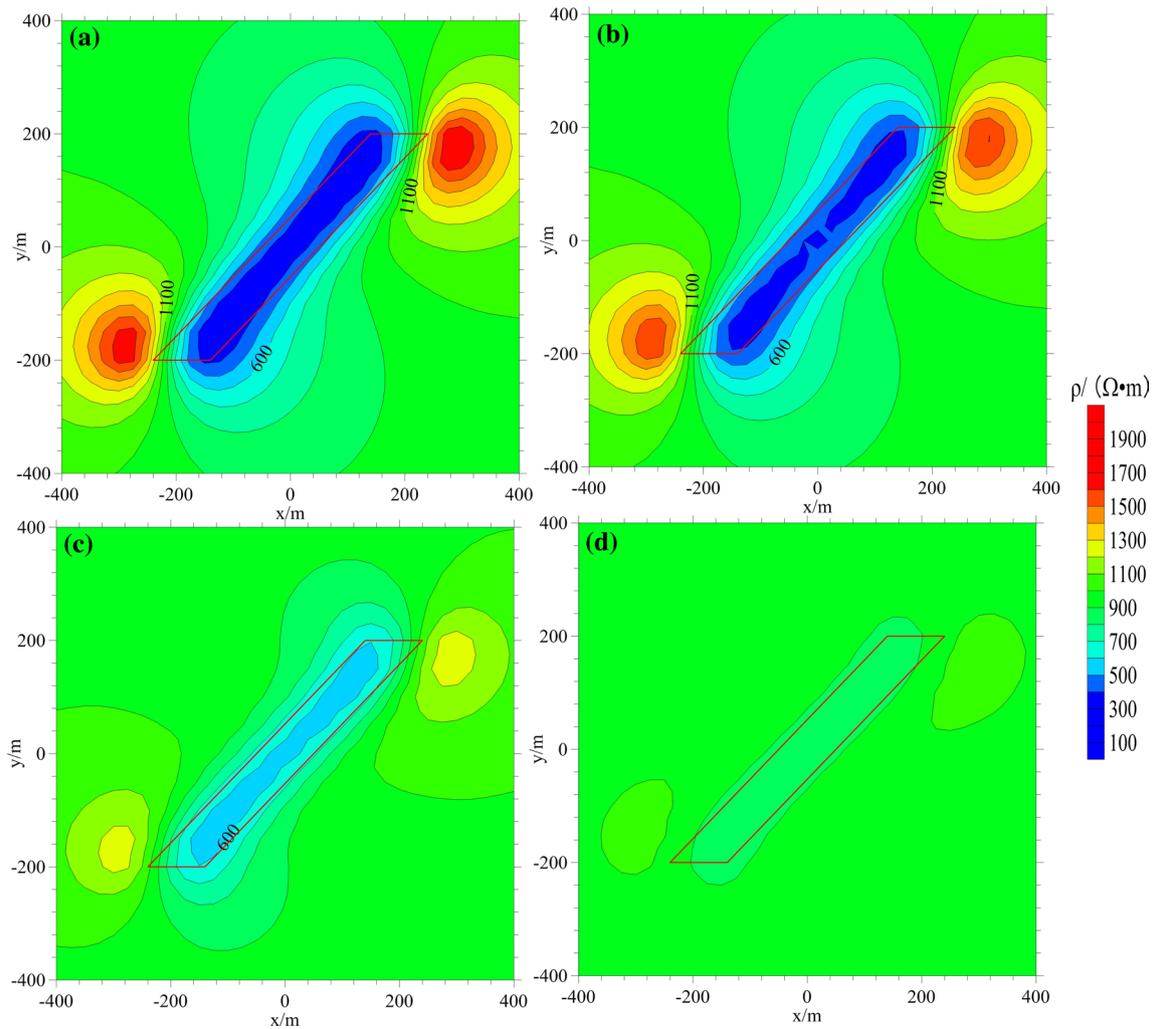
Fig. 11 Comparison of apparent resistivity plane at $z=0$ m when abnormal body resistivity changes (model C). **a** when the resistivity of conductive body is 1 Ω.m, **b** when the resistivity of conduc-tive body is 10 Ω.m, **c** when the resistivity of conductive body is 100 Ω.m, **d** when the resistivity of conductive body is 500 Ω.m

Because MT numerical simulation generally is simpler than CSAMT method, we often process CSAMT data with the MT inversion method. With the MT algorithm, we handle CSAMT data and select the offset 10 km and 50 km, respectively. For the inversion, we divide the volume of interest into 40×31 cells, respectively, in the x- and z-directions. Horizontal cell size is 25 m; the first vertical cell size is 10 m and is growing at the rate of 1.02. When the offset is 10 km, the inversion result is shown in Fig. 17c. There are resistive false anomalies on both sides of the target in the inversion results, which are caused by the distortion introduced in our article. At the same time, the resistive false anomaly will also appear in the lower part of the inversion profile, which is caused by the near-field effect. When the offset is 50 km, the inversion result is obviously better (Fig. 17d). The inversion resistivity is closer to the true model. The false anomalies resistivity in inversion profile is

significantly reduced, which is due to the reduced distortion effect as the offset increases.

Figure 17e illustrates inversion results using 2D CSAMT inversion method. The 2D inversion results are in good agreement with the design model, and distortion effects are removed during the inversion process. This is encouraging result because a correct 2D interpretation of 3D geological structure data can lead to a reasonable geological interpretation (Ledo 2005).

Through this numerical model, we can conclude that 1D inversion cannot remove the distortion caused by the 2D or 3D body, and the 2D inversion can effectively remove the distortion caused by the 2D or 3D geological body and improve the interpretation accuracy. When only using 1D inversion, the distortion needs to be removed before the inversion. Otherwise, its interpretation of reliability must involve the lower. Identification and recognition of the
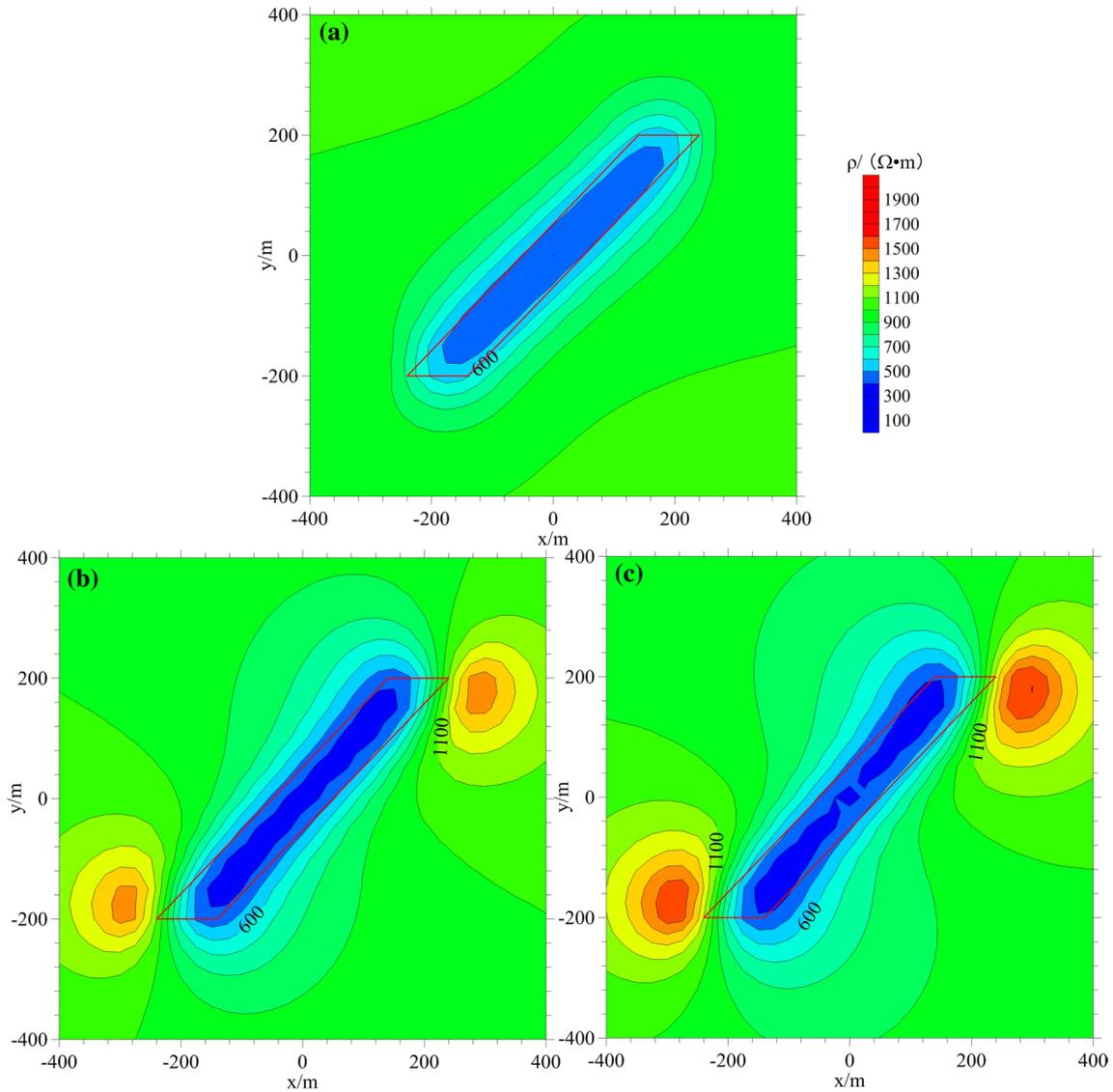
Fig. 12 Comparison of apparent resistivity plane at $z = 0$ m when frequency changes (model C). **a** The plane view of CSAMT apparent resistivity at frequency of 10,000 Hz. **b** The plane view of CSAMT apparent resistivity at frequency of 1000 Hz. **c** The plane view of CSAMT apparent resistivity at frequency of 100 Hz
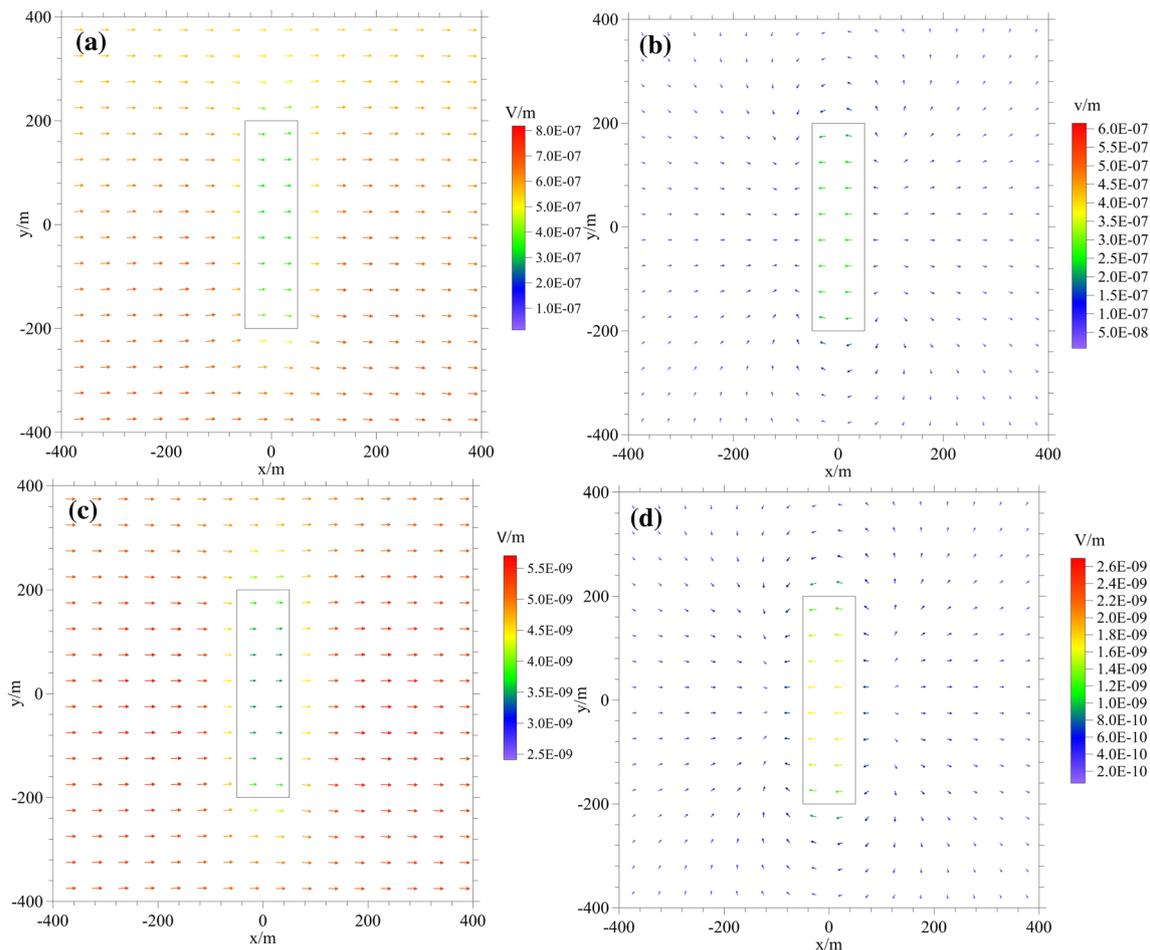
Fig. 13 The distribution of electric field for 100 Hz when $z = 0$ m (model C). **a** The total electric field, **b** the secondary electric field
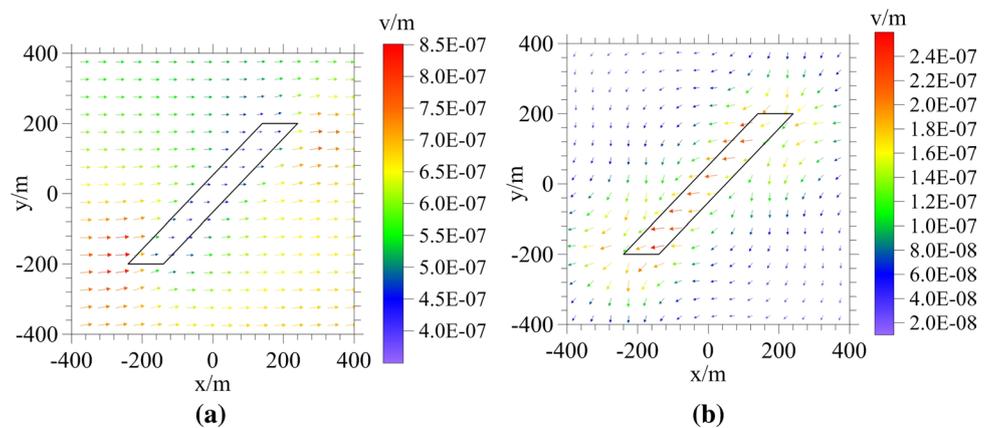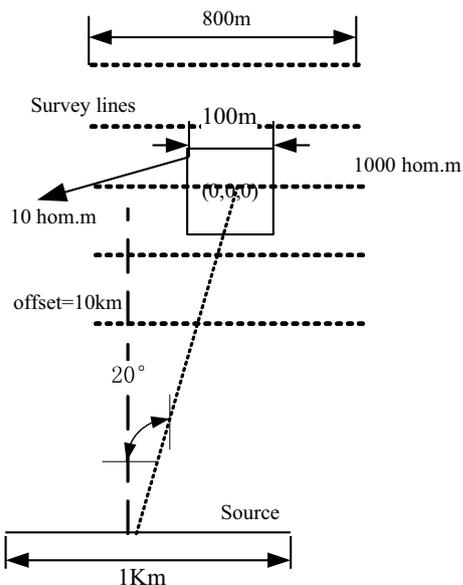
Fig. 14 Plane views of observation device for model D. The figure is out of scale

distortion effect in inversion result are acceptable options in interpretation.

Conclusions

In the CSAMT survey, the distortion effects caused by the target anomalous body are expressed as the abnormal body is stretched in the source direction, and false anomalies with opposite polarity of the target appear on both sides. It is due to the fact that the primary current is along the source direction, causing the induced charges to accumulate on both sides of the anomalous body and result in local currents. In addition, as the offset increases, this distortion effect will weaken.

In MT method, the accumulation position of induced charge is mainly related to geological structure. In CSAMT exploration, the distortion effects are not only related to geological structure, but also relate to the direction of the source, the relative relationship between the source and the geological structures, the offset and frequency.

When the anomalous body strike is inconsistent with the source, the position of accumulation of the induced charges will change. The strike of apparent resistivity and phase do not coincide with the abnormal body, and the deviation decreases with the decrease in frequency and the resistivity difference between the anomaly and the background.

When the abnormal body is not on the central axis of the source, the primary current direction is not horizontal. The direction of apparent resistivity and phase do not coincide with the abnormal body due to the change of the current direction.

When inverting 2D or 3D geological structures by 1D inversion method, it is necessary to remove the distortion effect. Otherwise, it will affect the accuracy of interpretation. When we process CSAMT data with the MT 2D inversion algorithm, it is still difficult to remove this distortion effect caused by artificial sources in the process of inversion. CSAMT 2D inversion algorithm can better restore the model and remove distortion effects.
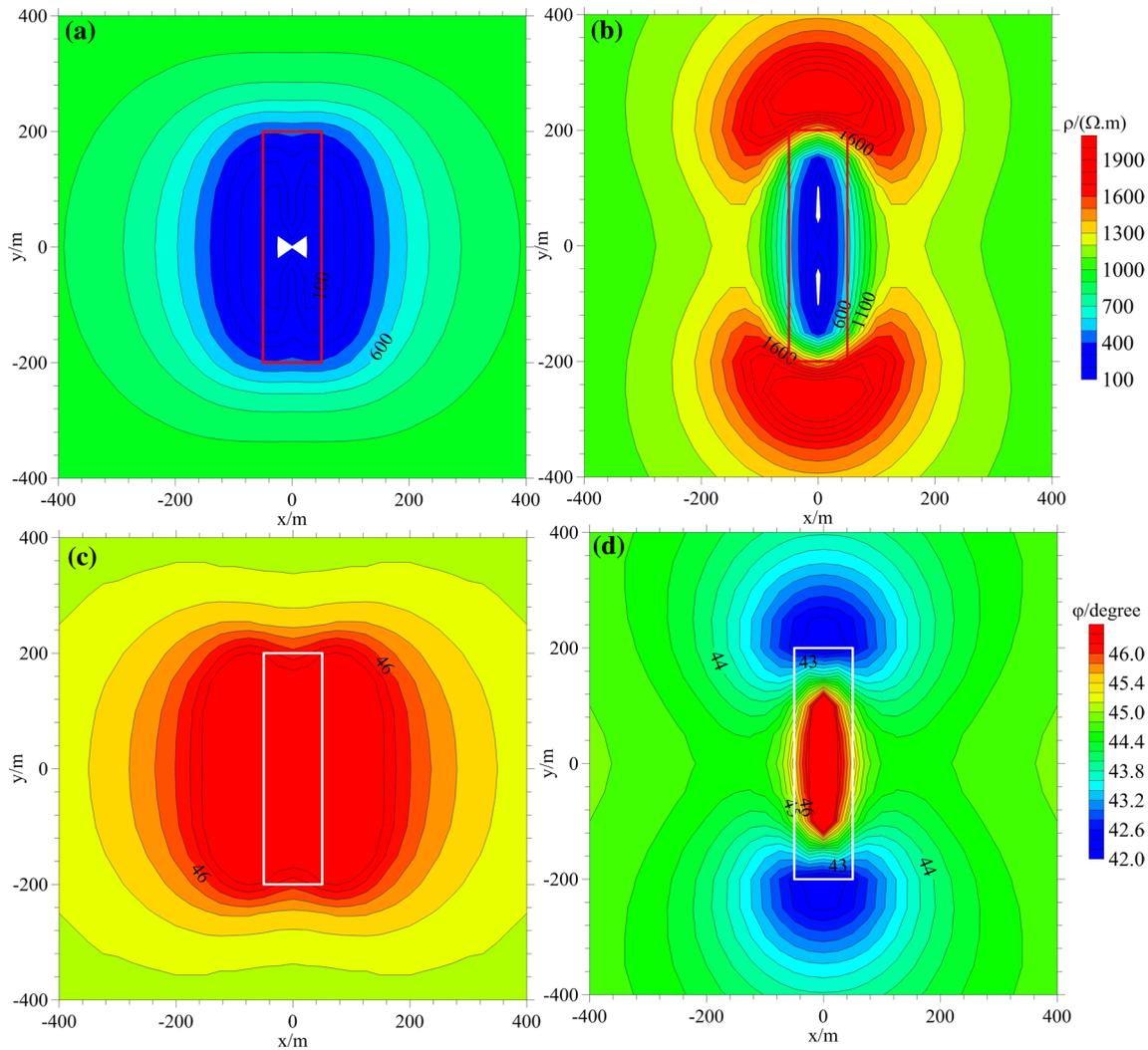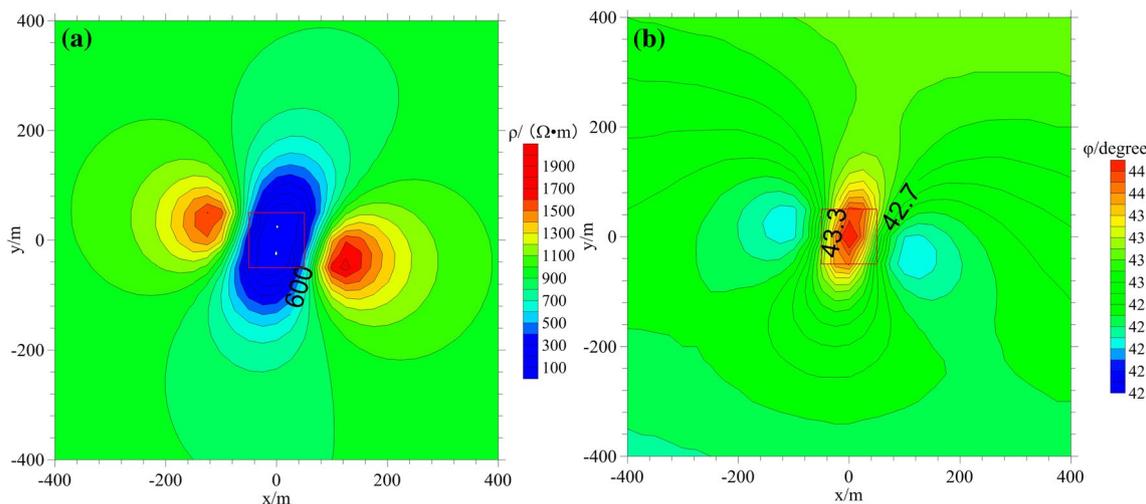

Fig. 15 The plane view of apparent resistivity and phase for 100 Hz when $z = 0$ m (model D). **a** The plane view of apparent resistivity, **b** the plane view of phase plan

Fig. 16 The distribution of electric field for 100 Hz when $z = 0$ m (model D). **a** The total electric field, **b** the secondary electric field

(a)

(b)
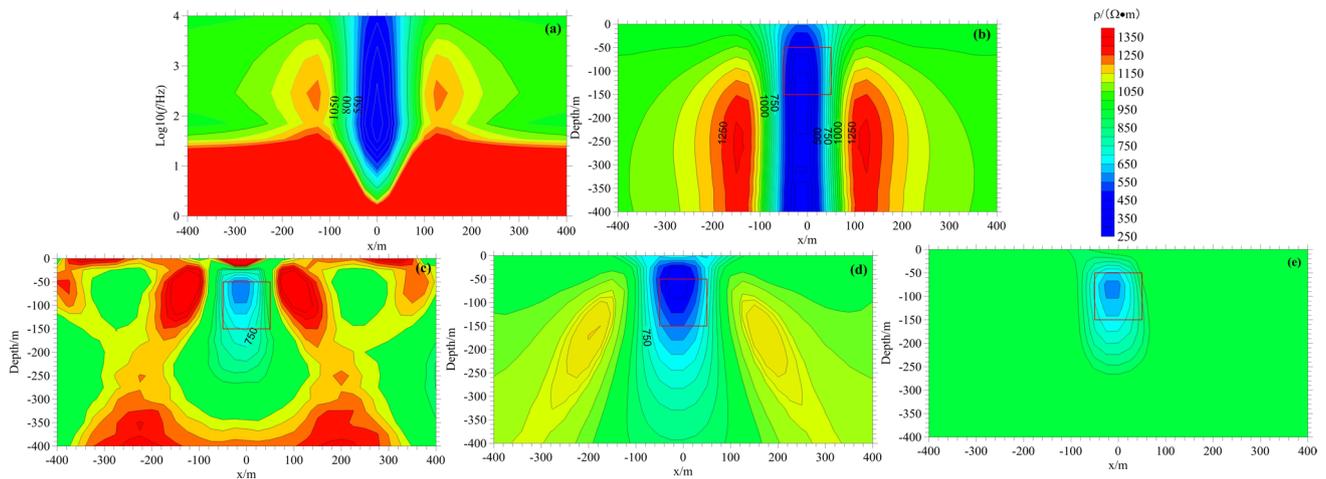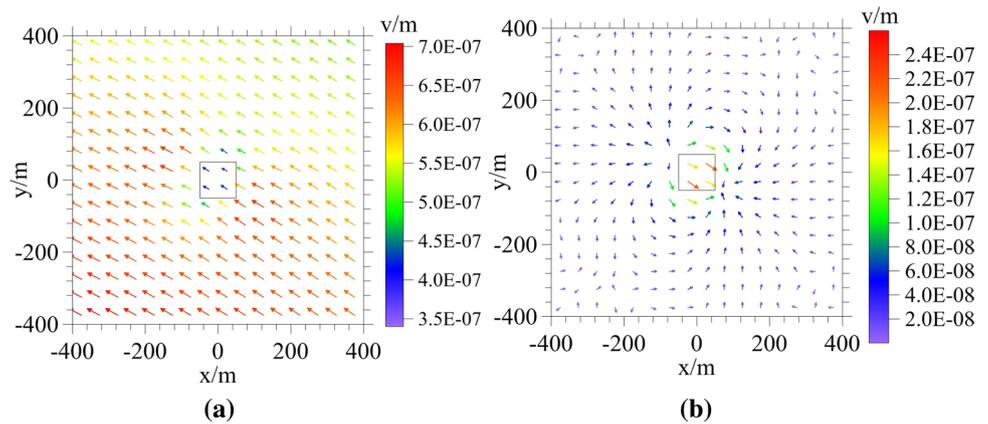
Fig. 17 The comparison of 1D and 2D inversion results. **a** CSAMT apparent resistivity pseudosection, **b** inversion results for CSAMT data when offset = 10 km using 1D CSMT inversion method, **c** inversion results for CSAMT data when offset = 10 km using 2D

MT inversion method, **d** inversion results for CSAMT data when offset = 50 km using 2D MT inversion method, **e** inversion results for CSAMT data when offset = 10 km using 2D CSAMT inversion method

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

An Z, Di Q (2016) Investigation of geological structures with a view to HLRW disposal, as revealed through 3D inversion of aeromagnetic and gravity data and the results of CSAMT exploration. J Appl Geophys 135:204–211

An Z, Di Q, Fu C (2013a) Geophysical evidence through a CSAMT survey of the deep geological structure at a potential radioactive

waste site at Beishan, Gansu, China. J Environ Eng Geophys 18(1):43–54

An Z, Di Q, Wang R (2013b) Multi-geophysical investigation of geological structures in a pre-selected high-level radioactive waste disposal area in Northwestern China. J Environ Eng Geophys 18(2):137–146

Avdeeva A, Moorkamp M, Avdeev D, Jegen M, Miensopust M (2015) Three-dimensional inversion of magnetotelluric impedance tensor data and full distortion matrix. Geophys J Int 202(1):464–481

Berdichevsky M, Dmitriev V (1976) Basic principles of interpretation of magnetotelluric sounding curves. Geoelectric and geothermal studies. Akademiai Kiado, Budapest, pp 165–221

Bibby H, Caldwell T, Brown C (2005) Determinable and non-determinable parameters of galvanic distortion in magnetotellurics. Geophys J Int 163(3):915–930

Di Q, Wang M, Shi K, Zhang G (2002) An applied study on prevention of water bursting disaster in mines with the high resolution V6 system. Chin J Geophys 45:787–792 **(in Chinese with English abstract)**

Di Q, Unsworth M, Wang M (2004) 2.5D CSAMT modeling with the finite element method over 2D complex earth media. Chin J Geophys 47:825–829 **(in Chinese with English abstract)**

Di Q, An Z, Fu C, Wang Z (2018) Imaging underground electric structure over a potential HLRW disposal site. J Appl Geophys 155:102–109

Fu C, Di Q, An Z (2013) Application of the CSAMT method to groundwater exploration in a metropolitan environmentGroundwater exploration with CSAMT. Geophysics 78(5):B201–B209

Goldstein M, Strangway D (1975) Audio-frequency magnetotellurics with a grounded electric dipole source. Geophysics 40(4):669–683

Groom R, Bailey R (1989) Decomposition of magnetotelluric impedance tensors in the presence of local three-dimensional galvanic distortion. J Geophys Res Solid Earth 94(B2):1913–1925

Jiracek G (1990) Near-surface and topographic distortions in electromagnetic induction. Surv Geophys 11(2–3):163–203

Jones A (2012) Distortion of magnetotelluric data: its identification and removal. The magnetotelluric method: theory and practice. Cambridge University Press, Cambridge, pp 219–302

Kuznetzov A (1982) Distorting effects during electromagnetic sounding of horizontally non-uniform media using an artificial field source. Earth Phys 18(1):130–137

Ledo J (2005) 2-D versus 3-D magnetotelluric data interpretation. Surv Geophys 26(5):511–543

Lei D, Fayemi B, Yang L et al (2017a) The non-static effect of near-surface inhomogeneity on CSAMT data. J Appl Geophys 139:306–315

Lei D, Di Q, Wu J, Wang X, Liu Y, Fayemi O, Luan X, Zhang W (2017b) Anti-interference test for the new SEP instrument: CSAMT study at Dongguashan Copper Mine, China. J Environ Eng Geophys 22(4):339–352

Li X, Bai D, Yan Y (2016) Three-dimensional inversion of magnetotelluric resistivity model with galvanic distortion. Chin J Geophys 59(6):2302–2315 **(in Chinese with English abstract)**

MacLennan K, Li Y (2013) Denoising multicomponent CSEM data with equivalent source processing techniques. Geophysics 78(3):E125–E135

Nabighian M (1991) Electromagnetic methods in applied geophysics. Society of Exploration Geophysicists, Oklahoma

Neukirch M, Rudolf D, Garcia X, Galiana S (2019) Amplitude-phase decomposition of the magnetotelluric impedance tensorMT amplitude-phase decomposition. Geophysics 84(5):E301–E310

Neukirch M, Galiana S, Garcia X (2020) Appraisal of magnetotelluric galvanic electric distortion by optimizing amplitude and phase tensor relations. Geophysics 85(3):E79–E98

Sternberg B, Washburne J (1988) Correction for the static shift in magnetotellutics using transient electromagnetic soundings. Geophysics 53(11):1459–1468

Swift C (1967) A magnetotelluric investigation of an electrical conductivity anomaly in the southwestern United States. Massachusetts Institute of Technology, Princeton University. Ph.D. Dissertation

Tang W, Li Y, Oldenburg D, Liu J (2018) Removal of galvanic distortion effects in 3D magnetotelluric data by an equivalent source technique. Geophysics 83(2):E95–E110

Wang R, Wang M, Di Q et al (2009) 2D numerical study on the effect of conductor between the transmitter and survey area in CSEM exploration. Appl Geophys 6(4):311–318

Wang R, Yin C, Wang M, Di Q (2015) Laterally constrained inversion for CSAMT data interpretation. J Appl Geophys 121:63–70

Wynn J, Mosbrucker A, Pierce H, Spicer K (2016) Where is the hot rock and where is the ground water–Using CSAMT to map beneath and around Mount St. Helens. J Environ Eng Geophys 21(2):79–87

Yan S, Junmei F (2004) An analytical method to estimate shadow and source overprint effects in CSAMT sounding. Geophysics 69(1):161–163

Zhdanov M, Lee S, Yoshioka K (2006) Integral equation method for 3D modeling of electromagnetic fields in complex structures with inhomogeneous background conductivity. Geophysics 71(6):G333–G345

Zhou N, Hou D, Xue G (2018) Effects of shadow and source overprint on grounded-wire transient electromagnetic response. IEEE Geosci Remote Sens Lett 15(8):1169–1173

Zonge K, Hughes L (1991) Controlled source audio-frequency magnetotellurics. Electromagnetic methods in applied geophysics. Society of Exploration Geophysicists, Oklahoma, pp 713–810

Generalisation and improvement of the compact gravity inversion method

Wenwu Zhu[1,2] · **Junhuan Peng[1]** · **Sanming Luo[2]** · **Xiangang Meng[2]** · **Jinzhao Liu[2]** · **Chuandong Zhu[2]**

Abstract

Compact gravity inversion (CGI) is widely used to invert gravity data following the principle of minimising the volume of the causative body due to its simplicity, high efficiency, and sharp-boundary inversion results. In this study, the compactness weighting function is generalised and the depth weighting function is introduced to CGI to obtain the reweighted CGI (RCGI) method. Although RCGI exhibits better flexibility than CGI, selecting an appropriate compactness factor α and depth weighting function β is difficult, and we design a parameter selection rule to search the proper α and β quantitively. Furthermore, we improve RCGI for boasting superior computational efficiency by gradually eliminating the model blocks that reach the designated boundaries in the iterative algorithm of inversion. This approach is termed the reweighted and element-elimination CGI (REECGI) method. The inversion results show that both RCGI and REECGI result in better inversion accuracy than CGI, and REECGI has higher computational efficiency than RCGI and CGI, which increases with the number of iterations.

Keywords Compact gravity inversion · Inversion theory · Compactness factor · Weighting function · Inversion accuracy · Computational efficiency

Introduction

Gravity inversion is a practical method that has been extensively applied in hydrology, oil and gas exploration, mineral exploration, and geological surveys (Blakely 1995; Chen et al. 2008; Karaoulis et al. 2014; Li and Oldenburg 1998; Mendonca and Silva 1994, 1995; Pilkington 1997, 2009; Portniaguine and Zhdanov 1999; Roland et al. 2013; Zhdanov 2015). Owing to the instability and non-uniqueness of inversion, which is known as the ill-posed problem (Hadamard 1902), it is difficult to obtain inversion information that accurately reflects the real geological conditions below the Earth's surface.

✉ Wenwu Zhu
   fmccea@sina.com

[1] China University of Geosciences (Beijing), Beijing 100083, China

[2] The First Crust Monitoring and Application Centre, China Earthquake Administration, Tianjin 300180, China

Researchers have made numerous attempts to overcome the ill-posed problem; one solution is to find the appropriate stability function combined with the error function. Constable et al. (1987) and Smith et al. (1991) introduced the minimum norm of the Laplace operator of the model, which can produce a smooth inversion solution; however, this method usually fails to accurately describe the true massive geological structure. Rudin et al. (1992) proposed a method based on total variation (TV) to rebuild noisy and blurred images, but the relative functions are not differentiable at zero. To solve this problem, Acar and Vogel (1994) improved the TV method by introducing a small real number.

However, these methods all attempt to smooth out the inversion results; therefore, they do not represent the actual situation, especially in the case of mineral exploration, where it is desirable to obtain inversion results with sharp boundaries. To obtain sharp-boundary images and overcome the smoothness problem, Last and Kubik (1983) developed the compact gravity inversion (CGI) method based on the minimum area (2D inversion) or volume (3D inversion) of the model blocks, resulting in sharp-boundary inversion images when the blocks are located at shallow depths. This method received widespread attention; however, the effects of non-convergence and concentration near the surface often

occur during actual gravity inversion. Portniaguine and Zhdanov (1999) proposed a similar approach based on the CGI called the minimum gradient support (MGS) method; however, a similar problem remains (i.e., the near-surface concentration effect). Therefore, the inversion results do not reflect the actual mineral distribution when the model blocks are located at great depths. To obtain accurate inversion results, other scholars used additional a priori information to generalise the CGI method (Barbosa and Silva 1994; Guillen and Menichetti 1984; Silva and Barbosa 2006; Silva et al. 2009, 2011); however, these methods strongly depend on a priori information, including the centre of gravity of the model blocks and the direction of the dyke, which are often unknown in field cases. When there is little a priori information, such interactive and generalised inversion methods are no longer applicable; therefore, Williams (2008) proposed another method of performing inversion without geological constraints. However, this led to new problems such as the need to estimate a large amount of parameters, which is difficult when geologists are not familiar with the local geology and make the procedure more time-consuming.

Considering the advantages and disadvantages of CGI and other methods, this study aims to generalise the CGI method and obtain more reasonable gravity inversion results, while further improving the calculation efficiency. First, we generalise the compactness weighting function of the CGI method and introduce the depth weighting function (Li and Oldenburg 1998) to make the CGI method more flexible for gravity inversion. Second, we propose an approach for quantitatively obtaining the appropriate compactness weighting function factor α and the depth weighting function β. Lastly, we improve the computational efficiency by continuously eliminating the model blocks reaching the bounds, ensuring that the kernel matrix dimensions decreases continuously.

Materials and methods

The density model is assumed to be discretised into several rectangles in two-dimensional (2D) models or rectangular prisms in three-dimensional (3D) models. For simplicity, the inversion method is introduced for a 2D case here; however, the discussion applies equally to the 3D case. It has been established that gravity anomaly data, $\tilde{G}$, has the following relationship with the density model, $\tilde{V}$:

$$\tilde{G} = A\tilde{V}, \tag{1}$$

where $\tilde{G}$ is an adjusted $n \times 1$ vector, $\tilde{V}$ is an $m \times 1$ vector, and A is the kernel matrix, which was defined by Last and Kubik 1983; the dimensions of A are $n \times m$. In Eq. (1), $\tilde{G}$ and A are known values that can be measured or calculated in advance;

only $\tilde{V}$ is unknown and needs to be obtained in order to approximate the solution of the actual mineral distribution.

Last and Kubik (1983) used the principle of minimisation of the area (or volume in 3D) of the model to solve Eq. (1), which can be stated as follows:

$$\text{Minimize } Q = \sum_i \tilde{v}_i^2/(\tilde{v}_i^2 + \varepsilon), \quad i = 1, 2, \ldots, m$$

$$\text{Subject to} \quad \tilde{G} - A\tilde{V} = 0 \tag{2}$$

$$V_{\min} \le \tilde{V} \le V_{\max},$$

where $\tilde{v}_i$ is the approximate solution of the ith block, ε is a sufficiently small value [$\approx 10^{-11}$ in the case of Last and Kubik (1983)], and $V_{\min}$ and $V_{\max}$ denote the lower bound vector with the same entries, $v_{\min}$ (a scalar), and the upper-bound vector with the same entries, $v_{\max}$ (a scalar), respectively.

The following characteristic exists for a single block in Q of Eq. (2):

$$q = \lim_{\varepsilon \to 0} \tilde{v}_i^2/(\tilde{v}_i^2 + \varepsilon) = \begin{cases} 0 & \text{for } \tilde{v}_i = 0 \\ 1 & \text{for } \tilde{v}_i \ne 0. \end{cases} \tag{3}$$

According to Eq. (3), the optimal solution of Eq. (2) will be obtained when the number of blocks ($= 0$) reaches a maximum, which means that the other blocks ($\ne 0$) are at a minimum. This is why the CGI method follows the principal of the minimum area. In fact, Eq. (2) is a special case as follows:

$$\text{Minimize } F = \sum_i \tilde{v}_i^2/(|\tilde{v}_i^\alpha + \varepsilon), \quad \alpha \ne 0$$

$$\text{Subject to } \tilde{G} - A\tilde{V} = 0 \tag{4}$$

$$V_{\min} \le \tilde{V} \le V_{\max},$$

And the following characteristic exists for a single block in F of Eq. (4):

$$f = \lim_{\varepsilon \to 0} \tilde{v}_i^2/(|\tilde{v}_i|^\alpha + \varepsilon) = \begin{cases} 0 & \text{for } \tilde{v}_i = 0 \\ |\tilde{v}_i|^{2-\alpha} & \text{for } \tilde{v}_i \ne 0. \end{cases} \tag{5}$$

Thus, Eqs. (4) and (5) are equal to Eqs. (2) and (3), respectively, when $\alpha = 2$. Equation (5) also follows the same principal as Eq. (3), i.e., the optimal solution is positively correlated to the number of model blocks ($= 0$) produced in the inversion results. The only difference between Eq. (2) and (4) or Eqs. (3) and (5) is that the constant 2 changes to the variable α. The introduction of this variable provides more flexibility than the CGI method during gravity inversion by allowing variation of α, which is helpful for obtaining better inversion solutions. Actual experiments show that the inversion results are more reasonable when α is a positive real number other than 2.

Next, we continue to follow the approach of Last and Kubik ([1983](#)) whereby $1/\left(|v_i|^{\alpha}+\varepsilon\right)$ in Eq. (4) is considered to be the compactness weighting function, which should be positive. Therefore, |•| is introduced, which represents the absolute value of •, to guarantee that $|v_i|^{\alpha}$ is always a positive real number (v_i^{α} is probably a negative or an imaginary number when $v_i < 0$). Thus, the compactness weighting function has the following matrix form:

$$\tilde{W} = \text{diag}\left(\tilde{w}_i\right), \tag{6}$$

with $\tilde{w}_i = \frac{1}{|\tilde{v}_i|^{\alpha}+\varepsilon}$, $i = 1, 2, \ldots, m, \alpha \neq 0$.

The iteration algorithm is introduced because $\tilde{W}$ is not a constant weighting matrix. In one iteration, $\tilde{W}$ is replaced by the inversion solutions obtained by the last iteration so they become constant matrices in the current iteration, and the iterative procedure is reduced to a classical least L_2-norm problem. The optimal solution of the problem in the kth iteration is as follows (Menke [1989](#)):

$$\tilde{V}^{(k)} = \left[\tilde{W}^{(k-1)}\right]^{-1} A^T \left\{ A \left[\tilde{W}^{(k-1)}\right]^{-1} A^T + \mu I \right\}^{-1} \tilde{G}, \quad k = 2, 3, \ldots\ldots, \mu > 0 \tag{7}$$

where $\tilde{W}^{(k-1)}$ represents the compactness weighting function calculated by Eq. (6) in the $(k-1)$th iteration and $\tilde{W}^{(1)}$ is an $m \times m$ identity matrix. μ is a damping factor (Levenberg [1944](#); Marquardt [1963](#)) used to improve the condition number of the matrices and I is an $n \times n$ identity matrix.

Experiments show that the near-surface concentration effect cannot be overcome when the blocks ($\neq 0$) are located at great depths when only considering the compactness weighting function of the model. Therefore, the depth weighting function (Li and Oldenburg [1998](#)) or the kernel weighting function (Zhdanov [2015](#)) must be added into Eq. (6) and (7) (Ghalehnoee et al. [2017](#); Rosas-Carbajal et al. [2017](#)). Both these functions have an equivalent effect in overcoming the near-surface concentration effect in the vertical direction (Pilkington [2009](#)); thus, the depth weighting function is used in this study. Accordingly, $\tilde{W}$ changes to a new form:

$$\tilde{W} = \text{diag}\left(\tilde{w}_i\right), \tag{8}$$

with $\tilde{w}_i = \frac{1}{\left(|\tilde{v}_i|^{\alpha}+\varepsilon\right)z_i^{\beta}}, \alpha > 0, \beta \geq 0, z_i > 0, i = 1, 2, \ldots, m,$

where z_i is the depth of the ith model and β is the depth weighting function factor. The reason for the condition of $\alpha > 0$ is that the value of $\tilde{v}_i$ obtained in the kth iteration is larger; therefore, the weight of $\tilde{v}_i$ should be larger in the next iteration according to Last and Kubik ([1983](#)). If $\alpha < 0$, a larger $\tilde{v}_i$ will correspond to a lower weight of $\tilde{v}_i$, which contradicts the earlier assumption; hence, $\alpha > 0$. We refer to this CGI method with the depth weighting function as the reweighted compact gravity inversion (RCGI)

method because RCGI is equivalent to CGI when $\alpha = 2$ and $\beta = 0$.

As α and β are variables, determining their appropriate values is another problem. Combining Eq. (4), (7), and (8), we propose two strategies for determining appropriate α and β values in a given interval according to whether the prior model v_i (i = 1, 2, ..., m) is known:

$$(\alpha, \beta) = \begin{cases} \arg\min \sum_i \left(\tilde{v}_i - v_i\right)^2, v_i \text{ is known } (a) \\ \arg\min (\gamma \delta), v_i \text{ is unknown } (b) \end{cases}, \tag{9}$$

$\alpha > 0, \beta \geq 0, i = 1, 2, \ldots, m.$ where

$$\begin{cases} \gamma = \text{number of } \tilde{v}_i(\neq 0) \\ \delta = \frac{\tilde{G} - A\tilde{V}_2}{\tilde{G}_2} \end{cases},$$

If the prior model is known, the optimal α and β values are obtained by minimising the sum of the squared residuals between the inversion results $\tilde{v}_i$ and the prior model v_i.

Otherwise, the α and β values are obtained by minimising the product of γ (i.e., the number of $\tilde{v}_i(\neq 0)$ that follows the principal of minimum area, allowing the inversion model to be compact) and δ (the relative l_2-norm of the observation data $\tilde{G}$ and the forward result of $\tilde{V}$ that prevent excessively large residuals). The appropriate α and β values for obtaining the optimal solution generally lie within a small interval (i.e., $0.9 \leq \alpha \leq 1.2$, $2 \leq \beta \leq 6$ in the inversion of synthetic data). Because the inversion results usually cannot be convergent if α and β are too large, as the best interval is unknown at the initial trial, we can instead search for α and β in a slightly larger interval with a slightly larger step to estimate the approximate interval, then gradually narrow down the interval and step to determine the final values of α and β.

Next, we try to improve the already high computational efficiency of the RCGI or CGI methods and obtain REECGI: We find that the weights of the models reaching the lower or upper bounds in the last iteration are all set to a very large value (i.e., 10^{11}); therefore, these blocks can be frozen in the next iteration. In fact, these blocks can be removed in the next iteration because their solutions have been obtained and there is no need to involve them in the calculations any longer. If so, the dimensions of the kernel matrices A and the weighting matrices $\tilde{W}$ in Eq. (7) will be reduced, which then leads to a smaller number of calculations in the next iteration. The specific process for this concept is as follows:

1. Calculate the inversion values using Eqs. (6) and (7);
2. Ascertain whether any inversion results reach the boundary values $v_{\min}$ and $v_{\max}$. If v_i reaches the boundaries:

$$A^{(k)} = \left[\, A(1,:)^{(k-1)},\, A(2,:)^{(k-1)},\, \cdots,\, A(i-1,:)^{(k-1)},\, \overline{A(i+1,:)^{(k-1)}},\, \cdots,\, A(n,:)^{(k-1)} \,\right],$$
$$\tilde{W}^{(k)} = \text{diag}\left(\tilde{w}_1^{(k-1)},\, \tilde{w}_2^{(k-1)},\, \ldots,\, \tilde{w}_{i-1}^{(k-1)},\, \tilde{w}_{i+1}^{(k-1)},\, \ldots,\, \tilde{w}_m^{(k-1)} \right), \qquad (10)$$
$$k \geq 2$$

where $A(i,:)^{(k-1)}$ represents the ith column of A in the $(k-1)$th iteration and $\left[\,\bullet\,\right]$ represents the matrices constituted by $\bullet$;

3. Repeat step (1) until convergence.

As we can see from Eq. (10) above, the calculation efficiency of REECGI is faster due to the reduction of $A^{(k-1)}$ and $\tilde{W}^{(k-1)}$.

Solving Eq. (7) is an essential step required for the CGI, RCGI, and REECGI methods. Now, we approximately count the smallest reduction of multiplications involved in solving Eq. (7) in one iteration. Assuming that there are t ($0 \leq t \leq m$) blocks reaching the boundary values in one iteration, the reduction of RCGI has $mn(3m + 2n)$ multiplications (this is the same as CGI as it is equal to RCGI when $\alpha = 2$, $\beta = 0$, and the choice of α and β does not change the matrix dimensions for multiplication), whereas REECGI only has $n(m - t)(3(m - t) + 2n)$ multiplications. Therefore, the reduction of multiplications for solving Eq. (7) is $nm(3m + 2n) - n(m - t)(3(m - t) + 2n) = nt(6m - 3t + 2n)$ in one iteration. This leads to significant time saving when Eq. (9) is used to find the appropriate α and β for more iterations.

Finally, a relative l_2-norm is used to terminate the iteration process in the synthetic data as follows:

$$e_1 = \frac{\tilde{V} - V_2}{V_2}, \qquad (11)$$

where $\tilde{V}$ is the recovered model and V is the true model. The threshold and maximum iteration times for the termination of the iteration need to be determined according to the actual situation. Using field data, the difference between two iterations can be considered as the principle for terminating the iteration process as follows:

$$e_2 = \frac{\tilde{V}^{(k)} - \tilde{V}_2^{k-1}}{\tilde{V}_2^{k-1}}, \qquad (12)$$

Results and discussion

Inversion of synthetic data

We used the model shown in Fig. 1 to compare the CGI, RCGI, and REECGI methods. The gravity anomaly dataset is obtained by forward modelling with equidistant observations (5 m spacing) in the horizontal direction, and no noise is added into the gravity anomaly data to better compare the three methods and prevent interference caused by noise. The corresponding model blocks are constructed for the same square with a length of 5 m; the example includes three separated bodies with one negative (blue rectangle) and two positive (yellow rectangles) blocks, distributed along the vertical direction. The left and middle models both measure 70 m × 20 m at a depth of 95 m, and the right model is 30 m × 30 m at a depth of 155 m. The lower and upper boundaries are set to − 1000 and 1000 kg m^{-3}, respectively.

First, we use RCGI to conduct the inversion calculations shown in Fig. 2, where α is a fixed value ($= 2$) for comparison with the CGI method. The blue dashed rectangles indicate the locations of the true model blocks. The left column shows the inversion results, and the right shows the relative calculation residuals. In Fig. 2a, b, $\alpha = 2$ and $\beta = 0$, which indicates that the CGI method is used for inversion of the
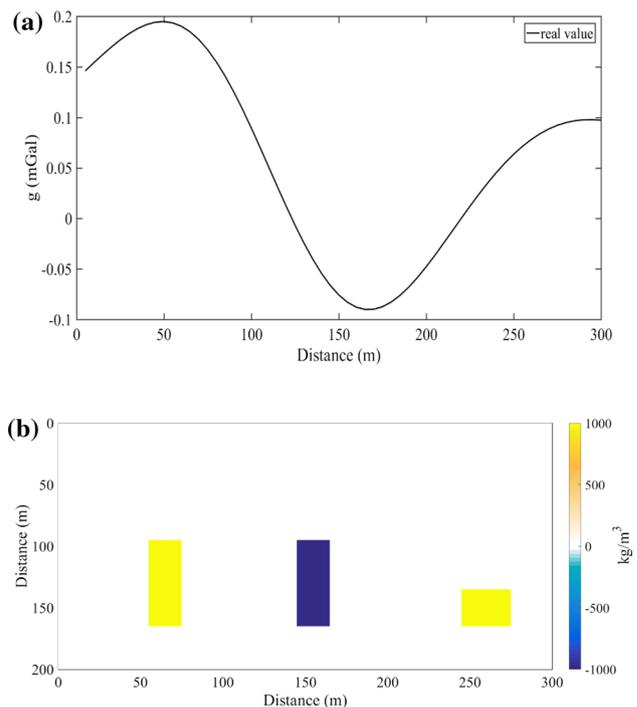
Fig. 1 **a** Actual gravity anomaly data reproduced in **b** by forward modelling of the true model with a positive density value of 1000 kg m^{-3} (yellow rectangles) and a negative value of − 1000 kg m^{-3} (blue rectangle)

synthetic test; i.e., an obvious near-surface concentration of the inversion results exists and the iterative algorithm is convergent. In Fig. 2c, d, $\alpha = 2$ and $\beta = 2.35$. Because β has a positive correlation with the depth of the inversion results, we gradually increase β from zero in order to determine a better value for β, which can improve the accuracy of the inversion results with respect to the real model; this resulted in a β value of 2.35. The location shown in Fig. 2c is closer to the real model than that in Fig. 2a, although the residual error e_1 shown in Fig. 2d is larger than that for CGI shown in Fig. 2b, and the iteration is also convergent. In Fig. 2e, f, the inversion results are no longer convergent, even if β ($= 2.40$) is only slightly larger than 2.35. Thus, this synthetic test shows that we cannot obtain reasonable inversion results

for the case where α is a fixed value (2). In other words, CGI is not applicable in this case.

Figure 2 shows that we cannot obtain good inversion results with a fixed α; therefore, we now use the RCGI and REECGI methods, in which α is also a variable, to obtain better inversion solutions and compare their computational efficiency. The maximum iteration times and e_1 in Eq. (11) are set to 200 and 0.20, respectively, for terminating the iteration algorithm. First, we determine the best α and β in the given interval ($0.9 \le \alpha \le 1.2$, $2 \le \beta \le 6$) for both methods according to Eq. (9a) shown in Fig. 3. In fact, we set a larger interval in the initial test, but the inversion results are too smooth or not convergent when α and β exceed this interval. The left column (Fig. 2a, c) shows the results obtained by
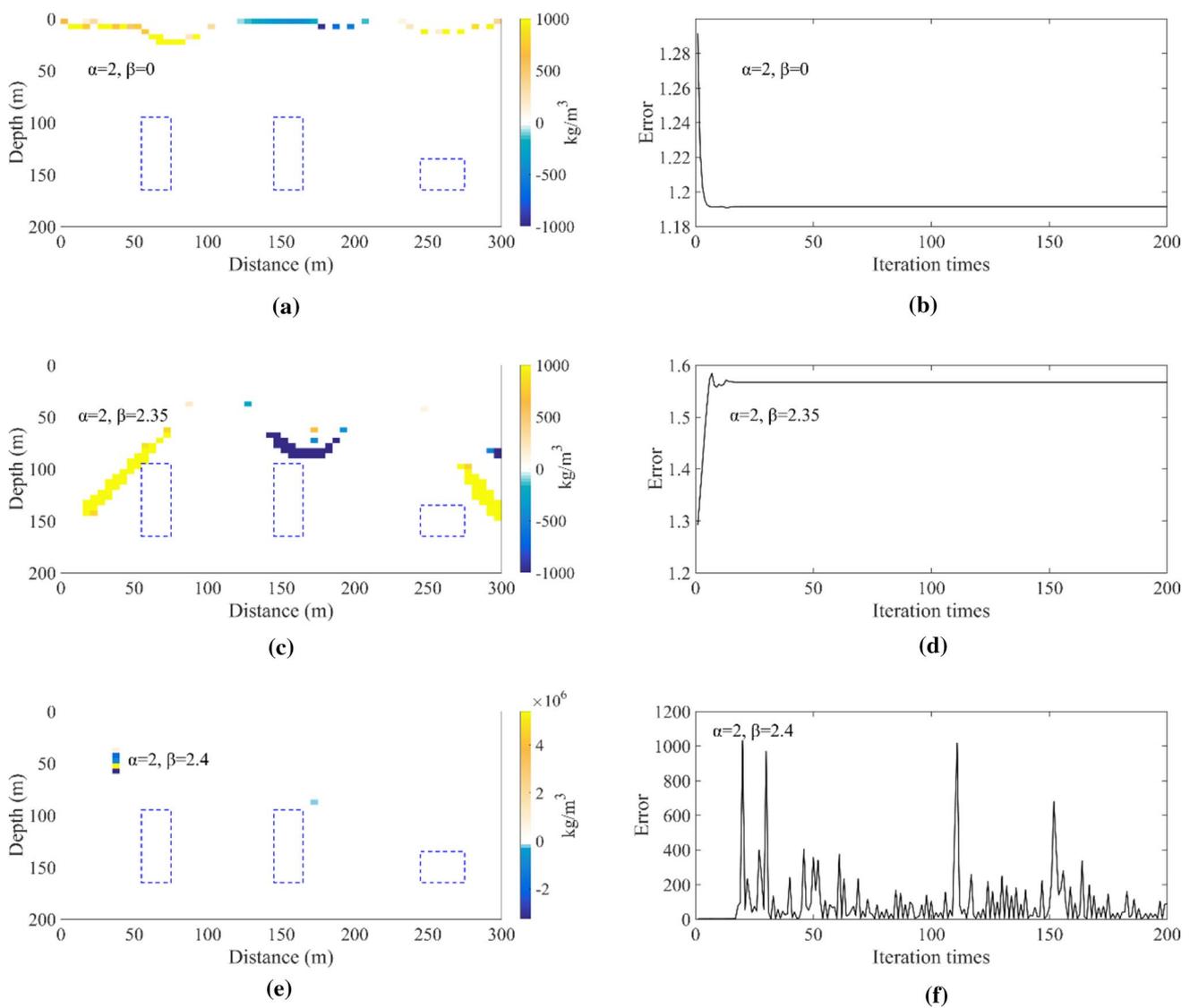
Fig. 2 Inversion results and relative residual errors using the RCGI with a fixed α value of 2 and different β. The blue dashed rectangles in **a**, **c**, and **e** indicate the locations of the true model blocks. The maximum iteration times and e_1 from Eq. (11) are set to 200 and 0.20, respectively, for terminating the iteration algorithm

the RCGI and the right column (Fig. 2b, d) shows the results obtained by the REECGI. The same α and β are obtained by both methods, i.e., $\alpha = 1.0$ and $\beta = 4.0$, as shown in Fig. 2a, b. Moreover, similar inversion results are obtained with the residual errors: 0.2120 and 0.2203 for RCGI and REECGI, respectively. However, there is a certain difference in the computational time between RCGI (697.73 s) and REECGI (650.58 s) when using a 2.2 GHz processor. The saved calculation time of REECGI is approximately 6.7% compared to RCGI.

Inversion of field data

In this section, we apply the CGI, RCGI, and REECGI methods to the inversion of gravity anomaly field data measured in an iron ore mine at Qian'an, Hebei Province, China. The iron ore is predominantly located in the strata of the Archean Santunying Formation with a background density of 2670 kg m^{-3}. There are three tectonic boundaries shown in Fig. 4: F1, F2, and F3. F1 is composed of north-west F1–1, F1–2, and F1–3 tectonic boundaries, F2 is composed of south-west F2–1 and F2–2 tectonic boundaries, and F3 is a north-east stratigraphic lithology interface. We follow the principle of cutting structures to design three profiles, P201, P202, and P203, placed 200 m from each other.

The CG-5 relative gravimeter (No. 1095) produced by Scintrex in Canada is used to measure the gravity anomaly with a measurement accuracy of better than 5 μGal. The field data is further processed by normal field, Bouguer, and terrain corrections. As shown in Fig. 4, the high gravity anomaly area is predominantly distributed to the west of F3, whereas the low gravity anomaly area is to the east. The background density information is removed from the gravity measurements by forward modelling.

According to existing geological data shown in Fig. 5 (Fig. 5a shows the 3D model structure of the ore body in the study area, Figs. 5b–d shows the 2D model structure beneath profiles P201, P202, and P203, respectively), the composition of the ore minerals is as follows. The dominant mineral type is hematite, followed by a small amount of martite with a band-shaped structure, represented as yellow polygons in the figure. Both ore minerals are treated as having the same density and the corresponding residual density is 660 kg m^{-3}. In addition, some low-density rock masses exist that constitute a broken complex (residual density: -470 kg m^{-3}), represented as dark blue polygons, as well as sedimentary rocks (residual density: -70 kg m^{-3}), which are represented as light blue polygons. Based on the available data, the entire inversion area is set to 1200 m × 1000 m, and the corresponding
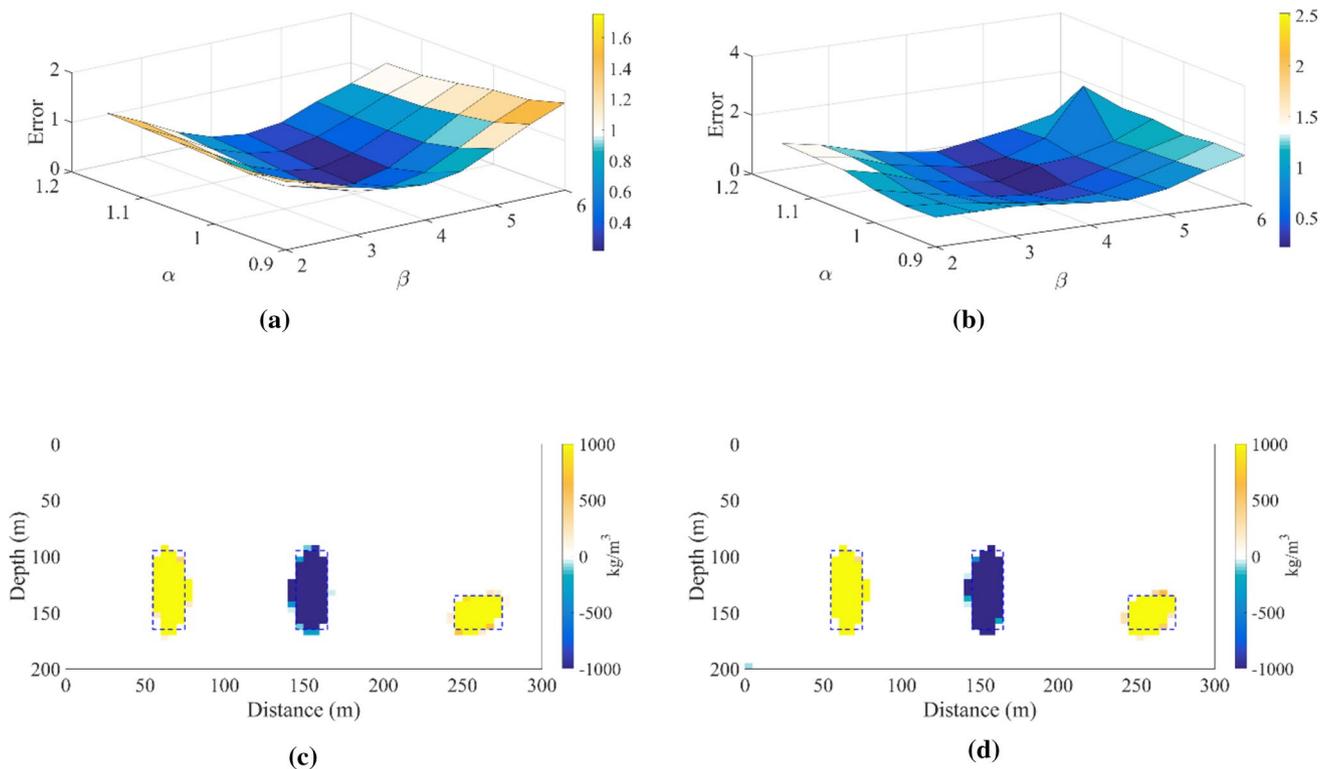
Fig. 3 Optimal α and β values determined by **a** RCGI and **b** REECGI. The same α and β values ($\alpha = 1.0$, $\beta = 4.0$) are obtained by both methods. Inversion results from **c** RCGI and **d** REECGI

Fig. 4 Study area of an iron ore mine in Qian'an, Hebei Province, China. There are three tectonic boundaries: F1 (composed of north-west F1–1, F1–2, and F1–3), F2 (composed of south-west F2–1 and F2–2), and F3 (north-east stratigraphic lithology interface). The three profiles P201, P202, and P203 are located 200 m from each other
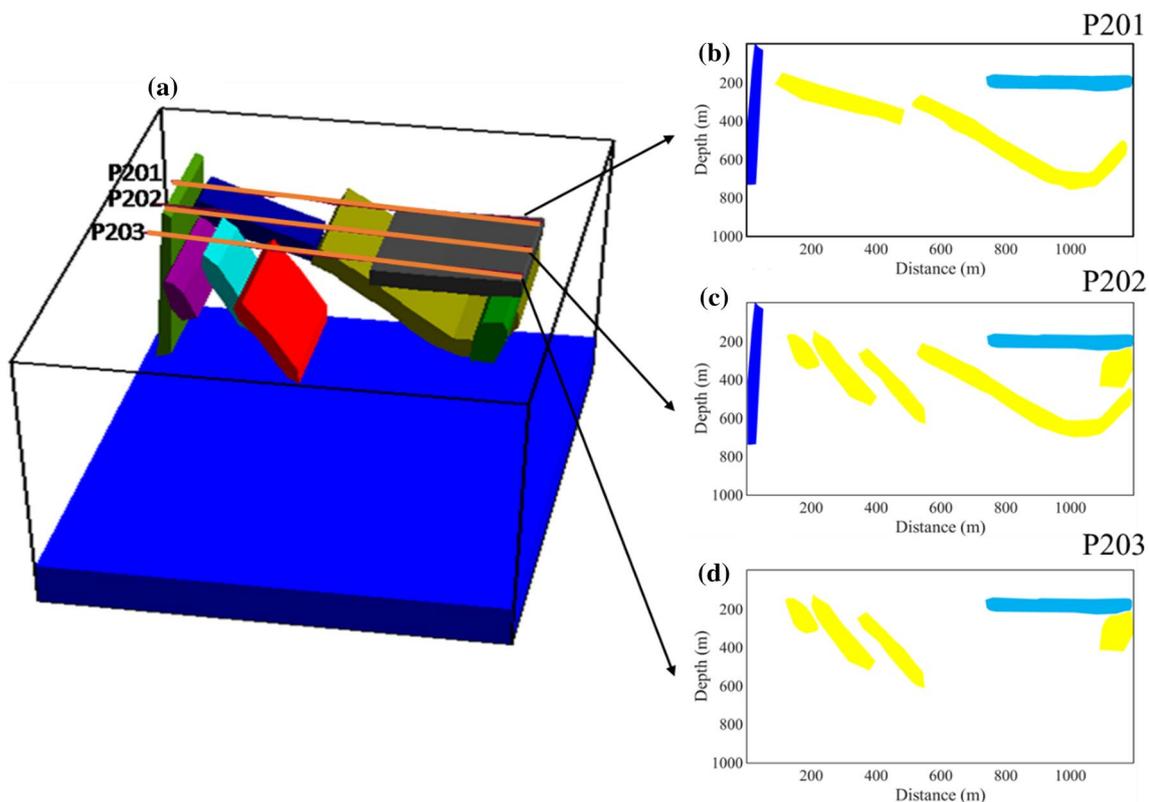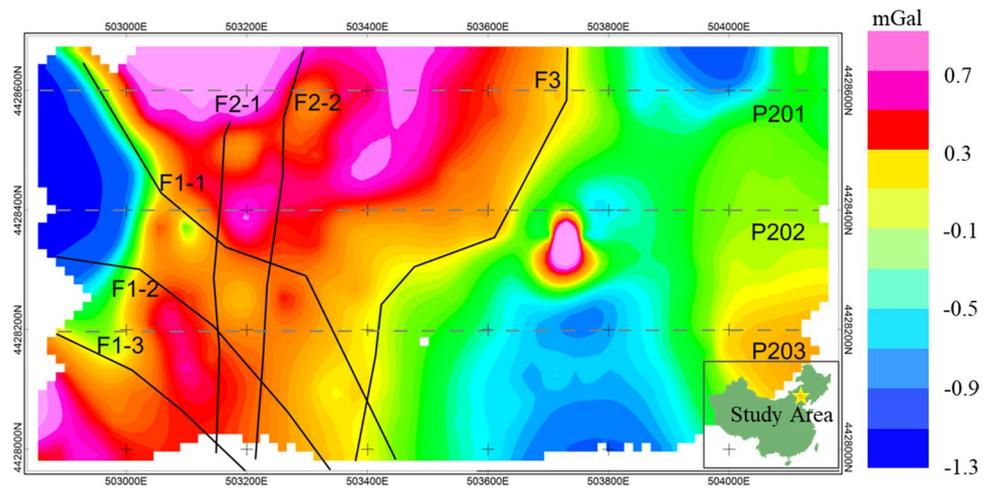
Fig. 5 Mineral distribution beneath the three profiles, P201, P202, and P203: **a** Overall distribution of minerals. Simplified mineral distribution beneath **b** P201; **c** P202; and **d** P203. Three types of miner-als are shown: hematite (residual density 630 kg m^{-3}) represented by yellow polygons, broken complexes (residual density -470 kg m^{-3}) in dark blue polygons, and sedimentary rocks in light blue polygons

model blocks consist of a 120 m × 100 m grid with dimensions of 10 m × 10 m in the horizontal and vertical directions, respectively.

Now, we conduct inversions of the three profiles using CGI, RCGI, and REECGI methods. The CGI method with fixed α (=2) and β (=0) is first used to confirm its unsuitability at great depths in the inversion area. As we can see from Fig. 6, the inversion results of all three profiles show

significant near-surface concentration effects, although they are compact (Fig. 6a–c, e) and the data residuals are small (Fig. 6b, d, f).

As CGI is deemed unsuitable for this study, we apply RCGI and REECGI for the three profiles. The best choice of α and β should naturally be considered; thus, the step sizes are set to 0.01, and the corresponding α and β values of the three profiles obtained by Eq. (9b) are shown in
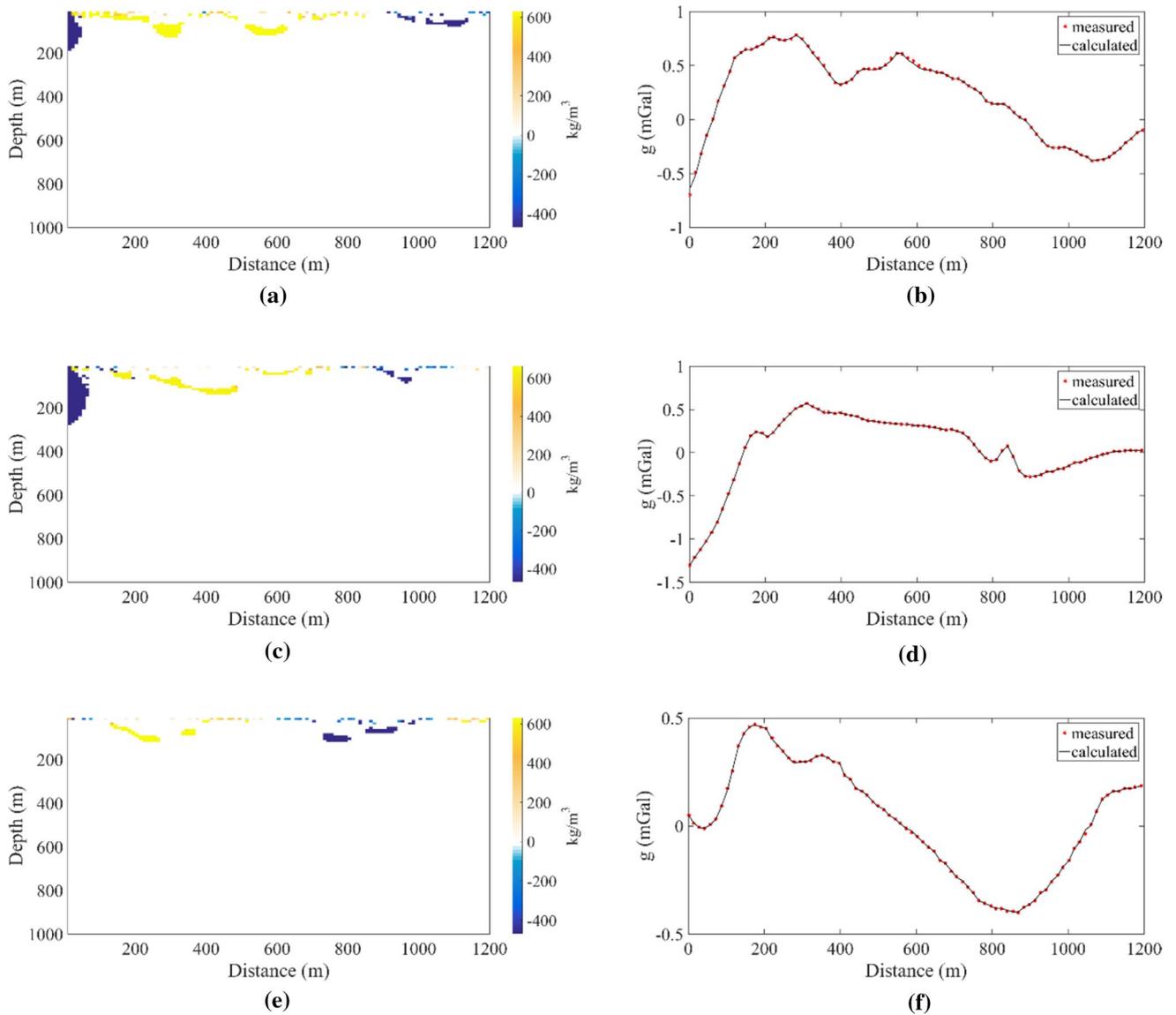
Fig. 6 Inversion results (left) and corresponding data misfits (right) determined by CGI for **a**, **b** P201; **c**, **d** P202; and **e**, **f** P203. Near-surface concentration effects are observed in the inversion results of all three profiles

Table 1. As the data fitting results are good, we do not show the data fitting charts for both methods as shown in Fig. 6b, d, f. The inversion results of the three profiles obtained with the appropriate α and β values using the two methods are shown in Fig. 7 (results in the left column are obtained by

RCGI and those in the right are obtained by REECGI). Both methods yield almost the same results for P201, P202, and P203, as shown in Fig. 7a–f, in which the hematite, broken complex, and sedimentary rocks are shown as yellow, dark blue, and light blue polygons, respectively.

Table 1 Optimal α and β values of RCGI and REECGI methods used for the inversion of three gravity anomaly profiles, P201, P202, and P203. Both step sizes are set to 0.01 for determining α and β values within the same interval

Method	P201		P202		P203	
	α	β	α	β	α	β
RCGI	1.60	1.70	1.49	1.13	1.70	1.70
REECGI						

For P201, compared to the prior geological information shown in Fig. 5b, the inversion results of the broken complex are closer to the actual situation than those of hematite and sedimentary rocks. The worst inversion solution is obtained where the hematite (positive density) and the sedimentary rocks (negative density) are vertically overlapped. Only a small amount of sedimentary rocks and hematite is obtained, as shown in Fig. 7a, b. For P202, compared to the results shown in Fig. 5c, the left broken complex and the shallow hematite are obtained relatively accurately, whereas the very deep minerals, especially when overlapped with sedimentary rocks, are difficult to obtain, as seen from Fig. 7c, d. For

P203, compared to results shown in Fig. 5d, the hematite and sedimentary rocks in the middle of Fig. 7e, f are obtained relatively accurately, unlike those at other locations.

Finally, we test the computational times of the RCGI and REECGI methods using P201, as shown in Fig. 8. In order to observe the calculation efficiency of the two methods more intuitively, the number of iterations are set to 100, 300, 600, and 1000 times, respectively. According to Fig. 8, the computational time of REECGI is lower than that of RCGI, and the time saved becomes more significant as the number of iterations increases.

(a)　　　　　　　　　　　　　　　　　　　　　(b)

(c)　　　　　　　　　　　　　　　　　　　　　(d)

(e)　　　　　　　　　　　　　　　　　　　　　(f)
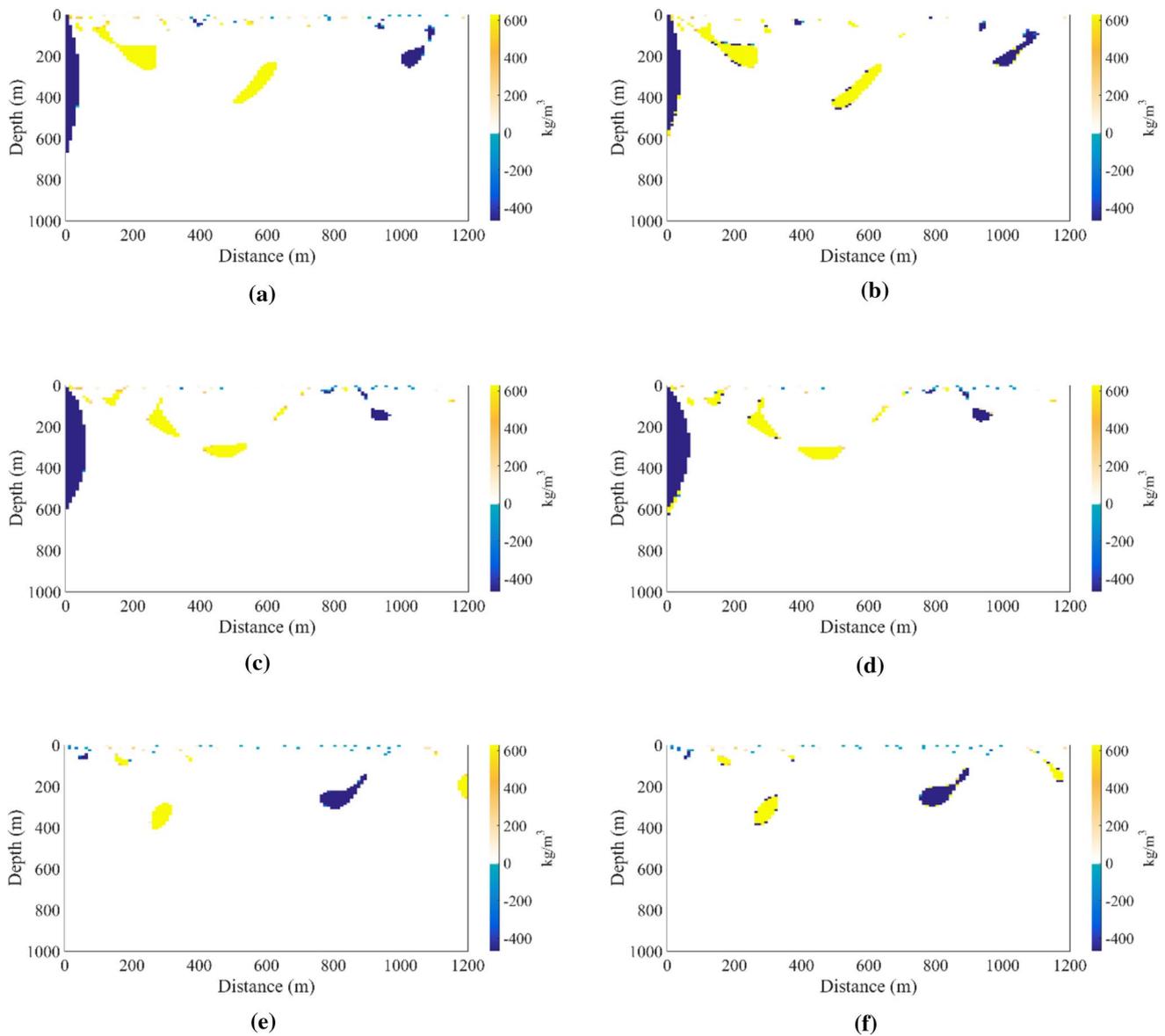
Fig. 7 Inversion results of profiles P201, P202, and P203 by RCGI and REECGI. The left column shows the inversion results of RCGI for **a** P201, **c** P202, and **e** P203. The right column shows the inversion results of REECGI for **b** P201, **d** P202, and **f** P203. The inversion results of the two methods are almost the same
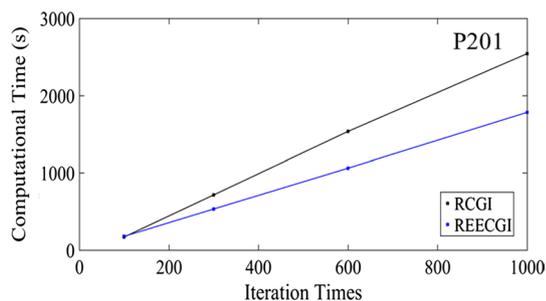
Fig. 8 Comparison of the computational time required by RCGI and REECGI methods using profile P201. The black and blue lines represent the computational time of RCGI and REECGI, respectively. REECGI is more efficient that RCGI, and the efficiency increases with the number of iterations

Conclusion

We developed the RCGI method to ensure high flexibility for gravity anomaly inversion, especially for compact mineral models located at great depths, by generalising the classic CGI method. We then developed the REECGI method by improving the RCGI method to have higher computational efficiency. In addition, we proposed a quantitative method for determining the two key parameters (the compactness weighting factor α and the depth weighing factor β) required to obtain the optimal solution of RCGI or REECGI methods. Synthetic and field tests both indicate that RCGI and REECGI result in more accurate inversions of positive and negative model blocks than CGI, and that REECGI is more efficient than RCGI. The REECGI method can be easily run on a computer in only a few seconds, and can be applied to the inversion of any gravity data to obtain more acceptable resolution in the vertical direction.

Data Availability Applicable.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Code availability Applicable.

References

Acar R, Vogel CR (1994) Analysis of total variation penalty methods. Inverse Prob 10:1217–1229

Barbosa VCF, Silva JBC (1994) Gen Compact Gravity Invers Geophys 59(1):57–68

Blakely RJ (1995) Potential theory in gravity and magnetic applications. Cambridge University Press, Cambridge

Chen S, Zhang J, Shi YL (2008) Gravity inversion using the frequency characteristics of the density distribution. Appl Geophys 5(2):99–106

Constable SC, Parker RC, Constable GG (1987) Occam's inversion: a practical algorithm for generating smooth models from EM sounding data. Geophys 52:289–300

Ghalehnoee MH, Ansari A, Ghorbani A (2017) Improving compact gravity inversion using new weighting functions. Geophys J Int 28(1):546–560

Guillen A, Menichetti V (1984) Gravity and magnetic inversion with minimization of a specific functional. Geophys 49:1354–1360

Hadamard J (1902) Sur les problèmes aux derivées partielles et leur signification physique. Princeton Univ Bull 13: 49–52. Reprinted in his Oeuvres, Centre Nat Rech Sci, Paris, 3:1099–1105

Karaoulis K, Revil A, Minsley B et al (2014) Time-lapse gravity inversion with an active time constraint. Geophys J Int 196:748–759

Last BJ, Kubik K (1983) Compact gravity invers Geophys 34:65–74

Levenberg K (1944) A method for the solution of certain nonlinear problems in least squares. Q Appl Math 2:164–168

Li YG, Oldenburg DW (1998) 3-D inversion of gravity data. Geophys 63(1):109–119

Marquardt DW (1963) An algorithm for least squares estimation of nonlinear parameters. SIAM J 11:431–441

Mendonca CA, Silva JBC (1994) The equivalent data concept applied to the interpolation of potential field data. Geophys 59(5):722–732

Mendonca CA, Silva JBC (1995) Interpolation of potential field data by equivalent layer and minimum curvature: a comparative analysis. Geophys 60(2):399–407

Menke W (1989) Geophysical data analysis: discrete inverse theory, Revised edn. Academic Press, Cambridge

Pilkington M (1997) 3-D magnetic imaging using conjugate gradients. Geophys 62:1132–1142

Pilkington M (2009) 3D magnetic data-space inversion with sparseness constraints. Geophys 74(1):L7–L15

Portniaguine O, Zhdanov MS (1999) Focusing Geophys Invers Images Geophys 64(3):874–887

Roland M, Vadim M, Dimitri K et al (2013) Gravity inversion using wavelet-based compression on parallel hybrid CPU/GPU systems: application to southwest Ghana. Geophys J Int 195:1594–1619

Rosas-Carbajal M, Jourde K, Marteau J et al (2017) Three-dimensional density structure of La Soufrière de Guadeloupe lava dome from simultaneous muon radiographies and gravity data. Geophys Res Lett 44(13):6743–6751

Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation-based noise removal algorithms. Phys D 60:259–268

Silva FJS, Barbosa CF, Silva JBC (2009) 3D gravity inversion through an adaptive-learning procedure. Geophys 74(3):I9–I21

Silva FJS, Barbosa CF, Silva JBC (2011) Adaptive learning 3D gravity inversion for salt-body imaging. Geophys 76(3):I49–I57

Silva JBC, Barbosa VCF (2006) Interactive gravity inversion. Geophysics 71(1):J1–J9

Smith RT, Zoltani CK, Klem GJ, Coleman MW (1991) Reconstruction of the tomographic images from sparse data sets by a new finite element maximum entropy approach. Appl Opt 30:573–582

Williams NC (2008) Geologically-constrained UBC-GIF gravity and magnetic inversion with examples from the Agnew-Wiluna greenstone belt, Western Australia. PhD thesis, The University of British Colombia, Faculty of Geophysics

Zhdanov MS (2015) Inverse theory and applications in geophysics. Elsevier, Amsterdam

RESEARCH ARTICLE - APPLIED GEOPHYSICS

Viscoelasticity expression and extension of seismic dispersion and attenuation in porous media with multiple fracture sets

Jinwei Zhang[1] · Renwei Ding[1] · Lihong Zhao[1] · Deying Wang[1]

Abstract

Intensive studies have been conducted on fluid-related seismic dispersion and attenuation in saturated anisotropic media. Most of the studies are concentrated on the transversely isotropy media. However, the fractures distribution in subsurface reservoirs is often complex. When there are multiple fracture sets developing in a porous background, the signatures of seismic dispersion and attenuation remain unclear. In this paper, we propose a method to calculate the frequency-dependent stiffness matrix of a porous medium with multiple fractures sets from a perspective of viscoelasticity. Due to the favorable approximation performance of the generalized standard linear solid model and Chapman model, we use a modified form of generalized standard linear solid model to simulate the frequency-dependent stiffness tensor of porous media with multiple fracture sets. The representation of the stiffness tensor utilizes the modulus defect to denote the effects the fractures including fracture density and geometry. With the procedure of calculating the stiffness tensors at low- and high-frequency limits, we can easily calculate the frequency-dependent stiffness tensor for media with multiple fracture sets with arbitrary orientations and directions. We then analyze the effects of the fracture parameters on the viscoelasticity characteristics taking orthotropic medium as an example. The results can help to understand the viscoelasticity and the mesoscopic seismic attenuation associated with fractures and fluids and can provide a practical rock physics model when dealing with reservoirs with complex fracture patterns.

Keywords Fractured reservoirs · Seismic rock physics · Dispersion and attenuation · Viscoelasticity

Introduction

Many theoretical mechanisms have been proposed to explain the fluid-related seismic dispersion and attenuation in porous media, among which wave-induced fluid flow is one of the predominant mechanisms. Commonly, there are three scales of fluid flow generating intrinsic attenuation in different frequency bands. Biot theory which laid the foundation for most studies of wave-induced fluid flow describes a macroscopic flow in the wavelength scale (Gassmann 1951; Biot 1956a, b). The squirt flow (Mavko and Nur 1975; Dvorkin et al. 1995) due to fluid pressure gradient between microcracks and pores occurs in a grain scale. Both Biot global flow and squirt flow cannot generate significant attenuation in

the seismic frequency band (Pride et al. 2004). It's believed that the mesoscopic flow due to mesoscopic heterogeneities is a primary source of the observable dissipation in seismic frequency band. For porous media, there are mainly two kinds of mesoscopic heterogeneities: patchy saturation and lithology variation. The patch geometries and distributions have a significant influence on the natures of the seismic dispersion and attenuation (White 1975; Johnson 2001). The heterogeneity due to lithology variation is often modelled as the case of two solid phases saturated with a single fluid (Pride and Berryman 2003a, b).

The seismic dispersion and attenuation in fractured porous media have also attracted considerable attentions (Shen et al. 2015; Shuai et al. 2017; Yuan et al. 2019). For fractured porous media, the strong level of heterogeneity from the introduction of mesoscopic fractures is the important source of the mesoscopic attenuation due to the distinctions between compliant mesoscopic fractures and stiff equant pores (Jakobsen 2004; Gurevich et al. 2009; Brajanovski et al. 2005, 2010; Müller et al. 2010).

✉ Jinwei Zhang
zhangjinwei.1987@163.com

1 College of Earth Science and Engineering, Shandong University of Science and Technology, No. 579, Qianwangang Road, Huangdao District, Qingdao, China

Chapman ([2003]) developed a mesoscopic fracture model by introducing a set of aligned mesoscale fractures into the microscale squirt flow model (Chapman et al. [2002]) and analyzed the frequency-dependent seismic characteristics as well as the anisotropy. While Chapman model is essentially based on a superposition principle, it is a good first-order approximation of accurate poro-elastic model and adaptive for actual applications. Most of the studies on frequency-dependent dispersion and attenuation for anisotropy media mainly focus on the transverse isotropy media, namely there is only one set of fractures. The mesoscale squirt flow model considering two sets of fractures has been evaluated by Chapman ([2009]). As is often the case, more than one set of fractures exist in subsurface formations (Shi et al. [2018]). Under the circumstance of tectonic stresses and diagenetic environments, there are often multiple sets of orthogonal or oblique fractures underground. Therefore, discussion on frequency-dependent characteristics of porous medium with multiple fracture sets becomes necessary and meaningful. And it's nontrivial to extend Chapman double fracture model to the case with three or more sets of fractures merely using the wave-induced flow mechanism. Ali and Jakobsen ([2011]) analyzed frequency-dependent seismic velocity and attenuation anisotropy in reservoirs with multiple fracture sets using T-matrix approach.

The seismic attenuation due to wave-induced flow mechanism exhibits the relaxation behavior, so the fluid-saturated media can be seen as a kind of viscoelastic media which inspires us to use viscoelastic models to approximate the fractured porous media. There are many kinds of viscoelastic models, such as Maxwell model, Kelvin–Voigt model, and the standard linear solid (SLS) model and so on. As a special case of Cole–Cole model, the SLS model also called Zener model is a simple realistic representation of viscoelastic media (Dvorkin and Mavko [2006]) and has been used for representation of the seismic attenuation in porous media (Picotti et al. [2010]). However, studies related the viscoelastic representation for the seismic attenuation in fractured porous media are rare. According to the results of Carcione and Gurevich ([2011]), the SLS model has a satisfactory fitting performance for Biot loss mechanism and squirt flow dissipation mechanism. The kernel of Chapman mesoscale model is based on the squirt flow model. Therefore, in this paper, we aim to express and extend the Chapman model from the perspective of viscoelasticity. We attempt to calculate the frequency-dependent stiffness tensor as well as the Q^{-1} matrix for porous media with multiple fracture sets through building the relationship between Chapman model and SLS model. The paper can provide a novel perspective when dealing with reservoirs with complex fracture patterns.

Methodology

Review of Chapman double fracture model

Chapman's mesoscale squirt flow model contains two scales of pore spaces: the pores and microcracks with a grain size and the fractures with radii much greater than the grain size but smaller than the seismic wavelength. The random distribution of microcracks and pores constitutes an isotropic background and the aligned fractures make the media anisotropic. The mesoscale model will degenerate to the microscale squirt flow model with the absence of the fractures. The density of microcracks is often set to be zero for simplicity and practicability in field application. Figure [1] shows a simple sketch of Chapman model.

According to Chapman ([2009]), the effective complex stiffness tensor for a fractured porous medium with two sets of aligned fractures can be expressed as:

$$C(\omega) = C^0 - \varphi_p C^1(\omega) - \varepsilon_f^1 C^2(\omega) - \varepsilon_f^2 C^3(\omega), \tag{1}$$

where C^0 is the isotropic elastic tensor associated with the properties of solid minerals; C^1 is the pore correction term; C^2 and C^3 are the correction terms of two sets of fractures. φ_p is the porosity; ε_f^1 and ε_f^2 are the densities of the two sets of aligned fractures. C^1, C^2 and C^3 in complex form containing attenuation information are related with the background properties, fracture parameters, fluid types, and frequency. The detailed parameterization process is rather complicated and the concrete form is given by Chapman ([2009]). When the directions and geometries of the two sets of fractures are identical, the results of Eq. [1] are consistent with those of the model presented by Chapman ([2003]).

There are two crucial parameters in Eq. [1]: the microscale relaxation time τ_m describing the grain scale fluid flow and the mesoscopic relaxation time τ_f describing the fracture scale fluid flow. The relationship between τ_m and τ_f is:

$$\tau_f = \frac{a_f}{\varsigma} \tau_m, \tag{2}$$

where a_f and ς are the radius of the fractures and the grain size, respectively. τ_m is function of the fluid mobility (defined as k/η, k denoting permeability and η denoting the fluid viscosity) and τ_f is also proportion to the fracture radius. As fracture radius increases (assuming the fracture density and aspect ratio keep constant), the specific surface area decreases, and more volume of fluids have to move through an element of surface area to equalize the induced pressure, which requires more time (Maultzsch et al. [2003]). The reciprocal of the relaxation time called characteristic frequency separates high-frequency behavior from low-frequency behavior and specifies the frequency band in which velocity changes most rapidly. The microscale squirt-flow

Fig. 1 A simple sketch of Chapman model: the white color represents solid minerals; the blue circles represent pores; the green ellipses represent microcracks; the red ellipses represent mesoscopic fractures and the black solid lines represent the pore throats. [Modified from Lan (2014)]
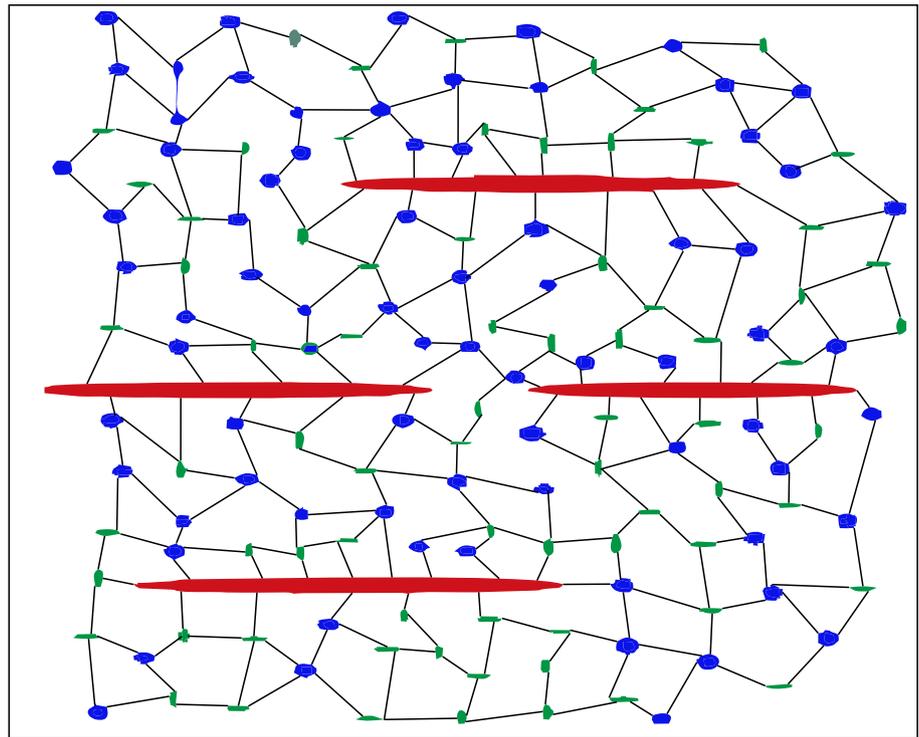
frequency usually lies between the sonic and ultrasonic frequency range. The incorporation of fractures makes the characteristic frequency be within the seismic frequency band.

Equation 1 contains information about fracture geometry and directions, so we can calculate the frequency-dependent stiffness tensor for a porous medium with two sets of aligned fractures with arbitrary directions even arbitrary seal conditions. When the directions and geometries of the two sets of fractures are identical, the results of Eq. 1 are consistent with those of the model presented by Chapman (2003). When the fractures are sealed, the fluids flow between fractures and background is blocked. We can model this case by setting the timescale to be infinity and correspondingly the characteristic frequency is 0.

The generalized standard linear solid model

Picotti and Carcione (2017) studied the attenuation characteristics of seismic velocity of patchy saturation model, using Cole–Cole model, and pointed out that the SLS model or called Zener model is suitable for the attenuation behavior of Biot model and squirt flow model while Cole–Cole model has a better fitting performance to fit the patchy saturation model. The standard linear solid (SLS) model, which is a series combination of a spring and a Kelvin–Voigt model, is a realistic representation of viscoelastic media (Dvorkin and Mavko 2006; Picotti et al. 2010). When there are several Zener elements parallel connected, a generalized SLS (GSLS) model is then constructed and the effective moduli of the GSLS model

is relevant with moduli and characteristic frequencies of each Zener element. As the asymptotic behavior of the inverse quality factor (Q^{-1}) of SLS model is as same as that of Chapman model ($Q^{-1} \propto \omega$ at low frequencies and $Q^{-1} \propto 1/\omega$ at high frequencies), we can use the SLS model to approximate the Chapman model with single fracture set and use GSLS model to approximate the Chapman model with multiple fracture sets. There are several representation forms of the complex modulus. We use the relaxed moduli (low-frequency limit) and the unrelaxed moduli (high-frequency limit) to calculate the complex modulus. For anisotropic viscoelastic model, the stiffness coefficients can be expressed as:

$$C_{ij}(\omega) = \frac{C_{ij}^{\infty}\left[C_{ij}^{0} + i\frac{\omega}{\omega_c}\sqrt{C_{ij}^{0}C_{ij}^{\infty}}\right]}{C_{ij}^{\infty} + i\frac{\omega}{\omega_c}\sqrt{C_{ij}^{0}C_{ij}^{\infty}}}, \tag{3}$$

where C_{ij}^{0} and C_{ij}^{∞} are the stiffness coefficients at low- and high-frequency limits, respectively. ω_c is the characteristic frequency. We set up a model to verify the fitting effect of SLS model on Chapman model. The properties of the solid minerals are $\lambda = 51$ Gpa, $\mu = 29$ Gpa and $\rho = 2710$ kg/m^3. The porosity is 10%. The fluid properties are $K_f = 25$ Mpa and $\rho_f = 65$ kg/m^3. We assume the microscale relaxation time is 4×10^{-7}s. Figure 2a shows the fitting results of Eq. 3 for Eq. 1 for the case of only one set of fractures (only one characteristic frequency). The fracture radius is 1 m and fracture density is 0.02. The grain size is 200 μm. The fractures are horizontal, constituting a VTI medium. Figure 2b

Fig. 2 **a** The approximation of SLS model on Chapman model (single fracture set). **b** The absolute difference and relative difference between C_{33} calculated from Eqs. 1 and 3

displays the absolute difference and relative difference between the real part of C_{33} calculated by Eqs. 1 and 3. We can see that the difference is negligibly small, which shows the fitting performance is desirable.

For the case of two sets of fractures, there are two characteristic frequencies in the model. We use the GSLS model for the representation. The original form of GSLS model gives the expression of the stiffness coefficients as:

$$c_{ij}(\omega) = \sum_{l=1}^{L} c_{ij}^{l}(\omega).$$ (4)

However, the conditions for Eq. 4 are that the elements in GSLS model are independent and parallel connected, which is not satisfied in our model with two sets of fractures embedded in a porous background medium. The fractures and pores interact under the effects of seismic waves on fluids. Therefore, we adopt a modified form of Eq. 4 as we find that the modulus defect can denote the contribution of each set of fractures including the fracture density and fracture orientations. The stiffness coefficients are calculated as below:

$$
\begin{aligned}
C_{ij}(\omega) = {} & \frac{D_{ij}^{1}}{D_{ij}^{1} + D_{ij}^{2}} \frac{C_{ij}^{\infty}\left[C_{ij}^{0} + i\frac{\omega}{\omega_{c}^{1}}\sqrt{C_{ij}^{0}C_{ij}^{\infty}}\right]}{C_{ij}^{\infty} + i\frac{\omega}{\omega_{c}^{1}}\sqrt{C_{ij}^{0}C_{ij}^{\infty}}} \\
& + \frac{D_{ij}^{2}}{D_{ij}^{1} + D_{ij}^{2}} \frac{C_{ij}^{\infty}\left[C_{ij}^{0} + i\frac{\omega}{\omega_{c}^{2}}\sqrt{C_{ij}^{0}C_{ij}^{\infty}}\right]}{C_{ij}^{\infty} + i\frac{\omega}{\omega_{c}^{2}}\sqrt{C_{ij}^{0}C_{ij}^{\infty}}}.
\end{aligned}
$$ (5)

The meanings of C_{ij}^{0} and C_{ij}^{∞} are same as those in Eq. 3. Here there are two sets of fractures embedded in the medium. D_{ij}^{1} and D_{ij}^{2} are the modulus defects of C_{ij} when individual fracture set is introduced into the porous background medium, defined as $D_{ij}^{k} = \left(C_{ij}^{\infty,k} - C_{ij}^{0,k}\right)/\sqrt{C_{ij}^{\infty,k}C_{ij}^{0,k}}$ where k is 1 or 2. Notably, $C_{ij}^{0,k}$ and $C_{ij}^{\infty,k}$ are the coefficients at low- and high-frequency limits when the kth set of fractures is embedded in the porous background. When the two set of fractures have same parameters, we then obtain $D^{1} = D^{2}$, and Eq. 5 degenerates to Eq. 3. Figure 3a shows the approximation results of Eq. 5 for Eq. 1 for the case of two sets of fractures. The parameters of the background medium and fluid are as same as those in Fig. 2. The fracture radii are 1 m and 0.02 m, respectively. The fracture densities are both 0.02. Both set of fractures are horizontal, constituting a VTI medium. Although the absolute and relative difference between the real part of C_{33} calculated by Eqs. 1 and 5 (Fig. 3b) is larger than those in Fig. 2b, it is still insignificant. From Figs. 2 and 3, we can see that the SLS model has a good fitting performance of Chapman model, and our new expression of GSLS model is adaptive to the Chapman model with two sets of fractures. Notice here we just display the real part of C_{33} in Figs. 2 and 3. Actually, the approximations of other stiffness coefficients also work very well. Since the modulus defect contains the effects of fracture densities and fracture orientations, Eq. 5 is adaptive for the case of two sets of fracture with arbitrary radii, orientations and geometries. Extending Eq. 5 to case of multiple fracture sets obtains:
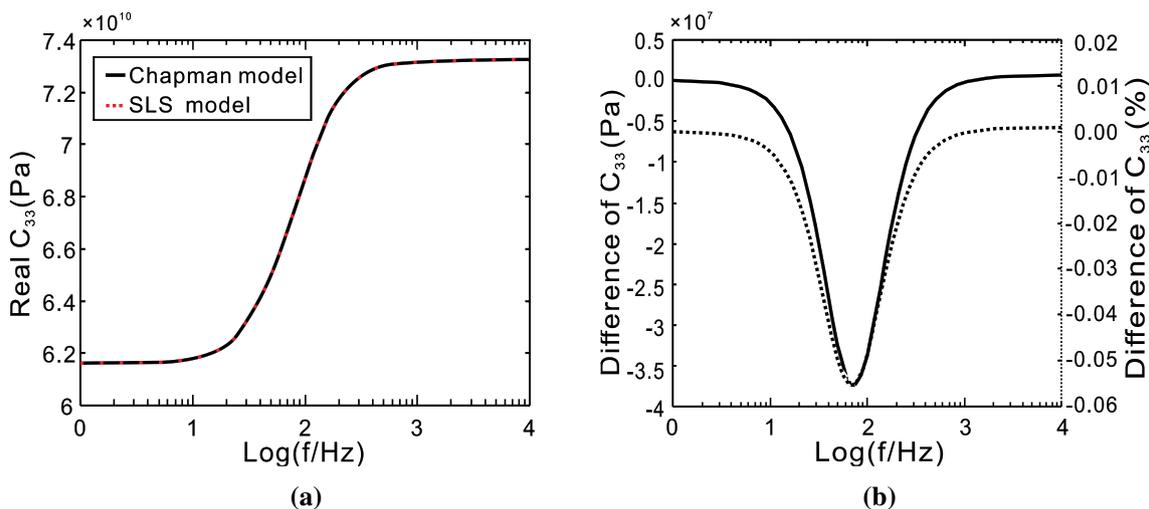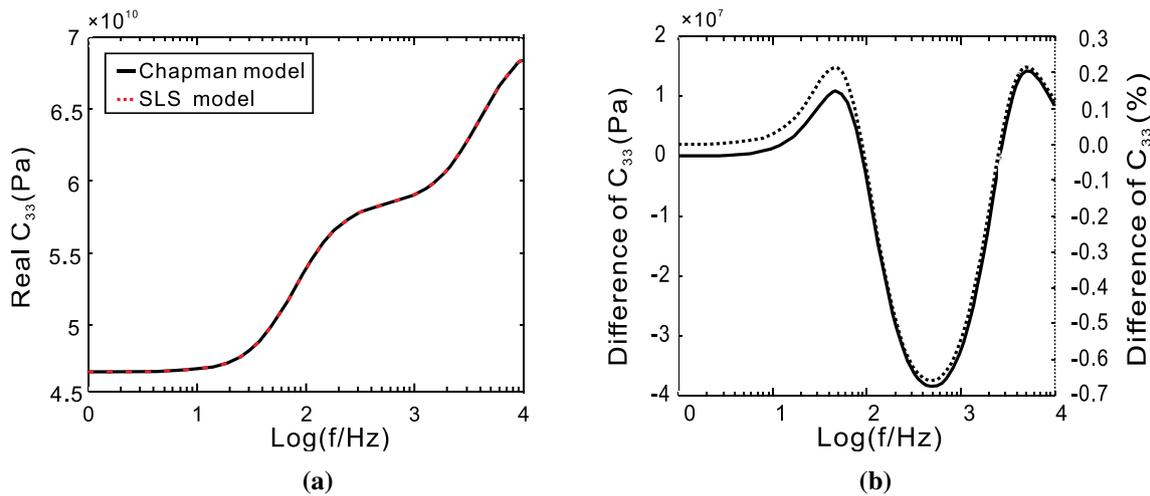
Fig. 3 **a** The approximation of SLS model on Chapman model (double-fracture sets). **b** The absolute difference and relative difference between C_{33} calculated from Eqs. 1 and 5

$$C_{ij}(\omega) = \sum_{k=1}^{K} \frac{D_{ij}^{k}}{\sum_{k=1}^{K} D_{ij}^{k}} \frac{C_{ij}^{\infty} \left[C_{ij}^{0} + i \frac{\omega}{\omega_{c}^{k}} \sqrt{C_{ij}^{0} C_{ij}^{\infty}} \right]}{C_{ij}^{\infty} + i \frac{\omega}{\omega_{c}^{k}} \sqrt{C_{ij}^{0} C_{ij}^{\infty}}} \qquad (6)$$

where D_{ij}^{k} is the modulus defect of C_{ij} induced by the kth set of fractures and ω_{c}^{k} is the characteristic frequency corresponding with the kth set of fractures.

The low- and high-frequency-limiting stiffness tensors

For a porous medium with three or more fracture sets, the stiffness tensors at low- and high-frequency limits (C_{ij}^{0} and C_{ij}^{∞} in Eq. 6) are unknown and hard to obtain through existing models. Here is a strategy to calculate the stiffness tensors at low- and high-frequency limits. Chapman's squirt flow model is consistent with the Gassmann equations in the low-frequency limit and consistent with the Hudson's isolated fracture model in the high-frequency limit. Therefore, we can use anisotropic Gassmann equation to obtain the low-frequency limit modulus and use Hudson's fracture model to get the high-frequency limit modulus. As long as we know the characteristic frequencies, we can calculate the frequency-dependent velocity and attenuation in the whole frequency band.

At low-frequency limit, we can't directly extract the pure contribution of fractures as the correction terms of the pores and fractures are coupled together when saturated with fluids in Chapman's model. But for the case of dry fractured matrix, the correction terms of pores and fractures are decoupled at zero frequency. Therefore, we firstly use Chapman's model to obtain the zero frequency

limit stiffness matrix C for dry matrix including one set of fractures by setting the inclusion moduli to zero and separate the background stiffness matrix including the pore correction term $C_{b} - C_{p}$ and fracture correction term C_{f}. Such scheme is applied repeatedly to acquire the corrections for each set of fractures, C_{f}^{i}, by controlling the parameters of the fractures. Then the corrections C_{f}^{i} are added together to get the whole correction for multiple sets of fractures $C_{F}^{low} = \sum C_{f}^{i}$ since the corrections for each set of fractures are independent and linear additive. The correction term C_{F}^{low} contains the effects of fracture orientations and densities. The effective stiffness matrix at zero frequency for dry rock C^{dry} is then obtained by $C^{dry} = C_{b} - C_{p} - C_{F}^{low}$. Finally, anisotropic Gassmann function is used to calculate the saturated stiffness matrix:

$$C^{sat} = C^{dry} + \alpha_{i} \alpha_{j} M, \quad i, j = 1, 2, \dots, 6, \qquad (7)$$

$$\alpha_{m} = 1 - \frac{\sum_{n=1}^{3} C_{nm}^{dry}}{3 K_{g}}, \qquad (8)$$

$$M = \frac{K_{g}}{\left(1 - \frac{K^{*}}{K_{g}}\right) - \Phi\left(1 - \frac{K^{*}}{K_{f}}\right)}, \qquad (9)$$

$$K^{*} = \frac{1}{9} \sum_{i=1}^{3} \sum_{j=1}^{3} C_{ij}^{dry}, \qquad (10)$$

where K_{g} and K_{f} are the bulk moduli of grain and fluid, and Φ is the total porosity $\Phi = \varphi_{p} + \varphi_{f}$. K^{*} is known as the effective bulk modulus.

At the high-frequency limit, fluid flow between pores and fractures can be ignored, and the fractures can be treated as isolated fractures in an isotropic background. Thus, the corrections of pores and fractures are decoupled for both drained and undrained situations. Consequently, we can directly derive the fracture correction term C_f^{high} for each set of fractures at saturated conditions and add them together after coordinate rotation to obtain the overall effect of fractures C_F^{high}. Just as C_F^{low}, C_F^{high} contains the effects of fracture orientations and densities. Finally, the effective stiffness matrix at high-frequency limit can be calculated by $C^{high} = C_b - C_p - C_F^{high}$.

After we obtain the stiffness matrix, the Q^{-1} matrix can be calculated by $Q_{ij}^{-1} = \mathrm{Im}(C_{ij})/\mathrm{Re}(C_{ij})$ (Zhu and Tsvankin 2006).

Numerical example

In order to give an example of the method above, we consider a model including three sets of fractures. The properties of solid minerals and fluids are as same as those in Figs. 2 and 3. Shown as Fig. 4, The parameters of fracture sets are defined as $[a_f^1 = 0.01\ \mathrm{m},\ \varepsilon_f^1 = 0.01]$ (normal to x_1 axis), $[a_f^2 = 0.1\ \mathrm{m}, \varepsilon_f^2 = 0.02]$ (normal to x_2 axis) and $[a_f^3 = 1\ \mathrm{m}, \varepsilon_f^3 = 0.04]$ (normal to x_3 axis). In order to manifest the effects of the fractures explicitly, the fracture radii are set of orders of magnitude difference. The mutual orthogonal fractures constitute an orthotropic medium. Therefore, the stiffness coefficient C_{11} is mainly controlled by the first set of fractures, and so as C_{22} corresponding with the second set of fractures, and C_{33} corresponding with the third set of fractures. The values of stiffness coefficients at low- and high-frequency limits are shown in Table 1. We can see that the values of relaxed C_{11}, C_{22} and C_{33} are smaller than values of the unrelaxed C_{11}, C_{22} and C_{33}, while C_{44}, C_{55} and C_{66} are equal at low- and high-frequency limits, meaning that they are frequency independent. Next, we calculate the modulus defects of the stiffness coefficients for each set of fractures. The results are shown in Table 2 (the modulus defects for C_{44}, C_{55} and C_{66} are not shown since they are all 0). The characteristic frequency is calculated by $\omega_c = 1/2\pi\tau_f$ and τ_f is obtained from Eq. 2 for each set of fractures.

We show the frequency-dependent C_{11}, C_{22} and C_{33}, calculated by Eq. 6 and corresponding Q_{11}^{-1}, Q_{22}^{-1} and Q_{33}^{-1} in Fig. 5. In orthotropic media, C_{11}, C_{22} and C_{33}, correspond to the quasi-P wave moduli when the P-wave propagates along these three axes, respectively. Due to the differences of fracture orientations and fracture densities, the magnitudes and shapes of C_{11}, C_{22} and C_{33}, (Q_{11}^{-1}, Q_{22}^{-1} and Q_{33}^{-1}) are different.

Fig. 4 The schematic diagram of fractured medium. Three sets of fractures are embedded in the background medium. The pores and solid minerals are not shown in the figure. The orthogonal fractures constitute an orthotropic medium
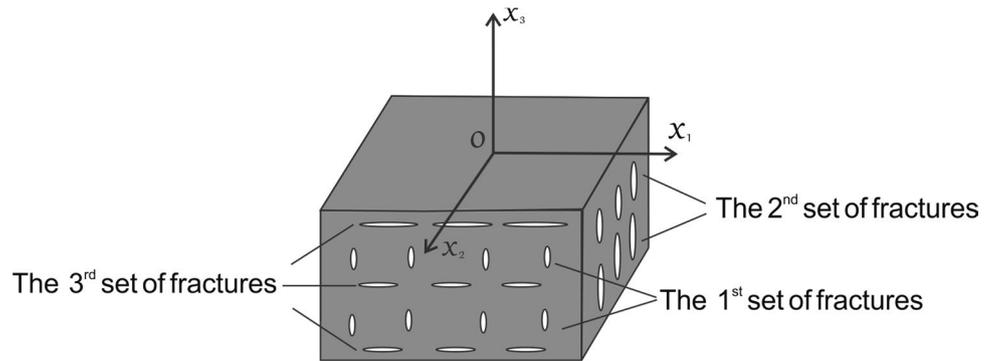
Table 1 The stiffness coefficients at low- and high-frequency limits for a fractured medium with three sets of fractures

Unit (Gpa)	C_{11}	C_{12}	C_{13}	C_{22}	C_{23}	C_{33}	C_{44}	C_{55}	C_{66}
Low	59.5	12.7	9.04	53.7	7.21	42.2	19.8	20.4	21.6
High	72.8	25.8	25.0	71.6	24.6	69.1	19.8	20.4	21.6

Table 2 The modulus defects of the stiffness coefficients for individual fracture set

	D_{11}	D_{12}	D_{13}	D_{22}	C_{23}	C_{33}
$D1$	0.0816	0.1005	0.1005	0.0170	0.0450	0.0170
$D2$	0.0344	0.2188	0.0934	0.1745	0.2188	0.0344
$D3$	0.0706	0.2018	0.5406	0.0706	0.5406	0.4093
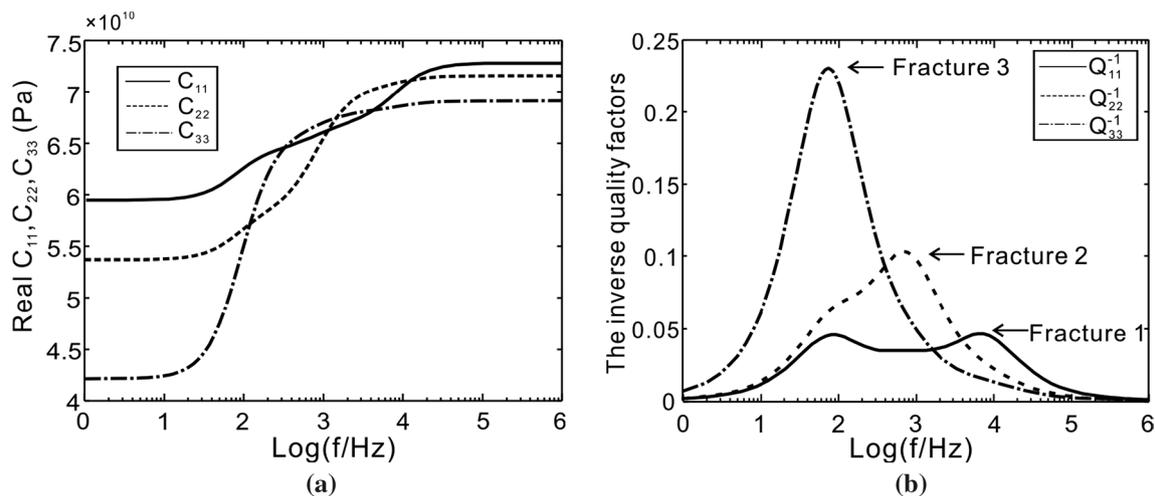
Fig. 5 The variations of C_{11}, C_{22} and C_{33} (**a**) and Q_{11}^{-1}, Q_{22}^{-1} and Q_{33}^{-1} (**b**) with frequency. Three sets of orthogonal fractures with different radii are embedded in the porous background. The properties of the background medium are same as those in Figs. 2 and 3

At low- and high-frequency limits, the magnitudes of C_{11}, C_{22} and C_{33}, remain unchanged and they are controlled by fracture density. As $\varepsilon_f^1 < \varepsilon_f^2 < \varepsilon_f^3$, the relationship between the moduli is $C_{11} > C_{22} > C_{33}$ at both low- and high-frequency limits. In the middle frequency band, the complex relationship between C_{11}, C_{22} and C_{33} is due to the effects of the other two sets of fractures. For C_{11} and C_{22}, we can easily see the effects of the other two sets of fractures. But for the variation of C_{33}, the dominant factor is the third set of fractures. The roles of the other two sets fractured are less prominent. From the frequency-dependent Q^{-1}, we can see that the three sets of fractures contribute to the performance of Q_{11}^{-1}, Q_{22}^{-1} and Q_{33}^{-1}, respectively. The values of inverse quality factors are mainly handled by the corresponding fracture set and affected by the other two sets of fractures. As

the density of the third fracture set is larger than other two fracture sets, its effects are considerable in Q_{11} (C_{11}) and Q_{22} (C_{22}) as well.

The model we construct is essentially orthotropic, which may be the simplest realistic symmetry for many geophysical problems. We can calculate the frequency-dependent velocity and inverse quality factor of quasi-P wave for any incident angles through solving the Christoffel equation. The quasi-P wave velocity and inverse quality factor are shown in Fig. 6 at the incident angles of 30° and 60°. Due to the difference of seismic wave propagation direction, the shapes and magnitudes of the two curves in the P wave velocity curve (Fig. 6a) and the Q^{-1} curve (Fig. 6b) vary obviously. When the incident angle is 30°, the curves of P-wave velocity and Q^{-1} are similar with C_{33} and Q_{33}^{-1} as the seismic wave
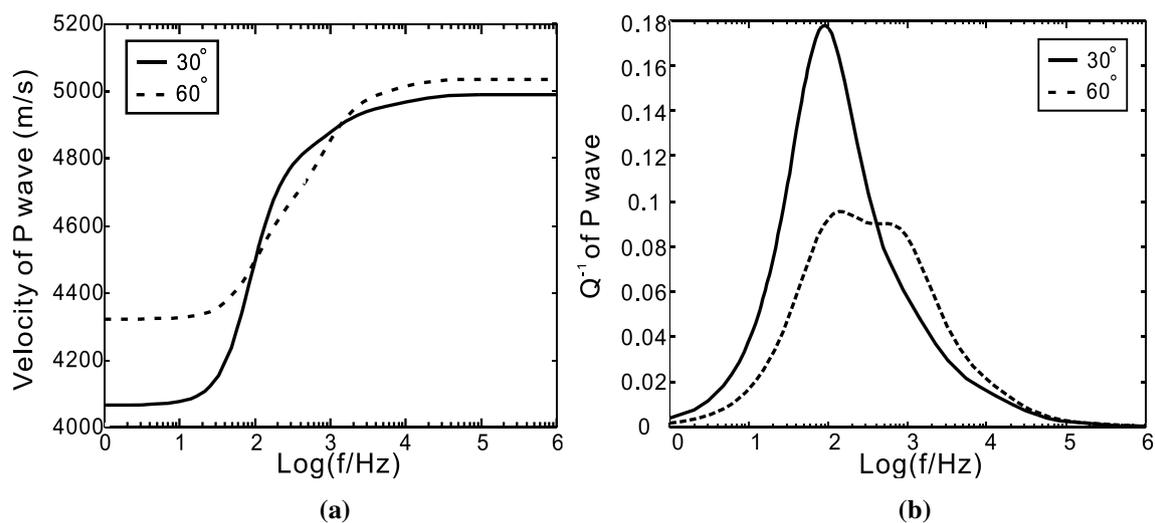
Fig. 6 The velocity (**a**) and Q^{-1} (**b**) variation with frequency for the case of three sets of fractures

is mainly affected by the third set of fractures. When the seismic wave propagates at 60°, the effects of fluid flow concerned with the second and third sets of fractures emerged. The wave propagation direction and fracture orientations only affect the magnitude of velocity and attenuation due to different compression strength of seismic wave on fractures.

Discussions

We try to express and extend Chapman model from the perspective of viscoelasticity. The behaviors of dissipation factor (inverse quality factor, Q^{-1}) versus frequency are often described by viscoelastic models. The relaxation function which is used to represent the viscoelasticity can be obtained by measuring the stress after imposing a rapidly constant unit strain in a relaxed sample of the medium. The SLS model is series combination of a spring and a Kelvin–Voigt model (Fig. 7). The stress (σ)–strain (ε) relation of SLS model is described as

$$\sigma + \tau_\sigma \partial_t \sigma = M_R \left(\varepsilon + \tau_\varepsilon \partial_t \varepsilon \right), \tag{11}$$

where τ_σ and τ_ε are the relaxation times of stress and strain. M_R is the relaxed modulus. The τ_σ, τ_ε and M_R are decided by the properties of the elements. The complex modulus can be obtained by performing a Fourier transform of Eq. 11 and is a function of the relaxation times. However, the relaxation times of stress and strain are hard to measure in field application. The relation between the crucial parameter τ_m in Chapman model and τ_σ and τ_ε is $\tau_m = \sqrt{\tau_\sigma \tau_\varepsilon}$. And fortunately, τ_m can be measured in laboratory experiments, so we can use SLS model to simulate Chapman model using Eq. 3.

When several sets of fractures are embedded in the background medium, SLS model will be extended to GSLS
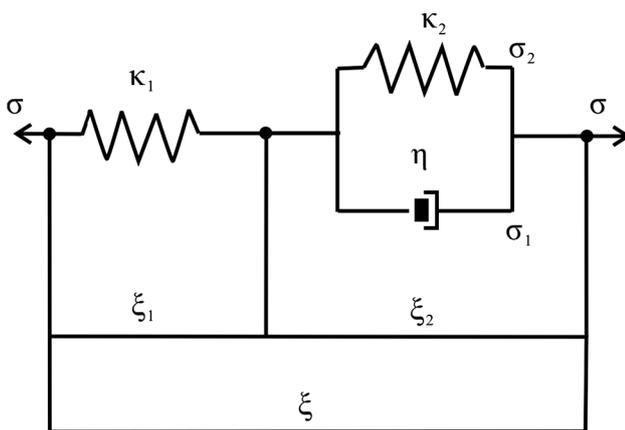
Fig. 7 The mechanical model for a Zener material: κ_1 and κ_2 means the stiffness coefficients of the springs; η means the viscosity coefficient

model. The magnitude of the seismic dispersion and attenuation are controlled by the fracture densities, fracture orientations, and seismic wave propagation direction besides the properties of background medium. We find that the modulus defect can denote the contribution of each set of fractures. Therefore, we adopt Eqs. 5 and 6 to represent the dynamic modulus.

Another advantage of the method is that the stiffness tensors at low- and high-frequency limits require minimum assumptions about pore and fractures geometries and are relatively easy. In comparison with the procedure by Galvin and Gurevich (2015), we propose a simpler method to estimate the stiffness tensors at low- and high-frequency limits. We do not use the linear slip model, since the excess compliance tensor due to fractures is hard to estimate. Baird et al. (2013) proposed the extraction of excess compliance tensor of the fractures using linear slip model in Chapman's model, but we find that the pores correction and fracture correction are coupled together when saturated with fluid. Thus, we cannot directly extract the frequency-dependent excess compliance tensor from the effective compliance tensor. Generally, the dynamic modulus should be consistent with Gassmann–Wood result at low frequencies and be consistent with Gassmann–Hill result at high frequencies. Due to the features of Chapman model, we use Chapman's model and anisotropic Gassmann function to calculate stiffness tensor at the low-frequency limit and use Chapman model only to calculate the stiffness tensor at high-frequency limit. Although the linear superposition theory is adopted in dealing with effects of multiple fracture sets, the strategy can also represent the interactions between seismic waves and fluids in fractures and pores.

The numerical test verifies the practicability and rationality of our method. The method is effective for fractured porous medium with different radii, different orientations and even different seal conditions. In Chapman model, microcracks and pores are interconnected and connected to at most single fracture. Fractures are not connected with other fractures. Therefore, a deficiency of our method is that the influences of fracture connection are not taken into consideration. From the results of Rubino et al. (2013, 2015), the fracture intersection will induce fluid flow between fractures which corresponds to another characteristic frequency. This characteristic frequency is related with the intersection area and intersection angle, so it is more complicated. We just investigate the model of three sets of fractures as an example. If more sets of fractures with different lengths are added into the background media, the Q^{-1} will be a nearly constant value (constant Q model) in a relative wide frequency band, which is convenient and practical in oil prospecting. Correspondingly, the P-wave velocity presents a linear trend with the frequency. There are also some other restrictions in Chapman's model, such as the fractures are

penny-shaped, the model is adaptive to the media with low porosity and so on. Moreover, the characteristic frequency is hard to estimate, which causes some uncertainty of the results when the method is applied to the field rock. The model we build is fully saturated with on kind of fluids. However, partial saturation is common in real reservoirs and partial saturation is also one of the major sources of seismic wave attenuation as well as modulus defect and velocity dispersion. Zhang et al. (2018) make an extension of the original Chapman model so that it can adapt to the case where the fluids in pores and fractures are different. We will expand the multiple fracture model in this frame in the future.

Conclusions

Multiple sets of fractures are often developed in reservoirs. The paper proposes a viscoelastic representation method to calculate the frequency-dependent stiffness tensor of the porous rocks with multiple fracture sets. We use the modified GSLS model to approximate the Chapman mesoscopic fracture model. The effective stiffness coefficients can be described by the following key parameters: low-frequency limiting stiffness coefficients, high-frequency limiting stiffness coefficients and characteristic frequencies corresponding to different sets of fractures. In addition, the modulus defects of the stiffness coefficients due to individual fracture sets are used to denote the different contributions of the fractures. The methodology should be useful for the simulation of wave propagation and quantitative prediction for reservoirs with complex fracture patterns.

Code availability The codes concerned in the manuscript are available.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of data and material The data concerned in the manuscript are available.

References

Ali A, Jakobsen M (2011) Seismic characterization of reservoirs with multiple fracture sets using velocity and attenuation anisotropy data. J Appl Geophys 75(3):590–602. https://doi.org/10.1016/j.jappgeo.2011.09.003

Baird AF, Kendall JM, Angus DA (2013) Frequency-dependent seismic anisotropy due to fractures: fluid flow versus scattering. Geophysics 78(2):WA111–WA122. https://doi.org/10.1190/geo2012-0288.1

Biot MA (1956a) Theory of propagation of elastic waves in fluid-saturated porous solid. I. Low-frequency range. J Acoust Soc Am 28(2):168–178. https://doi.org/10.1121/1.1908239

Biot MA (1956b) Theory of propagation of elastic waves in a fluid-saturated porous solid. II. Higher frequency range. J Acoust Soc Am 28(2):179–191. https://doi.org/10.1121/1.1908241

Brajanovski M, Gurevich B, Schoenberg M (2005) A model for P-wave attenuation and dispersion in a porous medium permeated by aligned fractures. Geophys J Int 163(1):372–384. https://doi.org/10.1111/j.1365-246X.2006.03068.x

Brajanovski M, Müller TM, Parra JO (2010) A model for strong attenuation and dispersion of seismic p-waves in a partially saturated fractured reservoir. Sci Chin Phys Mech Astron 53(8):1383–1387. https://doi.org/10.1007/s11433-010-3205-0

Carcione JM, Gurevich B (2011) Differential form and numerical implementation of Biot's poroelasticity equations with squirt dissipation. Geophysics 76(6):N55–N64. https://doi.org/10.1190/geo2010-0169.1

Chapman M (2009) Modeling the effect of multiple sets of mesoscale fractures in porous rock on frequency-dependent anisotropy. Geophysics 74(6):D97. https://doi.org/10.1190/1.3204779

Chapman M (2003) Frequency-dependent anisotropy due to meso-scale fractures in the presence of equant porosity. Geophys Prospect 51(5):369–379. https://doi.org/10.1046/j.1365-2478.2003.00384.x

Chapman M, Zatsepin SV, Crampin S (2002) Derivation of a microstructural poroelastic model. Geophys J Roy Astron Soc 151(2):427–451. https://doi.org/10.1046/j.1365-246X.2002.01769.x

Dvorkin J, Mavko G (2006) Modeling attenuation in reservoir and nonreservoir rock. Lead Edge 25(2):194–197. https://doi.org/10.1190/1.2172312

Dvorkin J, Mavko G, Nur A (1995) Squirt flow in fully saturated rocks. Geophysics 60(1):97–107. https://doi.org/10.1190/1.1443767

Galvin R, Gurevich B (2015) Frequency-dependent anisotropy of porous rocks with aligned fractures. Geophys Prospect 63(1):141–150. https://doi.org/10.1111/1365-2478.12177

Gassmann F (1951) Uber die elastizitat poroser Medien. Vier der Natur Gesellschaft in Zurich 96:1–23

Gurevich B, Brajanovski M, Galvin R et al (2009) P-wave dispersion and attenuation in fractured and porous reservoirs-poroelasticity approach. Geophys Prospect 57(2):225–237. https://doi.org/10.1111/j.1365-2478.2009.00785.x

Jakobsen M (2004) The interacting inclusion model of wave-induced fluid flow. Geophys J Int 158(3):1168–1176. https://doi.org/10.1111/j.1365-246X.2004.02360.x

Johnson DL (2001) Theory of frequency dependent acoustics in patchy-saturated porous media. J Acoust Soc Am 110(2):682–694. https://doi.org/10.1121/1.1381021

Lan H (2014) Wave field modelling in fractured porous media and frequency-dependent AVO reservoir parameters inversion. Jinlin University, Jinlin

Maultzsch S, Chapman M, Liu E, Li XY (2003) Modelling frequency-dependent seismic anisotropy in fluid-saturated rock with aligned fractures: implication of fracture size estimation from anisotropic measurements. Geophys Prospect 51(5):381–392. https://doi.org/10.1046/j.1365-2478.2003.00386.x

Mavko G, Nur A (1975) Melt squirt in the asthenosphere. J Geophys Res 80(11):1444–1448. https://doi.org/10.1029/JB080i011p01444

Müller TM, Gurevich B, Lebedev M (2010) Seismic wave attenuation and dispersion resulting from wave-induced flow in porous rocks - a review. Geophysics 75(5):75A147–75A164. https://doi.org/10.1190/1.3463417

Picotti S, Carcione JM, Rubino JG et al (2010) A viscoelastic representation of wave attenuation in porous media. Comput Geosci 36(1):44–53. https://doi.org/10.1016/j.cageo.2009.07.003

Picotti S, Carcione JM (2017) Numerical simulation of wave-induced fluid flow seismic attenuation based on the Cole-Cole model. J Acoust Soc Am 142(1):134–145. https://doi.org/10.1121/1.4990965

Pride SR, Berryman JG (2003a) Linear dynamics of double-porosity dual-permeability materials. I. Governing equations and acoustic attenuation. Phys Rev E 68(3):036603. https://doi.org/10.1103/PhysRevE.68.036603

Pride SR, Berryman JG (2003b) Linear dynamics of double-porosity dual-permeability materials. II. Fluid transport equations. Phys Rev E 68(3):036604. https://doi.org/10.1103/PhysRevE.68.036604

Pride SR, Berryman JG, Harris JM (2004) Seismic attenuation due to wave-induced flow. J Geophys Res Solid Earth. https://doi.org/10.1029/2003JB002639

Rubino JG, Caspari E, Milani M et al (2015) Seismic anisotropy in fractured low-permeability formations: the effects of hydraulic connectivity. Seg Tech Progr Expand. https://doi.org/10.1190/segam2015-5844460.1

Rubino JG, Guarracino L, Müller TM et al (2013) Do seismic waves sense fracture connectivity? Geophys Res Lett 40(4):692–696. https://doi.org/10.1002/grl.50127

Shen B, Siren T, Rinne M (2015) Modelling fracture propagation in anisotropic rock mass. Rock Mech Rock Eng 48(3):1067–1081. https://doi.org/10.1007/s00603-014-0621-x

Shi PD, Yuan SY, Wang TY et al (2018) Fracture identification in a tight sandstone reservoir: a seismic anisotropy and automatic multisensitive attribute fusion framework. IEEE Geosci Remote Sens Lett 15(10):1525–1529. https://doi.org/10.1109/LGRS.2018.2853631

Shuai D, Wei JX, Di BR et al (2017) Experimental study of fracture size effect on elastic-wave velocity dispersion and anisotropy. Geophysics 83(1):1–47. https://doi.org/10.1190/geo2016-0639.1

White JE (1975) Computed seismic speeds and attenuation in rocks with partial gas saturation. Geophysics 40(2):224–232. https://doi.org/10.1190/1.1440520

Yuan SY, LiuY ZZ et al (2019) Prestack stochastic frequency-dependent velocity inversion with rock-physics constraints and statistical associated hydrocarbon attributes. IEEE Geosci Remote Sens Lett 16:140–144. https://doi.org/10.1109/LGRS.2018.2868831

Zhang J, Huang H, Wu C et al (2018) Influence of patchy saturation on seismic dispersion and attenuation in fractured porous media. Geophys J Int 214:583–595. https://doi.org/10.1093/gji/ggy160

Zhu Y, Tsvankin I (2006) Plane-wave propagation in attenuative transversely isotropic media. Geophysics 71(2):T17–T30. https://doi.org/10.1190/1.2187792

Seismic attributes via robust and high-resolution seismic complex trace analysis

Mohsen Kazemnia Kakhki[1,4] · Kamal Aghazade[3] · Webe João Mansur[1,2] · Franciane Conceição Peters[1,2]

Abstract

Seismic attribute analysis has been a useful tool for interpretation objectives; therefore, high-resolution images of them are of particular concern. The calculation of these attributes by conventional methods is susceptible to noise, and the conventional filtering supposed to lessen the noise causes the loss of the spectral bandwidth. The challenge of having a high-resolution and robust signal processing tool motivated us to propose a sparse time–frequency decomposition which is stabilised for random noise. The procedure initiates by using sparsity-based, adaptive S-transform to regularise abrupt variations in the frequency content of the non-stationary signals. An adaptive filter is then applied to the previously sparsified time–frequency spectrum. The proposed zero adaptive filter enhances the high-amplitude frequency components while suppressing the lower ones. The performance of the proposed method is compared to the sparse S-transform and the robust window Hilbert transform in the estimation of instantaneous attributes through studying synthetic and real data sets. Seismic attributes estimated by the proposed method are superior to the conventional ones, in terms of robustness and high-resolution imaging. The proposed approach has a detailed application in the interpretation and classification of geological structures.

Introduction

Data interpretation in signal analysis can be better accomplished if a distinct aspect of the data is accessible. This aim can be achieved by transforming the data from one domain to another. The Fourier transform is one of the common transformations which empower us to survey the average properties of a remarkably vast portion of a trace, although it does not represent local variations. The complex trace was first introduced to seismology by Taner et al. (1979); it

resolved this problem by maintaining the local significance and providing a new perspective.

Traditional seismic interpretation methods are incapable of deciphering subtle geological features; this fact has been investigated by researchers, who have explored various techniques to resolve this challenge. Instantaneous seismic attributes take advantage of the complex trace in order to elongate the definitions of simple harmonic oscillation, and they have been adopted in the interpretation of structural features. Seismic attributes are analysed to determine stratigraphic and geological properties (Taner et al. 1979; Verma et al. 2018); they provide quantitative measures of phase, frequency, and reflector amplitude (e.g. the distribution of reef complexes, which can be explained by the instantaneous phase (Zheng et al. 2007)). Yuan et al. 2019 benefited the instantaneous phase and proposed a 3D geosteering phase attribute to recognise the geometry of subsurface features, such as discontinuities related to structural anomalies. Thin-bed tuning is the other challenging structure which is detectable by instantaneous frequency (Chopra and Marfurt 2005). The reservoir characterisation and limestone formations were delineated via seismic instantaneous amplitude,

✉ Mohsen Kazemnia Kakhki
   kazemnia@coc.ufrj.br

1   Modelling Methods in Engineering and Geophysics
    Laboratory (LAMEMO), COPPE, Federal University of Rio
    de Janeiro, Rio de Janeiro 21941-596, Brazil

2   Department of Civil Engineering, COPPE, Federal
    University of Rio de Janeiro, Rio de Janeiro, Brazil

3   Institute of Geophysics, University of Tehran, Tehran, Iran

4   Ferdowsi University of Mashhad, Mashhad, Iran

frequency and phase by imaging various target units (Farfour et al. 2015). Ali et al. (2019) used the dominant frequency attribute to define the characterisation of a hydrocarbon-bearing reservoir. Verma et al. (2018) inferred the dunal and interdunal deposits in 3D seismic data volume by combining coherence attributes and inverted *P*-impedance. Texture and edge attributes were used by Asjad and Mohamed (2015) to extract a salt dome.

Another significant stratigraphic exploration issue is when porous rocks are bound in a non-porous matrix. Bedi and Toshniwal (2019) estimated the porosity of a reservoir from the seismic attributes that have a reasonable correlation with the porosity properties of a rock (energy, mean, instantaneous amplitude, homogeneity, autocorrelation, cosine phase, contrast, dissimilarity and instantaneous frequency). Takam Takougang et al. (2019) applied coherence and similarity attributes when delineating fault and fractures from reverse time-migrated seismic sections. Liu et al. (2014) used a large variety of seismic attributes to characterise the turbidity channel. The channel boundaries are also detectable using seismic coherence and other edge-sensitive attributes, although their thickness cannot be defined via these attributes. Hence, spectral decomposition, which is sensitive to channel thickness, is used to complement the coherence and edge-sensitive attributes (Anees et al. 2019). Obiadi et al. (2019) applied spectral decomposition, integrated with seismic attributes, to identify the geometry and structural discontinuities of hydrocarbon reservoirs within complex tectonic settings. Cichostępski et al. 2019 have also benefited the spectral analysis in detection of bright spots in the presence of thin beds. Recently, Qi and Wang (2020) have employed local polynomial Fourier transform to tackle the nonlinear variation of the signal properties and obtain instantaneous frequency.

Although complex seismic attributes are applicable when defining complex structures, they are problematic in noisy data due to their sensitivity to noise. To alleviate this defect, Luo et al. (2003) presented a generalised version of the Hilbert transform (HT). Liu and Marfurt (2007) outlined the efficiency of time–frequency representation (TFR) in achieving cleaner instantaneous frequency in thin-bed and channel detection. Lu and Zhang (2013) introduced the windowed Hilbert transform (WHT), a TFR form of HT accompanied by a zero-phase adaptive filter to enhance instantaneous complex attributes. Despite the efficiency of filtering in order to remove the undesired frequency components in complex trace analysis, loss of the original data is the primary concern. Concerning this fact, Sattari (2017) proposed a fast sparse S-transform (SST) to achieve sparse WHT by applying the optimised windows in the frequency domain. Although the resolution of the seismic attributes improved via the SST, the presence of random noise remains unsolved. Therefore, in

order to achieve stable and high-resolution instantaneous spectral attributes, the SST is improved by the robust adaptive WHT (RAWHT) to achieve robust sparse S-transform (RSST), which concerns the abrupt changes in the frequency content of the signal and is less sensitive to the noise. The role of the adaptive filter is to suppress the lower-amplitude frequency components and improve the higher amplitudes.

In this study, a modified calculation of analytic signals is presented to provide a robust Hilbert transform which is of higher resolution, less sensitive to noise and provide a better estimation of instantaneous attributes rather than traditional HT. The main aim of the proposed method is to use sparsity-based window-parameter optimisation to improve the resolution of the seismic attributes while taking advantage of a zero-phase adaptive filter to stabilise the contaminated data when calculating the analytic signal in the time–frequency domain. We begin with the explanation of complex trace analysis followed by the calculation of instantaneous attributes. The analytic signal is then improved by using SST and a zero-phase adaptive filter and ending with synthetic and real examples to compare the performance of the proposed method to the SST and the RWHT.

Methodology

Calculation of the complex trace

If we assume that the real signal $x(t)$ is $x(t) = A(t)\cos\theta(t)$ and the imaginary part is $y(t) = A(t)\sin\theta(t)$, the complex trace $z(t)$, or the analytic signal, is computed as:

$$z(t) = x(t) + iy(t) = Ae^{j\theta(t)} \tag{1}$$

where $y(t)$ is the HT of the input signal $x(t)$ derived from the convolution of $x(t)$ with the function $-(1/\pi t)$. Hilbert transform is considered as a linear, time-invariant system with impulse response,

$$h(t) = -\left(\frac{1}{\pi t}\right) \tag{2}$$

Signal $x(t)$ can be analysed by applying the HT in the frequency domain. Considering $X(w)$ as the Fourier spectrum of $x(t)$ then $Z(w)$, the spectrum of the analytic signal, is calculated as follows:

$$Z(\omega) = X(\omega)[1 + iH(\omega)] = \begin{cases} 2X(\omega) & \omega > 0 \\ X(\omega) & \omega = 0 \\ 0 & \omega < 0 \end{cases} \tag{3}$$

where H(ω) is the filter as:

$$H(\omega) = \begin{cases} -i & \omega > 0 \\ 0 & \omega = 0 \\ i & \omega < 0 \end{cases} \qquad (4)$$

The amplitude spectrum of the complex trace $z(t)$ is double for positive frequencies, while it is zero for negative ones. Therefore, the complex trace can be formed by taking the Fourier transform of the real trace, making the amplitude of negative frequencies zero and doubling the amplitude of positive frequencies, then applying the inverse Fourier transform. After this, the instantaneous frequency and phase are easily achievable via the analytic signal.

Seismic instantaneous attributes (Taner et al. 1979) can be derived from the analytic signal. $A(t)$ and $\theta(t)$ denote the instantaneous amplitude and the instantaneous phase in Eq. (1), respectively. The instantaneous frequency can be obtained by taking the derivative of the instantaneous phase

$$\varphi(t) = \frac{d\theta(t)}{dt} \qquad (5)$$

The real seismic signal can have abrupt changes and interference (both in terms of time and frequency) because it carries information about the heterogeneous subsurface. Therefore, it can affect the obtained analytic signal, especially in terms of resolution. Sattari (2017) attempted to address this problem via optimised windows and proposed sparse ST.

Analytic signal using sparse ST

The main difference between the standard ST proposed by Stockwell et al. (1996) and the adaptive SST proposed by Sattari (2017) is that the latter uses frequency-dependent window parameters that are reversely proportional to the amplitudes of various frequency components, while the former uses a window length that is inversely proportional to frequency, while windowing the frequency domain input signal. The strategy used to obtain the adaptive SST relies on the fact that frequency components with higher amplitude are forced to dominate the time–frequency lattice by being localised through translation using high and short windows. The low-amplitude harmonics need to be smeared in the time–frequency domain by using low and wide windows, while being translated. Sattari (2017) applied this strategy by exploiting the matrix formula of ST as

$$TFR_{ST}^{[l]}[l, k] = IFT\left\{ A_{s[l]}[l, m] \odot \hat{X}[l, m] \right\}, \qquad (6)$$

in which $A_{s[l]}[l, m] \in \mathbb{R}^{N \times N}$ is the desired adaptive windows matrix. The columns of this matrix are constructed by the shifted random windows and the rows via the standard Gaussians window function, moved alongside the frequency axis employing $m = 0, ..., N$-1. Moreover, $1 \leq k \leq N$

and $1 \leq l \leq N$ are the time and frequency indices, respectively, and $s[l]$ denotes the standard deviation of the window function. $\hat{X}[l, m] \in C^{N \times N}$ stands for the frequency domain of the input signal $\hat{x}[l]$ with N elements repeated m times. Furthermore, operators $\odot$ and IFT represent the Hadamard product of the 2 matrices and the inverse Fourier transformation, respectively.

Taking advantage of the valuable information included in the input signal amplitude spectrum to distinguish between high- and low-amplitude frequency components according to their known positions (and by changing the optimisation direction from frequency to frequency shift) enabled us to use the amplitude spectrum to create the above-mentioned sparsity under the matrix formulation. According to the linear program provided by Sattari (2017), the change in the optimisation direction is performed by a simple transpose in the algorithm of the ST. This modification results in the optimised standard voice Gaussians along frequency shift (rows of $A_{s[l]}[l, m]$ where smooth and differentiable random windows are automatically attained along frequency (columns of $A_{s[l]}[l, m]$)). The Gaussians' window-length $s[l]$ is supposed to be a curve reversely proportional to the smoothed amplitude spectrum $\left| \hat{x}_{\varsigma u}[l] \right|$ with the scale between zero and one (Eq. 7).

$$s[l] = \frac{L}{2} * \frac{1}{\left\{ r * \left| \hat{x}_{\varsigma u}[l] \right| \right\} + 1} \qquad (7)$$

where L is the length of the input signal and $\left| \hat{x}_{\varsigma u}[l] \right|$ is its smoothed and scaled spectrum amplitude. The range of the curve is varied from 1 to $r + 1$ using $\left\{ r * \left| \hat{x}_{\varsigma u}[l] \right| \right\} + 1$; thus, the reciprocal term is in the range $1/(r + 1)$ and 1. The range of $s[l]$ is controlled by the parameter r, which is readily adjustable owing to its linear behaviour. For example, according to their bandwidth and spectral diversity, we can select the r equals to 1 or 2 for band-limited signals, 3 or 4 for seismic signals and 5–10 for wide-band signals.

The arbitrary windows of SST follow the criteria of the partition of unity (Lamoureux et al. 2003), which means that the superposition of them on the columns of $A_{s[l]}[l, m]$ onto the frequency axis adds up to 1. This characteristic makes the adaptive SST invertible similar to the conventional ST by a simple projection of the SST map onto the frequency (Eq. 8).

$$x[k] = \sum_{l=1}^{N} TFR_{ST}^{[l]}[l, k] \qquad (8)$$

Figure 1 compares the performance of the adaptive SST and the ST as well as the adaptive arbitrary windows applied to a non-stationary signal.
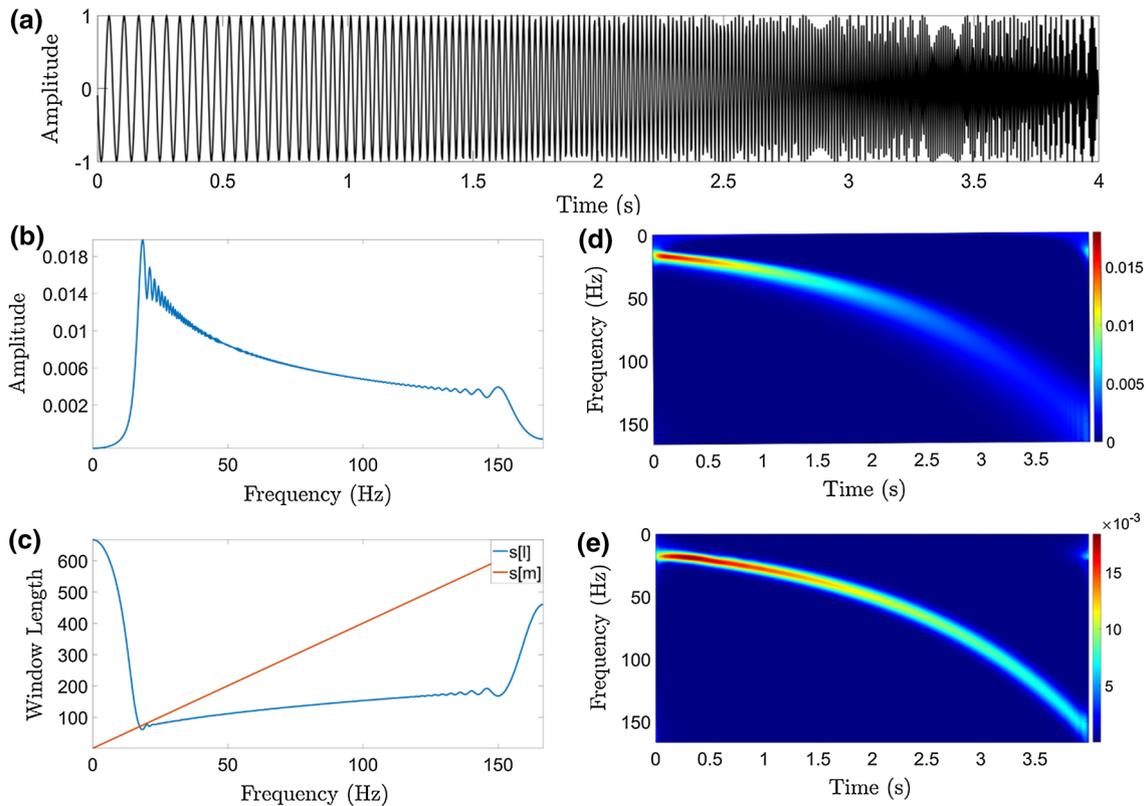
Fig. 1 **a** A non-stationary logarithmic chirp signal with sinusoid variations in amplitude along the time axis, **b** the correspondent amplitude spectrum, **c** the window length variation with frequency for standard ST (red) and adaptive sparse ST (blue). Time–frequency map of the non-stationary signal computed by **d** the conventional ST and **e** adaptive sparse

As a result, the adaptive SST is not only superior to the standard ST in terms of adaptivity and higher resolution, but also it is very efficient in that it adds no extra computation to the translation and modulation processes required for the spectral decomposition. This means it even performs better than the alternative energy concentration (ECM) methods used for adaptivity enhancement of Fourier-based spectral decomposition (Sattari et al. 2013). These methods are computationally heavy as they require computation of several time–frequency decompositions with different window-lengths, among which the sparsest result is searched for, while in the adaptive SST, the window parameters are optimally set to create sparsity. This makes the ECM methods impractical for real-world applications. In addition to the complexity, Sattari (2017) also showed the superiority of the adaptive SST over the standard ST and STFT optimised by ECM methods, in terms of robustness to noise, temporal and spectral interference resolution and the fact that it has only one free parameter to set (which is linear and well-behaved). However, under the low SNR, the SST is not stable. For these reasons, in this paper the TFR obtained via adaptive SST is filtered in the time–frequency domain.

We used the SST method proposed by Sattari (2017) to calculate the analytic part of a signal to have higher resolution compared to the other known methods. The windowed HT can be defined in the time–frequency domain as:

$$Z(\omega.\tau) = X(\omega.\tau)[1 + iH(\omega)] = \begin{cases} 2X(\omega.\tau) & \omega > 0 \\ X(\omega.\tau) & \omega = 0 \\ 0 & \omega < 0 \end{cases} \qquad (9)$$

Improved Hilbert transform

To resolve the problem of noise in the signal, we employ a time–frequency adaptive filter to the TFR obtained via adaptive SST. This filter is based on the assumption that the higher-amplitude spectrum has more signal content and is formed as

$$g(\omega.\tau) = \frac{|X(\omega.\tau)|^{N-1}}{\arg\max\limits_{\omega} \left(|X(\omega.\tau)|^{N-1} \right)}, \qquad (10)$$

where N is a weighting factor and $N \geq 1$, $|X(\omega.\tau)|$ is the amplitude spectrum of $X(\omega.\tau)$. The dominator in Eq. 10 is actually an optimisation subproblem to find the maximum value of $|X(\omega.\tau)|^{N-1}$ for each frequency, ω. The term argmax denotes this process.

The analytic signal can be constructed as:

$$z(t) = \int_{-\infty}^{\infty} F^{-1}\big(\tilde{Z}(\omega.\tau)\big) d\tau, \tag{11}$$

where

$$\tilde{Z}(\omega.\tau) = X(\omega.\tau)g(\omega.\tau)[1 + iH(\omega)], \tag{12}$$

and $F^{-1}\big(\tilde{Z}(\omega.\tau)\big)$ is the inverse Fourier transform of $Z(\omega.\tau)$.

Increases in the value of N, result in amplification of the frequencies with the maximum amplitude. The value of N depends on the signal-to-noise ratio (SNR); the higher the SNR, the lower N. By applying N greater than one, the SST develops into the RSST with enhanced higher-amplitude frequency components and suppressed lower ones. Although the SST is supposed to render less noisy results, it fails to suppress the noise when the SNR is low. On the other hand, the adaptive filter proposed by Lu and Zhang (2013) cannot distinguish the discrepancy between the desired signal and undesired noise, if applied directly to the TFR. The weight factor proposed by Lu and Zhang (2013) reduces the noise at the cost of losing the signal and conclusively losing the subsurface information. Therefore, applying a weighting order to the obtained TFR in SST can not only result in high-resolution TFR as SST but can also suppress the noise, as well as RWHT with the difference that the signal maintained. The main purpose of the applied adaptive filter is the enhancement of instantaneous attributes estimation in noisy data, although it has an application in improving the other seismic attributes.

Results

The performance of the proposed method is validated by applying the synthetic and real data set. We compare the RSST in obtaining seismic attributes with the SST and RWHT method to observe the discrepancies in their performance.

Numerical examples

The choice of adaptive Fourier-based time–frequency decomposition is user-dependent, which is relevant to whether the type of analysis focusses on the time or frequency domain and, more important than that, the characteristics of the signal (Radad et al. 2015). Seismic data are

narrow in frequency and wide in time; therefore, adaptive SST can be the best analytical choice, owing to the higher sparsity of the input frequency domain signal. Moreover, decomposition of the sparser version of the input signal can suppress the scattered random noise more efficiently in both the time and frequency domains, although not completely. The aforementioned facts are the reasons for applying SST in the first step to obtain high-resolution spectral attributes. In the following steps, a weight factor is added to the sparse TFR to suppress almost all of the noise available in the TFR; according to our scope of analysis, this results in a robust, high-resolution spectral amplitude.

To diagnose the superiority of the RSST over the SST and RWHT for decomposing narrowband signals, we compare the TFR of 5 non-stationary signals in Fig. 2. The signal, taken from Andrade et al. (2018), is a sum of 5 signals generated by a sampling interval of 0.003 s:

$$\begin{cases} x_1(t) = 0.8\cos(30\pi t) & 0 \leq t \leq 6s \\ x_2(t) = 0.6\cos(70\pi t) & 0 \leq t \leq 6s \\ x_3(t) = 0.7\cos(130\pi t + 5\sin(2\pi t)) & 4s \leq t \leq 8s \\ x_4(t) = \sin\left(\frac{8\pi 100^{t/8}}{log(100)}\right) & 6s < t \leq 10s \\ x_5(t) = 3e^{-1250(t-2)^2}\cos(710(t-2)) & 0 \leq t \leq 10s \end{cases} \tag{13}$$

where x_1 is a harmonic component of 15 Hz, x_2 is another harmonic with 35 Hz, x_3 is a frequency-modulated harmonic of about 65 Hz, x_4 is a sliding harmonic from 35 to 158 Hz, and x_5 is a Morlet wavelet with a central frequency of 113 Hz. The signal is shown in Fig. 2, along with its corresponding TFR obtained by SST, RWHT and RSST.

As can be seen, the TFRs obtained by RSST give spectra with higher resolution and more stability, in terms of time or frequency or both. In this example, the values of the input parameters are the same. We set the value of N and r to be 2 and 10, respectively. The filtering effect of the RWHT method can be observed in the time–frequency panel, as indicated by arrows. In these regions, the energy of the signal is attenuated, while the result of RSST proves its capability in deriving a high-resolution time–frequency spectrum without losing the effective energy of the signal. This is the result of employing the sparsity properties of SST in the RWHT algorithm. Furthermore, in comparison with SST, the RSST method generates higher resolution.

In Fig. 3, we examined the performance of different approaches for a double-linear chirp signal. In this example, the time–frequency panels of the signal are estimated via 3 methods with a sampling interval of 2 ms and the maximum time of 2 s. For RWHT, the central part of the panel, in which the spectra of 2 chirp signals cross each other, has lost its information (the ellipse in Fig. 3). The amplitude of the 2 signals is distinct due to the filtering effect of the RWHT method (the arrows in Fig. 3). Of the 3 methods we
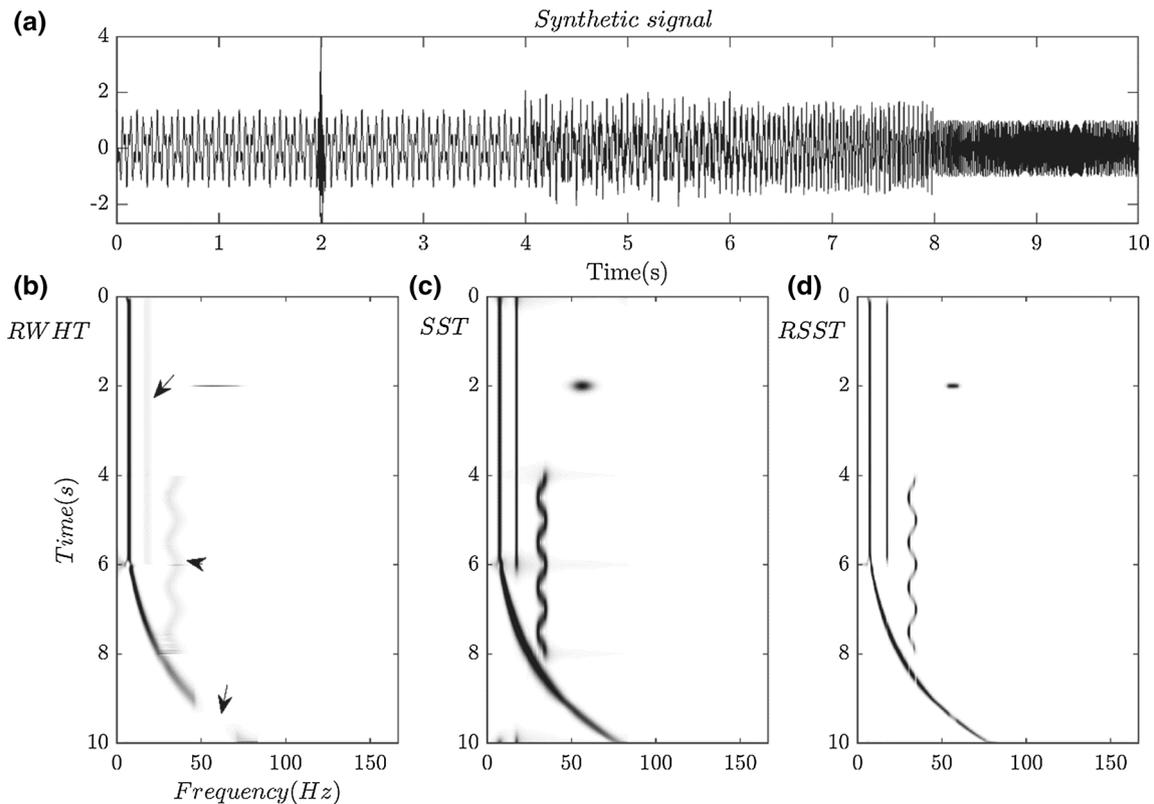
Fig. 2 a Synthetic signal and **b–d** time–frequency panels for methods RWHT, SST and RSST, respectively. The arrows indicate regions suffering from losing valuable information

compared, the resolution of TFR obtained by the RSST is noticeably higher.

The adaptive filter $g(\omega.\tau)$ defined in Eq. (10) tries to maintain signal information and avoid noise leakage into time–frequency panel. In Fig. 4, we added random noise to the chirp signal (Fig. 3) to demonstrate the role of the $g(\omega.\tau)$. It can be seen that unlike the TFR obtained via the SST method, $g(\omega.\tau)$ is similar to the desired TFR of the chirp signal (Fig. 4b). Meanwhile, multiplying the TFR$(\omega.\tau)$ by $g(\omega.\tau)$ causes that the random noise filtering carries out automatically, which results in higher resolution of the final TFR.

Although the SST reversely scales the global trend of the amplitude spectrum to be indifferent to the added noise, it is still susceptible to the random noise since the weight that the windows are applying to the signal is the same for the noise. Contrary to this, the RWHT method is robust in detecting the noise but at the cost of losing the signal. This point is depicted in Fig. 5 for a double-linear chirp signal with additive random noise of 1 dB SNR. The parameters taken in the example are the same as for the three TFRs for smoothing ($r = 10$) in SST and RSST and de-noising ($N = 7$) in RSST and RWHT. As can be seen, the RWHT method tries to de-noise the TFR panel, while most of the

signal's information has been lost in this domain. The resolution of the SST method suffers from the presence of noise which is not the case in the RSST method.

The value of the weighting factor (i.e. N in Eq. 10) is a crucial parameter in either RWHT or RSST. There is a trade-off between resolution resulted from applying the proposed filter and the SNR value. This trade-off is controlled by the value of N. In the case of noise-free signals (or signals with high SNR), our goal is to increase the resolution of the time–frequency panel. Overall, $5 \leq N \leq 10$ works well for high SNR values. However, in the presence of noise (i.e. low SNR), the noise filtering requires a high value of N at the cost of losing the resolution. In this regard, selecting the proper value of N is essential and requires setting a balance between resolution and de-noising criteria. Therefore, Fig. 6 shows the effects of this parameter on the results of the RSST method obtained for the noisy signal in Fig. 3. Based on the results, using high values of N results in losing some of the useful information in the signal (in the case of RWHT), while low values of N are incapable of dealing with noise (similar to SST).

The numerical examples show that the RSST outlines a higher resolution and is more robust to the noise when compared to the SST and RWHT. The input signal in this
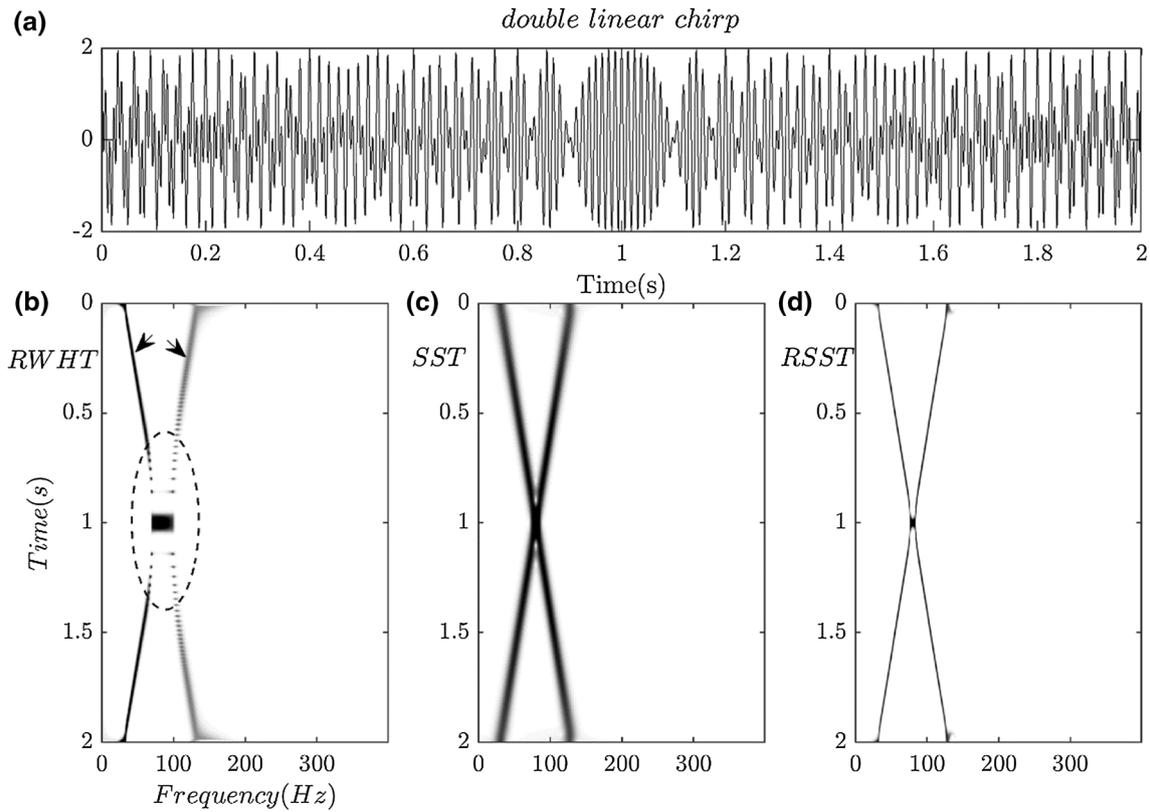
Fig. 3 **a** Double chirp signal and **b–d** its time–frequency panel for methods RWHT, SST, RSST, respectively. The RWHT method has failed in the central part of the panel indicated by an ellipse. Also, 2 arrows demonstrate the effect of filtering the higher frequencies of the signal
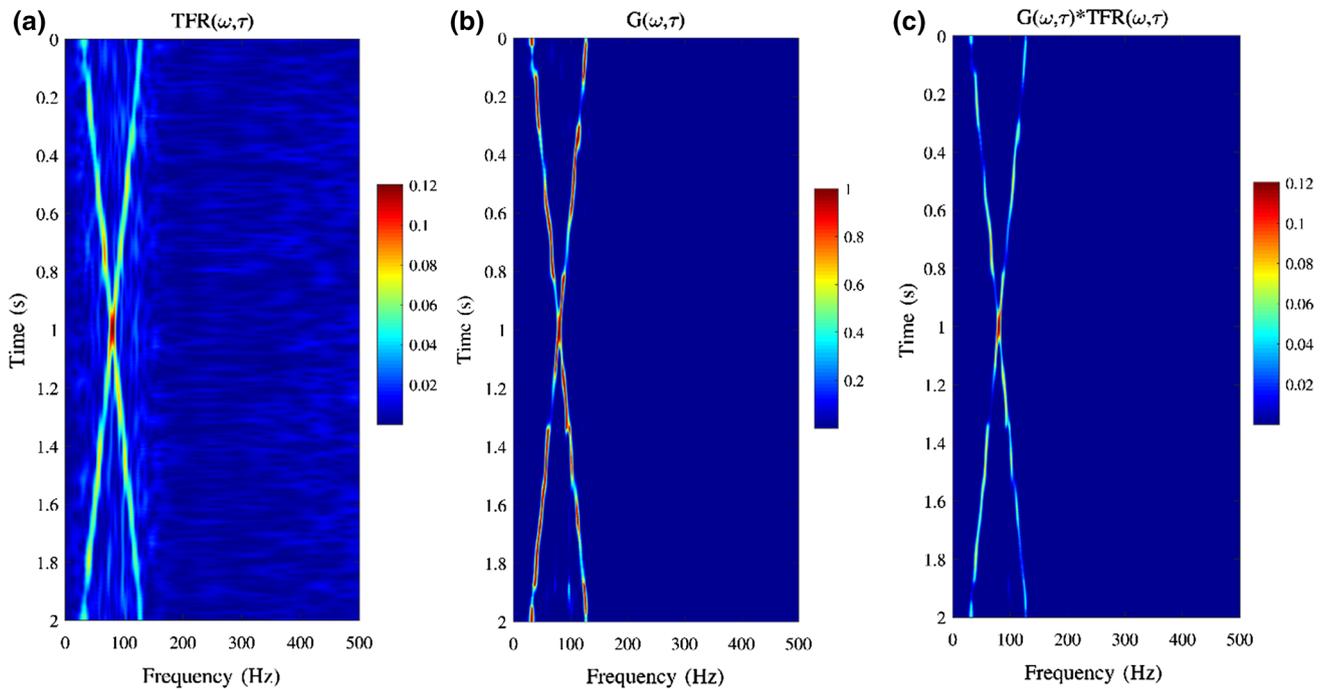
Fig. 4 The role of adaptive filter **a** TFR panel, **b** the adaptive filter, **c** final TFR panel after applying adaptive filter
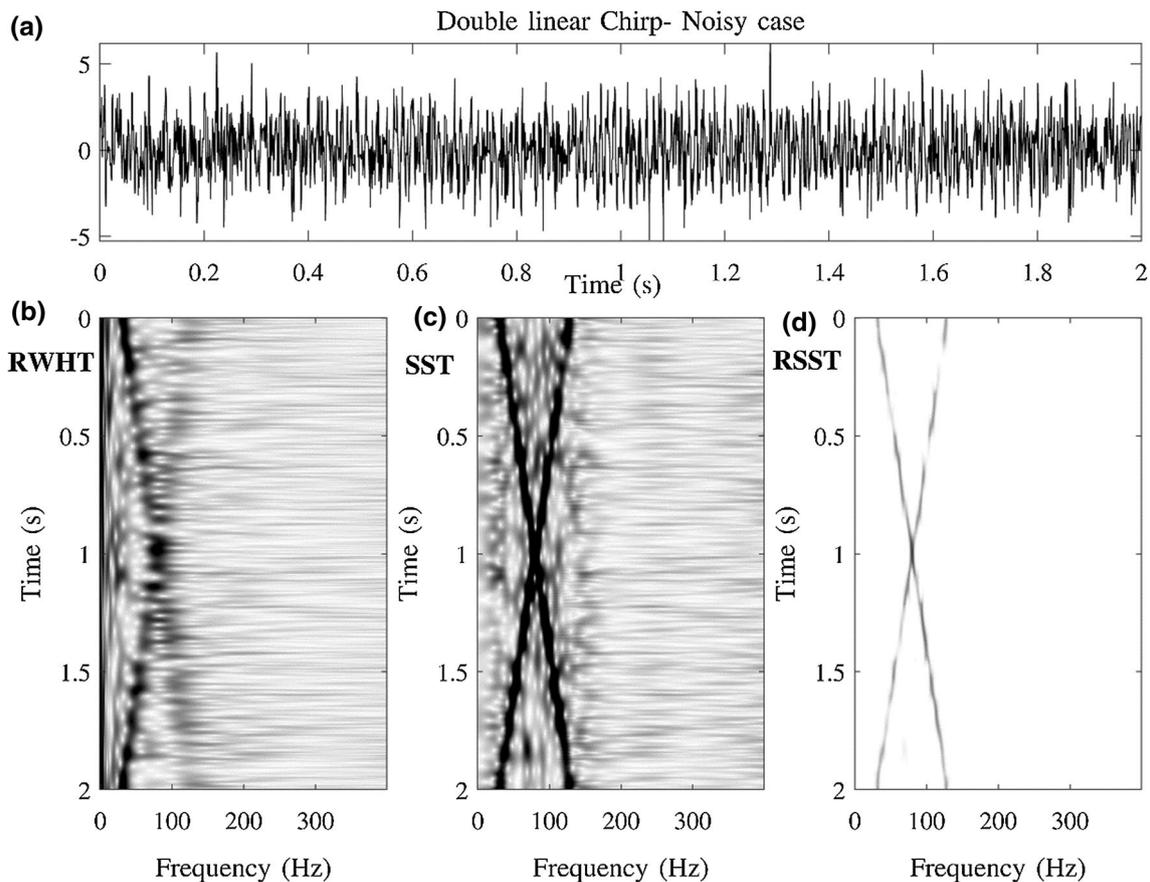
Fig. 5 The effect of noise for the double-linear chirp in Fig. 3. **a** the double-linear chirp with noise and the TRF obtained via **b** the RWHT method, **c** the SST method and **d** the RSST method

method can be regularised in time–frequency, while estimating the analytic signal owing to its low susceptibility to random noise. It is worth mentioning that regularising the true positions and amplitudes of different components in the time–frequency domain results in more accurate instantaneous attributes because the scattered energy in the time–frequency domain can cause fake complex indices.

Wedge model

To evaluate the performance of different methods, a synthetic wedge model is used in the next example (Fig. 7). For this case, we initiated the SNR value at 2 dB. After running the algorithms, the results of TFR for 3 single-frequency components are shown in Fig. 8. It can be seen that the RSST method outperforms the other methods. For low-frequency components (i.e. 5 Hz), the RWHT has failed to maintain the valuable information, while for higher-frequency components, the resulting TFRs are satisfactory. On the other hand, the SST method still suffers from noise. For a better comparison, a part of the TFR panel (the red box) is selected for a frequency of 20 Hz

from all 3 methods, and this shows that the RSST method has dealt with the tuning effect successfully (Fig. 9).

In Fig. 10, the results of instantaneous amplitude and the cosine of phase attributes for the noisy wedge model are shown. As can be seen, the SST method gives higher-resolution results compared to the RWHT method. However, with respect to accuracy, it can be observed that the layers of the noisy model achieved by the RSST are of higher resolution and of lower noise. As can be seen in the instantaneous amplitude, the non-filtered SST ignores the presence of noise, while filtering via RSST smoothens them. The filtering applied in RWHT enhances the peak frequency contribution to the reconstructed analytic signal while causes the loss of the initial spectral bandwidth. Of the 3 methods compared, the instantaneous attribute obtained by the RSST is more robust and efficient in resolving interfered wavelets under the random noise while maintaining the preliminary bandwidth. The cosine of instantaneous phase achieved by the RSST proves the power of proposed method in resolving interfered wavelets with details although the SST represents almost the same result. The resolution in RSST has improved via

Fig. 6 Evaluating the effect of weighting parameter N in the RSST method. High values of N result in losing useful TFR information, while low values fail to deal with noise
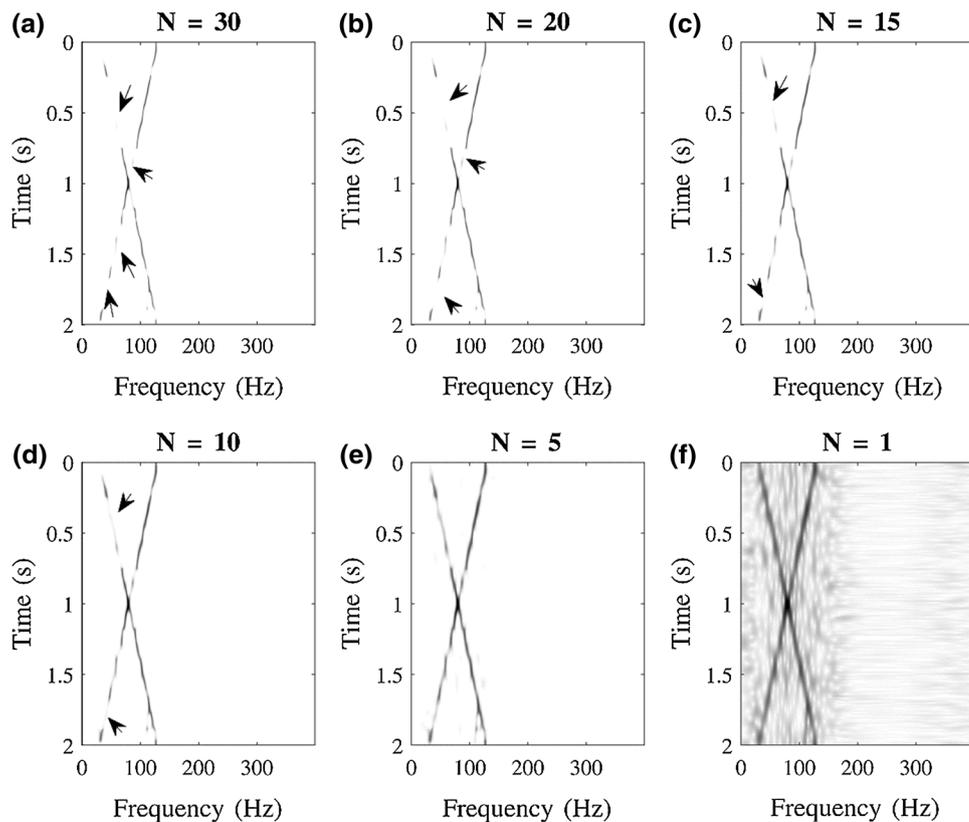
Fig. 7 Seismogram of a synthetic wedge model, **a** without noise and **b** with noise
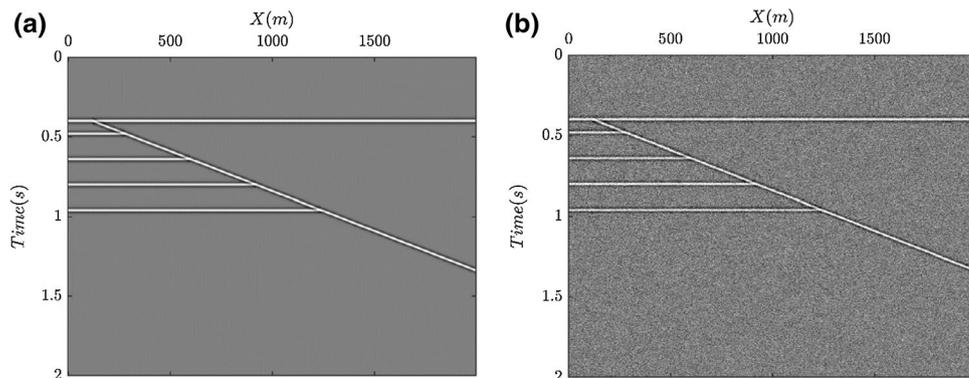
optimising window parameters and rendering the weight order to the signal in each selected window.

Real data

In this section, we consider the performance of 3 methods on both a 1D trace and 2D data set. In the first example, a real trace is chosen with a sampling interval of 2 ms and a recording time of 7 s (Fig. 11a). The related TFRs for different methods are also calculated with the same input parameters ($N = 10$ and $r = 15$).

Although an apparent trend is detectable in the time–frequency spectrum of the RWHT method, it is stretching along

the frequency coordinate. On the other hand, the resolution of TFR obtained by the SST is not satisfactory. The RSST method successfully derived a meaningful trend with a sparse nature in the time–frequency panel.

The next example is an excerpted part of a 2D section with a sampling interval of 4 ms from the F3 block of the North Sea (Fig. 12). A bright spot, which is an anomaly caused by biogenic gas pockets, and gas chimneys in the deeper parts are shown in the data.

For these data, both the SST and RSST methods were run and the resulting instantaneous frequency is shown in Fig. 12. In the deeper parts of the data, which are related to the presence of gas chimneys, the low frequencies are

Fig. 8 The results of the RWHT (**a–c**), SST (**d–f**) and WSST (**g–i**) methods applied on the wedge model in Fig. 7 for the single-frequency components of 5, 20 and 30 Hz. The red rectangles in 20 Hz are used to compare the results
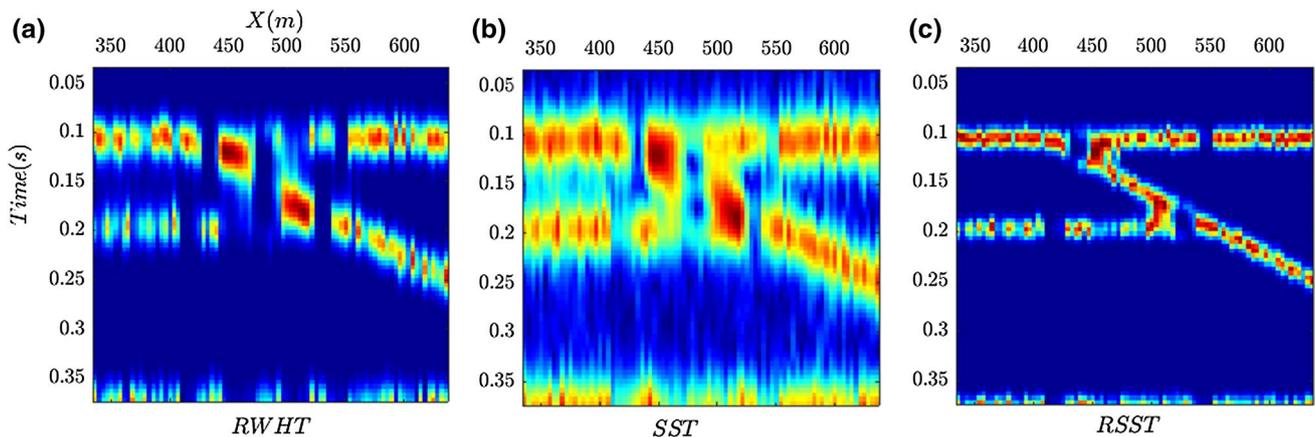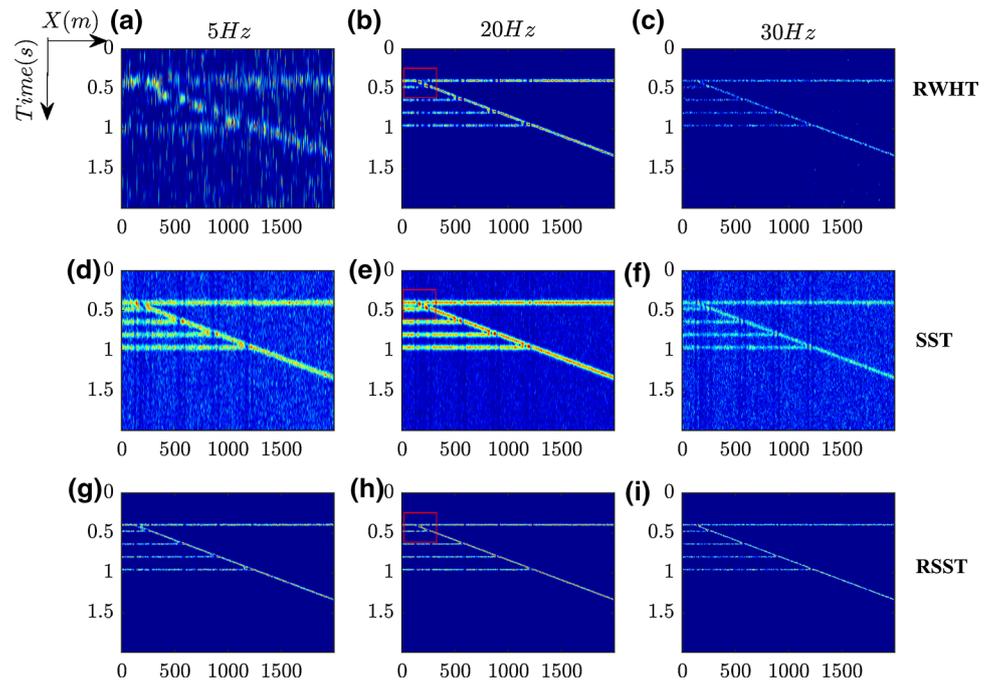
Fig. 9 The zoomed view of the red box in Fig. 8 for a better comparison of the results for **a** the RWHT, **b** the SST and **c** the RSST methods, respectively

visible due to the seismic energy absorption. This abrupt reduction in the instantaneous frequency is more obvious in the results obtained by RSST than the SST counterpart. Furthermore, the bright spot is successfully detectable from instantaneous frequency information derived by the RSST method than the SST.

Conclusion

In this paper, we have looked at the problem of estimating a stable, high-resolution complex trace analysis within the framework of sparsity-based optimisation and

time–frequency spectrum weighting orders. Poor-resolution and noise problems in the time–frequency domain illustrate the necessity for a stronger method to deal with these shortcomings. Therefore, the sparsity-based, adaptive S-transform was proposed as a spectral decomposition tool to enhance the resolution of the time–frequency WHT. The optimised windows satisfied the requirements for regularisation of abrupt frequency changes and were superior to the previous methods, in terms of computational cost and interference removal without leading to fake indices, although they were vulnerable to random noise. Therefore, the proposed spectral decomposition is improved via a zero-phase adaptive filter
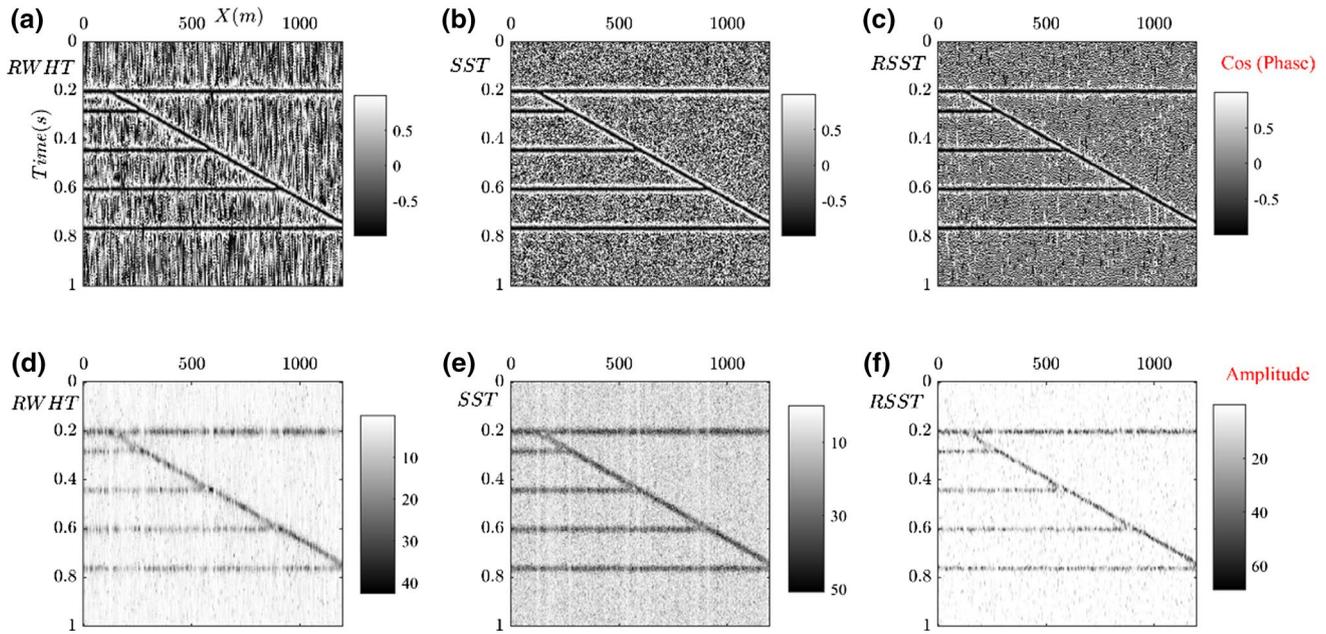
Fig. 10 Complex trace analysis for the noisy wedge model. Cosine of phase attribute (**a–c**) and instantaneous amplitude (**d–f**)

Fig. 11 Time–Frequency analysis of **a** a single trace obtained via **b** the RWHT, **c** the SST and **d** the RSST
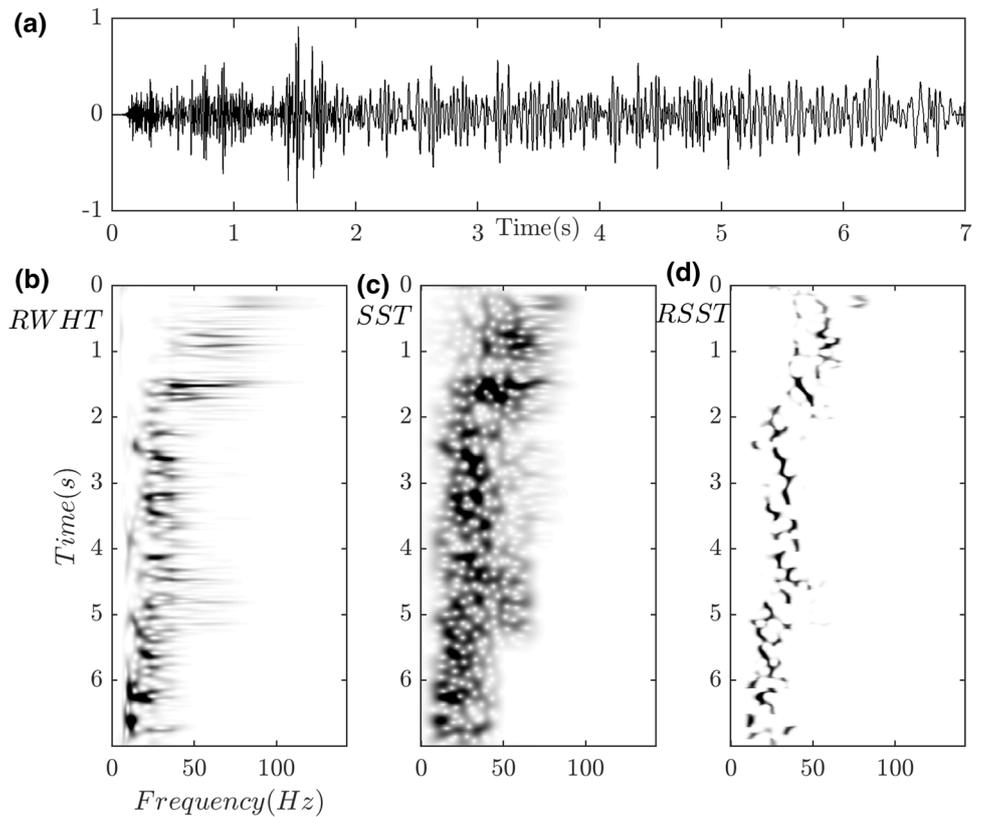
to suppress the residual noise by enhancing the frequency components with larger amplitudes. As for the computational cost, the proposed method was slightly slower than SST because of extra de-noising in TFR; however, it was faster than RWHT due to an additional inverse Fourier transform.

The proposed robust spectral decomposition approach was conclusively used to implement complex trace
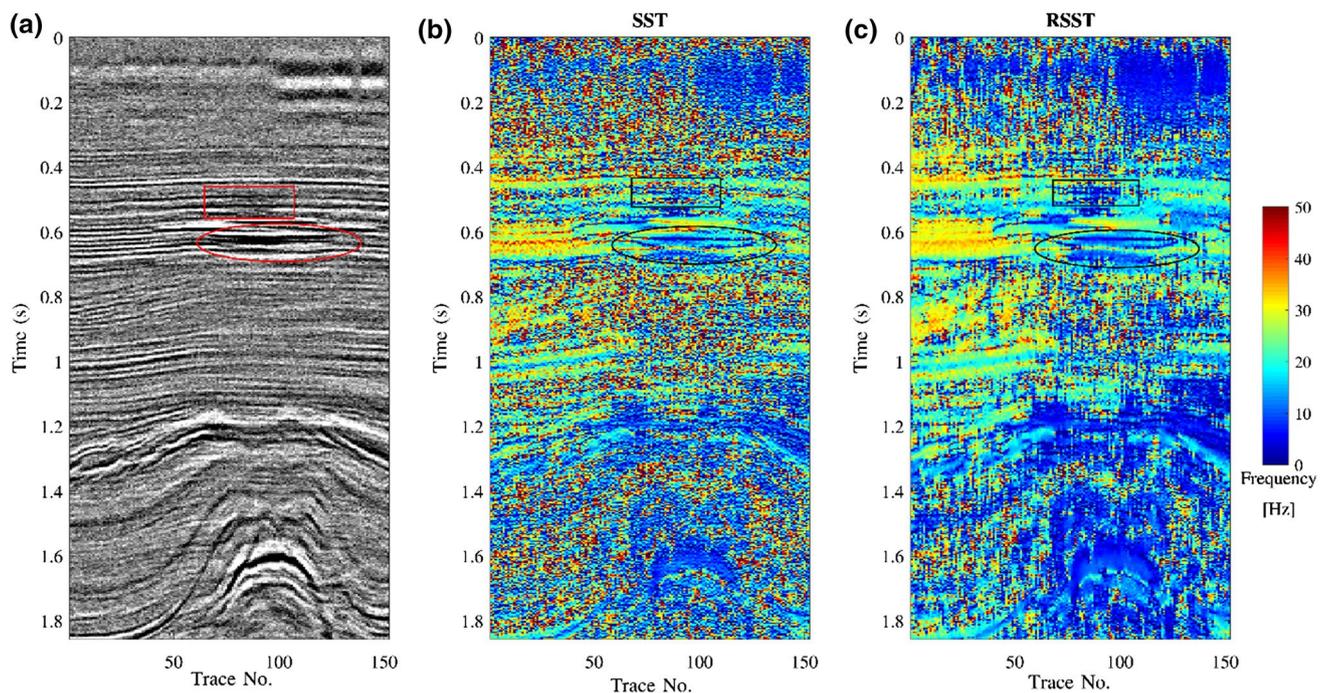
Fig. 12 **a** The real data section **b** instantaneous frequency panel of the real 2D data obtained via **b** the SST and **c** the RSST methods. The ellipses and rectangles are used for the comparison of the results

analysis of synthetic and real data sets. The results proved that the power of robust, adaptive ST in regularising the abrupt frequency changes and suppression of random noise resulted in high-resolution and robust instantaneous attributes, compared to the conventional methods that ignore these changes. Indeed, the proposed method regularised the entire frequency content of the signal by setting only one window parameter and suppressed the noise spread in both the time and frequency domains by adjusting the weighting order N. The proposed method is an adaptive, high-resolution, invertible and frequency-dependent time–frequency decomposition approach that has vast implications for interpreting complex trace analyses. It should be mentioned that the reason for using ST in this procedure (as a spectral decomposition) was relevant to the input data and corresponding applications and other transforms (like STFT) could also be used.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by MKK, KA, WJM and FCP. The first draft of the manuscript was written by MKK, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability Data are available under the reasonable request.

Code availability Code is available under the reasonable request.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ali A, Younas M, Ullah M et al (2019) Characterization of secondary reservoir potential via seismic inversion and attribute analysis: a case study. J Pet Sci Eng. https://doi.org/10.1016/j.petrol.2019.03.039

Andrade MCB, Porsani MJ, Ursin B (2018) Complex autoregressive time-frequency analysis: estimation of time-varying periodic signal components. IEEE Signal Process Mag. https://doi.org/10.1109/MSP.2017.2783942

Anees A, Shi W, Ashraf U, Xu Q (2019) Channel identification using 3D seismic attributes and well logging in lower Shihezi formation of Hangjinqi area, northern Ordos Basin. J Appl Geophys. https://doi.org/10.1016/j.jappgeo.2019.02.015

Asjad A, Mohamed D (2015) A new approach for salt dome detection using a 3D multidirectional edge detector. Appl Geophys. https://doi.org/10.1007/s11770-015-0512-2

Barnes AE (2007) A tutorial on complex seismic trace analysis. Geophys 10(1190/1):2785048

Bedi J, Toshniwal D (2019) PP–NFR: an improved hybrid learning approach for porosity prediction from seismic attributes

using non-linear feature reduction. J Appl Geophys. https://doi.org/10.1016/j.jappgeo.2019.04.015

Berthelot A, Solberg AHS, Gelius LJ (2013) Texture attributes for detection of salt. J Appl Geophys. https://doi.org/10.1016/j.jappgeo.2012.09.006

Chopra S, Marfurt KJ (2005) Seismic attributes A historical perspective. Geophys 70(5):3SO

Cichostępski K, Kwietniak A, Dec J (2019) Verification of bright spots in the presence of thin beds by AVO and spectral analysis in Miocene sediments of Carpathian Foredeep. Acta Geophys. https://doi.org/10.1007/s11600-019-00324-z

Farfour M, Yoon WJ, Kim J (2015) Seismic attributes and acoustic impedance inversion in interpretation of complex hydrocarbon reservoirs. J Appl Geophys. https://doi.org/10.1016/j.jappgeo.2015.01.008

Jones DL, Parks TW (1990) A high resolution data-adaptive time-frequency representation. IEEE Trans Acoust 10(1109/29):61539

Lamoureux MP, Gibson PC, Grossman JP, Margrave GF (2003) A fast, discrete Gabor transform via a partition of unity. J Fourier Anal Appl

Liu H, Lei X, Mao C, Li S (2014) Improving reservoir thickness prediction using seismic attributes and attributes fusion. Acta Geophys. https://doi.org/10.2478/s11600-013-0174-5

Liu J, Marfurt KJ (2007) Instantaneous spectral attributes to detect channels. Geophys 10(1190/1):2428268

Lukai, Zhang , WCK (2013) Robust estimation of instantaneous phase using a time-frequency adaptive filter. Geophys. https://doi.org/10.1190/GEO2011-0435.1

Luo Y, Al Dossary S, Marhoon M, Alfaraj M (2003) Generalized hilbert transform and its applications in geophysics. Lead Edge 10(1190/1):1564522

Obiadi II, Okoye FC, Obiadi CM et al (2019) 3-D structural and seismic attribute analysis for field reservoir development and prospect identification in Fabianski Field, onshore Niger delta, Nigeria. J African Earth Sci. https://doi.org/10.1016/j.jafrearsci.2019.103562

Partyka G, Gridley J, Lopez J (1999) Interpretational applications of spectral decomposition in reservoir characterization. Lead Edge 18(3):353

Qi P, Wang Y (2020) Seismic time–frequency spectrum analysis based on local polynomial Fourier transform. Acta Geophys. https://doi.org/10.1007/s11600-019-00377-0

Radad M, Gholami A, Siahkoohi HR (2015) S-transform with maximum energy concentration: application to non-stationary seismic deconvolution. J Appl Geophys. https://doi.org/10.1016/j.jappgeo.2015.04.010

Sattari H (2017) High-resolution seismic complex trace analysis by adaptive fast sparse S-transform. Geophys. https://doi.org/10.1190/GEO2015-0425.1

Sattari H, Gholami A, Siahkoohi HR (2013) Seismic data analysis by adaptive sparse time-frequency decomposition. Geophys. https://doi.org/10.1190/geo2012-0550.1

Stockwell RG, Mansinha L, Lowe RP (1996) Localization of the complex spectrum: The S transform. IEEE Trans Signal Process 10(1109/78):492555

Takam Takougang EM, Bouzidi Y, Ali MY (2019) Characterization of small faults and fractures in a carbonate reservoir using waveform inversion, reverse time migration, and seismic attributes. J Appl Geophys 161:116–123

Taner MT, Koehler F, Sheriff RE (1979) Complex seismic trace analysis. Geophys 10(1190/1):1440994

Verma S, Bhattacharya S, Lujan B et al (2018) Delineation of early Jurassic aged sand dunes and paleo-wind direction in southwestern Wyoming using seismic attributes, inversion, and petrophysical modeling. J Nat Gas Sci Eng. https://doi.org/10.1016/j.jngse.2018.09.022

Yuan S, Su Y, Wang T et al (2019) Geosteering phase attributes: a new detector for the discontinuities of seismic images. IEEE Geosci Remote Sens Lett. https://doi.org/10.1109/LGRS.2018.2866419

Zheng X, Li Y, Li J, Yu X (2007) Reef and shoal reservoir characterization using paleogeomorpology constrained seismic attribute analysis. In: Society of Exploration Geophysicists - 77th SEG International Exposition and Annual Meeting

Study on transmitted channel wave-based, horizontal multilayer 3-D velocity model inversion and quantitative coalbed thickness detection method

Zean Hu[1,2] · Pingsong Zhang[1,2] · Guangzhong Ji[1,2] · Xiaoyun Su[3]

Abstract

Most methods using transmitted channel wave (TCW) prospecting to quantitatively detect the thickness of coal seams based on the statistic relationship of group velocity in certain wave bands to the thickness of coal seams cannot be applied universally. To establish a universal applicable method, we first obtained the theoretical dispersion curve of TCW using the generalized reflection–transmission coefficient method and the 1-D horizontal multilayer velocity model, performed iteratively match calculation using the inversion model and the genetic algorithm and analyzed the distributive characteristics of shear wave velocity of coal and rock formations at a certain depth. We then obtained the 3-D velocity images of the coal seam working face based on TCW data using the 3-D back-projection technology. According to the changes of shear wave velocity at the coal–rock interface and the rate of inversion velocity change, we further proposed the quantitative discriminant model for coalbed thickness. Based on the model, we quantitatively interpreted the thickness of the coal seam by computing the depths corresponding to the extremes of the positive and negative rate of the shear wave velocity change and obtained the distribution characteristics of the coal thickness in the working surface. To verify the feasibility and validity of the proposed model for coalbed thickness, we conducted a 3-D physical similarity model experiment and subjected the collected two-component TCW data to inversion calculation and compared the obtained coal seam thickness with the known model parameters. Overall, our study achieved the universal 3-D quantitative detection of coalbed thickness and provided technical supports for intelligentized coalbed mining.

Keywords Channel wave dispersion curve · Genetic algorithm · 3-D velocity inversion · Quantitative coalbed thickness detection

Introduction

With the rapid development of science and technology, the intelligentized or unmanned coal mining technology has become the main developmental trend of coal mining toward the dual guarantee for coal production and personnel safety

✉ Pingsong Zhang
   pszhang1971@163.com

1 State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines, Anhui University of Science and Technology, Huainan, China

2 School of Earth and Environment, Anhui University of Science and Technology, Huainan, China

3 Xi'an Research Institute of China Coal Technology & Engineering Group Corp, Xi'an 710077, China

(Yuan 2017; Peng et al. 2019). The accurate and efficient detection of the thickness of coal seams in the coal mining face can directly guarantee the realization of intelligentized coal mining. Currently, two main methods have been used to predict the thickness of coal seams. The first one is a direct measurement that uses the drilling data in the ground as the constraint condition and does the transverse control on the 3-D seismic section (Du and Peng 2010). This method cleverly combines sparse drilling data with dense seismic data and effectively improves the accuracy of predicting the thickness of coal seams both within the inter-mine boundary and its extrapolated areas (Wang et al. 2017). Although the technique has continuously advanced (Chen and Wang 2016; Zou et al. 2018; Li et al. 2017), it requires higher borehole density in the prediction area, and its prediction accuracy is greatly affected by the quality of the 3-D seismic data. The second one is an indirect measurement from the geophysical perspective of coal

mines. Currently, it mainly utilizes three techniques including roadway radio wave prospecting, seismic body wave transmission tomography and channel wave prospecting. The roadway radio wave prospecting defines areas with abnormal coalbed thickness mainly based on the theory that a change in coalbed thickness will result in abnormal radio wave energy attenuation (Xiao et al. 2017). The seismic body wave transmission tomography defines the area with thinner coalbed in the working face based on the relationship of coalbed thickness to the seismic wave velocity effectively (Hu et al. 2017). However, limited by both detection conditions in the working face and the inversion methods, the influences of both coalbed thickness and its structures are not interpreted accurately enough. Recently, the channel wave prospecting has drawn great attention (Schott and Waclawik 2015; Wang et al. 2012; 2016a, b, c; Lei et al. 2017; Wang et al. 2016a, b, c). It directly carries coal seam information and has the potential to become the effective means for quantitative detection of coal seam thickness. Currently, its application for predicting coal thickness is mostly achieved by first constructing a quantitative coalbed thickness prediction model based on both channel wave velocity of a certain frequency and the actual geological results (Zhu et al. 2019). However, most of the models are established based on the empirical formula corresponding to a certain mining area and lack of applicability in other areas due to differences in geological conditions of the coal seam in different mining areas and differences in the lithology and thickness of the roof and floor. Therefore, making full use of the response characteristic of channel wave dispersion curve to the change in coalbed thickness is the key to quantitatively predict the thickness of coal seam and effectively overcome the shortcomings of the above methods.

The dispersion inversion of TCW phase velocity has been theoretically explored. Although the thickness of coal seams can be directly calculated using a complete phase velocity–dispersion curve based on the genetic algorithm (Hu et al. 2018), utilizing the channel wave in coal seam is a more promising breakthrough for the quantitative prediction of coal seam thickness. Aiming at the problem currently existing in the quantitative detection of coal seam thickness, in this study, we explored the characteristic mechanisms of TCW dispersion, developed a 3-D quantitative detection method of coalbed thickness and verified its feasibility and validity using a 3-D physical similarity model experiment. Overall, the method provided highly accurate, quantitative coal thickness information for the intelligentized coal mining and is of theoretical and applicable significance.

Methods

Dispersion curve

The channel wave dispersion curve can be obtained by analyzing the actually measured TCW signals. A variety of methods have been used to this purpose including short-time Fourier transform and multiple narrow-band filtering. In addition, the moving time-window method has been used to find the velocity dispersion curve of the direct channel wave group (Cox and Mason 1988) and multiple filtering has been used to treat coalbed TCW (Krajewski et al. 1987). According to Heisenberg's uncertainty principle, both the time resolution and frequency resolution are mutually constrained. The S-transform in time–frequency analysis can balance these two resolutions to achieve the best result (Stockwell 2007). Since the extraction of the TCW dispersion curve depends on the energy of each order based on the resolution consideration, we used the fundamental-order dispersion curve for inversion. The fundamental-order phase velocity–dispersion of a single-channel TCW can be found as follows (Wang et al. 2016a, b, c):

(a) According to the shot/receiver coordinates, the propagation path length L of the channel wave is obtained using the ray-tracing method. The time–frequency spectrum, $H(t_i, f_j)$, can be determined by performing time–frequency analysis on the measured single-channel TCW $X(t_i)$, where $i = 1, 2, ..., I$, I is the number of sampling points; $j = 1, 2, ...J$, J is the maximum value of the frequency window in time–frequency analysis.

(b) The group velocity of the TCW can be found from the relationship of distance to time:

$$v_g = L/t_i \tag{1}$$

(c) The dispersion spectrum of the TCW group velocity $H(v_{gi}, f_j)$ can be expressed by using the TCW time–frequency spectrum $H(t_i, f_j)$ and relation (1), and by changing $H(t_i, f_j)$ into $H^{'}(L/t_i, f_j)$.

(d) In the channel wave frequency–group velocity spectrum, the dispersion curve of the channel wave group velocity $C(v_g, f)$ is manually picked by applying the energy distribution law. Then, the TCW phase velocity–dispersion curve $C(v_c, f)$ can be obtained based on the relationship between group velocity v_g and phase velocity v_c,

$$v_g = v_c - \frac{dv_c}{d\lambda}\lambda \tag{2}$$

where λ is the wavelength.

1-D S-wave velocity model computation based on genetic algorithm

The coal thickness inversion of the channel wave exploration includes the following steps: (1) perform forward method, (2) obtain the theoretical dispersion curve corresponding to the predicted velocity model, (3) perform the iterative fitting to the actual data dispersion curve and (4) obtain the thickness information of the coal seam. In the solution to the theoretical dispersion curve of TCW in the multilayer horizontal medium, algorithms that are mainly derived from the algorithm for solving the theoretical dispersion of the surface wave are often applied, including Thomson–Haskell algorithm (Thomson 1950; Haskell 1953), Schwab–Knopoff algorithm (Knopoff 1964; Schwab 1970), δ matrix algorithm (Buchen and Ben-Hador 2007) and generalized reflection–transmission coefficient methods (Chen 1993). Among them, the latter one is used either to solve the problem of high-frequency precision loss in previous algorithms so as to make the calculation results more accurate or stable or to calculate the theoretical dispersion curve of TCW in a 1-D multilayered, horizontally stratified media (Ji et al. 2019). Both of them help realize the programmed computation and provide a theoretical support for the inversion calculation of the 1-D multilayer TCW dispersion curve.

Through the above methods, we can find the theoretical dispersion curve of the TCW in the 1-D multilayer, horizontal stratified medium with various parameters and carry out the sensitivity analysis of the model parameters. By changing the shear wave velocity, density and stratum thickness of the coal seam and its surrounding rocks in the multilayer, horizontally stratified model, we can obtain the theoretical dispersion curves with different phase velocities. Further sensitivity analyses find that the shear wave velocities of coalbed and its surrounding rocks as well as the coal thickness are the main factors influencing the dispersion characteristics of the channel wave, while the influence of the density parameter on the phase velocity is relatively weak (Hu et al. 2018). Therefore, the inversion parameters of the model can only consider the layer thickness Δh and the shear wave velocity V_s.

By constructing the multilayer inversion initial velocity model, we can obtain the theoretical dispersion curve corresponding to the predicted velocity model and further perform iterative fitting computation of the actual data dispersion curve. Utilizing the dispersion curve of the single-channel TCW data can find the shear wave velocities of both the rockbed and coalbed at different depths and different propagation paths. The inversion methods of the dispersion curve mainly include the damped least squares (Forbriger 2003), singular value decomposition (Xia et al. 2015), simulated annealing method (Calderónmacías and Luke 2007) and genetic algorithm (Yamanaka and Ishida,

1996; Feng et al. 2005; Buyuk et al. 2017). Among them, the genetic algorithm is adopted in this study due to its good convergence, stable computation and less influence by the initial model. The globally optimized and stochastically retrieval algorithm is formed by mimicking biological inheritance and evolutionary processes. It uses genetic operations such as mutation, crossover and selection to make the fittest data survival. After rounds of repeat iterations, the data set is continuously updated to gradually meet the set conditions and eventually find the optimal solution. The specific steps are as follows:

(a) Construct a 1-D multilayer, horizontally stratified medium inversion model with its layer thickness and number of layers meeting the vertical resolution requirement (Δh, N). In this study, the initial model parameters are set to $\Delta h = 1$ m, and $N = 100$; the shear wave velocity of the rockbed is 2000 m/s, and the shear wave velocity of the coalbed is approximately 1000 m/s. It should be noted that the velocity parameters and coal thickness information of the initial model refer to the tomographic results and geological data of the TCW velocity. Although the initial model is rough, it is sufficient to mark the search range of the required model parameters in the genetic algorithm.

(b) Set the amount of change in formation thickness, the amount of change in velocity and the numbers of populations of both thickness and velocity as 0.1 m, 10 m/s and 64, respectively, work out the first-round 64×64 theoretical dispersion curve $C(v_c, f)^1$ using the above method and compare it with the actually measured data curve $C(v_c, f)$ to find the mean square difference set $\Delta \varepsilon^1$, where the superscript is the times of iterative calculation.

(c) Take the random and universal selection method to screen the results obtained in the previous step, where small variance has a big selected probability. Then, perform the crossover and mutation processing and make them as the second round of model parameters. Extract the theoretical dispersion curve of the updated model $C(v_c, f)^2$. Compare it with the actually measured data curve $C(v_c, f)$ to find the new mean variance set $\Delta \varepsilon^2$;

(d) Repeat many rounds of iterative calculation and continuously update the data set. If the error continuously reduces and reaches the set threshold P, that is, $\Delta \varepsilon \leq P$, end the calculation; select the model parameter (H_n, V_{sn}) corresponding to the minimum error in this round as the inversion calculation result of the dispersion curve, where H_n and H_{sn} are the relative height and S-wave velocity of the formation of the inversion model, respectively. Set the relative heights of the detector and the seismic source point to 0, that is

$$H_n = \sum_{n=1}^{N} \Delta h_n dn, \qquad n = 1, 2, \ldots, N \qquad (3)$$

3-D S-wave velocity model computation

With the help of the above process, the inversion result of the single-channel TCW dispersion curve is obtained, that is, the shear wave velocities of the top and bottom rock formation and coal seam within a certain range, (H_n, V_{Sn}). For example, the 1-D inversion result of the channel wave of the jth ray can be obtained through the above procedure, $(H_n, V_{Sn})_j$, where $j = 1, 2, \ldots, J$, J is the number of rays. We believe that the average velocity of the shear wave at a certain relative height H_n in the vertical section of the jth TCW ray propagation path is V_{Sn} (horizontal stratified model).

Taking the 1-D multilayer, horizontally stratified velocity model $(H, V_S)_j$ as the initial model and doing 3-D meshing can obtain a 3-D grid velocity model. Assuming in a stratified model of a certain height, the average velocity of a total of I grids in the plane of the jth channel wave at the height $H = H_n$ is V_j in the inversion model, we have $S_{ij} = \frac{1}{V_j}$, where S_{ij} is the slowness. The propagation distance within the grid d_{ij} can be calculated according to the intersection of the projection of the TCW ray-tracing path on the plane. Furthermore, the travel time T_j corresponding to the jth channel wave can be obtained as

$$T_j = \sum_{i=1}^{I} d_{ij} S_{ij} \ (i = 1, \ldots, I) \qquad (4)$$

On this basis, we can obtain the matrix equation of the relative height H_n as follows:

$$\begin{bmatrix} d_{11} & d_{12} & d_{13} & \ldots & d_{1i} \\ d_{21} & d_{22} & d_{23} & \ldots & d_{2i} \\ d_{31} & d_{32} & d_{33} & \ldots & d_{3i} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ d_{j1} & d_{j2} & d_{j3} & \ldots & d_{ji} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ \ldots \\ S_i \end{bmatrix} = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ \ldots \\ T_j \end{bmatrix} \qquad (5)$$

Since T_j can be calculated from Eq. (4), the slowness matrix (S_{ij}) can be found from Eq. (5). In order to solve such a large sparse matrix as Eq. (5), the back-projection method is an effective algorithm. The principle of the method is that the data of all the rays are allocated by the ratio of the path length d_{ij} in each grid along the propagation path to the total length of the propagation paths $\sum_i d_{ij}$.

After all the operations end, we can obtain the slowness vector of the grid as follows:

$$\hat{S}_i = \sum_{j=1}^{J} \left(d_{ij}/v_{sj} \right) / \sum_{j=1}^{J} d_{ij} \qquad (6)$$

where J is the total number of TCW rays. Performing the above inversion calculation of each of the TCW signals in sequence can find a 1-D horizontal velocity distribution at a certain relative height (H, V_{Sn}). Conducting the above processing of the 1-D data of each height in turn can find the 3-D spatial distribution of the shear wave velocities in the working face of the coal seam. In the 3-D velocity body, the rate of velocity change can be calculated along the depth direction (Z-axis). If the shear wave velocities of both the coal seam and rock stratum have a great difference, the extreme of the rates will be generated at the interface of coal and rock. With the subtraction of two relative heights corresponding to the two extremes, we can get the spatial distribution of the studied coal seam thickness.

From the above description, we conclude the procedure of the 3-D coalbed thickness detection method based on the TCW dispersion curve inversion in Fig. 1.

Algorithm validation at laboratory scale

Physical model preparation

According to the similarity theory, the similarity criterion of the simulation model is mainly reflected in its scale factor. The model needs to meet the geometric similarity ratio for simulating kinematics and the dynamic similarities for simulating the dynamic characteristics. In the consideration of the difference in size between the simulation model and the measured model, the parameters of the seismic wave such as the propagation distance, frequency and wavelength change. The transmission seismic wave tomographic detection of the actual coalbed face showed that the coalbed thickness is about 4 m and the main frequency of the transmitted wave is about 200 Hz (Hu et al. 2018). In accordance with the scale factor, the coalbed thickness and the detection frequency are set as 40 mm and 20 kHz, respectively. The determination of these parameters is the principle for subsequent model design and test probe frequency selection in laboratory scale. Table 1 shows the specific scale factor of the simulation model for this numerical simulation.

Coal measure strata consist mainly of coal seam and sandstone or limestone components with their lithology varying greatly. Mixing similar materials of aggregate and cement according to a certain mass ratio can simulate the physical properties of real rock mass. Because this experiment mainly focuses on the elastic wave parameters of coal and rock, similar materials such as river sand as aggregate as well as cement and gypsum as cement are used (Wang et al. 2016a, b, c). In order to find a suitable material mix ratio, we tested a series of mix ratio of the above similar materials (Table 2). The mix ratio scheme is aimed to gradually increase the mass ratio of aggregate and cement while
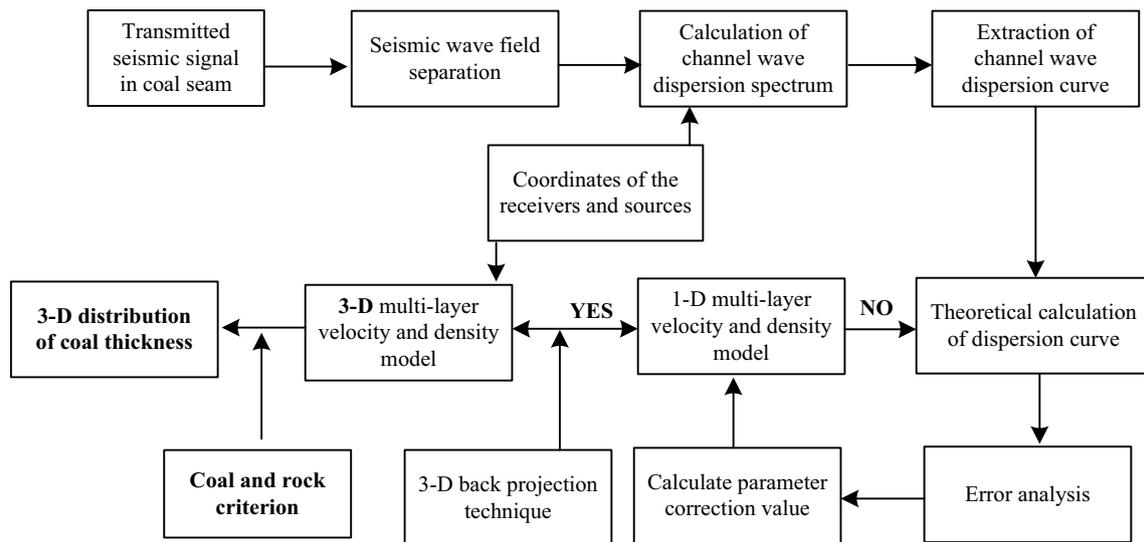
Fig. 1 Roadmap on 3-D quantitative coalbed thickness prediction inversion technique

Table 1 Scale factor of similarity simulation

Seismic parameters	Model scale factor
Space length	1:100
Velocity	1:1
Frequency	100:1
Sampling interval	1:100

ensuring the properties of the cement to be invariant so as to make the physical parameters such as longitudinal and shear wave velocities as well as density close to reality.

In the condition of strictly following the mix ratio set in Table 2, the raw materials were uniformly mixed, filled into the 10-mm standard sample boxes and placed in a cool area. After 24 h, they were demolded and cured in an aqueous solution for 7 days. One day before the ultrasonic test, they were taken out and dried in a cool area. A total

of 8 samples with 2 repeat specimens for each ratio were prepared (Fig. 2). The length, width, height and weight of each specimen were measured to obtain the average density of each sample (Table 2). During the experiments, ultrasound as a widely used method for nondestructive testing of dynamic mechanical parameters of rock, soil and concrete was utilized.

Under the existing experimental conditions, the shear wave and longitudinal wave cannot be effectively separated. This phenomenon will lead to unreliable results in the direct shear wave speed test. Therefore, we instead use the longitudinal wave velocity to estimate the shear wave velocity (Peng et al. 2004). In this study, all samples were ultrasonically tested three times in their transverse X-, longitudinal Y- and vertical Z-axes using the transmission method with grease-coupled longitudinal wave probes (Fig. 3) to obtain the average longitudinal wave velocity of each sample (Table 2). According to the ultrasonic test results and the actual field detection experience, No.5 sample was selected

Table 2 Composition of similar materials for physical models

Sample no.	Modeled stratum	Sand (kg)	Concrete (kg)	Water (kg)	Gypsum (kg)	Mean density (kg/m^3)	Mean longitudinal wave velocity (m/ms)
1#	Rock	4	1	0.5	–	1.52	2.04
2#	Rock	4	1.2	0.6	–	1.63	2.19
3#	Rock	4	1.4	0.7	–	1.8	2.54
4#	Rock	4	1.6	0.8	–	1.98	2.66
5#	Rock	4	1.8	0.9	–	1.94	2.78
6#	Coal	10	1	2	2	1.58	1.10
7#	Coal	10	1.5	3	3	1.64	1.29
8#	Coal	10	2	4	4	1.67	1.69

Fig. 2 Specimens with different mixing ratios

as the similar material of the roof and floor rock strata and No.8 sample as the similar material of coal seam. The ultrasonic longitudinal wave velocities of No.5 and No.8 samples were measured to be 2.78 m/ms and 1.69 m/ms, respectively.

According to the actual situation of the working face, a symmetric, horizontally stratified, three-layered physical model with dramatic changes in coalbed thickness at the transverse level was designed, as shown in Fig. 4.

Figure 5 shows the images of the physical model designed in Fig. 4 with dimensions consistent with the design parameters. At $X = 1050$ mm, the thickness of the low velocity interlayer changes sharply from 80 to 40 mm and its shape meets the expected design requirements.

Fig. 3 Scheme for ultrasonic testing of specimens
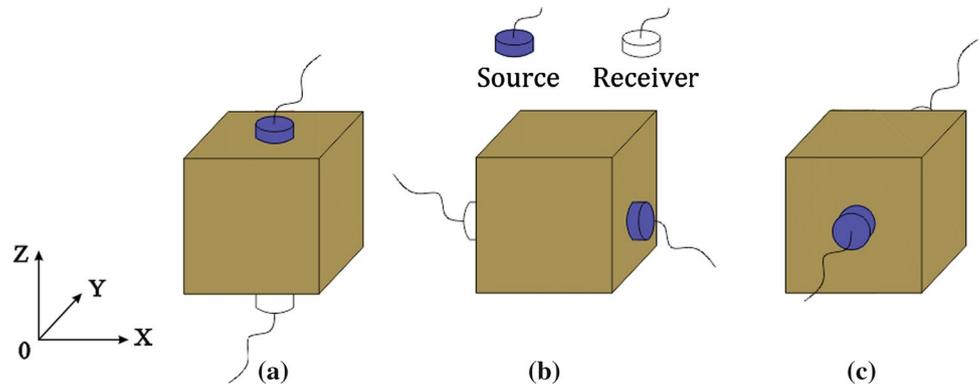
Fig. 4 Design diagram of horizontal coal seam physical model with thickness catastrophe
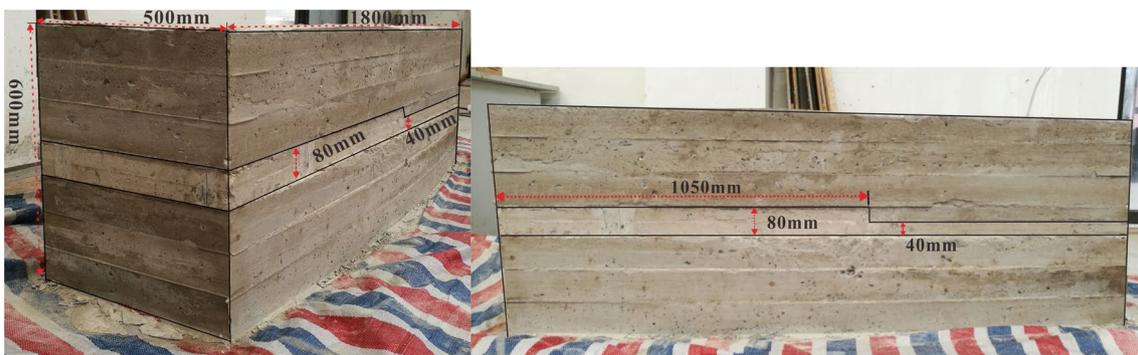
Fig. 5 Photographs of the actual physical model

Measurements

Figure 6 shows the ultrasound transmission monitoring system with one source corresponding to multiple receivers used in the experiment. In order to simulate the actual situation using seismic transmission wave to prospect the coalbed at the working face, a total of 18 sources in 18 channels with spacing of 100 mm were placed in the soft interlayer in the middle of one side of the model and a total of 18 receivers with track pitch of 100 mm were placed in the corresponding soft interlayer on the other side of the model. The source and receiver numberings are the same from small to large in the X-axis direction on their own side. When the source emits in turn from channel 1 to 18, at each emitting, the 18 receivers on the opposite site will detect the transmitted wave.

Data processing

In the experiment, the receiver used a two-component sensor with its main frequency of 20 kHz, Y-component, i.e., the sensitivity direction, parallel to the Y-direction of the model, and X-component parallel to the X-direction of the model, as shown in Fig. 7. The Love-type TCW signals were, respectively, identified from the received X- and Y-component signals and combined according to Eq. (7),

$$P = X \cos \alpha + Y \cos \beta \tag{7}$$

where α and β denote the angle of the X- and Y-components with the polarization direction of the Love-type TCW, respectively. They can be determined from the relative coordinates of the seismic source point (source) and the detector (receiver).

Figure 8a shows the randomly selected data from No.3 source and No.8 receiver for dispersion analysis. Figure 8b shows the dispersion spectrum obtained based on the data from Fig. 8a using the method described in Sect. 1.1.
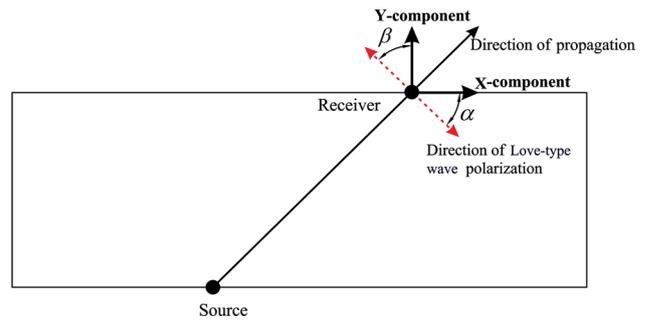
Fig. 7 Scheme of Love-type channel wave extraction from horizontal and vertical components

From the maximal energy distribution characteristics in the dispersion spectrum, the group velocity curve of TCW is manually identified. According to the theoretical relationship between the group and phase velocities, the dispersion curve of the phase velocity can be computed. By inversion calculation of the phase velocity dispersion curve, the distribution characteristics of the shear wave velocity and density in the depth direction can be found, as shown in Fig. 9.

Figure 10 shows the 3-D shear wave velocity distribution of the coalbed roof and floor obtained by performing the TCW inversion for each channel and back-projecting the velocity in the depth direction. From the results, it is clear that the velocity model has better stratification with the shear wave velocities of the roof and floor being ~1600 m/s, the shear wave velocity of the coal seam being ~700 m/s and little change in the velocity on the X–Y section. All of these indicate that the inversion has better stability.

Figure 11 shows the 3-D shear wave velocity body of the same grid element whose rate of velocity change in the Z-axis can be calculated from $(x_i, y_j, z_k, (V_{k2}-V_{k1})/V_{k1})$. In the data set of the rate of 3-D velocity change, the rate of velocity change reaches the negative extremum at the lower interface of the coal seam (in the negative Z-axis direction),
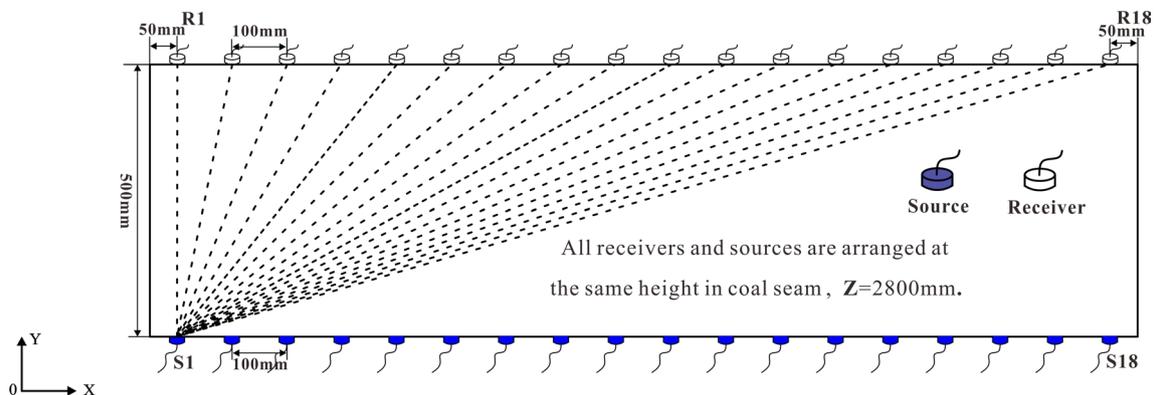
Fig. 6 Schematic of the ultrasonic CT test on physical model
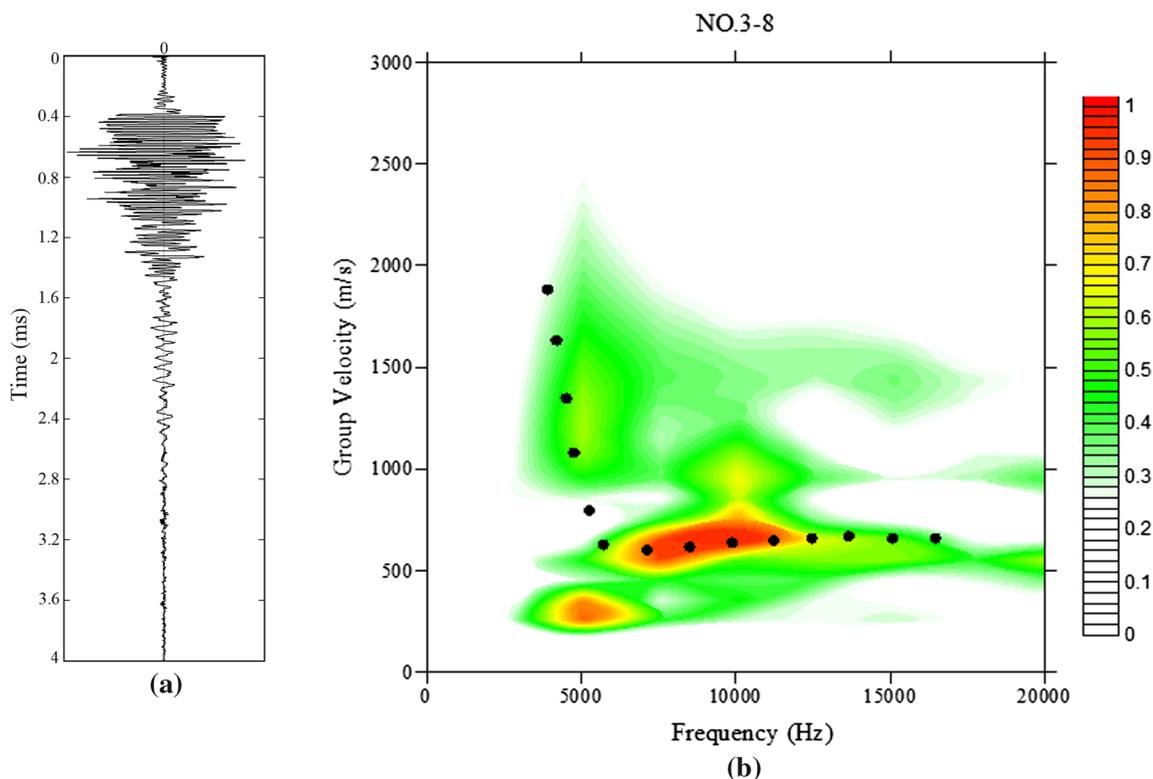
Fig. 8 Signals obtained from No.8 receiver when transmitted from No.3 source **(a)** and its dispersion spectrum **(b)**
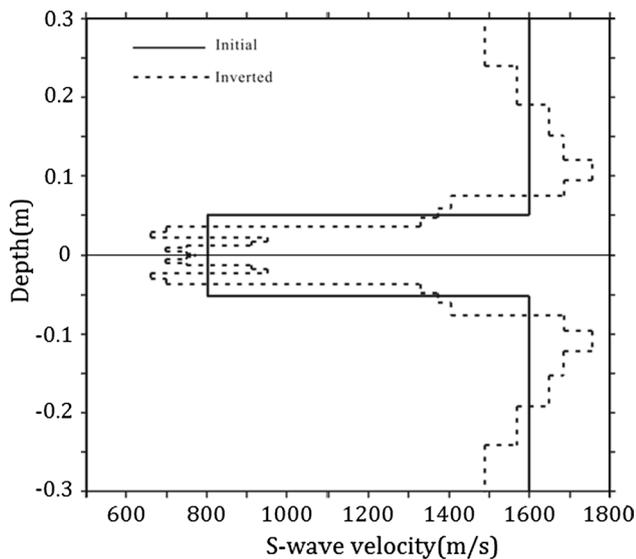
Fig. 9 Inversion results of single Love-type channel wave dispersion curve

and positive extremum at the upper interface of the coal seam (in the positive Z-axis direction). Finding the difference between the two z_k values corresponding to the two extrema (x_i, y_j, $|z_{kmax} - z_{kmin}|$) can obtain the spatial distribution of coalbed thickness, as shown in Fig. 11. In the

presence of the soft interlayer in the roof and floor, there may be multiple extrema. In this case, the upper and lower extreme points closest to the coal seam can be selected as the coal–rock interface.

In Fig. 11, the black line denotes the actual abruptly changed position of the coal seam thickness. The average coal seam thickness is in the range of 7–9 cm on the left side of the black line and 3–5 cm on the right side with the inversion error controlled at about 10%. Meanwhile, because the influence of the test system on the left and right ends of the model results in fewer rays, the inverted effect is poorer. In general, the distribution of coalbed thickness is in a good agreement with the actual physical model, indicating that the quantitative detection method for coalbed thickness presented in this study is accurate and effective. Furthermore, the velocities of coal and rock obtained using this inversion method are different from the comprehensive velocity response obtained by the transmission body wave velocity tomography. This kind of characteristic has important application significance for ensuring safe mining of coal mines.

Discussions

Our study has some limitations. First, the inversion model is an idealized 1-D up–down symmetrical multilayer velocity model, while the actual detected coal measures
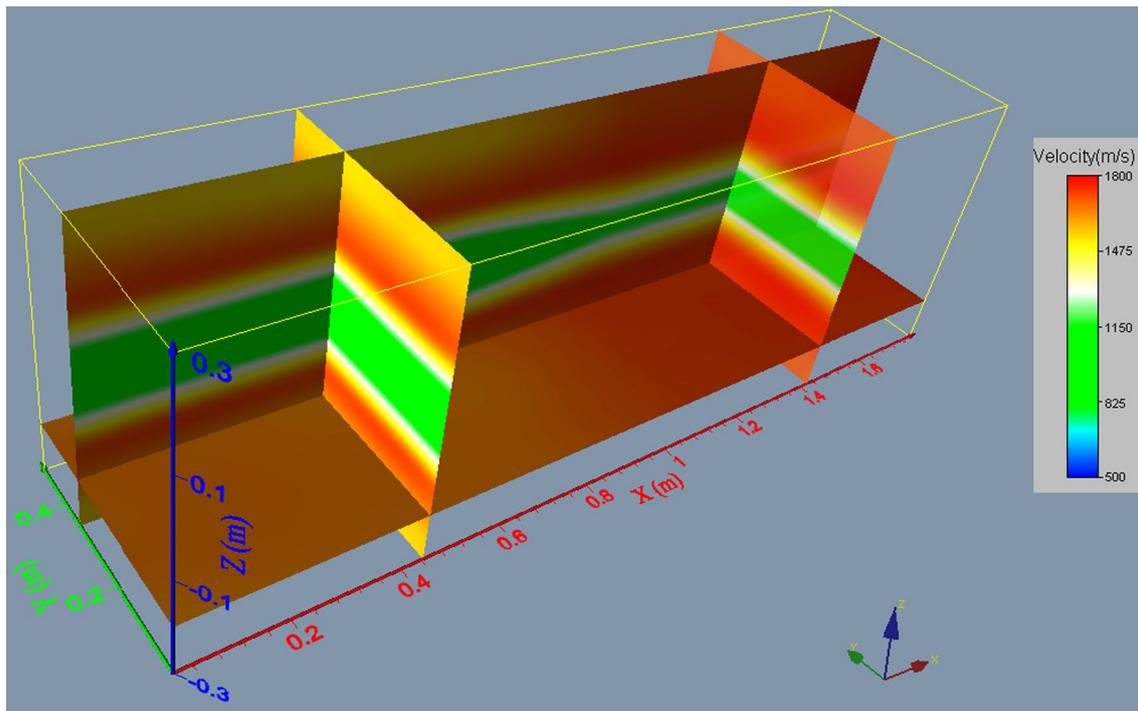
Fig. 10 3-D display of velocity inversion results in ultrasonic CT test of the physical model
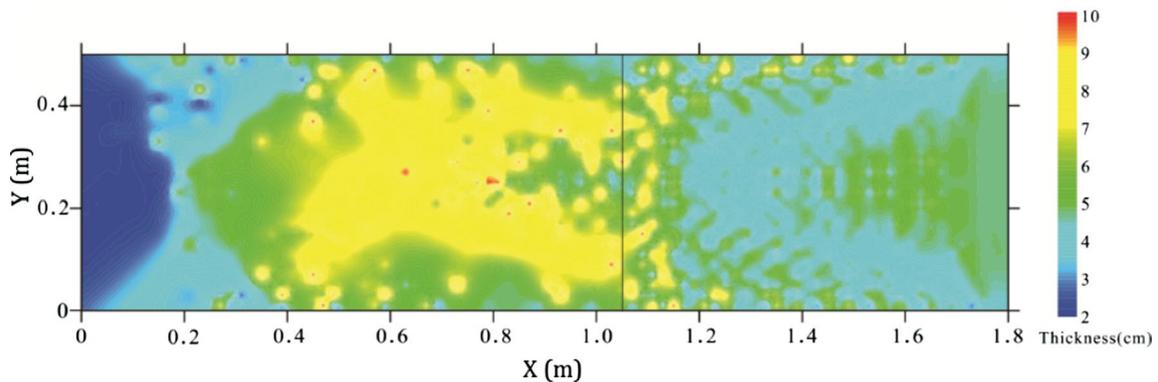
Fig. 11 Inversion results of the coal seam from ultrasonic CT test of the physical model

strata are generally not up–down symmetrical stratigraphic combinations. The effects of parameter change in coal seam roof and floor strata on the dispersion curve are indistinguishable, which will affect our explanation about the directional anomaly problem of the roof and floor strata. This study focuses on the quantitative detection of coal thickness. Thus, the direction problem of the velocity anomaly has little effect on the interpretation of coal thickness. Second, the inversion model is an idealized 1-D horizontal velocity model, while in reality, the coal strata often have topographic fluctuations, which will cause errors in TCW ray-tracing and dispersion curve extraction and further affect the precision of the quantitative inversion results of coalbed thickness. For the coal seam with thickness fluctuations, the studies on the inversion method of both coal thickness and geological structure need to be further deepened. At last, the inversion method used in this study relies on the dispersion curve of the TCW, but the presence of dirt bands in the coal seam will seriously affect the shape of the dispersion curve. The subsequent study should include the response of the coal seam with dirt bands so as to make clear their effect on quantitative interpretation of coal thickness.

Conclusions

In this study, we explored the inversion algorithm for quantitative measurement of coalbed thickness by adopting both theoretical analysis and physical model experiments. By constructing a 3-D velocity model and inversion calculation, we obtained the high-precision inversion results of complete dispersion curve of TCW. By analyzing the distribution characteristics of coal–rock interface inversion velocity, we proposed a novel coal–rock recognition mode and realized the quantitative detection and interpretation of coal seam thickness.

Our results can be summarized as follows:

1. Constructed the velocity inversion model for a 1-D horizontal multilayer medium, achieved high-precision inversion of the complete dispersion curve of TCW by using genetic algorithm and obtained the distribution characteristics of the shear wave velocity in the depth direction.
2. Established the calculation method of the TCW-based 1-D inversion velocity in the 3-D space by combining the back-projection technology, obtained the 3-D inversion velocity body of the TCW, analyzed the distribution characteristics of inversion velocity on the coal–rock interface, proposed the coal thickness calculation method based on the rate of velocity change and realized the quantitative identification of coal seam thickness.
3. Performed 3-D physical similarity model experiment in the working face and simulated and analyzed the propagation characteristics of the 3-D TCW in coal seam with different thicknesses. Based on the inversion results of the 3-D physical similarity simulation experiment, objectively and quantitatively interpreted the coal seam thickness. The error rate in the coalbed thickness obtained using the quantitative inversion method in the physical similarity simulation experiment and the actual coalbed thickness is within 10%.

References

Buyuk E, Zor E, Karaman A (2017) Rayleigh wave dispersion curve inversion by using particle swarm optimization and genetic algorithm [C]// In: EGU General Assembly Conference. EGU General Assembly Conference Abstracts

Buchen PW, Ben-Hador R (2007) Free-mode surface-wave computations. Geophys J Int 124(3):869–887

Calderónmacías C, Luke B (2007) Improved parameterization to invert Rayleigh-wave data for shallow profiles containing stiff inclusions. Geophysics 72(1):1–10

Chen T, Wang X (2016) Thickness prediction of tectonically deformed coal using calibrated seismic attributes: a case study. ASEG Ext Abstr 1:1–5

Chen X (1993) A systematic and efficient method of computing normal modes for multilayered half-space. Geophys J Int 115(2):391–409

Cox KB, Mason IM (1988) Velocity analysis of SH channel waves in the Schwalback seam of Ensdorf Colliery [J]. Geophys Prospect 36(3):298–317

Du W, Peng S (2010) Coal seam thickness prediction with geostatistics [J]. Chin J Rock Mechan Eng 29(s1):2762–2767

Feng S, Sugiyama T, Yamanaka H (2005) Effectiveness of multi-mode surface wave inversion in shallow engineering site investigations. Explor Geophys 58(1):26–33

Forbriger T (2003) Inversion of shallow-seismic wave fields: II. inferring subsurface properties from wavefield transforms. Geophys J Royal Astron Soc 153(3):719–734

Haskell NA (1953) The dispersion of surface waves on multilayered media. Bull Seismol Soc Am 43:17–34

Hu Z, Zhang P, Xu G (2018) Dispersion features of transmitted channel waves and inversion of coal seam thickness. Acta Geophys 66(5):1001–1009

Hu Z, Zhang P, Xu G (2017) Research advances of seismic tomography technology in coal seam. Prog Geophys 32:2451–2459

Ji G, Li H, Wei J et al (2019) Preliminary study on wave field and dispersion characteristics of channel waves in VTI coal seam media. Acta Geophys 4:1–12

Knopoff L (1964) A matrix method for elastic wave problems. Bull Seismol Soc Am 54(1):431–438

Krajewski P, Dresen L, Schott W et al (1987) Studies of roadway modes in a coal seam by dispersion and polarization analysis: a case history. Geophys Prospect 35:767–786

Lei F, Wang W, Li S, Yao X, Teng J, Gao X (2017) Research on the channel wave field characters of goaf in coal mine and its application. In: Di Q, Xue G, Xia J (eds) Technology and Application of Environmental and Engineering Geophysics. Springer Geophysics. Springer, Singapore

Li Q, Guo L, Sun Y et al (2017) Seismic attributes fusion and its research in predicting thickness of coal [J]. Prog Geophys 32(5):2014–2020

Peng S, Gao Y, Peng X et al (2004) Study on the rock physic parameters of coal bearing strata in Huainan Coalfield. J China Coal Soc 29(2):177–181

Peng S, Feng J et al (2019) Automation in U.S. longwall coal mining: A state-of-the-art review. Int J Mining Sci Technol 2:151–159

Schott W, Waclawik P (2015) On the quantitative determination of coal seam thickness by means of in-seam seismic surveys. Can Geotech J 52:1496–1504

Schwab I (1970) Surface-wave dispersion computations: Knopoff's method. Bull Seismol Soc Am 60(2):321–344

Stockwell RG (2007) A basis for efficient representation of the S-transform. Digital Signal Process 17(1):371–393

Thomson WT (1950) Transmission of elastic waves through a stratified solid medium. J Appl Phys 21(2):89–93

Wang W, Gao X, Li S (2012) Channel wave tomography method and its application in coal mine exploration: An example from Henan Yima Mining area. Chinese J Geophys 55(3):1054–1062

Wang W, Xue G, Gao X, et al (2016) Channel wave tomographic imaging method and its application in detection of collapse column in coal// In: International Conference on Environment and

Engineering Geophysics & Summit Forum of Chinese Academy of Engineering on Engineering Science and Technology.

Wang B, Liu S, Zhou F et al (2016) Dispersion characteristics of SH transmitted channel waves and comparative study of dispersion analysis methods. J Comput Theor Nanosci 13(2):1468–1474

Wang C, Si J, Zhang X (2016) Research on similar material proportioning test based on orthogonal design. Coal Technol 35(8):21–23

Wang X, Li Y, Chen T et al (2017) Quantitative thickness prediction of tectonically deformed coal using extreme learning machine and principal component analysis: a case study [J]. Comput Geosci 101:38–47

Xiao Y, Wu R, Yan J et al (2017) Field strength propagation law of radio wave penetration and effective perspective width for coal face. J China Coal Soc 42(3):712–718

Xia J, Gao L, Pan Y et al (2015) New findings in high-frequency surface wave method [J]. Chinese J Geophys (in Chinese) 58(8):2591–2605

Yamanaka H, Ishida H (1996) Application of genetic algorithm to an inversion of surface-wave dispersion data. Bull Seismol Soc Am 86(2):436–444

Yuan L (2017) Scientific conception of precision coal mining. J China Coal Soc 42(1):1–7

Zhu M, Cheng J, Cui W et al (2019) Comprehensive prediction of coal seam thickness by using in-seam seismic surveys and Bayesian kriging. Acta Geophys 67(3):825–836

Zou G, Xu Z, Peng S et al (2018) Analysis of coal seam thickness and seismic wave amplitude: A wedge model. J Appl Geophys 148:245–255

RESEARCH ARTICLE - APPLIED GEOPHYSICS

Numerical simulation of shear wave attenuation in borehole inserted by a horizontal fracture

Binpeng Yan[1] · Weiming Ou[2] · Xingguo Huang[3] · Nuno Vieira da Silva[4,5]

Abstract

The amplitude of shear waves is attenuated when passing through horizontal fractures crossing a borehole. In this study, we investigate the amplitude attenuation of shear waves throughout simulation of full-wave acoustic logging with the finite-difference method. As the fracture aperture is very small, it needs to be represented in a very fine gird when carrying out finite-difference simulation. Therefore, the variable-grids finite-difference method is adopted to avoid over-sampling in the non-fracture regions, yielding substantial savings in computational cost. We demonstrate the accuracy of waveform modeling with the variable-grid finite-difference by benchmarking against that obtained with the real-axis integrating method. We investigated the effects of several important parameters including fracture aperture, distance from receiver to fracture, borehole radius and extended distance utilizing that benchmarked variable-grid finite difference code. We determined a good linear relationship between the attenuation coefficient of shear wave amplitude and the fracture aperture. Then, the effects of distance from receiver to fracture, the borehole radius and the extended distance of fracture on shear wave attenuation are also studied. The attenuation coefficient of shear wave becomes smaller with the increasing borehole radius. While, it increases as the distance from receiver to fracture and the extended distance of fracture increase. These effect characteristics are conducive to the use of shear wave to evaluate fractures.

Keywords Horizontal fracture · Shear wave attenuation · Variable grids · Acoustic logging · Finite-difference

Introduction

Fractured reservoirs are widely distributed in various geological formations, such as carbonate rocks, hard sandstones, igneous rocks and shales. Fractures play a crucial role as migration channels of petroleum from the source rock to the reservoir, and as storage space of petroleum. In addition, the

existence of fractures changes substantially the shape of the waveforms recorded in well-logging. Thus, mastering the effects of fractures on acoustic waveforms is beneficial to the detection and evaluation of fractures.

P-wave, S-wave, Stoneley wave and pseudo-Rayleigh wave are excited by monopole source in acoustic wave logging. As known, Stoneley waves are sensitive to permeable fractures (Hornby et al. 1989). The interaction between a Stoneley wave and a fracture can be calculated by analytical methods (Hornby et al. 1989; Tang and Cheng 1993; Kostek et al. 1998b; Bakku et al. 2013). Previous research on the interaction between Stoneley waves and fractures has been carried out. For example, Hornby et al. (1989) established the relationship between fracture aperture and the Stoneley wave propagation in frequency domain assuming rigid formations. Kostek et al. (1998a, b) studied the effect of fractures on the reflection coefficient of Stoneley waves in the elastic formation. Minato and Ghose (2017) studied Stoneley wave propagation in boreholes intersecting multiple closely spaced fractures. To study the interaction between fracture and Stoneley wave, the frequency of acoustic sources is

✉ Binpeng Yan
  yanbp@cupk.edu.cn

1 China University of Petroleum-Beijing At Karamay,
   Xinjiang, China

2 The 54th Research Institute of China Electronics Technology
   Group, Shijiazhuang, China

3 Department of Earth Sciences, University of Bergen,
   Allegaten 41, 5020 Bergen, Norway

4 Department of Earth Science and Engineering,
   Imperial College London, South Kensington Campus,
   London SW7 2AZ, UK

5 Present Address: Total E&P UK, Tarland Road, Westhill,
   Aberdeen AB32 6JZ, UK

generally set to be less than 4 kHz. However, the sources utilized in array acoustic logging source generally operate at central frequencies above 8 kHz. In the condition of high frequency source P-waves, S-wave and pseudo-Rayleigh waves are excited. Those modes, especially pseudo-Rayleigh waves, can be regarded as noise as they interfere with the study of the relationship between Stoneley waves and fractures. Like Stoneley waves, shear waves are also sensitive to fractures. The amplitude of shear-waves is also attenuated by scattering, fluid flow and wave mode conversion when crossing fracture. Therefore, we study the relationship between attenuation of shear-waves and fractures to adapt to the complex acoustic modes in full-wave acoustic logging.

The attenuation of shear-waves across fractures can be observed in recordings from acoustic logging as well as simulated numerically. Morris et al. (1964) determined the location of fractures according to shear wave amplitude reduction. Paillet (1980) investigated shear wave propagation in the vicinity of fractures based on the acoustic logging data and found shear amplitude attention in the vicinity of fractures. Zlatev et al. (1988) used physical experiments to simulate propagation of acoustic waves in boreholes crossing fractures reporting the impact of fractures on amplitude attenuation of shear waves. Goldberg and Gant (1988) reported on the use of attenuation of shear waves to identify and characterize fractures in geological formations. Matuszyk et al. (2013) used the finite element method to simulate the waveform of acoustic wave in the borehole, the simulation results showed that the shear wave attenuation increased with the increase in the distance between the fracture and the receiver after crossing the fracture. Those studies investigated the decay of the amplitude of shear waves in fractured formations. However, the relationship between the attenuation coefficient of shear waves and fracture aperture, the effects of borehole radius, bedrock, and extended distance of fracture from borehole axis in radial direction on the attenuation coefficient of shear wave, were not reported on those previous studies. Therefore, the main goal of this paper is to determine and discuss the effects of all those factors on the attenuation coefficient of shear waves.

We take a model with axial symmetry placing a horizontal fracture between two identical elastic formations. Borehole sonic waveforms can be computed numerical with either the finite-difference (FD) or the finite-element method. The latter is particularly suitable to handle irregular grid spacing or irregular geometry; however, it is computationally more demanding than FD method.

With the improvement of computer performance, the FD method (Chen et al. 1998; Guan et al. 2009; Yan et al. 2020) has become widely used to simulate propagation of elastic

waves in acoustic logging. Finite-difference grids can also be used to compute model responses of fractured media (Stephen et al. 1985). The fracture aperture is generally less than 1 mm. That typical length requires very fine grid discretization, imposing limitations on the FD time-stepping in order guarantee stability of computations. This problem is overcome using variable grids which avoid spatial over-sampling and consequently relaxing the time-step required for stability and saving computing time. Some previous work reported using FD with variable grids in the simulation of seismic wave scattering by hydraulic fractures (Groenenboom and Falk 2000), and in the simulation of borehole acoustic waves in a porous formation with tilted thin fractures (Yan et al. 2015).

To further understand the effect of fractures on shear wave attenuation, we investigate the relationship between the fracture and the attenuation coefficient of shear wave. In this paper, the variable grid FD method is utilized to simulate full-wave acoustic logging in a formation with horizontal fractures and obtain waveforms. The attenuation coefficient of shear wave is calculated according to the amplitude changes of waveform, and the quantitative relationship between fracture aperture and amplitude attenuation of shear wave is analyzed. We further investigate the effects of receiver-to-fracture distance; elastic parameters of the formation and borehole radius on the attenuation coefficient of shear wave are also investigated. Finally, the relationship between the extended distance of fractures and the attenuation coefficient of shear wave is studied by simulating full-wave acoustic logging with finite fractures.

Theory

The model of a borehole inserted by a horizontal fracture

We assume that a planar axis-symmetric fracture intersects a borehole. Figure 1 is a schematic diagram of a borehole which is inserted by a horizontal fracture. It is assumed that both the borehole and the fracture are filled with water. The radius of the borehole is a, and the fracture aperture is h. The velocity of acoustic wave in water is $v_f = 1500$ m/s, and the water density ρ_f is 1000 kg/m^3. The bedrock surrounding the borehole is assumed to be an elastic and isotropic formation. The extended distance of the fracture from borehole axis in radial direction is L, which ranges between zero and infinity. The solid circular and squares on the borehole axis represent a point source and receivers, respectively. The source is below the fracture depth, and the receivers are located
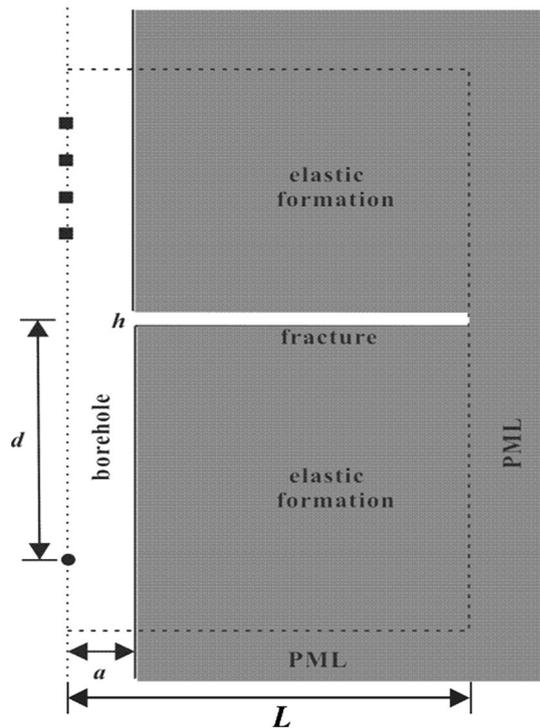
Fig. 1 A schematic diagram of the borehole inserted by a horizontal fracture. The perfectly matched layer (PML) absorbing boundary is added to the boundary of the model

above the fracture depth. The distance from the source to the fracture is $d = 2$ m.

Elastic wave equations

As the model shown in Fig. 1 has axial symmetry, we can use cylindrical coordinates for modeling and reduce the model dimensions to 2D. This has clear advantages to reduce computational overhead. In a cylindrical coordinate system, without considering the body force, the first-order velocity–stress equations for isotropic elastic media of the monopole source are (Randall et al. 1991)

$$\frac{\partial \tau_{rr}}{\partial t} = (\lambda + 2\mu)\frac{\partial v_r}{\partial r} + \lambda\left(\frac{v_r}{r} + \frac{\partial v_z}{\partial z}\right), \quad (1a)$$

$$\frac{\partial \tau_{zz}}{\partial t} = (\lambda + 2\mu)\frac{\partial v_z}{\partial z} + \lambda\left(\frac{v_r}{r} + \frac{\partial v_r}{\partial r}\right), \quad (1b)$$

$$\frac{\partial \tau_{\theta\theta}}{\partial t} = (\lambda + 2\mu)\frac{v_r}{r} + \lambda\left(\frac{\partial v_r}{\partial r} + \frac{\partial v_z}{\partial z}\right), \quad (1c)$$

$$\frac{\partial \tau_{rz}}{\partial t} = \mu\left(\frac{\partial v_r}{\partial r} + \frac{\partial v_z}{\partial z}\right), \quad (1d)$$
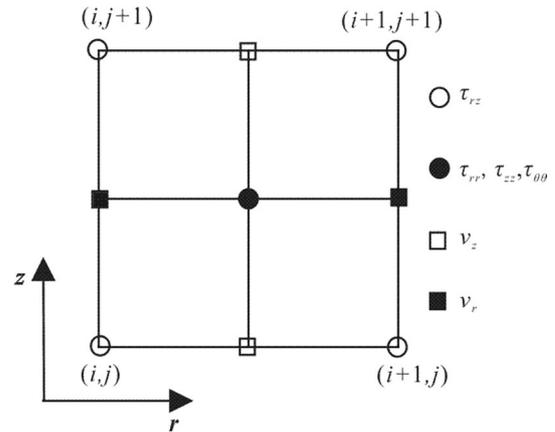
Fig. 2 Discretization cell with staggered grids in the r–z plan. The hollow circle represents shear stress, the solid circle represents normal stress, the hollow square represents v_z, and the solid square represents v_r. The parameters i and j in brackets denote the grid indexing along r and z directions, respectively

$$\rho\frac{\partial v_r}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}(r\tau_{rr}) + \frac{\partial \tau_{rz}}{\partial z} - \frac{\tau_{\theta\theta}}{r}, \quad (1e)$$

$$\rho\frac{\partial v_z}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}(r\tau_{rz}) + \frac{\partial \tau_{zz}}{\partial z}, \quad (1f)$$

where τ_{rr}, τ_{zz} and $\tau_{\theta\theta}$ are normal stress, τ_{rr} is the shear stress, ρ is the media density λ and μ are the Lame constants of the media, v_r and v_z are velocity components in r- and z-directions, respectively.

The physical parameters of the medium are discretized utilizing staggered grids (Virieux 1986), as shown in Fig. 2. In the r- and z-directions, the grid spacings are Δr and Δz, respectively. The discretization of Eqs. 1a–1f is formulated as the following second-order FD formulas:

$$(\tau^{n+1}_{rr(i+1/2,j+1/2)} - \tau^n_{rr(i+1/2,j+1/2)})/\Delta t$$
$$= (\lambda_{(i+1/2,j+1/2)} + 2\mu_{(i+1/2,j+1/2)})L_r v^{n+1/2}_{r(i,j+1/2)} \quad (2a)$$
$$+ \lambda_{(i+1/2,j+1/2)}(L_z v^{n+1/2}_{z(i+1/2,j)} + \sigma_i v^{n+1/2}_{r(i,j+1/2)}/r_{i+1/2}),$$

$$(\tau^{n+1}_{zz(i+1/2,j+1/2)} - \tau^n_{zz(i+1/2,j+1/2)})/\Delta t$$
$$= (\lambda_{(i+1/2,j+1/2)} + 2\mu_{(i+1/2,j+1/2)})L_z v^{n+1/2}_{z(i+1/2,j)} \quad (2b)$$
$$+ \lambda_{(i+1/2,j+1/2)}(L_r v^{n+1/2}_{r(i,j+1/2)} + \sigma_i v^{n+1/2}_{r(i,j+1/2)}/r_{i+1/2}),$$

$$(\tau^{n+1}_{\theta\theta(i+1/2,j+1/2)} - \tau^n_{\theta\theta(i+1/2,j+1/2)})/\Delta t$$
$$= \lambda_{(i+1/2,j+1/2)}(L_r v^{n+1/2}_{r(i,j+1/2)} + L_z v^{n+1/2}_{z(i+1/2,j)}) \quad (2c)$$
$$+ (\lambda_{(i+1/2,j+1/2)} + 2\mu_{(i+1/2,j+1/2)})\sigma_i v^{n+1/2}_{r(i,j+1/2)}/r_{i+1/2},$$

$$(\tau^{n+1}_{rz(i,j)} - \tau^n_{rz(i,j)})/\Delta t = \mu^H_{(i-1/2,j-1/2)}(L_z v^{n+1/2}_{r(i,j-1/2)} + L_r v^{n+1/2}_{z(i-1/2,j)}),$$

(2d)

$$\sigma_i \rho_{(i-1/2,j+1/2)}(v^{n+1/2}_{r(i,j+1/2)} - v^{n-1/2}_{r(i,j+1/2)})/\Delta t$$
$$= L_r \tau^n_{rr(i-1/2,j+1/2)} + L_z \tau^n_{rz(i,j)}$$
$$+ (\sigma_i \tau^n_{rr(i-1/2,j+1/2)} - \sigma_i \tau^n_{\theta\theta(i-1/2,j+1/2)})/r_i,$$

(2e)

$$\sigma_j \rho_{(i+1/2,j-1/2)}(v^{n+1/2}_{z(i+1/2,j)} - v^{n-1/2}_{z(i+1/2,j)})/\Delta t = L_z \tau^n_{zz(i+1/2,j-1/2)}$$
$$+ L_r \tau^n_{rz(i,j)} + \sigma_j \tau^n_{rz(i,j)}/r_{i+1/2},$$

(2f)

where the superscripts n of variables represent the time discrete index, the subscripts i and j refer to spatial position index of nodes in the r- and z-directions, respectively, Δr is the time sampling step, L_r and L_z refer to space difference operator in r and z directions. The forward arithmetic averaging operators σ_i and σ_j are defined as $\sigma_i f_i = (f_i + f_{i+1})/2$ and $\sigma_j f_j = (f_j + f_{j+1})/2$, respectively, and the harmonic mean $\mu^H_{(i-1/2,j-1/2)}$ is defined as

$$\mu^H_{(i-1/2,j-1/2)} = 4/(1/\mu_{(i+1/2,j+1/2)} + 1/\mu_{(i+1/2,j-1/2)}$$
$$+ 1/\mu_{(i-1/2,j+1/2)} + 1/\mu_{(i-1/2,j-1/2)}).$$

(3)

At the interface liquid–solid, $\mu^H_{(i-1/2,j-1/2)}$ is set to zero as formulated in Moczo (1989).

To reduce the numerical dispersion during the computation period, the grid space needs to satisfy the following inequality:

$$\Delta_{\max} \le v_{\min}/10f_{\max},$$

(4)

where $\Delta_{\max}$ is the max grid spacing, $v_{\min}$ is the minimum velocity of the model and $f_{\max}$ is the maximum frequency of the source function (Falk et al. 1996; Aoi Fujiwara 1999). The stability condition for a second-order FD discretization is given by

$$v_{\max}\Delta t \left[\frac{1}{\Delta r^2} + \frac{1}{\Delta z^2}\right]^{1/2} < 1,$$

(5)

where $v_{\max}$ is the maximum velocity of the model (Randall et al. 1991).

The expression of the source time function selected for all the examples outlined is defined as

$$s(t) = \begin{cases} \frac{1}{2}\left[1 + \cos\frac{2\pi}{T_c}\left(t - \frac{T_c}{2}\right)\right]\cos 2\pi f_0\left(t - \frac{T_c}{2}\right), & 0 \le t \le T_c \\ 0, & t > T \end{cases},$$

(6)

where f_0 is the center frequency of the source, and T_c represents the source pulse width, with $T_c = 2/f_0$.

Variable grids FD method

Generally, the fracture aperture in a formation is less than 1 mm, and it requires at least a few FD grids to represent the propagation of acoustic waves in that fracture. Therefore, the FD grid spacing needs to be very small in a region with fractures, and the grid spacing should not exceed a few hundred microns. That very small grid spacing imposes strong restrictions on the length of the FD time-stepping in order to satisfy numerical stability. In a region away from fractures, the grid spacing does not need be as small since the typical dimension of any heterogeneities is substantially larger than that of fracture aperture. That means that models become oversampled in over those regions, and the time-stepping is also unnecessarily small, introducing computational overhead. We overcome those limitations using variable grid FD. That allows utilizing fine grids over the fractured area while using a reasonable (satisfying the usual requirements to prevent grid-dispersion) grids spacing in the non-fractured area. The use of variable grid FD thus allows reducing the amount of redundant computations due to oversampling. The variable grid spacing utilized herein consists of tying together several regions with different grid spacing as reported in Aoi and Fujiwara (1999). When using such an approach, the discretization with a second-order stencil does not require interpolation of the variables in the overlapping regions.

Figure 3 is a schematic diagram of using the variable grids to divide the formation with a horizontal fracture. The
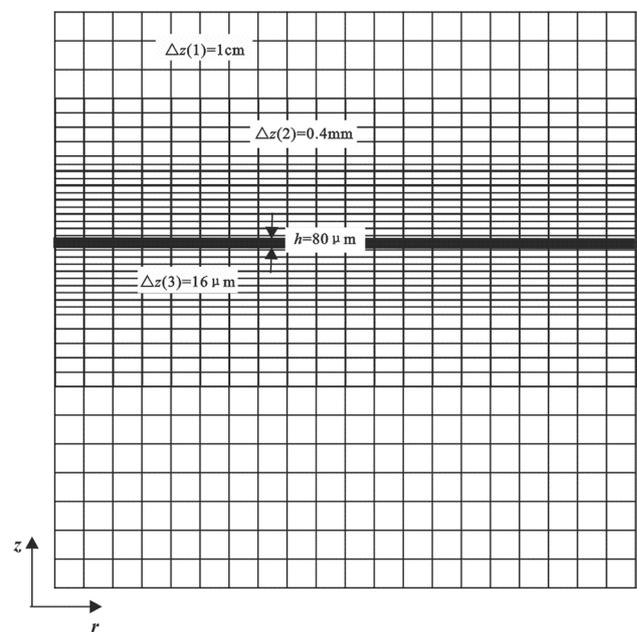
Fig. 3 The variable grids FD scheme in the formation with a fracture. The black thick lines represent the horizontal fracture

fracture extends along the *r*-direction, and the constant grid is enough to describe the fracture length. Therefore, the spacing is taken constant and with a value of 1 cm in the *r*-direction and it is variable in the *z*-direction. Depending on the grid spacing, the whole model is divided into three different computing regions, Zone 1, Zone 2 and Zone 3. In the *z* direction, the grid spacing of each of these three regions is denoted as $\Delta z(1)$, $\Delta z(2)$ and $\Delta z(3)$, respectively. The grid spacing $\Delta z(1)$ is set to 1 cm, satisfying the stability condition. The aperture *h* of the horizontal fracture is 80 μm. If five fine grid nodes are used to represent the fracture aperture, the grid spacing $\Delta z(3)$ is 16 μm. The grid spacing is adjusted according to the fracture aperture while meeting the best trade-off between spacing restrictions imposed by the stability condition and the efficiency of the computational program. In the vicinity of the fracture, another fine grid area is added, and its grid spacing $\Delta z(2)$ is 0.4 mm. In this way, the grid spacing is gradually increased from 1 cm to 16 μm. In gradually increased grid zones shown in Fig. 3, the finite difference coefficients in this paper are fixed to compute the wavefield records.

The computational regions with different grid steps have overlapping regions, which are used to avoid artificial reflections at the boundaries between the coarse and fine grid zones and to transfer numerical values between different computational regions. Because of the characteristics of staggered grids, the jump ratio of grid size is an odd integer. The time step Δt should be chosen according to the minimum grid step size. The detailed steps of the variable grids FD method are described as follows:

1. Add the monopole source to the normal stress component and use formula (2e) and (2f) to calculate the velocity component value of each computational region at the time $t = 1/2\Delta t$.
2. The velocity components of each overlapping part of coarse and fine grid regions are transferred to each other.
3. Use formula (2a), (2b), (2c) and (2d) to calculate the stress values of each region at the time $t = \Delta t$.
4. The stress component values of the overlapping parts of each computational region are transmitted to each other.

By repeating the above steps, we can calculate the wavefield at each time in the model.

Verifying the results of the variable grids FD method

In this part, the variable grids FD method is used to simulate the propagation of acoustic waves in a borehole, which is surrounded by a homogeneous elastic formation. Then, we verify the accuracy and effectiveness of the variable grid FD method by comparing waveforms computed with this method against those computed with real axis integration

(RAI) method (Tsang and Rader 1979). We select an elastic model utilizing the parameters labeled as Formation 1 and listed in Table 1. The outer boundaries are truncated with a non-splitting perfectly matched layer (PML) absorbing boundary (Wang and Tang 2003) to eliminate any spurious reflected waves generated by the model boundary, as shown in Fig. 1. The central frequency of the acoustic source is 10 kHz. The number of grid nodes in Zone 1 is 640 by 128, and the number of grids in both Zone 2 and Zone 3 is 112 by 128. The time step Δt we choose is 1/625 μs. Even if the variable grids FD method is adopted, the amount of computations is still very large. We alleviate that issue using GPU parallelism, taking advantage of the fact that the vast majority of computational tasks in our algorithm are array operations which are ideal for that type of computing architecture. For a matter of reference, we report the use of a GeForce GTX 1060 graphics card. The computation time for this example is 3.2 h. While, the RAI method analytically calculates the waveforms trace by trace and takes nearly 2 min for each trace.

Figure 4a compares the waveforms computed using the variable grid FD method against those obtained with the RAI method. The waveform recorded by the receiver represents the fluid pressure $p = -\tau_{rr}$ in the borehole. We placed eleven receivers located above the acoustic source. The distance between the first receiver and the source is 1 m. The interval between the receivers is 0.2 m. The black solid curves represent the waveform calculated by the variable grid FD method, and the red dashed curves represent the waveform simulated by the real axis integral method. As can be seen, the waveforms obtained by these two methods are very close. Effectively one can only observe a small difference in the amplitude with increasing propagation distances. Figure 4b shows the differences of the waveforms computed by two different methods. That difference (less than 10%) can be explained by the fact that the FD is a band-limited discrete representation of a medium, whereas the RAI method has spectral accuracy since it uses a discrete Fourier transform. This comparison demonstrates the validity of the variable grids FD method in the simulation acoustic waves propagating in a borehole.

Table 1 Physical parameters of the elastic formation

Formation	v_p m/s	v_s m/s	ρ kg/m³
Formation 1	3670	2170	2650
Formation 2	4000	2700	2650
Formation 3	6220	3455	2710
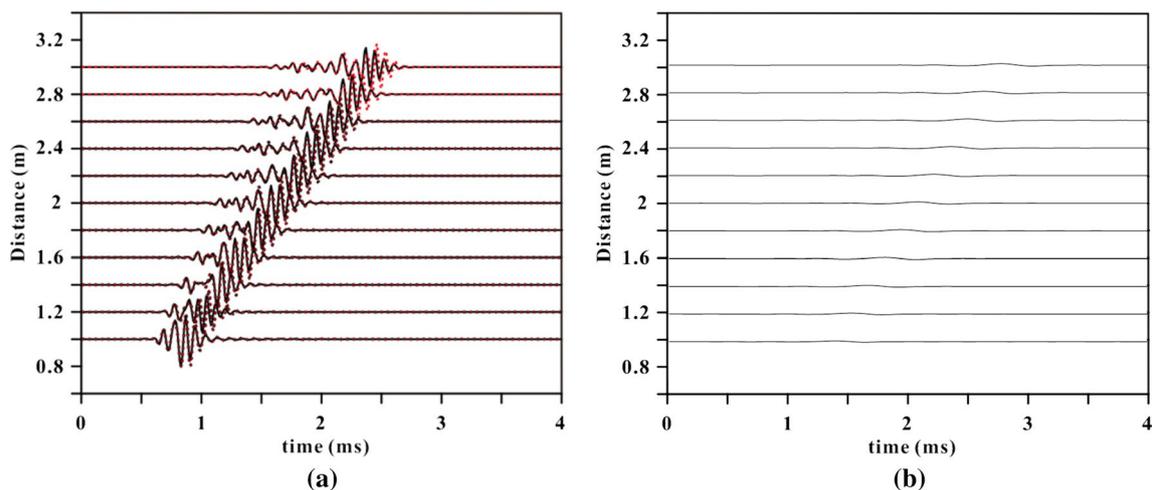
Fig. 4 a The comparisons of waveforms between the variable grid FD method (black solid curves) and the RAI method (red dashed curves). **b** The differences of the waveforms computed by two different methods

Attenuation of shear wave across the fracture

Wavefields and waveforms of waves propagating in a borehole intersecting an infinite horizontal fracture

The existence of fractures affects the shear wavefield and consequently recorded waveforms. We simulated the propagation of acoustic waves in a borehole crossing an infinite horizontal fracture, using the elastic parameters of Formation 1 (Table 1). Figure 5 shows the acoustic wavefield at different times in the formation. The wavefield is the field of the normal stress component τ_{rr}. The monopole source is located at $z = 500$ cm and at the axis of symmetry, i.e., $r = 0$ cm. The fracture aperture is 1 mm, which is located at $z = 300$ cm. In Fig. 5a, the acoustic wave is absorbed after it propagates along the elastic formation to the boundary of the model, showing that the PML boundary implemented in the program is effective. The refracted shear wave propagates upward along the borehole wall. As shown in Fig. 5b, the energy of shear waves is significantly damped by the existence of the fracture. Effectively, the fracture acts as a barrier for the propagation of shear-waves. For later times, the converted shear waves propagate within the medium, and above the fracture, resulting from waves propagating vertically along the borehole and refracting into the medium, as can be seen in Fig. 5c.

Figure 6 shows receiver records of the acoustic waveforms propagating in along borehole. The fracture is 2 m above the source. There are reflected waves in the waveform below the fracture, and the amplitude of shear wave above the fracture is very small. It can be seen that the fracture has a strong attenuation effect on the energy of the shear wave.

Moreover, the subsequent waveforms with large amplitudes are pseudo-Rayleigh wave and Stoneley wave in Fig. 6.

Relationship between the attenuation coefficient of shear wave and the fracture aperture

As shown in Fig. 6, the first break time of S-waves is about 1.3 ms at the source distance of 3 m. To show more clearly the amplitude variation of shear wave crossing fractures with different apertures, we set up a window selecting shear wave recordings between 1.2 ms and 1.45 ms, as depicted in Fig. 7. Even though the fracture aperture is very small ($h = 80$ μm), the amplitude of shear waves attenuates a lot, so the amplitude of the shear wave is very sensitive to the fracture. For different apertures of fractures, the amplitudes of shear waves are very close, but there are some differences between them. In this paper, we define the attenuation coefficient A of shear wave under the effect of fracture as

$$A = 1 - \frac{A_1}{A_2}, \tag{7}$$

where A_1 and A_2 are the maximum amplitudes of shear wave in the homogeneous formation and the fractured formation at the same source distance, respectively.

We repeated the numerical simulations for Formations 2 and 3 as listed in Table 1. Figure 8 shows the correlation between the fracture aperture and the attenuation coefficient of shear waves for each one of those formations. As can be seen, attenuation of shear waves depends linearly on the fracture aperture. In all the three cases, the correlation coefficient is greater than 0.993 which shows that a linear model is adequate to explain that relation. That linear relation between shear wave attenuation and fracture aperture is

Fig. 5 The acoustic wave-field at different times in the model containing a horizontal fracture. Figure **a**–**c** show the wavefield distribution at three instants of 0.9, 1.1 and 1.6 ms. The dashed line represents the borehole wall
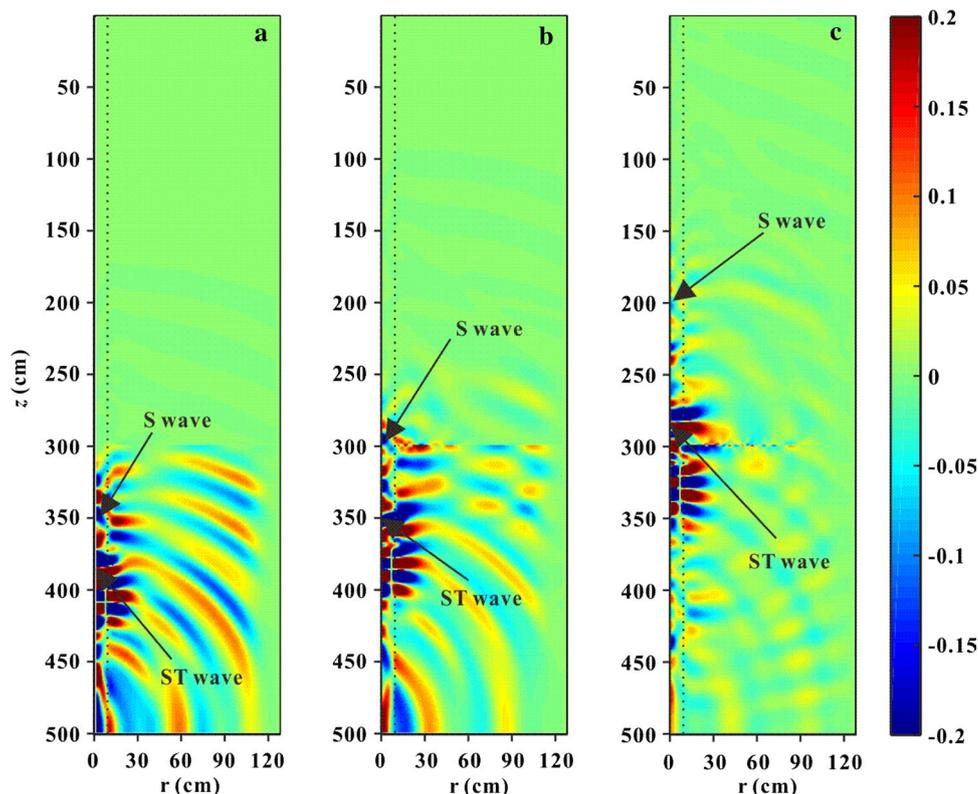
independent from the elastic parameters, i.e., the velocity of compressional and shear waves and the density. In addition, the larger the fracture aperture, the larger the stronger the attenuation is. However, the slope of the regression linear models is very small, in all the tested cases, which suggests that effectively the fracture aperture is a less dominant factor in the attenuation of shear waves (also observed on Fig. 7).

The intercept of the regressed linear equations is a positive number in all the three cases, showing that when the aperture of the fracture is zero, the fracture acts on the shear wave as an interface, causing attenuation of shear waves.

For different elastic formations, the wave impedance differs between elastic formations and the fracture fluid are different, leading to differences in shear wave scattering
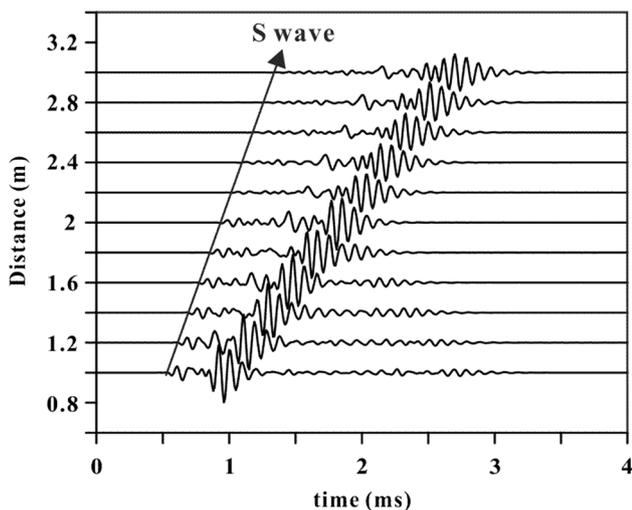
Fig. 6 The receiver recorded acoustic waveforms in the borehole. The receiver array has eleven receivers located above the acoustic source. The distance between the first receiver and the acoustic source is 1 m. The interval between receivers is 0.2 m
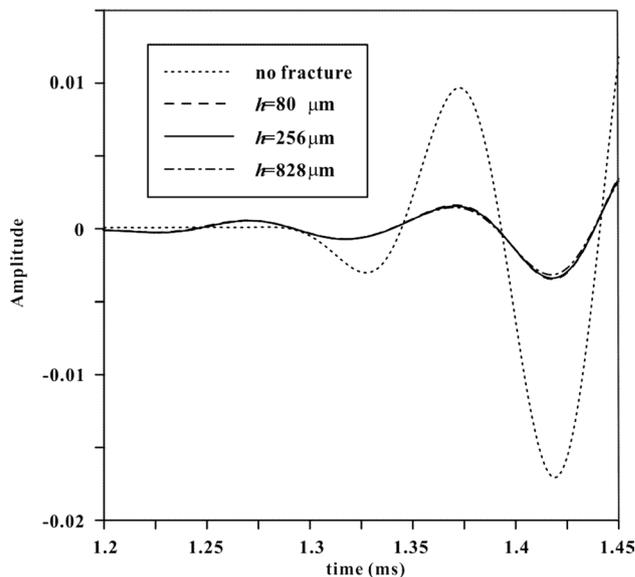
Fig. 7 The amplitudes of shear waves with different fracture apertures. We set up a window selecting shear wave recordings between 1.2 ms and 1.45 ms at the source distance of 3 m
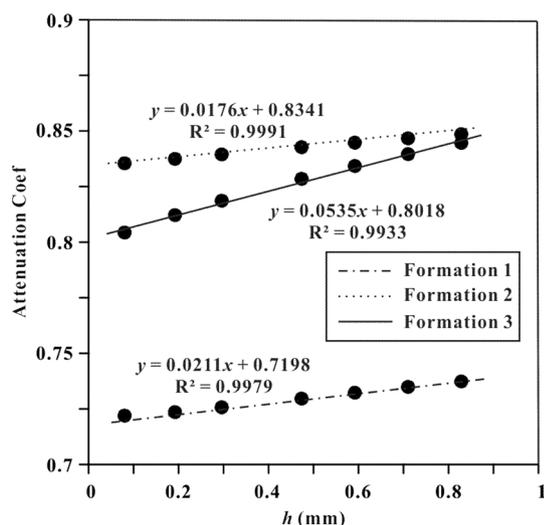
Fig. 8 The linear relationship between the fracture aperture and the attenuation coefficient of shear wave with different elastic formations. The symbol y represents the attenuation coefficient of shear wave, x represents the fracture aperture, and R^2 is the correlation coefficient

and mode conversion by the fracture. Therefore, the slope and intercept of linear equations are necessarily different for different formations. In these three different formations, the attenuation coefficient of shear wave exceeds 0.71. Consequently, one can conclude that identification of fractures through shear wave attenuation is very effective.

Attenuation coefficient of shear waves at different distances from fracture

The shear waves cannot propagate through fractures filled with water. When shear waves hit a fracture, a fraction of their energy will be converted into energy of other modes. Receivers above and close to a fracture, record waves that are superimposed by waves of other modes and converted shear waves. Paillet (1980) proposed that the shear wave mode conversion increases the shear wave amplitude recorded near fractures. The farther the distance to the source, the larger the decrease in the amplitude of shear waves (Matuszyk et al. 2013). In this section, we investigate the dependence of shear wave attenuation with distance from a fracture.

Figure 9 shows the amplitude of shear wave at different distances to the fracture opening. The aperture h of the single fracture is 512 μm. In this example, we take the elastic parameters defined for the Formation 2, listed in Table 1. In Fig. 9a–d, the vertical distance between the receiver and the fracture plane is 0.2 m, 0.3 m, 0.7 m and 0.8 m, respectively. It can be seen that the longer the distance from the receiver to the fracture, the more the

attenuation of shear wave amplitude. Therefore, the wave mode conversion can reduce the wave attenuation recorded by the shear wave window near the fracture.

Figure 10 shows the shear wave attenuation coefficient corresponding to the receiver with different source distances. Figure 10 is obtained utilizing the same parameters of the elastic formation and fracture as those used to compute the waveforms depicted in Fig. 9. As expected, the attenuation coefficient of shear wave increases with increasing distance between the receiver and the fracture up to a certain limit beyond which the attenuation coefficient is bounded. In our example, upper limit is around 0.85, corresponding to a distance of 2.6 m. Effectively that means, when the distance from the receiver to the fracture is greater than 0.6 m, the converted shear wave recorded by the receiver has been separated from waves of other modes.

Attenuation coefficients of shear waves with different borehole radii

The radius of borehole also affects acoustic propagation in the borehole. Therefore, we investigate the effect of borehole radius on the attenuation of shear wave in this part. Figure 11 shows the relationship between the attenuation coefficient of shear waves and fracture aperture for different borehole radius. The distance between source and receiver is 2.2 m, 2.4 m, 2.8 m and 3 m, respectively. As we discussed above, when the source is closer to the fracture, the shear wave window intercepts both shear wave and waves of other modes. In the model, the borehole radius is increased from 10 to 15 cm at a 1 cm interval. The elastic formation we chose is Formation 2 listed in Tab. 1. When the receiver is far away from the fracture, the shear wave attenuation coefficient calculated does not change much with the increase in borehole radius and are all around 0.87. This indicates that the variation of borehole radius has little effect on the proportion of converted shear wave energy. When the receiver is closer to the fracture, the attenuation coefficient of the shear wave decreases with the increase in the radius of the borehole. Especially when the distance between receiver and fracture is only 0.2 m, there is a good linear relationship between attenuation coefficient and borehole radius. With the increase in borehole radius, the difference of shear wave attenuation between the receivers near and far from the fracture increases. This indicates that the proportion of converted wave energy of other modes increases with the increase in borehole radius.
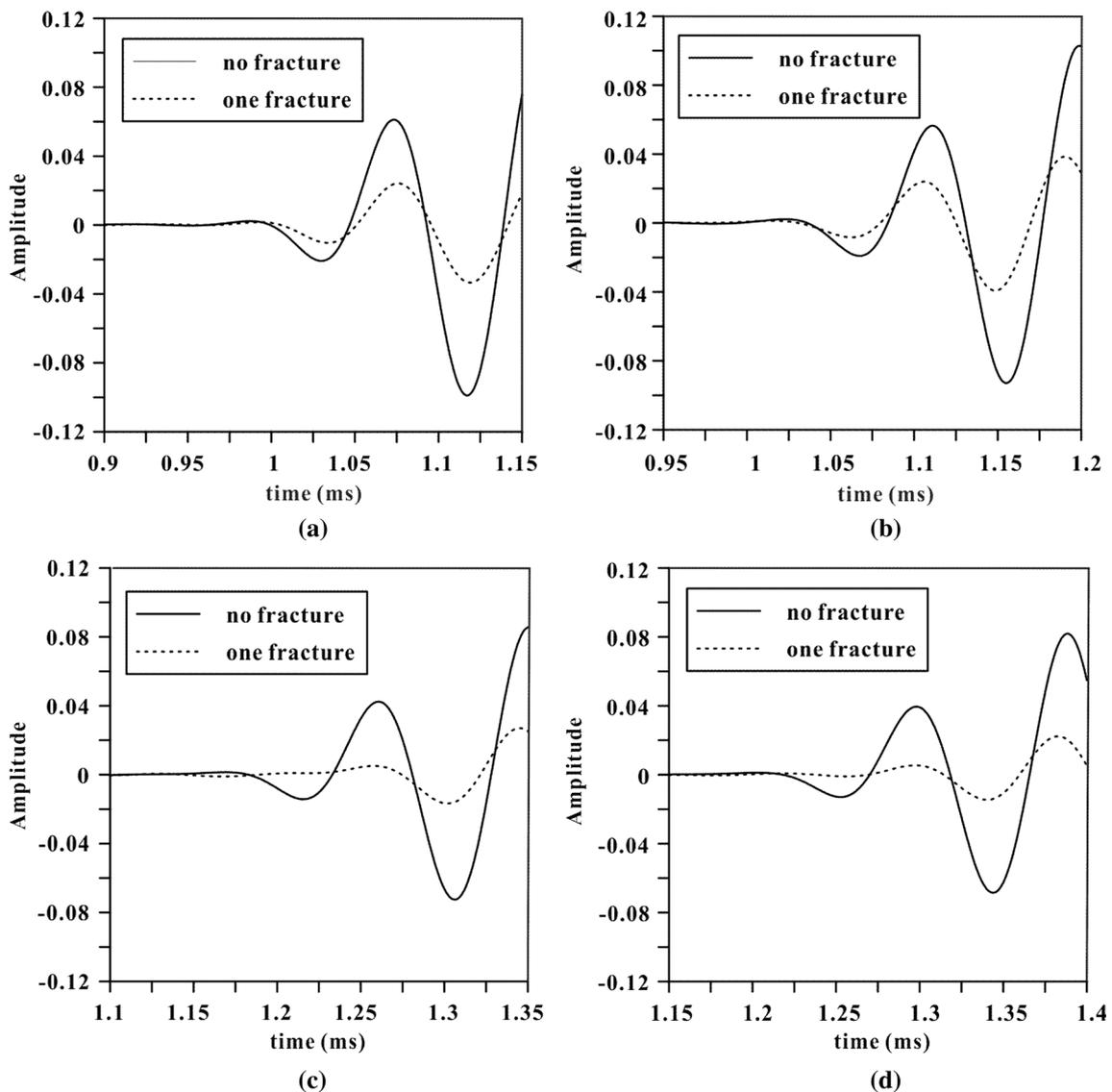
Fig. 9 Recordings of the shear wave at **a** 0.2 m, **b** 0.3 m, **c** 0.7 m and **d** 0.8 m from the fracture plane. The solid curve represents no fracture in the model, and the dashed curve represents a fracture inserting the borehole in the model

Attenuation of shear wave with different extended distances of fractures

Geological formations can present fractures caused by natural stress fields or caused by mechanical forces resulting from human-related activity. Conceptually, those fractures extend a finite distance L in the radial direction from the borehole axis. In particular, the extended distance of hydraulic fractures is generally very short. In this section, we simulate the propagation of shear wave in a finite fracture model, and we investigate the relationship between the attenuation coefficient of shear waves and the radial

extended distance L of fracture. Figure 12 shows waveforms of shear wave recordings computed for different horizontal extension of a fracture. The fracture aperture is set to 1 mm. The distance between the source and the fracture is 2 m, and the distance between the source and the receiver is 2.8 m. It is found that the amplitude of shear wave decreases with increasing L. When L is 0.6 m, the amplitude of shear wave is similar to that of an infinite fracture. Figure 13 shows the relationship between the attenuation coefficient of shear wave and the extended distance L of fracture. When L is 0.2 m, the attenuation coefficient of shear wave decreases to 0.12. It can be seen that
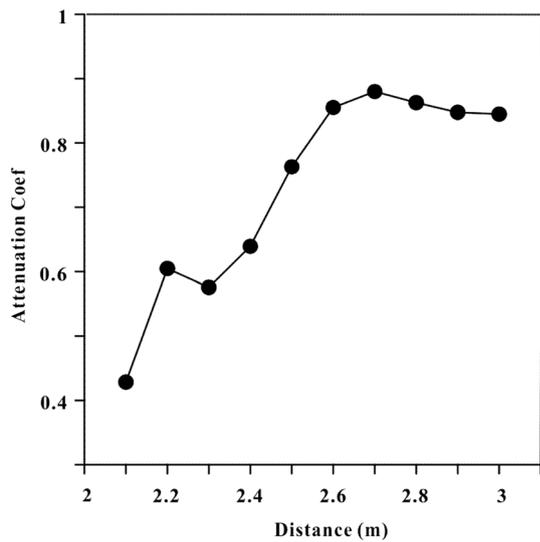
Fig. 10 Variation of the shear wave attenuation coefficient with distance between a receiver and the plane of a fracture. The abscissa represents the source distance of the receiver
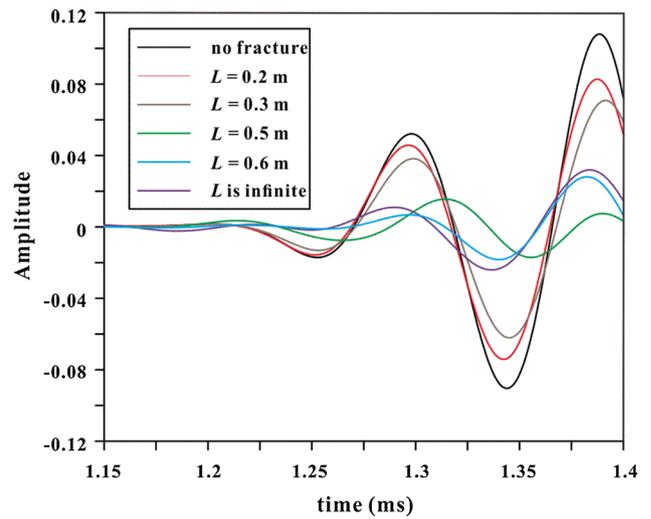
Fig. 12 Waveforms of shear wave corresponding to the fractures with different extended distances
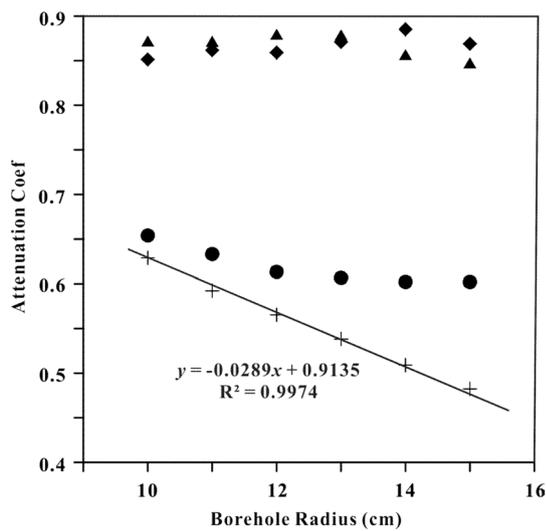
Fig. 11 The relationship between the attenuation coefficient of shear waves and borehole radius in the presence of a horizontal fracture. The cross, circle, triangle and diamond represent the attenuation coefficients corresponding to source-receiver distances of 2.2 m, 2.4 m, 2.8 m and 3 m, respectively
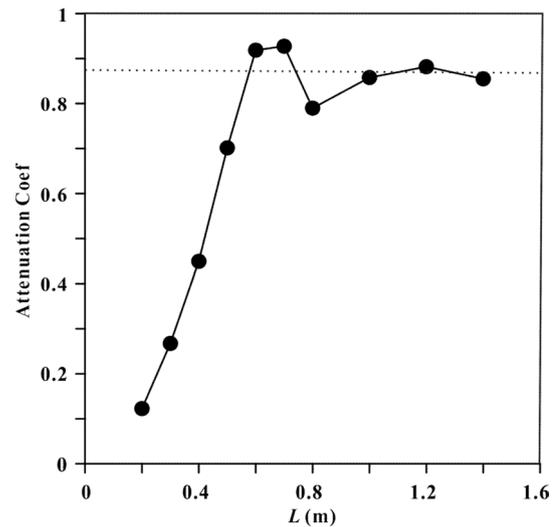
Fig. 13 Relationship between the attenuation coefficient of shear wave and the extended distance of fracture. The dashed line represents the attenuation coefficient corresponding to an infinitely extended fracture

Discussion and conclusion

the fracture has little effect on the amplitude of shear wave when the extended distance is very short. From Fig. 13, one can observe that the attenuation coefficient of shear waves for lateral extensions greater than 0.6 m converges to that of resulting from a fracture with finite extension.

In this paper, the variable grid FD method is used to simulate full wave acoustic logging in a formation with a horizontal fracture. The effects of fracture aperture, distance between a receiver and a fracture, borehole radius and extended distance of fracture on attenuation coefficient of shear waves are investigated. As demonstrated, the variable grid FD method is adequate to conduct our investigation as it accommodates easily the much smaller

dimensions of a fracture aperture without increasing computational overhead. The waveforms simulated by variable grid FD method match very closely the results calculated by RAI method very well, demonstrating the accuracy of the variable grid FD method. After a shear wave crosses a horizontal fracture, the amplitude of shear wave attenuates significantly due to the reflection losses and mode conversion of shear wave caused by fractures. We obtained numerically some relations of the effects of fracture aperture, source distance of receiver, borehole radius and extended distance of fracture on the attenuation coefficient of shear wave.

Our results show that there is a clear linear relationship between the attenuation coefficient of shear waves and the fracture aperture. In all the tested elastic formations, the intercepts of the linear equations are all very large, indicating that fractures cause strong attenuation of shear waves. In addition, slopes of the linear relations are relatively small, showing that the attenuation coefficient of shear wave is not sensitive to fracture aperture. Effectively, that means that it is difficult to determine the fracture aperture from the attenuation of shear waves. The elastic parameters of the formation also impact on the attenuation of shear waves caused by fractures. For receivers far away from the fracture, the shear wave attenuation is not affected by changing borehole radius. For an infinite horizontal fracture, shear waves are strongly attenuated regardless of the elastic parameters of the formation, borehole radius and fracture aperture. Therefore, strong attenuation of shear waves can be used to characterize fractures with acoustic logging.

Due to modes conversion, the attenuation of shear waves near the fracture is not observable as it is masked by the energy of the different modes. As the distance between the receiver and the fracture increases, the attenuation coefficient of shear wave increases gradually. Also because of the effect of mode conversion, the attenuation coefficient of the shear wave decreases increasing radius of the borehole. This aspect is particularly relevant, as it demonstrates that for carrying out fracture characterization from shear wave attenuation, the recordings need to be free from recorded converted modes. Hence, the recording system needs to be designed for that purpose.

A very significant finding, from this study, is demonstrating that attenuation of shear waves increases with increasing lateral extension of fractures. Therefore, the attenuation coefficient of shear wave can provide information about the extended distance of fracture. In the future, we plan to extend this work in the characterization of fractured formations envisioning applications to exploration of unconventional resources.

References

Aoi S, Fujiwara H (1999) 3D finite-difference method using discontinuous grids. Bull Seismol Soc Am 89(4):918–930

Bakku SK, Fehler M, Burns D (2013) Fracture compliance estimation using borehole tube waves. Geophysics 78:D249–D260

Chen Y, Chew W, Liu Q (1998) A three-dimensional finite difference code for the modeling of sonic logging tools. J Acoust Soc Am 103(2):702–712

Falk J, Tessmer E, Gajewski D (1996) Tube wave modeling by the finite-difference method with varying grid spacing. Pure Appl Geophys 148(1–2):77–93

Goldberg D, Gant WT (1988) Shear-wave processing of sonic log waveforms in a limestone reservoir. Geophysics 53(53):9

Groenenboom J, Falk J (2000) Scattering by hydraulic fractures: finite-difference modeling and laboratory data. Geophysics 65:612–622

Guan W, Hu H, He X (2009) Finite-difference modeling of the monopole acoustic logs in a horizontally stratified porous formation. J Acoust Soc Am 125(4):1942–1950

Hornby BE, Johnson DL, Winkler KW, Plumb RA (1989) Fracture evaluation using reflected Stoneley-wave arrivals. Geophysics 54:1274–1288

Kostek S, Johnson DL, Randall CJ (1998) The interaction of tube waves with borehole fractures. Part I Numer Models Geophys 63:800–808

Kostek S, Johnson DL, Winkler KW, Hornby BE (1998) The interaction of tube waves with borehole fractures. Part II Anal Models Geophys 63:809–815

Matuszyk PJ, Torres-Verdín C, Pardo D (2013) Frequency-domain finite-element simulations of 2D sonic wireline borehole measurements acquired in fractured and thinly bedded formations. Geophysics 78(4):D193–D207

Minato S, Ghose R (2017) Low-frequency guided waves in a fluid-filled borehole:simultaneous effects of generation and scattering due to multiple fractures. J Appl Phys 121(10):104902

Moczo P (1989) Finite-difference technique for SH-wave in 2-D media using irregular grids—application to the seismic response problem. Geophys J Int 99:321–329

Morris RL, Grine DR, Arkfeld TE (1964) Using compressional and shear acoustic amplitudes for the location of fractures. J Pet Technol 16(6):623–632

Paillet FL (1980) Acoustic propagation in the vicinity of fractures which intersect a fluid-filled borehole. SPWLA 21st Annual Logging Symposium, Society of Petrophysicists and Well-Log Analysts

Randall CJ, Scheibner DJ, Wu PT (1991) Multipole borehole acoustic waveforms: synthetic logs with beds and borehole washouts. Geophysics 56(11):1757–1769

Stephen RA, Pardo-Casas F, Cheng CH (1985) Finite-difference synthetic acoustic logs. Geophysics 50:1588–1609

Tang XM, Cheng CH (1993) Borehole stoneley wave propagation across permeable structures. Geophys Prospect 41:165–187

Tsang L, Rader D (1979) Numerical evaluation of the transient acoustic waveform due to a point source in a fluid-filled borehole. Geophysics 44:1706–1720

Virieux J (1986) P-SV wave propagation in heterogeneous media: velocity–stress finite-difference method. Geophysics 51:889–901

Wang T, Tang X (2003) Finite-difference modeling of elastic wave propagation: a nonsplitting perfectly matched layer approach. Geophysics 68(5):1749–1755

Yan SG, Xie FL, Gong D, Zhang CG, Zhang BX (2015) Borehole acoustic fields in porous formation with thin fracture. Chin J Geophys 58:307–317

Yan BP, Wang SX, Ji YZ, Huang XG, da Silva NV (2020) Frequency-dependent spherical-wave reflection coefficient inversion in acoustic media: theory to practice. Geophysics 85(4):R425–R435

Zlatev P, Poeter E, Higgins J (1988) Physical modeling of the full acoustic waveform in a fractured, fluid-filled borehole. Geophysics 53(9):1219–1224

RESEARCH ARTICLE - APPLIED GEOPHYSICS

Complex lithology prediction using mean impact value, particle swarm optimization, and probabilistic neural network techniques

Yufeng Gu[1] · Zhongmin Zhang[2] · Demin Zhang[2] · Yixuan Zhu[2] · Zhidong Bao[3] · Daoyong Zhang[1]

Abstract

Lithology prediction is a fundamental problem because the outcome of lithology prediction is the critical underlying data for some basic geological work, e.g., establishing stratigraphic framework or analyzing distribution of sedimentary facies. As the geological formation generally consists of many different lithologies, the lithology prediction is always viewed as a tough work by geologists. Probabilistic neural network (PNN) shows high efficiency when solving pattern recognition problem since learning data do not need to do any pre-training of learning data and calculation results are universally reliable, and then, this model could be considered as an effective solution. However, there are two factors that seriously limit the PNN's performance: One is existence of the interference variables of learning samples, and the other is selection of the window length of probability density distribution. In view of adverse impact of those two factors, two techniques, mean impact value (MIV) and particle swarm optimization (PSO), are introduced to improve the PNN's calculation capability. Thus, a new prediction method referred as MIV–PSO–PNN is proposed in this paper. The proposed method is validated by three well-designed experiments, and the corresponding experiment data are recorded by two cored wells of the LULA oilfield. For the three experiments, prediction accuracies of the results provided by the proposed method are 81.67%, 73.34% and 88.34%, respectively, all of which are higher than those provided by other comparative approaches including backpropagation (BP), PNN, and MIV-PNN. The experiment results strongly demonstrate that the proposed method is capable to predict complex lithology.

List of symbols

$p(\mathbf{y}\|\mathbf{X}_i)$	The probability density value of $\mathbf{y}$ for the cluster $\mathbf{X}_i$
n	The number of variables, or the number of samples in Eq. (2), or the number of output units in Eq. (6)
$\mathbf{x}_{ik}$	kth sample of $\mathbf{X}_i$
l	The dimension of sample, or the number of hidden units in Eq. (6) and this number is an integer
σ	The smooth factor, also named as "spread"
$\mathbf{B}_{fn}$	The test data
m	The number of learning samples, or the number of input units in Eq. (6)
f	The number of test samples
$\mathbf{D}_{fm}$	The matrix of the Euclidean distance
$\mathbf{P}_{fm}$	The matrix of the probability density
$\mathbf{S}_{fz}$	The matrix of the total probability density
$\mathbf{C}_{fz}$	The ratio matrix containing the pattern recognition ratio of each test sample for each cluster
β_1^1	The enlarged coefficient
β_1^2	The reduced coefficient
$\mathbf{H}$	The learning data without interference variables
$\mathbf{R}$	The vector of MIV results
$f_{\mathrm{MIV}}(\cdot)$	The MIV results
$\mathbf{\Gamma}$	The spread population
$\mathbf{\sigma}_i$	ith spread seed
$\mathbf{P}_{fm}^i$	The matrix of the probability density scaled by ith spread seed

✉ Yufeng Gu
aaaaa3388@126.com

1 Strategic Research Center of Oil and Gas Resources, Ministry of Natural Resources, Beijing, China

2 Sinopec Exploration & Production Research Institute, Beijing, China

3 College of Geosciences, China University of Petroleum (Beijing), Beijing, China

$\mathbf{\Delta}_i^j$ The iteration result of ith spread seed; its corresponding predicted result has the smallest error in the first j iteration times

$\mathbf{\sigma}_i^j$ The iteration result of ith spread seed; its corresponding predicted result has the smallest error in jth iteration time

$\mathbf{O}_{if}^j$ The predicted result corresponding to the ith spread seed in jth iteration time

$\mathbf{O}_{f\_std}$ The standard result

t The iteration time

$\mathbf{\Delta}_g^j$ The iteration result of one spread seed; its corresponding predicted result has the smallest error in the first j iteration times

$\mathbf{\sigma}_\Gamma^j$ The iteration result of one spread seed; its corresponding predicted result has the smallest error in jth iteration time

$\mathbf{O}_{\Gamma f}^j$ The predicted result corresponding to one spread seed and has the smallest error in jth iteration time

$\mathbf{W}_i^j$ The iteration step of ith spread seed in jth iteration time

ω The inertia weight

c_1 The acceleration coefficient

c_2 The acceleration coefficient

r_1 Random value limited in [0,1]

r_2 Random value limited in [0,1]

σ_{max_1} The left limit of spread seed

σ_{max_2} The right limit of spread seed

W_{max} The limit of iteration step

$\min(\cdot)$ The function acquiring the minimum

$\|\cdot\|_2$ The function calculating L2 norm

ω_{max} The maximum inertia

ω_{min} The minimum inertia

t_{max} The maximum iteration time

v_{nor} The normalized logging value

v The logging value

v_{max} The maximum value of one log

v_{min} The minimum value of one log

a An integer value and its taking range is [1,10]

$\mathbf{\sigma}_{\#1}$ The spread vector used for prediction of #1 well

$\mathbf{\sigma}_{\#2}$ The spread vector used for prediction of #2 well

Introduction

Lithology information is the hugely important analysis reference for some basic geological work such as establishing stratigraphic framework, reconstructing model of sedimentary facies, or plotting sand body distribution. Many geologists have emphasized the significance of acquiring lithology information in their cases (Bogdanov et al. 1998; Akhmanov et al. 2003; Kominz et al. 2011; Chetelat et al.

2013; Mirzaei-Paiaman et al. 2018, 2019a, b). Lithology prediction is an effective approach to gain the information of lithology distribution, since the lithology types along the well can be uncovered by the predicted results. Thus, how to create excellent lithology prediction models and improve their calculation capabilities has been a critical research topic in the field of geology (Roberts et al. 1993; Inaba et al. 2001; Mirzaei-Paiaman and Salavati 2012; Bosch et al. 2013; Mirzaei-Paiaman et al. 2015; Mirzaei-Paiaman and Saboorian-Jooybari 2016; Zhang et al. 2018; Zhu et al. 2018).

The classic lithology prediction model is the crossplot which is a two-dimensional coordinate (Porter et al. 1969). Each dimension represents a kind of log. The applied logs should have significant logging responses on all kinds of predicted lithologies. Because each kind of crossplot has its special prediction capability, any lithology type, to a certain extent, can be predicted in this way. Guo et al. (2015) have tried to predict lithologies of the gas and brine formations in the Sichuan Basin under analyzing the results provided by some different crossplots. However, the crossplot is only feasible when all kinds of predicted lithologies have their own distinct logging responses, and then, it is only suitable for the prediction of simple lithology rather than complex lithology. Simple lithology is a relative concept of complex lithology. If lithologies have their own distinct logging responses, they belong to the simple ones, e.g., conglomerate, sandstone, mudstone, limestone, and dolomite (Jacobson 1988; Jacobson and Wyatt 1993). However, most reservoirs presenting with strong heterogeneity are comprised of many different lithologies, and these lithologies are usually the subdivisions of one simple lithology. For example, the limestone, a simple lithology component of carbonate reservoir, can be divided into spherulite, laminate, grainstone, packstone, etc. (Thompson et al. 2015). These subdivision lithologies generally have similar logging responses, implying that the prediction could not be completed by the crossplot effectively. Thus, they are defined as the complex lithologies (Jacobson and Wyatt 1993; Djafarov et al. 2004; Kou et al. 2010). With more and more petroleum resources are discovered in the carbonate and tight sandstone reservoirs which widely consist of complex lithologies, the prediction of complex lithology gradually becomes a hot topic in recent years. Currently, the main tool for the complex lithology prediction is neural network.

Many highly efficient networks have been modified and applied in the logging interpretation research, including recurrent neural network (RNN), semi-supervised deep network, gradient boosting decision tree (GBDT), XGBoost, and LightGBM, whereas in this paper probabilistic neural network (PNN) is proposed to use to make a solution for complex lithology prediction (Specht 1990; Xie et al. 2018; Dev and Eden 2019; Sun et al. 2019;

Zhang et al. 2018; Zhu et al. 2018, 2020). The reason is that compared to those networks which universally employ a lot of empirical parameters to complete their training processes, the advantage of PNN is that it only has to set one empirical parameter (window length of probability density distribution) before modeling, extremely reducing the workload of parameter tuning. PNN utilizes probability density results to complete recognition. The used learning data do not need to finish some kinds of mandatory transformation before prediction such as normalization or coordinate mapping, and even allows few incorrect samples to exist due to the application of probability density distributions. Kakouei et al. (2014) used five neural networks to identify the lithological facies of four wells of South Pars field (Iran) and claimed that the PNN has the better performance. However, there are two drawbacks that limit computing efficiency of the PNN, which are the existence of inference variables of the learning samples, and the selection of window length of the probability density distribution. Interference variables involved in the learning samples can cause the size and the center position of some manually defined learning data clusters to be unreasonable. Chi et al. (2014) pointed out this issue in their study based on the experiments and discussion of simultaneous fault detection. The window length of the probability density distribution directly determines the probability density value, and therefore, the selection of the window length is a key for the PNN calculation. Chen et al. (2009) did three experiments regarding the pattern recognition capability of the PNN and proved that using single value to scale all data clusters of the PNN will lead to inaccurate or totally incorrect predicted results. In order to improve the PNN's prediction capability, two techniques, mean impact value (MIV) and particle swarm optimization (PSO), are introduced in this paper. MIV can rapidly find out the significant variables for computing object, and is simpler to be realized and modified on computer under the comparison with factor analysis (FA) and principal component analysis (PCA) (Barrow 2009; Ma 2011; Qi et al. 2016). Yong-Yan and Wang (2011) applied the MIV technique to enhance the calculation capability of support vector machine (SVM) and successfully obtained the expected results in the prediction. PSO is a favorable algorithm for the multi-objective optimization problem, since compared to simulated annealing algorithm (SAA) and genetic algorithm (GA), its iteration is simpler, implying that it is more feasible to be integrated with other models and could be more time saving on optimization (Kennedy and Eberhart 1995; Trelea 2003; de Almeida 2010; Sahoo and Jha 2016). Ahila et al. (2012) attempted to apply PSO-PNN model to solve a classification problem regarding power system disturbance and, according to the analysis of predicted results, demonstrated the superiority of PSO'

utilization. Therefore, with the applications of MIV and PSO, the problem of two drawbacks showing on PNN's prediction performance will be addressed perfectly.

Given the advantages of MIV and PSO for the calculation of PNN, an improved PNN model referred as MIV-PSO-PNN is proposed, which could provide a better solution for the complex lithology prediction. In the next chapter, methodology of the proposed method will be described in details. The topic of the third chapter includes source of the validation data, validation process, and discussion of the validation results. Finally, the features of the proposed method will be concluded in the last chapter.

Methodology

Theoretical equations

The pattern recognition mechanism of the PNN complies with the Bayesian classification rule (Goldszmidt and Moises 1997). If the original data $\mathbf{X}$ can be divided into z clusters $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_z$, the type of test sample $\mathbf{y}$ will be determined by the formula below:

$$\begin{cases} p(\mathbf{y}|\mathbf{X}_i) > p(\mathbf{y}|\mathbf{X}_j) \\ i \neq j \end{cases} \Rightarrow \mathbf{y} \in \mathbf{X}_i \tag{1}$$

where $p(\mathbf{y}|\mathbf{X}_i)$ is the probability density value of $\mathbf{y}$ for the cluster $\mathbf{X}_i$, $i = 1, 2, \ldots, z$.

The probability density value satisfies the following estimation equation (Parzen 1962):

$$p(\mathbf{y}|\mathbf{X}_i) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{(2\pi)^{l/2}\sigma^l} e^{\frac{-\|\mathbf{y}-\mathbf{x}_{ik}\|^2}{2\sigma^2}} \tag{2}$$

where n is the number of samples of $\mathbf{X}_i$, $\mathbf{x}_{ik}$ is kth sample of $\mathbf{X}_i$, l is the dimension of $\mathbf{y}$ or $\mathbf{x}_{ik}$, σ is the smooth factor, determining the window length of probability density distribution, and since MATLAB is selected as the computing environment in this paper, the program term "spread" will be used to describe this factor in the following paragraphs.

Based on Eq. (2), Specht (1990) proposed the PNN model that can classify patterns of the test data intelligently. The PNN contains four layers: the input, the pattern, the sum, and the output. The input is loading learning and test data. In the pattern, the learning data will be clustered firstly, then, probability density distributions for all the clusters will be established, and at last, probability density values of the test data will be calculated by Eq. (2). The function of the sum is to obtain the total probability value of each test sample for each cluster. Finally, under a competitive calculation, the predicted results of the test data will be provided by the output. Supposing that the learning data are $\mathbf{A}_{mn}$ and can

be divided into z clusters. The amount of samples of each cluster is assigned as v_i $(i = 1, 2, \ldots, z)$. Then, the predicted results $\mathbf{O}_f$ can be obtained (the relevant derivation is shown in "Appendix 1").

The pattern is considered as the most important layer for the PNN because all the probability density values which directly determine the accuracies of the predicted results are figured out here (Specht 1990; Zhang 2008). Equation (2) indicates that spread mainly controls the change of the probability density values. As such, the selection of spread is a key point for the PNN calculation. Figure 1a–c shows the different probability density distributions when the PNN adopts different spreads. Small spread will cause the distributions to be very small as shown in Fig. 1a. In this way, the calculated probability density values will be too small to enable the PNN to generate the reliable predicted results.

If a larger spread is applied, the formed distributions will become too large or even be integrated with each other as shown in Fig. 1c, and then, the PNN also will be ineffective to predict the test data (Trelea 2003; Chen et al. 2009; Chen and Chu 2009; Ahila et al. 2012). Figure 1b manifests that the PNN is capable to establish the reasonable distributions when using a suitable spread. Note that three spread values exampled in Fig. 1a–c are just used to make a demonstration for the significance of spread selection, having no practical implications.

Although the PNN is functional in recognizing patterns when utilizing a suitable spread, the reliability of the calculated result still cannot be guaranteed due to its two drawbacks. One is the interference variables involved in the learning data (Chi et al. 2014). Figure 1d shows an illustration. In this figure, all the original distributions are established
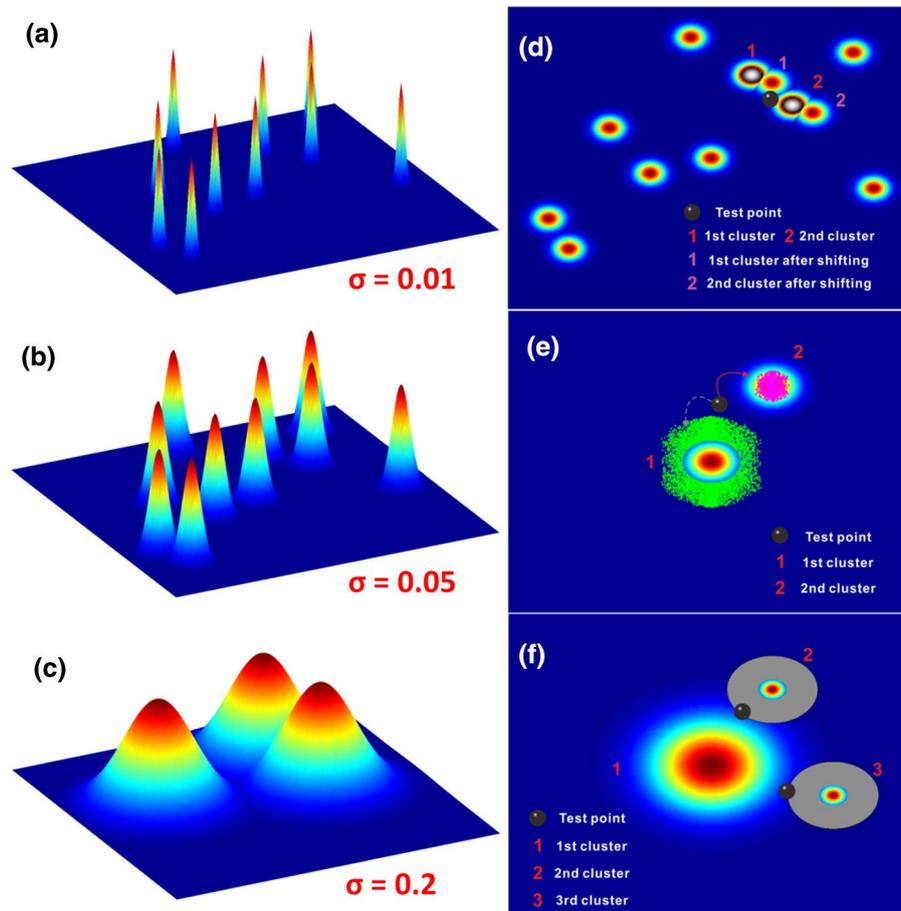
Fig. 1 Different situations of probability density distributions and some problems of PNN in data process. **a**–**c** Different probability density distributions, respectively, when PNN uses different spreads (σ) 0.01, 0.05, and 0.2. **d** Shift of two clusters, caused by interference variables of the learning data. **e** Misrecognition of test point, caused by the usage of unreasonably scaled probability density distributions. The green and pink learning data points belong to the 1st and 2nd clusters, respectively. The gray and red arrows, respectively, indicate

the test point should belong to the 1st cluster, but actually it will be classified into the 2nd cluster by PNN. **f** Misrecognition of two test points, caused by unbalance amount of the learning data in each cluster. The gray areas are the original probability density distributions of the 2nd and 3rd clusters, and the rainbow areas are the real distributions of the 1st, 2nd, and 3rd clusters. The change of the distributions of the 2nd and 3rd clusters is caused by a large reduction of the learning data of these two clusters

by the learning data. The learning data are assumed to be comprised of 8 logs (or 8 variables). Apparently, for the test point (or test sample), the 2nd cluster can generate the larger probability density value. Nonetheless, after a new log which can be viewed as an interference variable is added into the learning data, the locations of the 1st and the 2nd cluster are shifted, and then, the larger probability density value of the test point will be provided by the 1st cluster. The addition of that interference log results in an incorrect recognition of the test point. Therefore, the interference variables of learning data must be removed to ensure the reliability of the calculated results. As mentioned before, the MIV technique is introduced to solve this problem. This technique will treat one variable of the learning data as the calculation object firstly, then enlarging and reducing this variable's data with the same ratio so as to create two new kinds of learning data. One learning data contain the bigger learning data of that variable, and the other has the smaller learning data. After processing two new kinds of learning data, the PNN will work out two calculated results. If there is a standard result and the difference between the standard and the calculated result is named as Impact Value (IV) of that variable, two IVs will be obtained. Changing the learning data of that variable in the same way but with different ratios will produce different IVs. Thus, the mean of those IVs, which is called MIV, will be known. If the MIV is smaller than a defined threshold, the relevant variable will be considered as the interference one, and accordingly, its learning data will be removed (Yong-Yan and Wang 2011; Qi et al. 2016).

Now assuming that the first variable of learning data $\mathbf{A}_{mn}$ is the calculation object. Correspondingly, the enlarged and the reduced learning data will be determined, which are $\mathbf{A}'_{mn}$ and $\mathbf{A}''_{mn}$, respectively (the relevant derivation is shown in "Appendix 2"). Two calculated results corresponding to the two kinds of the above learning data are assigned as $\mathbf{O}'_f$ and $\mathbf{O}''_f$, respectively. If the standard result is $\mathbf{O}_f$, the two matching rates $(\sum (O'_i - O_i)/(O'_i - O_i))/f$ and $(\sum (O''_i - O_i)/(O'_i - O_i))/f$ $(i = 1, 2, \ldots, f)$ will be the IVs. Then, the MIV (the mean of two IVs) of the first variable is revealed. Define this MIV is R_1 and threshold is γ. If $R_1 > \gamma$, the first variable is viewed as the significant one, or is the interference one. Other variables can be evaluated in the same way. Therefore, based on the usage of the MIV, the new learning data without interference variables can be expressed as:

$$\mathbf{H} = \{\mathbf{A}_{mi} | R_i > \gamma, R_i \in \mathbf{R}, \mathbf{R} = f_{\mathrm{MIV}}(\mathbf{A}_{mn}, \beta_i^1, \beta_i^2), i = 1, 2, \ldots, n\} \tag{3}$$

where $\mathbf{H}$ is the learning data without interference variables, $\mathbf{R}$ is the vector of MIV results, and $f_{\mathrm{MIV}}(\cdot)$ stands for the MIV results of $\mathbf{A}_{mn}$ under the usage of coefficients β_i^1 and β_i^2.

The other drawback of the PNN is the selection of spread. Figure 1e shows an illustration. This figure indicates that the test point should belong to the 1st cluster which contains more learning data points. However, the PNN defines that all the probability density distributions are scaled equally. In this way, the 1st cluster will correspond to a smaller distribution, while the 2nd cluster which contains fewer learning data points will correspond to a bigger distribution. Consequently, the test point will be falsely classified into the 2nd cluster. This recognition situation demonstrates the importance of spread selection and also implies that the different cluster should be scaled with the different spread. Similarly, as mentioned before, the PSO is introduced to solve this problem. It can rapidly work out the approximate optimal solutions by iterating the random initial solutions (Kennedy and Eberhart 1995; Trelea 2003; Chen and Chu 2009; Ahila et al. 2012). Therefore, if the spreads are initialized randomly, the suitable spreads for the probability density distributions finally can be assigned by the PSO. For the PSO, the initial solutions are required to constitute a population, and each solution is regarded as a seed. If the number of seeds is q, the spread population can be written as $\mathbf{\Gamma} = \{\mathbf{\sigma}_i | \mathbf{\sigma}_i = (\sigma_{1i}, \sigma_{2i}, \ldots, \sigma_{zi}), \ i = 1, 2, \ldots, q\}$. Now, according to the expression of the spread population, the processing procedure of the PNN that starts from the pattern can be rewritten, and the new predicted result corresponding to the $\mathbf{\sigma}_i$ is assigned as $\mathbf{O}_{if}$ (the relevant derivation is shown in "Appendix 3"). According to the accuracy of the predicted result, the spread seeds need to be upgraded, and the corresponding iteration equation set is given below (Shi and Eberhart 1998, 1999; Eberhart and Shi 2001; Trelea 2003; Chen and Chu 2009; Ahila et al. 2012):

$$\begin{cases} \mathbf{\Delta}_i^j = \{\mathbf{\sigma}_i^j | \mathbf{\sigma}_i^j = \min(\left\| \mathbf{O}_{if}^j - \mathbf{O}_{f\_\mathrm{std}} \right\|_2), i = 1, 2, \ldots, z, j = 1, 2, \ldots t\} \\ \mathbf{\Delta}_g^j = \{\mathbf{\sigma}_\Gamma^j | \mathbf{\sigma}_\Gamma^j = \min(\left\| \mathbf{O}_{\Gamma f}^j - \mathbf{O}_{f\_\mathrm{std}} \right\|_2), j = 1, 2, \ldots, t\} \\ \mathbf{W}_i^{j+1} = \omega \mathbf{W}_i^j + c_1 r_1 (\mathbf{\Delta}_i^j - \mathbf{\sigma}_i^j) + c_2 r_2 (\mathbf{\Delta}_g^j - \mathbf{\sigma}_i^j) \\ \mathbf{\sigma}_i^{j+1} = \mathbf{\sigma}_i^j + \mathbf{W}_i^{j+1} \\ s.t. \begin{cases} \mathbf{\sigma}_i^j \in [\sigma_{\max\_1}, \sigma_{\max\_2}] \\ \mathbf{W}_i^j \in [-W_{\max}, W_{\max}] \end{cases} \end{cases} \tag{4}$$

where $\mathbf{\Delta}_i^j$ is the iteration result of ith spread seed and its corresponding predicted result has the smallest error in the first j iteration times, $\mathbf{\sigma}_i^j$ is the iteration result of ith spread seed and its corresponding predicted result has the smallest error in jth iteration time, $\mathbf{O}_{if}^j$ is the predicted result corresponding to the ith spread seed in jth iteration time, $\mathbf{O}_{f\_\mathrm{std}}$ is the standard result, $\mathbf{\Delta}_g^j$ is the iteration result of one spread seed and its corresponding predicted result has the smallest error in the first j iteration times, $\mathbf{\sigma}_\Gamma^j$ is the iteration result of one spread seed and its corresponding predicted result has the

smallest error in *j*th iteration time, $\mathbf{O}_{\Gamma f}^j$ is the predicted result corresponding to one spread seed and has the smallest error in *j*th iteration time, $\mathbf{W}_{t_i}^j$ is the iteration step of *i*th spread seed in *j*th iteration time, ω is the inertia weight, c_1 and c_2 are the acceleration coefficients, r_1 and r_2 are the random values limited in [0,1], $\sigma_{\max\_1}$ and $\sigma_{\max\_2}$ are the left and rights limit of spread seed, respectively, $W_{\max}$ is the limit of iteration step, $\min(\cdot)$ is the function acquiring the minimum, and $\|\cdot\|_2$ is the function calculating L2 norm.

If the difference between the standard and the predicted result is in acceptable error interval, the iteration will stop and the seed corresponding that predicted result will be considered as the approximate optimal solution which includes the suitable spreads for all the probabilistic density distributions.

Shi and Eberhart (1998, 1999), and Eberhart and Shi (2001) did some significant researches regarding calculation performance of the PSO. They found that using a larger inertia weight ω in the former iteration stage will facilitate global searching and a small inertia weight will be more suitable for a local searching in the later iteration. Therefore, lots of algorithms used for modifying the inertia weight are developed, such as linear decreasing inertia weight (LDIW) (Shi and Eberhart 1998, 1999; Eberhart and Shi 2001), exponential decreasing inertia weight (EDIW) (Pediwal et al. 2012), and sigmoid increasing inertia weight (SIIW) (Malik et al. 2007). In this paper, the LDIW is adopted to modify the inertia weight. There are two reasons: One is that the LDIW is easier to apply to improve the PSO's computing efficiency than the EDIW and the SIIW, and the other is that the study of comparing those algorithms is not the topic of this paper although other two algorithms could be more effective for the PSO iteration. The equation of the LDIW is:

$$\omega^t = \omega_{\max} - t(\omega_{\max} - \omega_{\min})/t_{\max} \tag{5}$$

where $\omega_{\max}$ is the maximum inertia, $\omega_{\min}$ is the minimum inertia, and $t_{\max}$ is the maximum iteration time.

The maximum and the minimum inertia weight are two empirical parameters, and according to the previous findings (Shi and Eberhart 1998, 1999; Eberhart and Shi 2001), the recommended values for these two parameters are 0.9 and 0.4, respectively. Thus, in the following validation part, 0.9 and 0.4 will be assigned as the default value of the maximum and the minimum inertia weight.

Except the two drawbacks mentioned above, there is one problem that still needs to be considered for the PNN calculation. As shown in Fig. 1f, gray areas are the original probability density distributions of the 2nd and the 3rd cluster, and obviously two test points belong to these two clusters. If a large amount of the learning data supporting for the 2nd and the 3rd cluster is cut off, meanwhile many data points added into the 1st cluster, two test points will be falsely

classified into the 1st cluster by the PNN although all the distributions (the rainbow areas) have been scaled reasonably by the PSO. As such, in order to ensure the reliability of the predicted results, the PNN should apply the reasonable learning data in which the amount of each kind of data should not be too large or too small.

Calculation workflow

Figure 2 presents calculation workflow of the MIV-PSO-PNN model. The first step is to prepare the basic learning data which is comprised of logs and the corresponding lithology data. Since the reasonability of learning data and probability density distributions will be ensured by the following applications of MIV and PSO, orthogonality of the logs is not necessary to be considered during computing. Because the lithology data are a kind of character data which cannot be processed by computer program directly, a transformation of such data is required. One way of realizing transformation is to record the character data by numbers. For example, four kinds of predicted lithologies including packstone, wackestone, laminate, and spherulite can be converted into four discrete numbers 1, 2, 3, and 4 orderly. The second step is to divide the basic learning data into two parts: the learning and the test data. After that, the MIV process will be conducted. But before that, the optimal taking range of the spread must be known. Thus, in the third step, the learning data are required to be divided into two parts: the small learning and the small test data. Based on the process of the small learning and the small test data, the optimal taking range can be determined by the PNN. Then, in the fourth step, after the MIV process which runs by the learning and the test data, the significant logs for the PNN will be determined. The significant logs will constitute the new learning and test data which will be utilized in the PSO process. Finally, under the advantage of the approximate optimal solution of the spread provided by the PSO, the reasonable probability density distributions will be established, and then, the formed MIV-PSO-PNN model can be used to predict the practical logging data that has no lithology information.

Validation and results

Data source

The validation data are recorded by the lacustrine carbonate reservoirs of the LULA oilfield. The work zone and study region within the oilfield are presented in Fig. 3a. The LULA oilfield is one of the giant petroleum-bearing units of the Santos Basin (Rosenfeld and Hood 2006; Nakano et al. 2010). Because Wertheim (2008, 2009), Nakano et al.

Fig. 2 Calculation workflow of hybrid network MIV-PSO-PNN
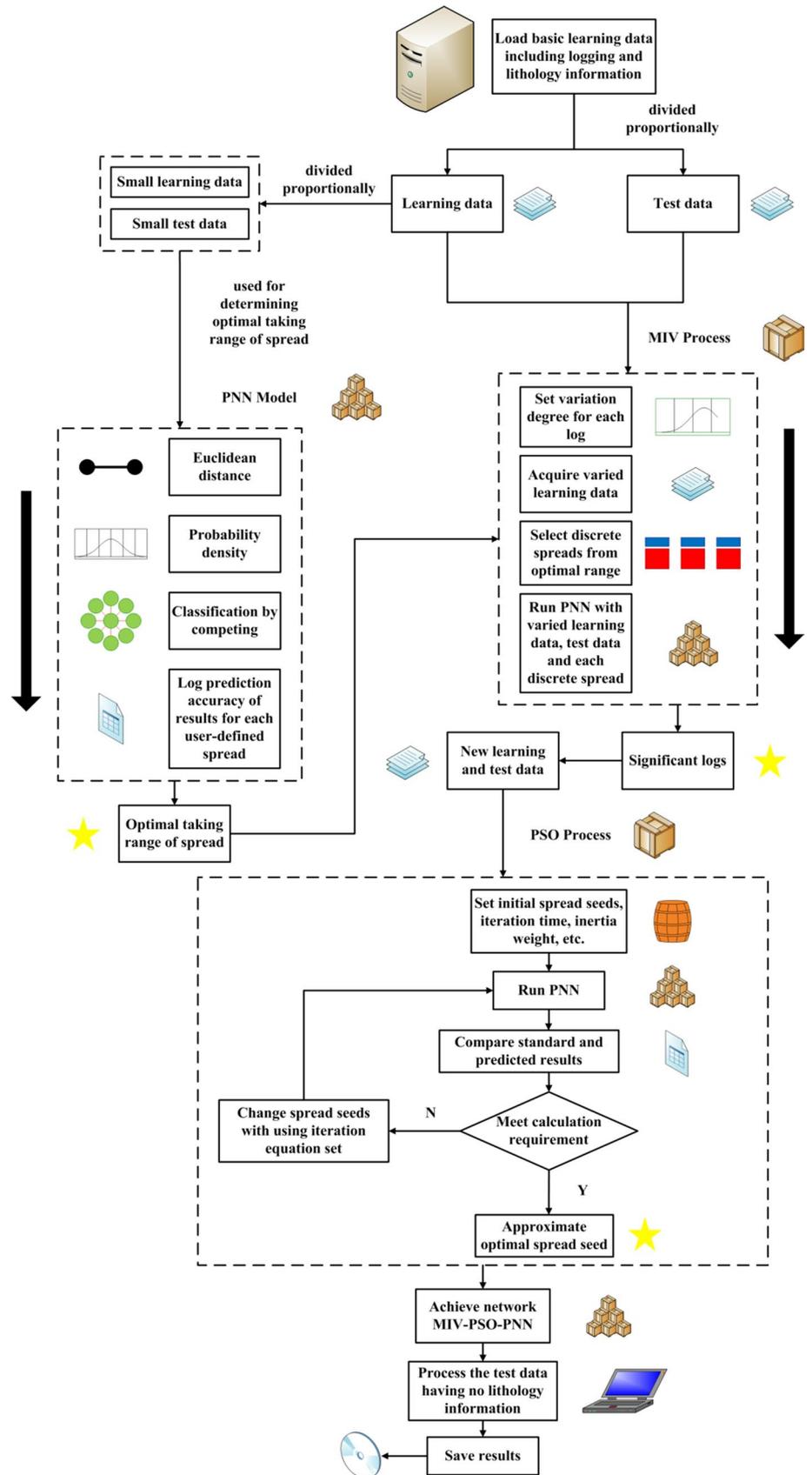
Fig. 3 Work zone and study region of the LULA oilfield, and a general sedimentary profile of the study region. **a** Work zone and study region of the LULA oilfield. The region includes 11 cored wells. **b** A general sedimentary profile of the study region. According to the results of core observation, 13 lithology types are discovered, all of which are listed in the lithology legend. For each kind of lithology as shown in the legend, the left and right pictures, respectively, are the core image, and the freehand lithology pattern used for explaining the sedimentary profile
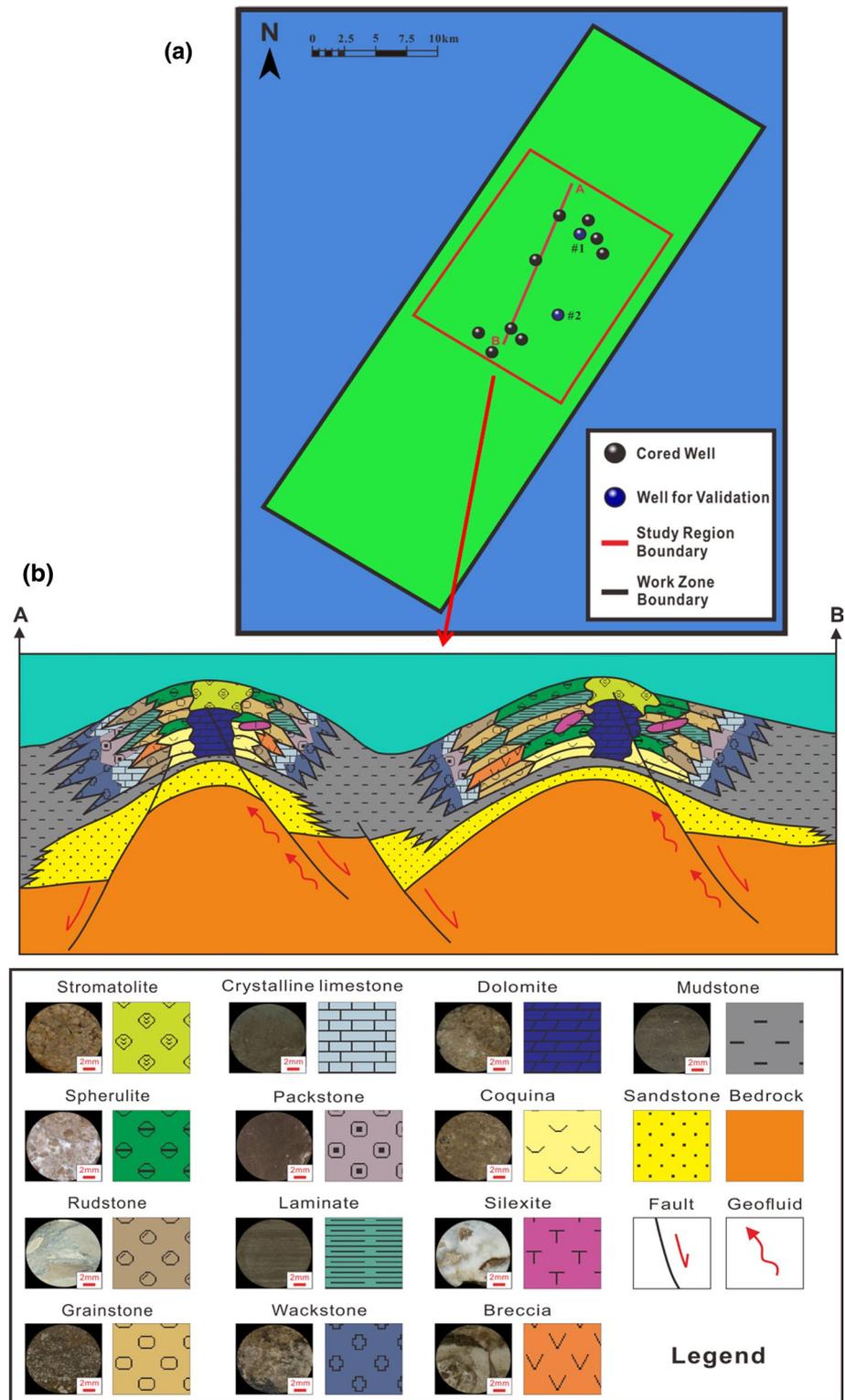
(2010), and Wang et al. (2013) indicated that according to their findings, the potential petroleum reserves of the LULA oilfield can reach up to 50–80 bboe (billion barrel of oil equivalent). Such huge petroleum reserves draw the attention of many geologists, resulting in a research hotspot

regarding the LULA oilfield and the Santos Basin in recent years (Nakano et al. 2010; Andrioni et al. 2012; Wang et al. 2013; D'Afonseca et al. 2013). Many studies, such as Cobbold et al. (2001), Huc (2004), Duarte and Viana (2007), Lentini et al. (2010), Nakano et al. (2010), Andrioni et al.

(2012), Guerra and Underhill (2012), Wang et al. (2013), Wu (2015), Boyd et al. (2015), Thompson et al. (2015), and Alves et al. (2017), have revealed that the major petroleum reserves of the Santos Basin are accumulated within the pre-salt lacustrine carbonate reservoirs. Furthermore, based on the analysis of research findings, those researchers also considered that the primary sedimentary process of the pre-salt carbonate reservoirs underwent three stages which are the rifting, the transition, and the depression. (Depression is also called subsidence.)

In the early rifting stage, the hydrodynamic force was relatively weaker, and then, a large amount of fine-grained sediments was deposited, especially in the places of platform margins and depressions where the crystalline limestone, wackestone, and packstone were developed. Subsequently, due to a regression and the rise of temperature, the salinity of the water increased gradually, finally creating the bioclastic bars which are mainly comprised of the coquina. When the sedimentary process came to the transition stage, the water level kept on dropping. Under the high-frequency impact and transportation of the water, the limestone and mudstone were blended irregularly; thus, many new kinds of carbonate rocks such as spherulite, grainstone, rudstone, and breccias formed and deposited within the platforms and platform margins. Meanwhile, the laminate and wackestone overlapped vertically in the depression places. After that, by the effect of transgression, abundant fine-grained sediments were deposited once again, which causes the platforms and platform margins preferred to accumulate the grainstone and packstone. The laminate progressively became the major part of the depressions. Due to a tectonic movement that occurred at the end of this stage, the geothermal fluid intruded into the formations via the generated faults and fractures. Therefore, the dolomite and silexite also can be found within the platforms (Andrioni et al. 2012; Guerra and Underhill 2012; Boyd et al. 2015; Thompson et al. 2015). During the depression period, many kinds of microbes such as algae appeared within the shallow areas of the platforms, then facilitating the generation of stromatolite. For the platform margins, most of the sediments were the spherulite, grainstone, and rudstone, and among the depression places, the packstone and wackestone were the major sediments. Thus, based on the sedimentary process described above, a general sedimentary profile of the pre-salt lacustrine carbonate reservoirs is obtained as displayed in Fig. 3b.

Nowadays, in order to discover more petroleum-bearing units in the Santos Basin, geologists should discover more information about the geological conditions of the pre-salt carbonate reservoirs. Then, some complicated and challenging geological work have to be accomplished, e.g., establishment of the high-resolution sequence framework and drawing of the high-resolution stratigraphic correlation. The prerequisite of doing those challenging work is to achieve the reliable lithology data of the target reservoirs (Freire et al. 2011; Michelon et al. 2013; Stanton et al. 2014). Therefore, lithology prediction becomes a key problem that should be solved firstly.

The lithology prediction problem for the study region (Fig. 3a) has two features: One is that due to the high cost of sampling cores, only few cored wells are available, and then, the amount of the learning data is small (the amount of cores for any well is less than 1000); and the other is that according to the results of core observation, there are 13 lithology types that have to be predicted. (All predicted lithology types are presented in the legend of Fig. 3b, and for each kind of lithology, the left and right pictures used for explaining Fig. 3b are the core image and the freehand lithology label, respectively.) Consequently, this is a complex lithology prediction problem since many lithology types, including spherulite, rudstone, grainstone, packstone, and crystalline limestone, are the subdivisions of the limestone. Besides, this prediction problem is difficult to solve because the method has to predict 13 lithology types, but only small learning data are available. As mentioned before, the PNN is an ideal method for the lithology prediction, but its calculation mechanism has two drawbacks. Then, the new model MIV-PSO-PNN is proposed to make a solution for such complex lithology prediction problem. In the next step, the proposed method and other methods will be validated contrastively to uncover whether the proposed method is more effective and robust for the complex lithology prediction.

Validation and results

In the following validation, four networks, MIV-PSO-PNN, MIV-PNN, PNN, and BP, will be validated so as to create a comparison in the aspect of computation efficiency. The aim of using PNN and BP is to reveal whether the PNN is more capable than the BP on complex lithology prediction, and the comparison of the MIV-PSO-PNN, MIV-PNN, and PNN is to check the PNN's prediction effect after the PNN is sequentially modified by the MIV and PSO. The validation data are recorded by two cored wells #1 and #2 (Fig. 3a). The learning data of each well consist of some samples. Each sample contains 10 kinds of logging data and one kind of character lithology data. The 10 logs are acoustic log (AC, µs/ft), gamma ray (GR, API), spontaneous potential (SP, mV), computed gamma ray (CGR, API), photoelectric absorption cross-sectional index (PE, b/e), densilog (DEN, g/cm^3), compensated neutron porosity log (CNL, %), resistivity of flushed zone (RXO, Ω m), resistivity of invaded zone (RI, Ω m), and formation true resistivity (RT, Ω m). The resolution of any log is 0.1 m. In other words, there are 10 logging points within 1 m for any log. As described in the calculation workflow, the character lithology data must be converted into the digital type in order to facilitate the PNN calculation on computer. Here, the numbers from

1 to 13 are used in turn to represent the 13 prediction objects which are stromatolite, spherulite, rudstone, grainstone, packstone, laminate, dolomite, crystalline limestone, wackstone, breccia, silexite, coquina, and mudstone (Fig. 3b). Hence, the applied lithology data of each sample actually are a number from 1 to 13.

Prediction for #1 well

The basic learning data of #1 well include 218 samples. Here, 158 samples are chosen randomly to constitute the learning data, and the test data are comprised of the other 60 samples. Note that in order to avoid the problem as shown in Fig. 1f, the number of samples for each predicted lithology type should be balanced. The number of samples for each predicted lithology type is 13 or 14, implying that the samples of the basic learning data have been evenly divided by 13 predicted lithology types. Similarly, the constituting of the learning and test data also should guarantee the balance of the number of samples for each predicted target. As the values of RXO, RI, and RT are much larger than those of other logs, normalization for all data should be done before prediction. Normalizing data can avoid submergence of the significant small values during the data process (Jain and Dubes 1988; Smyth and Speed 2003). Here, define that the normalization range of each log is from 0 to 1. For each log, after the maximum and minimum values are found, the normalization can be completed by the expression $v_{nor} = (1 - 0)(v - v_{min})/(v_{max} - v_{min}) + 0 = (v - v_{min})/(v_{max} - v_{min})$. Assume that the normalized learning and test data are **A** and **B**, and the sizes of **A** and **B** are 158×11 and 60×11.

For the BP, a pre-training of the learning data is required to do before data prediction (Paola and Schowengerdt 1993; Andrew 2001). The aim of doing pre-training is to establish the optimal mapping relationship between the logging data and the numerical lithology data. The frame of the BP network should be firstly determined. The essential calculation parameters, calculation functions, and learning method are all given in Table 1. Except for activation function of hidden layer, transition function of output layer, maximum iteration time, iteration accuracy and lower limit of iteration gradient, other parameters, and learning method must be ensured in accordance with the problem situation (Paola and Schowengerdt 1993; Andrew 2001). As the number of logs is 10 and the number of predicted lithology types is 13, the numbers of input and output units are 10 and 13, respectively. The number of hidden units is usually determined by an empirical equation (Paola and Schowengerdt 1993; Murata et al. 1994; Andrew 2001):

$$l = \sqrt{(m + n)} + a \qquad (6)$$

Table 1 Frame of BP network

Element	Value, function, or method
Number of input units	10
Number of hidden units	14
Number of output units	13
Activation function of hidden layer	$f(x) = 1/(1 + e^{-x})$
Transition function of output layer	$f(x) = x$
Learning method	RPROP
Maximum iteration time	1000
Iteration accuracy (expected level)	0.001
Lower limit of iteration gradient	0.001

where m is the number of input units, n is the number of output units, l is the number of hidden units and this number is an integer, and a is an integer value and its taking range is [1,10].

In this experiment, the a is set to 10. Then, the number of hidden units l is 14. The function of learning method is to evaluate the calculated results. If mean square error (MSE) of the calculated results does not meet the calculation demand, the learning method will adjust the weights and thresholds of the BP network under processing those calculated results so as to improve the accuracies of the calculated results in the next iteration. Resilient backpropagation (RPROP) is introduced as the learning method in this experiment because this method is suitable to deal with the pattern recognition problem (Riedmiller 1994; Wong and Nandi 2004).

Now, the pre-training of the learning data can be accomplished by the BP network. The pre-training actually is an iteration calculation of the learning data. The variations of the iteration accuracy and iteration gradient during the whole training process are presented in Figs. 4a and 5a, respectively. The iteration time is referred as the epoch in Figs. 4a and 5a. The MSE shown in Fig. 4a is used to evaluate the difference between the standard and calculated results. A smaller MSE indicates the calculated results are closer to the standard ones. The variation of the iteration gradient shown in Fig. 5a implies the correctness of the iteration direction. Generally, the downtrend of the gradient manifests the iteration direction has pointed to a global or a local minimum. If the global minimum is pointed, the gradient will keep on dropping, and then, the MSE will keep on decreasing, accelerating the iteration accuracy to reach the expected level. Nonetheless, if the local minimum is pointed, the iteration direction will change again and again around this minimum. In this situation, the gradient will present with a fluctuation during the iteration, and accordingly, the decreasing rate of the MSE will be rather slow (Paola and Schowengerdt 1993; Murata et al. 1994; Andrew 2001). Therefore, the iteration

information shown in Figs. 4a and 5a indicates that the iteration has trapped in a local minimum. In other words, as the MSE does not reach up to the expected level, the established mapping relationship between the logging data and numerical lithology data is not optimal.

After the pre-training process, the trained BP network can be used to predict the test data. The predicted results are presented in Fig. 8a. In this figure, series means the series number of the test data. Number 1 in the series axis stands for the first test sample. D-Value is the difference between the standard and the predicted result. For example, if the predicted result of a sample is 1 (number 1 corresponds to stromatolite) and the standard result is 3 (number 3 corresponds to rudstone), the D-Value of this sample is 2. Hence, the D-Value larger than 0 demonstrates that the corresponding predicted result is wrong. Then, from the statistics of the D-Value data, the prediction accuracy of the results can be known. This figure reveals that 19 samples which correspond to the 0 D-Value are predicted correctly. As such, the prediction accuracy (Match-D) is 31.67%, which is recorded in Table 2 (raw 3, col 2). Figure 9 shows some freehand predicted results. Thirteen standard results in the core column are listed as the references, all of which are determined by the observation of the 13 thin sections cited on the right side. Obviously, only three results, the 3rd, 6th, and 13th, meet with the standard results, manifesting a lower prediction accuracy.

For the PNN, the first task is to ensure the optimal taking range of the spread. According to the workflow presented in Fig. 2, the learning data **A** should be divided into two parts: a small learning and a small test data. Ninety-eight samples are selected randomly from the **A** to constitute the small learning data, and the other 60 samples constitute the small test data. The upper limit of the spread should not be larger than the maximum Euclidean distance of the learning data (Broomhead and Lowe 1988; Specht 1990; Chen et al. 2009; Shaw and Bagha 2012; Kakouei et al. 2014). Then, as the maximum Euclidean distance of the small learning data is 0.2168, the original taking range of the spread is set as (0, 22). Here, a try-and-error method is introduced to optimize this range. The calculation of this method has two steps. In the first step, the PNN will utilize the small learning data, the small test data, and some spreads which are taken from the original spread range to obtain some prediction accuracies. In the second step, the spreads that correspond to the highest prediction accuracies will be picked out to determine the optimal spread range. The spread vector used by the try-and-error method is expressed as $\boldsymbol{\sigma}_{\#1} = \{\sigma_i | \sigma_i = 0.01i, \ i = 1, 2, \cdots, 20\}$. This vector contains 20 spreads so that 20 prediction accuracies will be provided by the try-and-error method. All the prediction accuracies are presented in Fig. 6a. Apparently, the spreads corresponding to the highest accuracies start from 0.04 to

0.11. Therefore, the optimal taking range of the spread is [0.04, 0.11].

The next task is to determine the best spread. Although the 0.1 spread corresponds to the highest prediction accuracy as shown in Fig. 6a, it cannot be considered as the best spread immediately. That spread only can be viewed as the best spread for the PNN when the above small test data are processed. If another small test data are applied by the PNN, the best spread generated by the try-and-error method might be 0.11, 0.8, or other values. Thus, how to find out the real best spread is a question. The MIV process can give the right answer. The function of the MIV has been mentioned in "Methodology" section. In order to comprehensively evaluate the impact degree of each log on the prediction accuracy, the MIV is required to use three groups of enlarged and reduced coefficients, which are 1.1/0.9, 1.2/0.8, and 1.3/0.7. Actually, any group will be acceptable if two contained coefficients are enlarged and reduced with same ratio, respectively, so that these three groups here are just examples used for explaining the MIV calculation. 1.1 and 0.9 of the first group are the β_1^1 and β_1^2, respectively, and so do other two groups. Now, according to the MIV process, three MIVs of each log can be obtained under the application of three coefficient groups. Since it is hard to know which MIV is the most important, the mean of three MIVs is viewed as the real MIV of each log. Results of the real MIV for all logs are listed in Table 3. The larger MIV in the table implies the corresponding log is more significant for the PNN calculation. If the threshold γ involved in Eq. (3) is set to 2, the number of the significant logs that correspond to the 0.1 spread will be the highest. Generally, achieving more significant logs indicates that the corresponding spread is more suitable to scale the probability density distributions of the PNN. This view is explained in Fig. 1a–c. If the distributions scaled by a spread are rather small (Fig. 1a) or are integrated with each other (Fig. 1c), changing the learning data of some or whole logs will not cause obvious variations of those distributions because the locations and sizes of those distributions are not sensitive to the changes of the learning data. In this way, the prediction accuracy of the PNN will not be greatly affected by the changes of the logs, which does not comply with the calculation mechanism of the PNN. As a result, fewer or even no significant logs will be obtained from the MIV process. But if the distributions are scaled reasonably (Fig. 1b), changing the learning data of some logs could cause some distributions to be small or integrated, and then, the prediction accuracy of the PNN will be greatly affected by the changed probabilistic density values which are obtained from the changed probabilistic density distributions. Under this situation, those logs present significance on the variation of the PNN's prediction accuracy and consequently will be revealed by the MIV process. If the distributions are scaled by the spread better, the prediction accuracy of the PNN
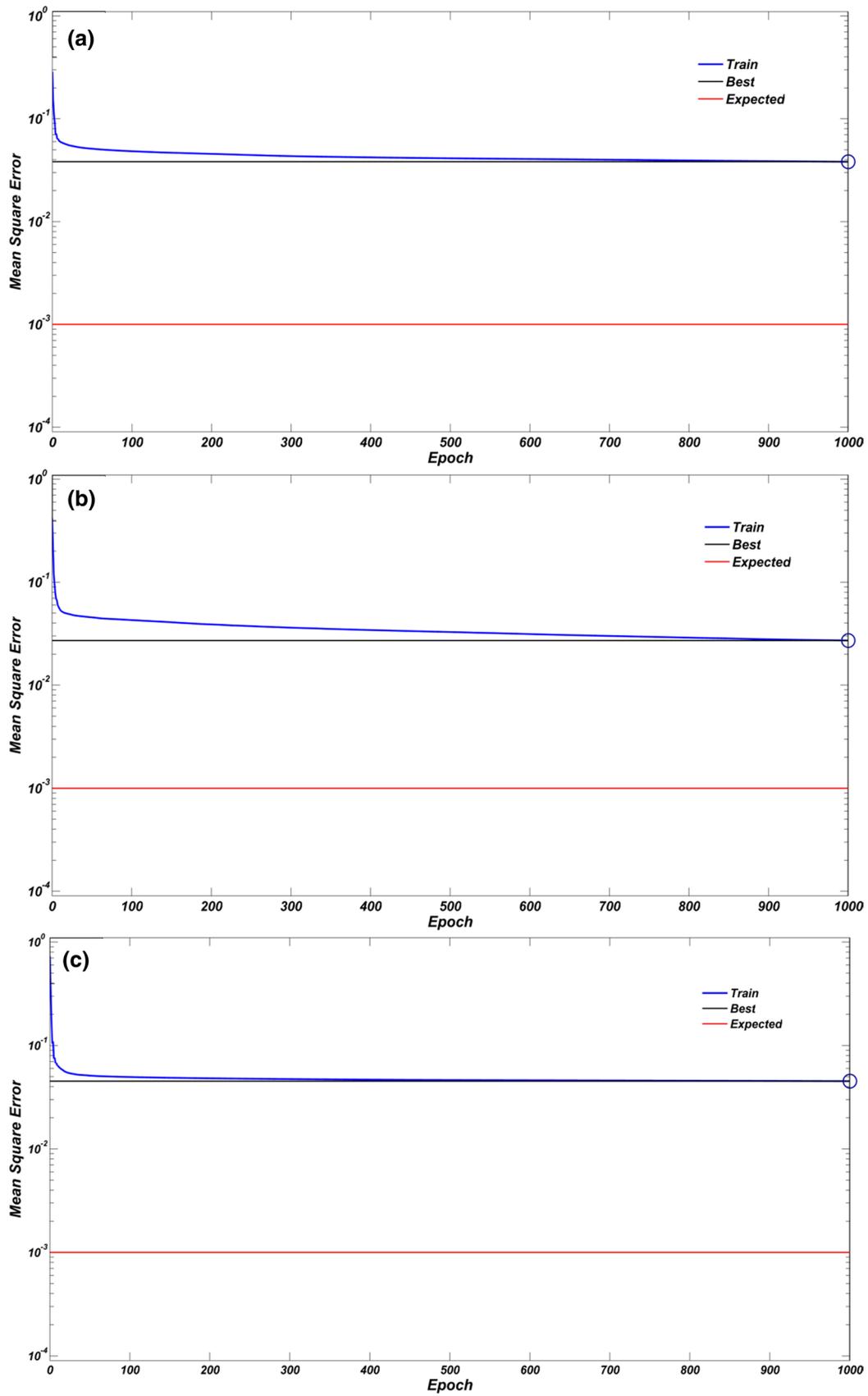
◀**Fig. 4** Variations of iteration accuracy of BP in the experiments of #1 well (**a**), #2 well (**b**), and the #2 well under application of the mixed learning data (**c**). Iteration time is referred as epoch here. Mean square error (MSE) is used to evaluate the difference between the standard and calculated results. The smaller MSE indicates the calculated results are closer to the standard results

will be sensitive to the changing of more logs, implying that more logs will be figured out by the MIV process. Now, according to the above analysis, the 0.1 spread can be treated as the optimal spread which is most suitable to scale the distributions of the PNN. Table 4 provides the information regarding optimal spread and significant logs of this experiment. After the MIV process, the new learning and test data which consist of the significant logs are produced. The new learning and test data are labeled as **C** and **D**, respectively. The sizes of **C** and **D** are 158×8 and 60×8, respectively.

Figure 8b, c shows predicted results of the PNN and MIV-PNN when the 0.1 spread is used. Prediction accuracies of those two models are 48.34% and 65%, respectively, which are listed in Table 2 (raw 3, cols 3 and 4). Figure 9 displays some freehand predicted results. The figure manifests that six samples, the 1st, 4th, 6th, 9th, 10th, and 13th, are labeled right by the PNN, while the MIV-PNN provides 8 correct results which are the 1st, 2nd, 4th, 6th, 7th, 9th, 10th, and 13th. Note that for the PNN, the learning and test data are **A** and **B**, and for the MIV-PNN, they are **C** and **D**.

In the next step, the PSO technique will be applied to optimize the spreads. All the essential calculation parameters of the PSO are given in Table 5. This parameter setting has been validated that its major change will not cause obvious variation on the prediction accuracy. In other words, the parameters of PSO are not much sensitive to the prediction accuracy of PNN. In view of this, an effective parameter setting of PSO is easy to be found out, implying that the application of PSO in fact can reduce the workload of parameter tuning. The size of population means the number of spread seeds. Since there are 13 kinds of lithology types that have to be predicted, the size of each seed is 13. The fitness functions are used to work out the predicted results which will be processed by Eq. (4) in a further step to optimize the spread seeds. It should be known that for the PSO process, the learning and test data are **C** and **D**. The PSO's calculation process is presented in Fig. 7a. D-Degree in this figure is the deviation degree that evaluates the difference between the calculated and the standard result. It is determined by the function $\|\cdot\|_2$ which appears in Eq. (4). Smaller D-Degree means the smaller difference. Thus, the figure illustrates that the calculated result is progressively closing to the standard result during the whole iteration, demonstrating the effective

optimization of the PSO. After the PSO process, the PNN can apply the approximate optimal spread seed to make a prediction. The predicted results are presented in Fig. 8d. Obviously, the samples corresponding to the 0 D-Value are more than those in Fig. 8c. Accordingly, the prediction accuracy is rising to 81.67% as shown in Table 2 (raw 3, col 5). Figure 9 displays some freehand predicted results. For the MIV-PSO-PNN, only two samples are labeled by mistake, which are the 5th and 12th.

Prediction for #2 well

The basic learning data of #2 well contain 219 samples. The number of samples for each predicted lithology type is 13 or 14, indicating that there is no problem concerning the balance of samples for the predicted targets. In total, 159 samples are chosen randomly to constitute the learning data, and the test data consists of the other 60 samples. Both of the learning and test data should be evenly divided by the predicted lithology types. Before the prediction, implementing normalization is still essential for all data.

For the BP, its frame still needs to be formed in Table 1. For the pre-training process, the variations of the iteration accuracy and iteration gradient during the whole iteration are presented in Figs. 4b and 5b, respectively. The iteration accuracy shown in Fig. 4b does not reach up to the expected level even at the end of the iteration, and Fig. 5b manifests that the iteration traps in a local minimum. Hence, the optimal mapping relationship is not established in this experiment. Figure 10a shows the predicted results of the BP. As there are only 17 correct results which correspond to the 0 D-Value, the prediction accuracy is just 28.34% (Table 2, raw 4, col 2). Figure 11 displays some freehand predicted results. From the BP column, it can be known that only two samples are recognized correctly, which are the 8th and 12th.

For the PNN, the optimal taking range of the spread should be firstly determined. Similarly, 99 samples are chosen randomly from the learning data to constitute the small learning data. The other 60 samples constitute the small test data. The maximum Euclidean distance of the small test data is 0.2329. Thus, the original taking range of the spread should be set as (0, 0.23). The spread vector used by the try-and-error method is expressed as $\boldsymbol{\sigma}_{\#2} = \{\sigma_i | \sigma_i = 0.01i, \ i = 1, 2, \cdots, 20\}$. The calculated results of the try-and-error method are shown in Fig. 6b. Apparently, since the spreads that correspond to the highest prediction accuracies start from 0.03 to 0.09, the optimal taking range should be [0.03, 0.09].

The next step is to determine the best spread. For the MIV process, the previous three groups of the enlarged and
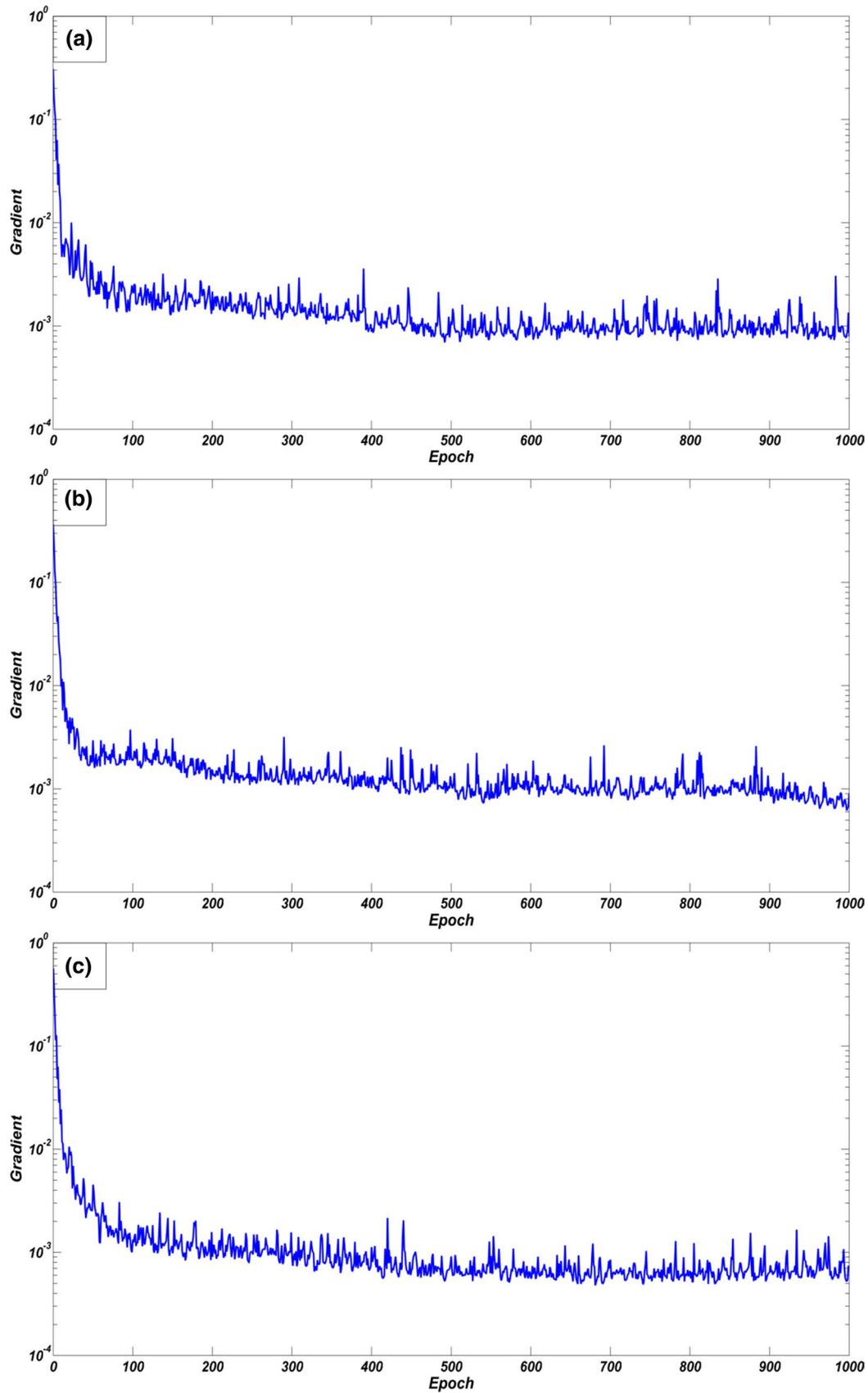
Fig. 5 Variations of iteration gradient of BP in the experiments of #1 well (**a**), #2 well (**b**), and the #2 well under application of the mixed learning data (**c**)

Table 2 Summary of prediction accuracies of four validated networks

Target well	Prediction accuracy (%)			
	BP	PNN	MIV-PNN	MIV-PSO-PNN
#1	31.67	48.34	65	81.67
#2	28.34	53.34	61.67	73.34
#2*	38.34	61.67	73.34	88.34

#2*: the well is validated by the mixed learning data

reduced coefficients are introduced to generate the MIV results in this experiment. Based on the analysis of MIV results, 0.04 is considered as the best spread. Therefore, all significant logs that correspond to the 0.04 spread can be known. Table 4 shows the information of the best spread and significant logs. Table 4 clearly shows that the new learning and test data are all comprised of 6 logs. Similarly, for the following prediction, the original learning and test data will be used by the PNN, and the new learning and test data will be used by the MIV-PNN, and the applied spread is 0.04. The predicted results of the PNN and MIV-PNN are presented in Fig. 10b, c, respectively. Based on the statistics of the predicted results, the PNN's prediction accuracy is worked out, which is 53.34% (Table 2, raw 4, col 3), and the MIV-PNN's prediction accuracy is 61.67% (Table 2, raw 4, col 4). Figure 11 displays some freehand predicted results. Five samples, the 1st, 4th, 8th, 10th, and 12th, are labeled right by the PNN, while the MIV-PNN figures out 7 correct samples which are the 2nd, 5th, 7th, 8th, 10th, 12th, and 13th.

For the next PSO process, the parameter setting provided in Table 5 is still used in this experiment. Figure 7b shows the calculation process. The downtrend of the deviation degree shown in Fig. 7c indicates that the spread seeds have been optimized by the PSO. Then, with utilizing the new learning data, the new test data, and the obtained approximate optimal spread seed, the PNN can make a new prediction. Figure 10d shows the predicted results of the MIV-PSO-PNN. In this figure, as most samples correspond to the 0 D-Value, the prediction accuracy is high, which is 73.34% (Table 2, raw 4, col 5). Figure 11 displays some freehand predicted results. It is obvious that in the MIV-PSO-PNN column only three samples are recognized mistakenly, which are the 4th, 6th, and 11th.

Prediction for #2 well using mixed learning data

The aim of designing this experiment is to disclose whether any validated network can create a higher prediction accuracy after processing the learning data with a larger volume.

Now, the basic learning data of two cored wells are mixed, which contains 437 samples. In total, 377 samples are chosen randomly to constitute the mixed learning data, and the mixed test data certainly consists of the other 60 samples. The number of samples for each predicted lithology type should be even in the learning and test data. Undoubtedly, data normalization should be accomplished before the prediction.

For the BP, the previous formed frame is still adopted. For the pre-training process, the variations of the iteration accuracy and iteration gradient during the whole iteration are presented in Figs. 4c and 5c, respectively. The information shown in those two figures indicates that the mapping relationship is not established well after the pre-training process. The predicted results provided by the trained BP are presented in Fig. 12a. Although more samples that correspond to the 0 D-Value are predicted correctly in comparison with those in Fig. 10a, the prediction accuracy of the BP is still low, which is only 38.34% (Table 2, raw 5, col 2). Figure 13 displays some freehand predicted results. Obviously, only five samples, the 2nd, 4th, 5th, 8th, and 12th, are labeled right.

For the PNN, 317 samples are chosen randomly to constitute the small mixed learning data, and the other 60 samples constitute the small mixed test data. Because the maximum Euclidean distance of the small mixed test data is 0.22, the original taking range of the spread is set as (0, 0.22). The spread vector used by the try-and-error method is $\boldsymbol{\sigma}_{\#2} = \{\sigma_i | \sigma_i = 0.01i, \ i = 1, 2, \cdots, 20\}$. Figure 6c shows the calculated results. The spreads that correspond to the relatively higher prediction accuracies start from 0.03 to 0.09, and then, the optimal taking range is [0.03, 0.09].

For the MIV process, the previous three coefficient groups are still applied in this experiment. The best spread and significant logs generated by the MIV process are listed in Table 4, manifesting the new mixed learning and test data are comprised of 6 logs. Similarly, for the following prediction, the original mixed learning and test data will be used by the PNN, and the new mixed learning and test data will be used by the MIV-PNN, and the applied spread is 0.04. The predicted results of the PNN and MIV-PNN are presented in Fig. 12b, c, respectively. Compared to Fig. 10b, c, Fig. 12b, c provides more correct samples that correspond to the 0 D-Value. Hence, in this experiment the prediction accuracies of those two models are both improved (Table 2, raw 5, cols 3 and 4). Figure 13 displays some freehand predicted results. Seven samples are recognized correctly by the PNN, while the MIV-PSO-PNN figures out more right results which are the 1st, 2nd, 3rd, 5th, 7th, 8th, 11th, 12th, and 13th.

Fig. 6 Variation trends between spread and prediction accuracy, and optimal taking ranges of spread in the experiments of #1 well (**a**), #2 well (**b**), and the #2 well under application of the mixed learning data (**c**). Match-D means prediction accuracy
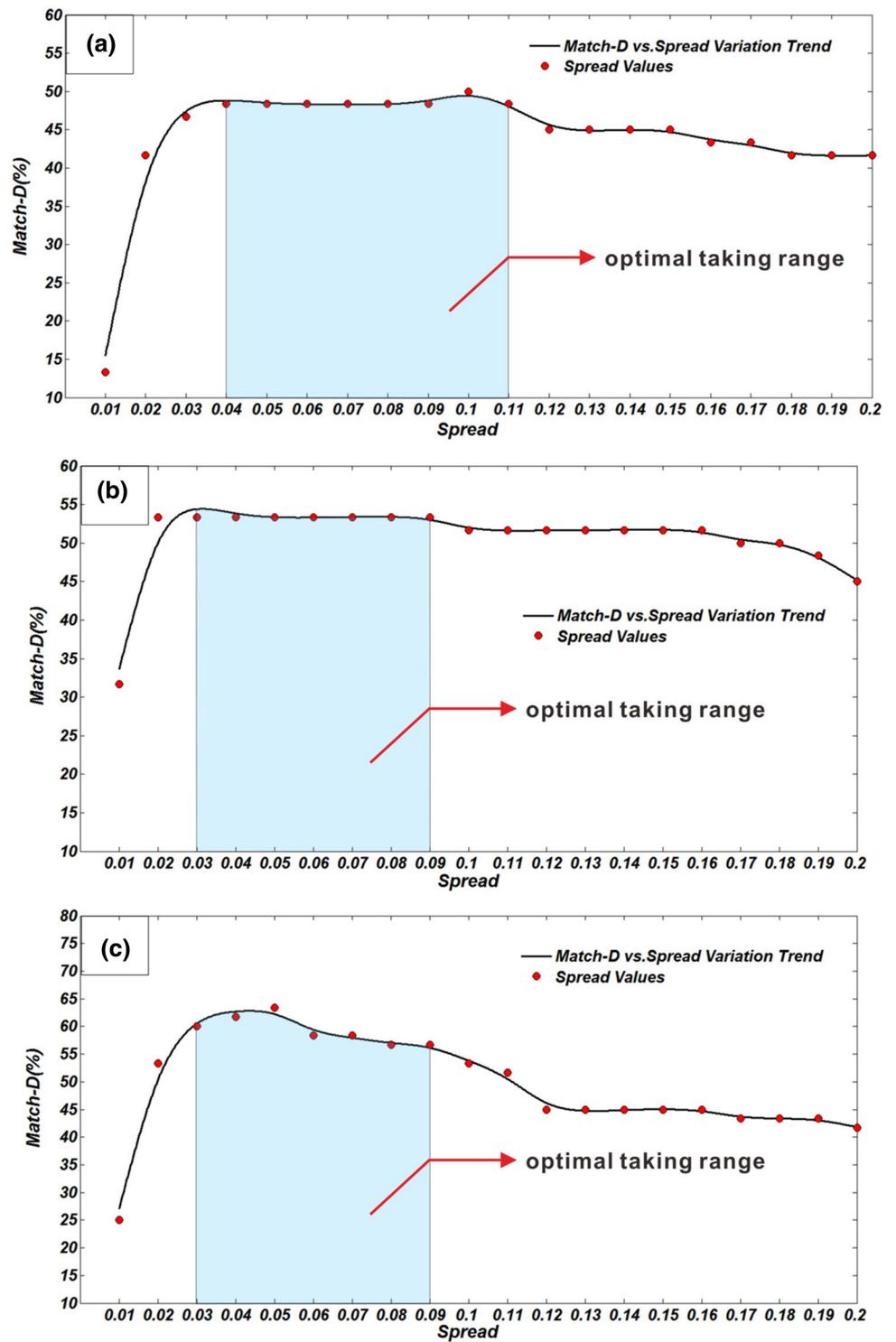
Table 3 MIV results of the prediction of #1 well. The larger MIV implies that the corresponding log is more significant for PNN calculation (%)

Log	Spread							
	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.11
AC	3.33	3.89	5	3.33	3.33	3.34	3.34	2.78
CGR	13.89	8.89	7.22	1.11	1.11	0.56	3.33	3.89
CNL	12.23	13.33	8.89	5.56	5	2.78	3.33	0.55
DEN	0	0	0	0	0	0.56	2.23	2.23
GR	3.37	5	5	8.33	7.78	8.89	8.33	5.56
PE	1.67	3.34	5	7.22	6.66	6.66	7.21	7.21
RT	0	0	0	0	0	0	0	0
RI	0	0	0	0	0	0	0	0
RXO	0	0	0	0	0	0	0	0
SP	6.11	6.67	7.78	7.78	5.56	6.11	3.89	4.45

Table 4 Summary of optimal spreads and significant logs

Target well	Optimal spread	Significant log
#1	0.1	AC CGR CNL DEN GR PE SP
#2	0.04	AC CGR CNL GR PE RXO
#2*	0.04	AC CGR CNL DEN GR SP

#2*: the well is validated by the mixed learning data

Table 5 Parameter setting of PSO

Parameter	Value, range, or function
Size of population	20
Size of spread seed	13
Iteration time	100
Fitness function	PNN
ω_{start}	0.9
ω_{end}	0.5
c_1	1.5
c_2	1.5
r_1	[0,1]
r_2	[0,1]
σ_{max_1}	0.04
σ_{max_2}	0.11
W_{max}	0.03

For the PSO process, the parameters are defined as same as before. Figure 7c shows the calculation process. The information in this figure indicates that the spread seeds have been effectively optimized. Now, with applying the new mixed learning data, the new mixed test data, and

the determined approximate optimal spread seed, the PNN is capable to make a new prediction. Figure 12d presents the predicted results of the MIV-PSO-PNN. Similarly, compared to Fig. 10d, Fig. 12d provides more samples that correspond to the 0 D-Value. Thus, the prediction accuracy of the MIV-PSO-PNN is also improved, which is 88.34% (Table 2, raw 5, col 5). Figure 13 displays some freehand predicted results and reveals that only the 11th sample is recognized mistakenly.

Discussion

The prediction accuracies of the BP in three experiments are 31.67%, 28.34%, and 38.34%, respectively, all of which are the lowest ones. Additionally, only fewer samples in the BP columns of Figs. 9, 11, and 13 are labeled correctly. Those experiment results demonstrate that the BP is ineffective to deal with the complex lithology prediction. The reason is shown in Figs. 4 and 5. As mentioned before, for any experiment, the gradient fluctuation presented in Fig. 5 implies that the iteration direction is always changing around a local minimum during the whole training process. Under this situation, the decreasing rate of the MSE will be rather slow, and then, the iteration accuracy will not reach up to the expected level even at the end of iteration. This explanation has been proved by the information shown in Fig. 4. Therefore, the pre-training process is not accomplished well in any experiment, and naturally without the optimal mapping relationship, the BP is incapable to provide the predicted results with desirable prediction accuracy.

The prediction accuracies of the PNN in three experiments are all higher than those of the BP. Moreover, for

Fig. 7 PSO calculation processes in the experiments of #1 well (**a**), #2 well (**b**), and the #2 well under application of the mixed learning data (**c**)

Prediction for #1 Well
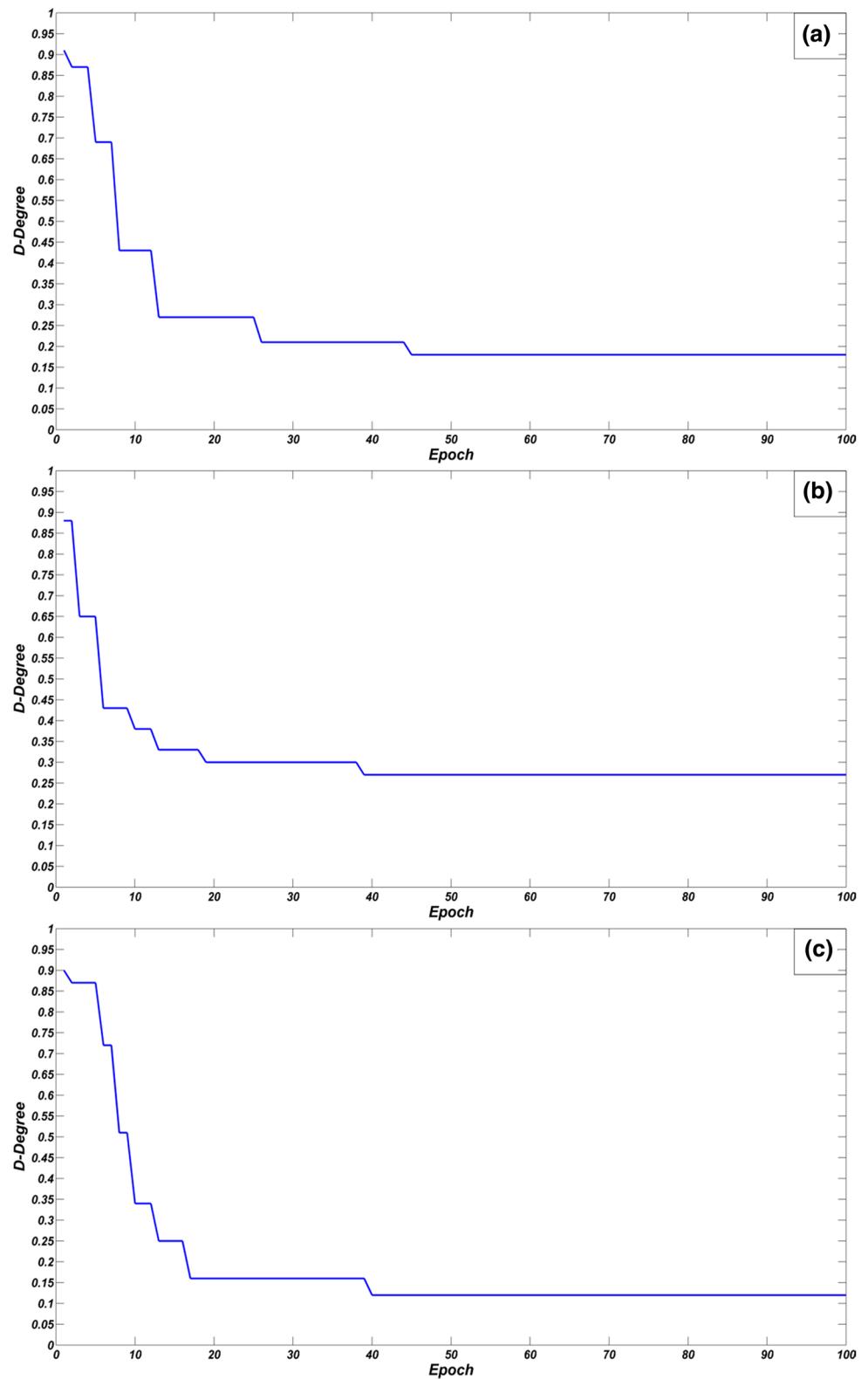

Fig. 8 Predicted results and prediction accuracy of BP (**a**), PNN (**b**), MIV-PNN (**c**), or MIV-PSO-PNN (**d**) in the experiment of #1 well. Series is the series number of test data. D-Value is the difference between the standard and predicted results

the freehand predicted results shown in Figs. 9, 11, and 13, more correct samples can be found in the PNN columns. Thus, the PNN is more effective than the BP in the aspect of complex lithology prediction. The better performance of the PNN can be explained by the PNN's calculation mechanism. The PNN takes advantage of statistical information of the learning data to predict the test data, and then, the PNN will be robust to work out the reliable results even though the processing learning data contain a few incorrect samples. In other words, for the PNN, the reliability of the predicted results will not be affected by a few incorrect samples involved in the learning data. Nonetheless, for the BP, those incorrect samples will greatly affect the prediction effect since the calculation mechanism is featured by the analytical operations. The learning data used for lithology prediction must contain some incorrect samples because the contained lithology data are generated empirically based on the core observation and naturally its correctness cannot be guaranteed. Hence, it is easy to understand why the PNN presents better prediction performance in the data validation than the BP.

The predicted results and prediction accuracies of the MIV-PNN in three experiments verify that the prediction capability of the PNN can be enhanced when the MIV technique is integrated in the PNN process. Figure 1d illustrates the adverse impact of interference logs on the PNN's

prediction effect. Then, in order to ensure the reliability of the predicted results, the PNN must apply the learning data without interference logs. The MIV technique is functional to remove the interference logs by evaluating the impact degree of each log. Evaluation is completed by the MIV results. The larger MIV means the corresponding log is more significant for the PNN calculation. Consequently, based on the analysis of the MIV results, all significant logs in the learning data will be picked out. Then, with utilizing the learning data comprised of the significant logs, the PNN is capable to provide more accurate predicted results and higher prediction accuracies in the validation.

Compared to the MIV-PNN, the prediction accuracies of the MIV-PSO-PNN in three experiments are all higher. Meanwhile, for the freehand predicted results shown in Figs. 9, 11, and 13, more correct samples can be found in the MIV-PSO-PNN columns. Therefore, it is clear that the MIV-PNN's prediction effect can be improved by the integration of the PSO technique. Figure 1e illustrates the different probabilistic density distribution needs to be scaled with the different spread. As such, acquiring a spread vector which includes optimal spreads for all the distributions is essential for the PNN. The PSO is efficient to achieve the approximate optimal solution for the multi-objective optimization problem. If spreads for all the distributions are viewed as the objects, the approximate optimal spreads will
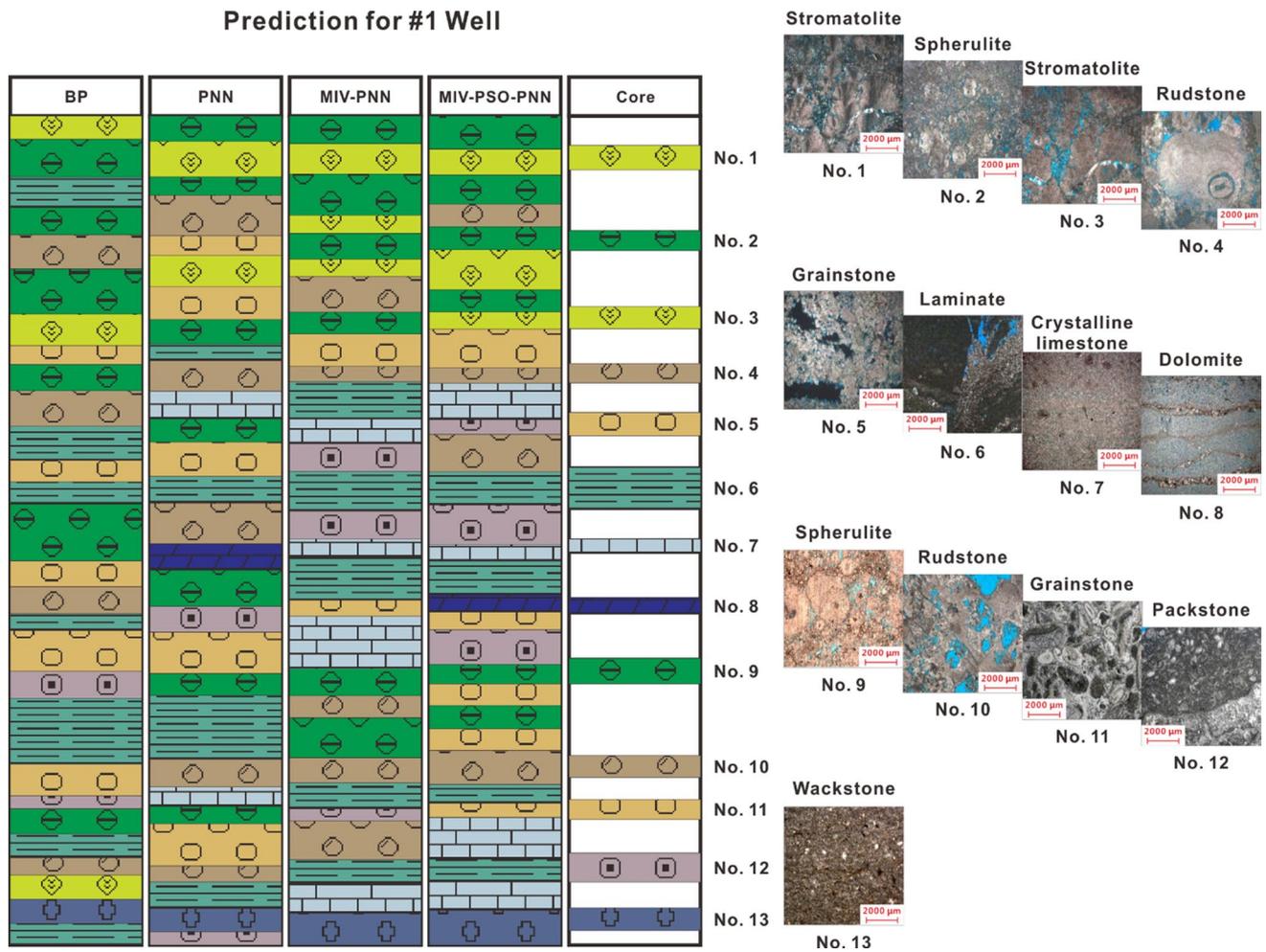
Fig. 9 Some freehand predicted results of BP, PNN, MIV-PNN, and MIV-PSO-PNN in the experiment of #1 well. Thirteen standard results in the core column are listed as references, all of which are determined by observation of the 13 thin sections cited on the right side

be produced by the PSO. The information shown in Fig. 7 has well demonstrated that the initialized spreads can be effectually optimized by the PSO. Thus, by the usage of the PSO, the MIV-PSO-PNN undoubtedly performs better than the MIV-PNN in the validation.

Compared to the predicted accuracies of the 2nd experiment, the higher accuracies of the 3rd experiment prove that the prediction effect of any validated model can be enhanced under the condition of processing the learning data with a larger volume, especially for the MIV-PSO-PNN which creates the perfect prediction accuracy 88.34%. Consequently, according to the validation results, it is known that the proposed method is effective to predict complex lithology, and can generate more accurate predicted results when processing the learning data with a larger volume.

Conclusion

According to the analysis and discussion of the above validation results, some features of the proposed hybrid network MIV-PSO-PNN are summarized as follows:

1. PNN takes advantage of the statistical information of the learning data to predict the test data, and then, it is robust enough to provide reliable results even though the learning data contain some incorrect samples.
2. MIV technique has the capability to remove the interference logs from the learning data by evaluating the impact degree of each log, and therefore, PNN will be functional to process the learning data that contain interference logs after integrated with the MIV.
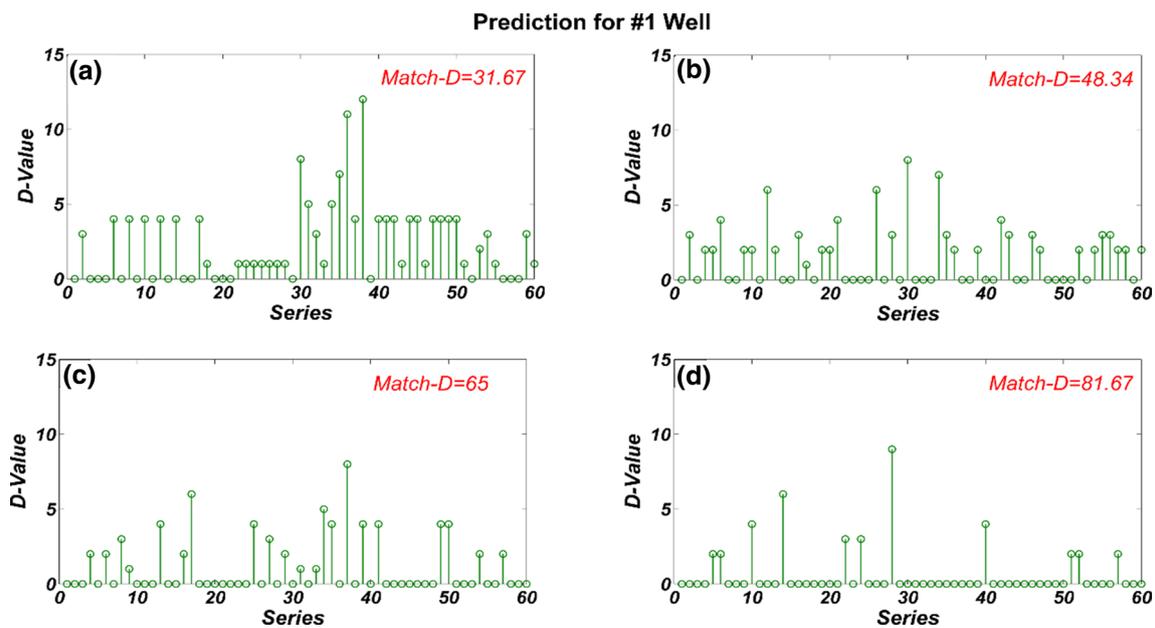
Prediction for #2 Well
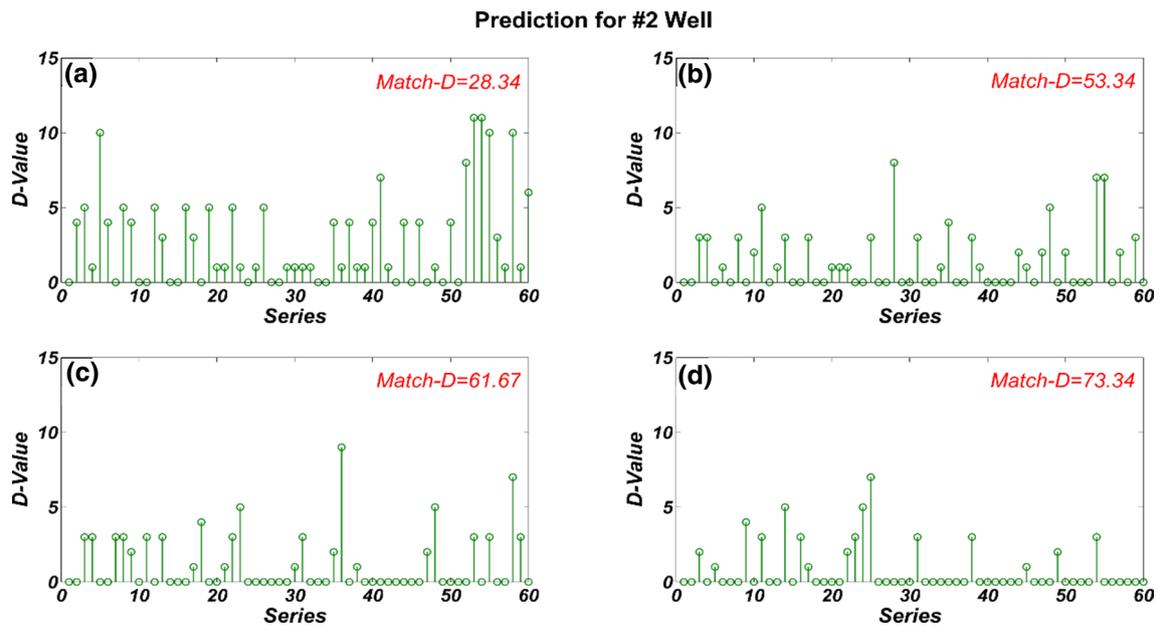
Fig. 10 Predicted results and prediction accuracy of BP (**a**), PNN (**b**), MIV-PNN (**c**), or MIV-PSO-PNN (**d**) in the experiment of #2 well
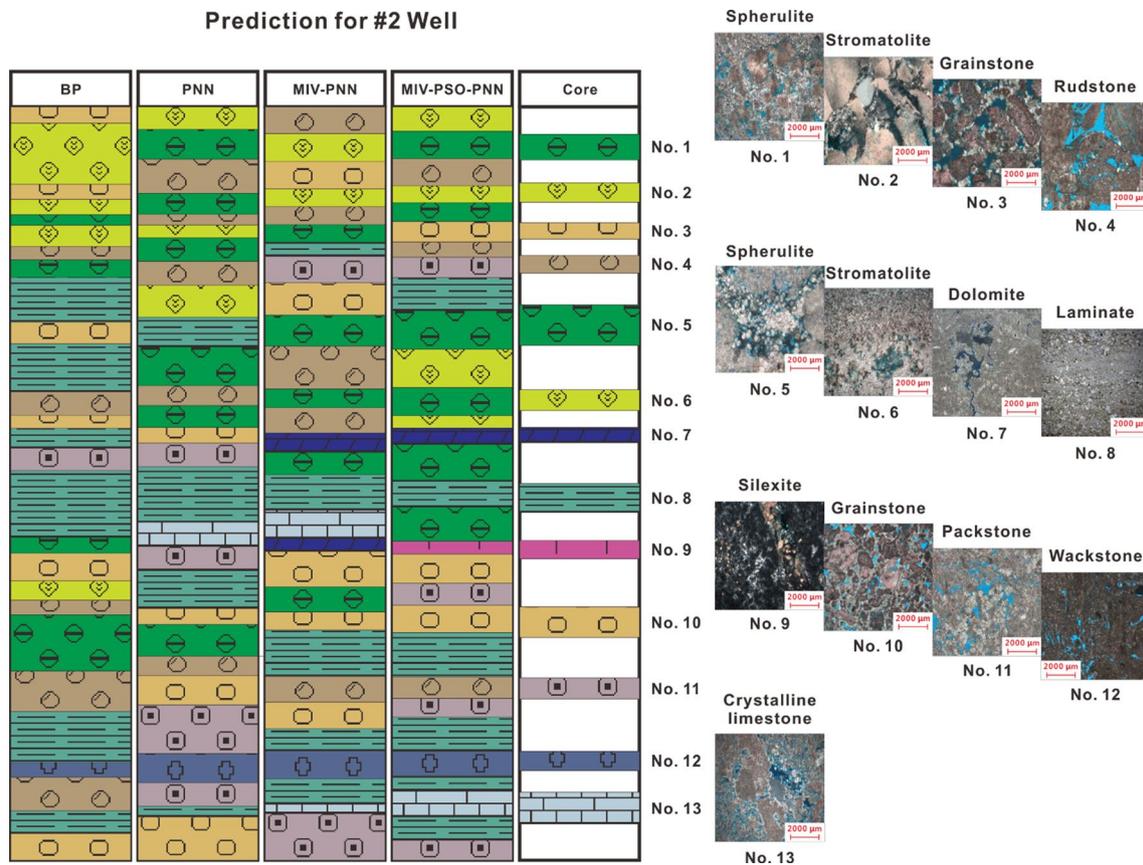
Fig. 11 Some freehand predicted results of BP, PNN, MIV-PNN, and MIV-PSO-PNN in the experiment of #2 well

Prediction for #2 Well
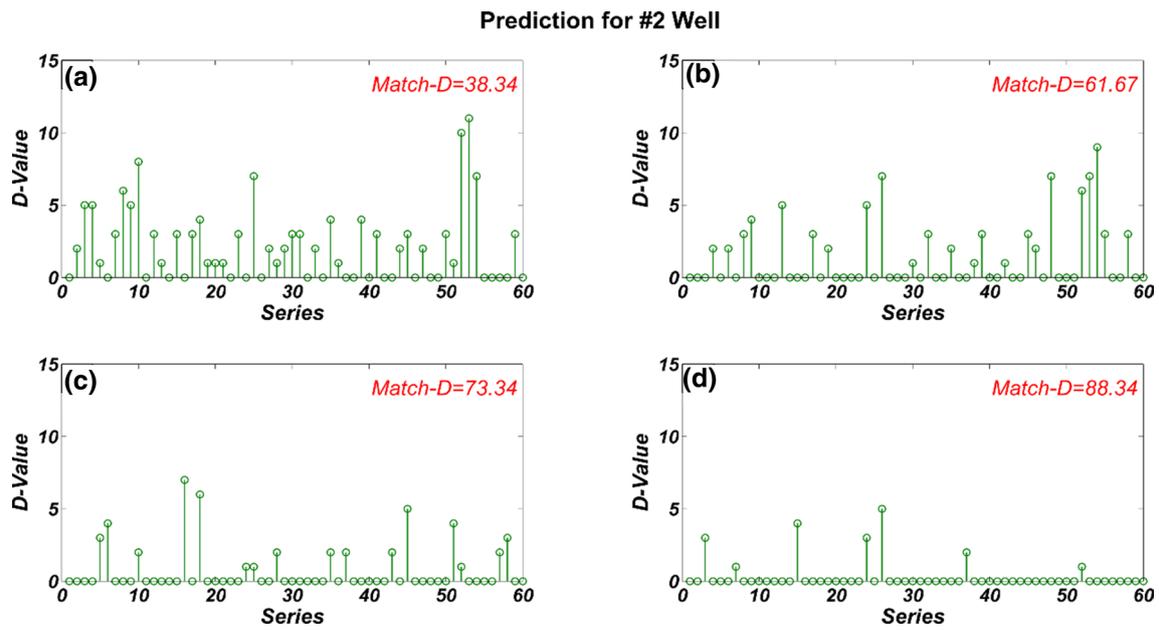
Fig. 12 Predicted results and prediction accuracy of BP (**a**), PNN (**b**), MIV-PNN (**c**), or MIV-PSO-PNN (**d**) in the experiment of #2 well
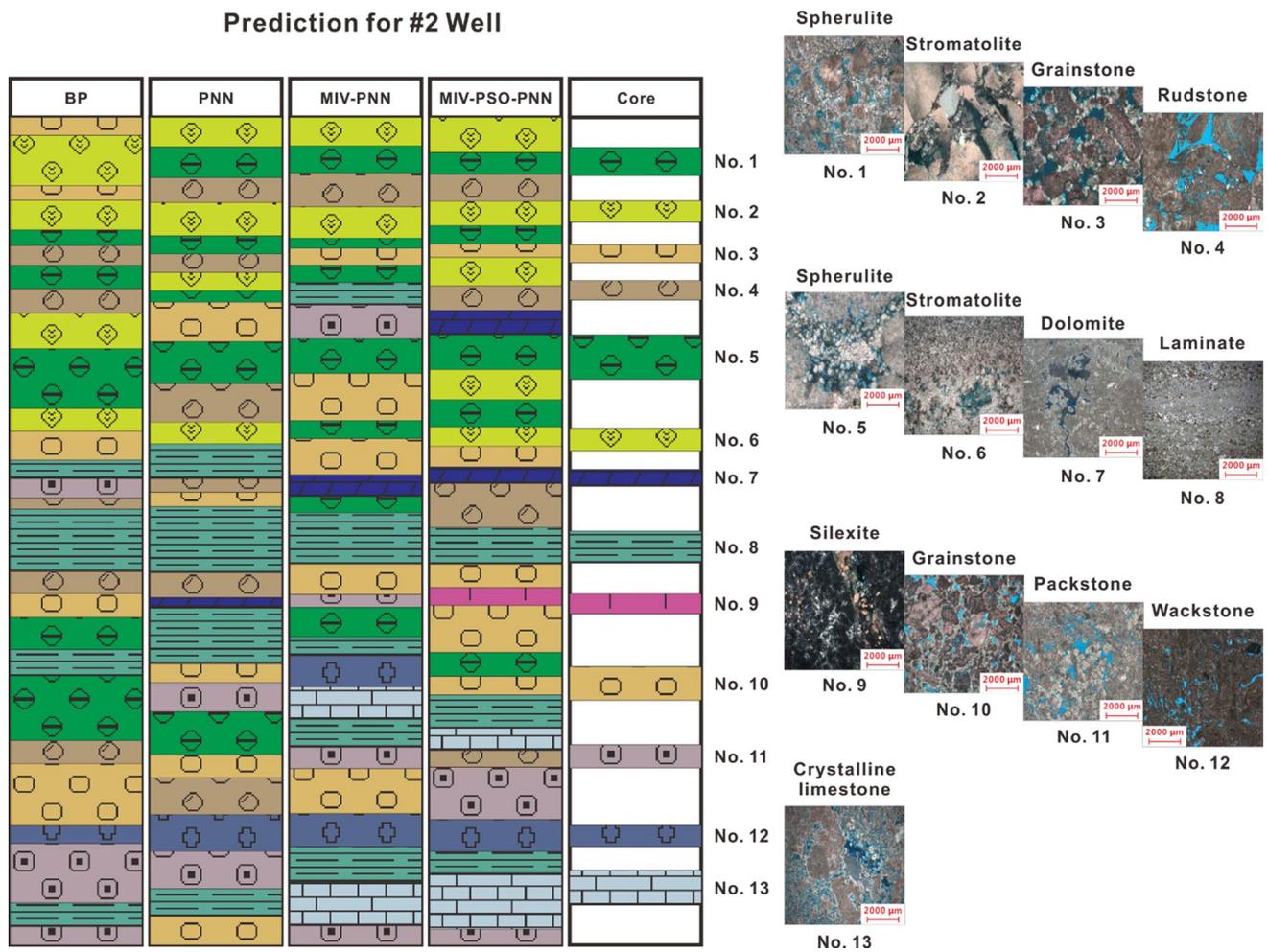
Fig. 13 Some freehand predicted results of BP, PNN, MIV-PNN, and MIV-PSO-PNN in the experiment of #2 well

3. PSO technique shows to be efficient for the multi-objective optimization problem, and obviously will assist PNN to increase reliability of the prediction results through scaling all probability density distributions reasonably.

4. MIV-PSO-PNN will be capable to produce the results with higher prediction accuracy under the condition of processing the learning data with a larger volume.

Compliance with ethical standards

Conflict of interest 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.

Appendix 1

Input layer

$$\mathbf{A}_{mn} = \begin{bmatrix} a_{11} & a_{12} & , & \cdots & , & a_{1n} \\ a_{21} & a_{22} & , & \cdots & , & a_{2n} \\ , & , & , & & , \\ \vdots & \vdots & & \ddots & & \vdots \\ , & , & & , & , \\ a_{m1} & a_{m2} & , & \cdots & , & a_{mn} \end{bmatrix} \text{ and } \mathbf{B}_{fn} = \begin{bmatrix} b_{11} & b_{12} & , & \cdots & , & b_{1n} \\ b_{21} & b_{22} & , & \cdots & , & b_{2n} \\ , & , & , & & , \\ \vdots & \vdots & & \ddots & & \vdots \\ , & , & & , & , \\ b_{f1} & b_{f2} & , & \cdots & , & b_{fn} \end{bmatrix} \longrightarrow \text{Euclidean distance calculation}$$

Pattern layer

$$= \begin{vmatrix} d_{11} & d_{12} & , & \cdots & , & d_{1m} \\ d_{21} & d_{22} & , & \cdots & , & d_{2m} \\ , & , & , & & , \\ \vdots & \vdots & & \ddots & & \vdots \\ , & , & & , & , \\ d_{f1} & d_{f2} & , & \cdots & , & d_{fm} \end{vmatrix} = \mathbf{D}_{fm} \xrightarrow{\text{Probability density calculation}} \begin{bmatrix} p_{11} & p_{12} & , & \cdots & , & p_{1m} \\ p_{21} & p_{22} & , & \cdots & , & p_{2m} \\ , & , & , & & , \\ \vdots & \vdots & & \ddots & & \vdots \\ , & , & & , & , \\ p_{f1} & p_{f2} & , & \cdots & , & p_{fm} \end{bmatrix} = \mathbf{P}_{fm} \xrightarrow{\text{sum}}$$

sum layer

$$\begin{bmatrix} \sum_{k=1}^{v_1} p_{1k} & \sum_{k=v_1+1}^{v_2} p_{1k} & , & \cdots & , & \sum_{k=v_{z-1}+1}^{v_z} p_{1k} \\ \sum_{k=1}^{v_1} p_{2k} & \sum_{k=v_1+1}^{v_2} p_{2k} & , & \cdots & , & \sum_{k=v_{z-1}+1}^{v_z} p_{2k} \\ , & , & , & & , \\ \vdots & \vdots & & \ddots & & \vdots \\ , & , & & , & , \\ \sum_{k=1}^{v_1} p_{fk} & \sum_{k=v_1+1}^{v_2} p_{fk} & , & \cdots & , & \sum_{k=v_{z-1}+1}^{v_z} p_{fk} \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & , & \cdots & , & s_{1z} \\ s_{21} & s_{22} & , & \cdots & , & s_{2z} \\ , & , & , & & , \\ \vdots & \vdots & & \ddots & & \vdots \\ , & , & & , & , \\ s_{f1} & s_{f2} & , & \cdots & , & s_{fz} \end{bmatrix} = \mathbf{S}_{fz} \xrightarrow{\text{compete}}$$

Output layer

$$= \frac{s_{ij}}{\sum_{k=1}^{z} s_{ik}} (i = 1, 2, \ldots, f) = \begin{bmatrix} c_{11} & c_{12} & , & \cdots & , & c_{1z} \\ c_{21} & c_{22} & , & \cdots & , & c_{2z} \\ , & , & , & & , \\ \vdots & \vdots & & \ddots & & \vdots \\ , & , & & , & , \\ c_{f1} & c_{f2} & , & \cdots & , & c_{fz} \end{bmatrix} = \mathbf{C}_{fz} \xrightarrow{\text{max in row}} \mathbf{O}_f$$

"Euclidean distance" is determined by the power exponent component of Eq. (2)."max in row" means extracting the maximum value of each row.

Appendix 2

$$\mathbf{A}_{mn} \times \begin{bmatrix} \beta_1^1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ , & , & , & , \\ \vdots & \vdots & \ddots & \vdots \\ , & , & , & , \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n} = \begin{bmatrix} a'_{11} & a_{12} & \cdots & a_{1n} \\ a'_{21} & a_{22} & \cdots & a_{2n} \\ , & , & , & , \\ \vdots & \vdots & \ddots & \vdots \\ , & , & , & , \\ a'_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \mathbf{A}'_{mn}$$

$$\mathbf{A}_{mn} \times \begin{bmatrix} \beta_1^2 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ , & , & , & , \\ \vdots & \vdots & \ddots & \vdots \\ , & , & , & , \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n} = \begin{bmatrix} a''_{11} & a_{12} & \cdots & a_{1n} \\ a''_{21} & a_{22} & \cdots & a_{2n} \\ , & , & , & , \\ \vdots & \vdots & \ddots & \vdots \\ , & , & , & , \\ a''_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \mathbf{A}''_{mn}$$

Appendix 3

$$\begin{vmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ , & , & , & , \\ \vdots & \vdots & \ddots & \vdots \\ , & , & , & , \\ d_{f1} & d_{f2} & \cdots & d_{fm} \end{vmatrix} \xrightarrow{\text{substituting } \sigma_i} \begin{bmatrix} p^i_{11} & p^i_{12} & \cdots & p^i_{1m} \\ p^i_{21} & p^i_{22} & \cdots & p^i_{2m} \\ , & , & , & , \\ \vdots & \vdots & \ddots & \vdots \\ , & , & , & , \\ p^i_{f1} & p^i_{f2} & \cdots & p^i_{fm} \end{bmatrix}$$

$$= \mathbf{P}^i_{fm} \xrightarrow{\text{sum and compete}} \mathbf{O}_{if}$$

References

Ahila R, Sadasivam V, Manimala K (2012) Particle swarm optimization-based feature selection and parameter optimization for power system disturbances classification. Appl Artif Intell 26(9):832–861

Akhmanov GG, Silva IP, Erba E, Cita MB (2003) Sedimentary succession and evolution of the Mediterranean Ridge western sector as derived from lithology of mud breccia clasts. Mar Geol 195(1–4):277–299

Alves TM, Fetter M, Lima C, Cartwright JA, Cosgrove J, Gangá A (2017) An incomplete correlation between pre-salt topography, top reservoir erosion, and salt deformation in deep-water Santos Basin (SE Brazil). Mar Pet Geol 79:300–320

Andrew AM (2001) Backpropagation. Kybernetes 30(9–10):85–104

Andrioni M, Lima JAM, Guerra LA, Ribeiro EO, Nunes LMP, Ceccopieri W (2012) Ocean eddies' influence on Lula field, Santos Basin, Brazil. In: ASME 2012, international conference on ocean, offshore and arctic engineering

Barrow J (2009) Lithology identification and prediction through amplitude variation with offset analysis in an area of the Malay basin, gulf of Thailand/Jerreh barrow. J Braz Chem Soc 20:57–63

Bogdanov YA, Gorbunova ZN, Serova VV (1998) Lithology and mineralogy of bottom sediments from the Bear island area, Norwegian sea. Oceanology 38(4):542–547

Bosch D, Ledo J, Queralt P (2013) Fuzzy logic determination of lithologies from well log data: application to the KTB project data set (Germany). Surv Geophys 34(4):413–439

Boyd A, Souza A, Carneiro G, Machado V, Trevizan W, Coutinho B, Netto P, Polinski R, Bertolini A (2015) Pre-salt carbonate evaluation for Santos Basin, offshore Brazil. Petrophysics 56(6):577–591

Broomhead DS, Lowe D (1988) Radial basis functions, multi-variable functional interpolation and adaptive networks. Adv Neural Inf Process Syst 4148:728–734

Chen CH, Chu CT (2009) High performance iris recognition based on 1-D circular feature extraction and PSO-PNN classifier. Expert Syst Appl 36(7):10351–10356

Chen N, Sun F, Ding L, Wang H (2009) An adaptive PNN-DS approach to classification using multi-sensor information fusion. Neural Comput Appl 18(5):455–467

Chetelat B, Liu CQ, Wang Q, Zhang G (2013) Assessing the influence of lithology on weathering indices of Changjiang river sediments. Chem Geol 359(6):108–115

Chi MV, Wong PK, Wong KI (2014) Simultaneous-fault detection based on qualitative symptom descriptions for automotive engine diagnosis. Appl Soft Comput 22(5):238–248

Cobbold PR, Meisling KE, Mount VS (2001) Reactivation of an obliquely rifted margin, Campos and Santos Basins, southeastern Brazil. AAPG Bull 85(11):1925–1944

D'Afonseca L, Lecerf D, Souza A, Sanchez F, Prigent H, Vidal T (2013) Multi-azimuth imaging for deep-water pre-salt reservoirs in Santos Basin, Brazil. In: EAGE conference and exhibition incorporating SPE EUROPEC

de Almeida JA (2010) Stochastic simulation methods for characterization of lithoclasses in carbonate reservoirs. Earth Sci Rev 101(3–4):250–270

Dev VA, Eden MR (2019) Formation lithology classification using scalable gradient boosted decision trees. Comput Chem Eng 128:392–404

Djafarov IS, Khafizov SF, Syngaevsky PE (2004) NMR application in reservoirs with complex lithology: a case study. Petrophysics 45(2):119–129

Duarte CSL, Viana AR (2007) Santos drift system: stratigraphic organization and implications for late Cenozoic palaeocirculation in the Santos Basin, SW Atlantic Ocean. Miner Pediatr 276(1):171–198

Eberhart R, Shi Y (2001) Particle swarm optimization: developments, applications and resources. In: Proceedings of 2001 IEEE congress on evolutionary computation, vol 1, pp 81–86

Freire EB, Lykawka R, Gabaglia GPR, Rodrigues EB, Terra GJS (2011) Searching for potential Analogues for the pre-salt Santos Basin, Brazil: high-resolution stratigraphic studies of microbialite-bearing successions from Salta Basin, Argentina. In: AAPG international conference and exhibition

Goldszmidt R, Moises M (1997) Bayesian network classifiers. Machine Learning 29(2–3):131–163

Guerra MCM, Underhill JR (2012) Role of halokinesis in controlling structural styles and sediment dispersal in the Santos Basin, offshore Brazil. Geol Soc Lond Spec Publ 363:175–206

Guo Y, Ma H, Ba J, Yu H, Long C (2015) Impact of data distribution on fluid sensitivity analysis: a quantitative investigation. J Appl Geophys 119:1–15

Huc AY (2004) Petroleum in the south Atlantic. Oil Gas Sci Technol 59(59):243–253

Inaba T, Suzuki N, Hirai A, Sekiguchi K, Watanabe T (2001) Source rock lithology prediction based on oil diacholestane abundance in the siliceous-clastic Akita sedimentary basin, Japan. Organ Geochem 32(7):877–890

Jacobson LA (1988) Macroscopic thermal neutron capture cross section measurements. IEEE Trans Nucl Sci 35(1):817–821

Jacobson LA, Wyatt DF (1993) Elemental yields and complex lithology analysis from the pulsed spectral gamma log. SPWLA Annu Log Symp 37(1):50–61

Jain AK, Dubes RC (1988) Algorithms for clustering data. Technometrics 32(2):227–229

Kakouei A, Masihi M, Sola BS, Biniaz E (2014) Lithological facies identification in Iranian largest gas field: a comparative study of neural network methods. J Geol Soc India 84(3):326–334

Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of IEEE international conference on neural networks, vol 4, pp 1942–1948

Kominz MA, Patterson K, Odette D (2011) Lithology dependence of porosity in slope and deep marine sediments. J Sediment Res 81(9–10):730–742

Kou Y, Shi YM, Boren LI, Qin XS (2010) The complex lithology rock-electricity features of volcanic rocks in Kelameili gas field. Acta Petrol Sin 26(1):291–301

Lentini MR, Fraser SI, Sumner HS, Davies RJ (2010) Geodynamics of the central south Atlantic conjugate margins: implications for hydrocarbon potential. Petrol Geosci 16(3):217–229

Ma YZ (2011) Lithofacies clustering using principal component analysis and neural network: applications to wireline logs. Math Geosci 43(4):401–419

Malik RF, Rahman TA, Hashim SZM, Ngah R (2007) New particle swarm optimizer with sigmoid increasing inertia weight. Int J Comput Sci Secur 1(2):43–52

Michelon D, Marques E, Figueiredo J, Ferraz H, Barros P (2013) Santonian-Campanian channelized systems of the Santos Basin, Brazil: stratigraphic framework and reservoir potential. In: GCSSEPM proceedings, pp 381–396

Mirzaei-Paiaman A, Saboorian-Jooybari H (2016) A method based on spontaneous imbibition for characterization of pore structure: application in pre-SCAL sample selection and rock typing. J Nat Gas Sci Eng 35(A):814–825

Mirzaei-Paiaman A, Salavati S (2012) The application of artificial neural networks for the prediction of oil production flow rate. Energy Sources A Recov Util Environ Eff 34(19):1834–1843

Mirzaei-Paiaman A, Saboorian-Jooybari H, Pourafshary P (2015) Improved method to identify hydraulic flow units for reservoir characterization. Energy Technol 3(7):726–733

Mirzaei-Paiaman A, Ostadhassan M, Rezaee R, Saboorian-Jooybari H, Chen Z (2018) A new approach in petrophysical rock typing. J Petrol Sci Eng 166:445–464

Mirzaei-Paiaman A, Sabbagh F, Ostadhassan M, Shafiei A, Rezaee R, Saboorian-Jooybari H (2019a) A further verification of FZI* and PSRTI: newly developed petrophysical rock typing indices. J Petrol Sci Eng 175:693–705

Mirzaei-Paiaman A, Saboorian-Jooybari H, Chen Z, Ostadhassan M (2019b) New technique of true effective mobility (TEM-Function) in dynamic rock typing: reduction of uncertainties in relative permeability data for reservoir simulation. J Petrol Sci Eng 179:210–227

Murata N, Yoshizawa S, Amari SI (1994) Network information criterion-determining the number of hidden units for an artificial neural network model. IEEE Trans Neural Netw 5(6):865–872

Nakano C, Pinto AC, Marcusso J, Minami K (2010) Pre-salt Santos Basin extended well test and production pilot in the Tupi area-the planning phase. J Petrol Technol 62(2):66–68

Paola JD, Schowengerdt RA (1993) A review and analysis of back-propagation neural networks for classification of remotely-sensed multi-spectral imagery. Int J Remote Sens 16(16):3033–3058

Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33(3):1065–1076

Pediwal J, Mahor A, Khatri N (2012) Exponential decreasing inertia weight particle swarm optimization in economic load dispatch. Int J Eng Innov Res 1(5):380–384

Porter CR, Pickett GR, Whitman WW (1969) A method of determining rock characteristics for computation of log data: the litho-porosity cross plot. Log Anal 6:16–24

Qi M, Fu Z, Chen F (2016) Research on a feature selection method based on median impact value for modeling in thermal power plants. Appl Therm Eng 94:472–477

Riedmiller M (1994) Advanced supervised learning in multi-layer perceptrons-from backpropagation to adaptive learning algorithms. Comput Stand Interfaces 16(3):265–278

Roberts HH, Sydow J, Bouma AH (1993) Seismic stratigraphy, sedimentology, and reservoir potential of a late Pleistocene shelf-edge delta. AAPG Bull 77(9):1598–1599

Rosenfeld, J.H., Hood, J.F., 2006. Play potential in the deepwater Santos Basin, Brazil. Offshore, 66(9)

Sahoo S, Jha MK (2016) Pattern recognition in lithology classification: modeling using neural networks, self-organizing maps and genetic algorithms. Hydrogeol J 25(2):311–330

Shaw L, Bagha S (2012) Online EMG signal analysis for diagnosis of neuromuscular diseases by using PCA and PNN. Int J Eng Sci Technol 4(10):4453–4459

Shi Y, Eberhart R (1998) A modified particle swarm optimizer. In: Proceedings of 1998 IEEE international conference on evolutionary computation, pp 69–73

Shi Y, Eberhart R (1999) Empirical study of particle swarm optimization. In: Proceedings of 1999 IEEE congress on evolutionary computation, pp 1945–1950

Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods 31(4):265–273

Specht DF (1990) Probabilistic neural networks. Neural Netw 3(1):109–118

Stanton N, Ponte-Neto C, Bijani R, Masini E, Fontes S, Flexor JM (2014) A geophysical view of the southeastern Brazilian margin at Santos Basin: insights into rifting evolution. J S Am Earth Sci 55(2):141–154

Sun J, Li Q, Chen M, Ren L, Huang G, Li C, Zhang Z (2019) Optimization of models for a rapid identification of lithology while drilling: a win-win strategy based on machine learning. J Petrol Sci Eng 176:321–341

Thompson DL, Stilwell JD, Hall M (2015) Lacustrine carbonate reservoirs from early cretaceous rift lakes of western Gondwana: pre-salt coquinas of brazil and west Africa. Gondwana Res 28(1):26–51

Trelea IC (2003) The particle swarm optimization algorithm: convergence analysis and parameter selection. Inf Process Lett 85(6):317–325

Wang X, Wu C, Guo Y, Meng Q, Zhang Y, Tao Y (2013) Accumulation feature of Lula oilfield and its exploratory implication for pre-salt reservoirs in Santos Basin, Brazil. China Pet Explor 18(3):61–69

Wertheim PH (2008) Pre-salt discoveries continue in Brazil. Offshore 68(7)

Wertheim PH (2009) Tupi extended well test sheds light on Brazil's subsalt bonanza. Offshore 69(7)

Wong MLD, Nandi AK (2004) Automatic digital modulation recognition using artificial neural network and genetic algorithm. Signal Process 84(2):351–365

Wu C (2015) Petroleum geology characteristics and exploration targets of pre-salt formations in Santos Basin, Brazil. Petrol Geol Exp 37(1):1–5

Xie Y, Zhu C, Zhou W, Li Z, Liu X, Tu M (2018) Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances. J Petrol Sci Eng 160:182–193

Yong-Yan LU, Wang WG (2011) Variable selection of financial distress prediction-the SVM method based on mean impact value. Syst Eng 29(8):73–78

Zhang S (2008) Application of probabilistic neural network technique to lithology inversion of heterogeneous stratum. Acta Petrol Sin 29(4):549–552

Zhang DX, Chen YT, Meng J (2018) Synthetic well logs generation via recurrent neural networks. Pet Explor Dev 45(4):629–639

Zhu LQ, Zhang C, Zhang CM, Wei Y, Zhou XQ, Cheng Y, Huang YY, Zhang L (2018) Prediction of total organic carbon content in shale reservoir based on a new integrated hybrid neural network and conventional well logging curves. J Geophys Eng 15:1050–1061

Zhu LQ, Zhang C, Zhang CM, Zhang ZS, Zhou XQ, Liu WN, Zhu BY (2020) A new and reliable dual model- and data-driven TOC prediction concept: a TOC logging evaluation method using multiple overlapping methods integrated with semi-supervised deep learning. J Petrol Sci Eng 188:106944

RESEARCH ARTICLE - APPLIED GEOPHYSICS

Periods of refracted P-waves in coal seams and their applications in coal thickness estimations

Zuiliang Liu[1,2] · Ji Wang[3]

Abstract

The direct and accurate estimations of coal thicknesses are prerequisites for intelligent mining practices. One of the most effective methods for detecting the distributions of coal thicknesses in coal mining panels is the in-seam seismic (ISS) method. In the present study, after examining the formation processes and propagation characteristics of refracted P-waves in ISS data, it was concluded that the refracted P-waves in coal seams are mainly formed by the multiple transmission and reflection of the P-waves between the coal and rock interfaces of roof and floor at critical angles. This results in the refracted P-waves having strong periodicity, and these periods are proportional to the coal thicknesses. This study adopted numerical simulation models with different coal thicknesses, and the aforementioned periodicity characteristics were examined. It was found that the coal seam thicknesses could be calculated using the periods of the refracted P-waves. However, in thin- or medium-thick coal seams, it was found that multiple transmitted P-waves overlapped and the periods could not be read directly. Therefore, in order to solve this problem, this study composed source wavelets with the main frequency of the source signals and then composite synthetic P-waves by convoluting the source wavelets with the sequences of various coal thicknesses. The suitable estimated coal thickness corresponded to the minimum value of the errors between the synthetic and actual refracted P-waves. An experiment was conducted in the No. 42224 panel of the Chaigou Coal Mine in order to validate the proposed method. The experimental results revealed that the estimated coal thicknesses from the refracted P-waves were consistent with the actual geologic conditions in the coal mine. Due to the fact that the refracted P-waves arrive earlier than other waves in seismic records, the refracted P-waves could be easily identified and processed. Overall, the proposed method was found to be a simple application process for accurate coal thickness estimations.

Keywords Refracted P-wave · Coal seam thickness · In-seam seismic

Introduction

The methods currently applied in coal mining processes are moving toward intelligent choices which require the shearing or tunneling systems to predict the altitudes of the interfaces

---

✉ Ji Wang
  wangji@cctegxian.com

  Zuiliang Liu
  liuzuiliang@163.com

1. School of Resources and Geosciences, China University of Mining and Technology, Xuzhou, China

2. Yangquan Coal Industry (Group) Co., Ltd., Yangquan, China

3. Xi'an Research Institute of China Coal Technology and Engineering Group Corp., Xi'an, China

between coal seams and rock strata in order to adjust cutting heights in advance (Wang et al. 2020a). During these processes, the accurate detections of the coal seam thicknesses are one of the preconditions of the interface predictions (Wang et al. 2011). At the present time, there are two geophysical methods which have the ability to underground detect the thicknesses of coal seams, geological radar (Liu et al. 2019) and in-seam seismic (ISS) methods (Zhu et al. 2019). The principle of ISS thickness detection is based on the differences in dispersions and amplitude attenuations when the in-seam waves propagate in coal seams with different thicknesses. Therefore, ISS methods can be subdivided into the tomographic methods based on amplitude attenuations (Ji et al. 2014) and the inversion methods based on dispersion curves (Hu et al. 2018).

The velocities of the P-waves and S-waves in coal seams are significantly lower than those of surrounding rock. When

seismic waves propagate in coal seams, they will reflect and interfere repeatedly between the coal and rock interfaces of roof and floor forming in-seam waves. In previous related studies, Krey (1963) theoretically examined the origins of in-seam waves and presented a dispersion formula for them. According to the formula, under the conditions of unchanged velocities of the coal seams and surrounding rock, the frequencies and velocities of Airy phases can be determined by the coal thicknesses (RÄDer et al. 1985). Therefore, the thicknesses can be estimated by calculating frequencies and velocities of Airy phases from the ISS data and comparing them with the theoretical dispersion curves of different coal thicknesses (Hu et al. 2018). On that basis, Schott and Waclawik (2015) used the correspondence between the thicknesses and velocities to convert the velocity tomography into the distributions of coal seam thicknesses in the panels. However, this method requires high data quality. The actual underground seismic data tend to be seriously interfered with many types of noise. As a result, it has proven difficult to accurately determine the frequency and velocity values of Airy phases from actual data, producing errors to the corresponding results of the thickness estimations. Another method commonly used to estimate the coal thicknesses involves examining the amplitudes of the in-seam waves (Buchanan 1978). This method determined the tomographic attenuation coefficient of the in-seam wave amplitudes and then interprets the thicknesses according to the correlations between the attenuation coefficients and the coal seam thicknesses (Ji et al. 2014). This method has been found to be suitable for actual data with relatively low SNR. However, the thicknesses can only be roughly estimated at the present time due to the many complex factors affecting the attenuation of in-seam wave amplitudes.

Underground seismic data include not only in-seam waves, but also P-waves and S-waves (Yancey et al. 2007). In this research study, only P-waves are discussed. Differing from the P-waves in one or two layers, the P-waves in the coal seams which are surrounded by rock have long-wave trains. This is due to the fact that the P-waves propagate along the coal–rock interfaces and are guided by the coal seams. For that reason, the P-waves in the coal seams are also referred to as P–P interference waves or P-guided waves (Zhang et al. 2019). It has been observed that when P-waves propagate along the coal–rock interfaces, they will simultaneously refract to the coal seams and can be received by detectors installed in coal mines. Consequently, these waves are also referred to as refracted P-waves (Krey 1963; Regueiro 1990). Refracted P-waves are characterized by highest velocities in ISS recordings. Due to the fact that they arrive first at the detectors and do not interfere with other types of waves, they are very easy identify and process. Similar to in-seam waves, refracted P-waves can also be used to detect abnormal structures. For example, changes in velocity or

amplitude can be used for the tomographic images of a panel (Gritto 2003; Wang et al. 2020b), or the reflected waves are identify to detect any structures located in front of tunnel faces (Liu et al. 2019).

In this study, the propagations of the P-waves in coal seams were discussed and the relationships between the coal seam thicknesses and the periods of refracted P-waves were investigated. This study's results were verified using numerical simulation models with different coal seam thicknesses. It was found that in thin- or medium-thick coal seams, multiple transmitted P-waves had overlapped, resulting in the periods not being able to be observed directly. Therefore, an alternative method was proposed in this study to estimate the thicknesses, and the results were verified by both simulation data and actual data.

Theoretical analysis

It has been observed that when P-waves propagate from coal seam to rock stratum, transmission and reflection of P-wave will occur at the interface. During those processes, the wave velocity in the coal v_1, wave velocity in the rock v_2, incident angle α_i, and transmission angle α_t will meet Snell's law as follows (Sheriff and Geldart 1995):

$$\frac{v_1}{\sin \alpha_i} = \frac{v_2}{\sin \alpha_t} = v, \tag{1}$$

where v is the apparent velocity of the P-wave propagating along the interface. As $v_2 > v_1$, so $\alpha_t > \alpha_i$, which means that the transmission angle is larger than the incident angle. Then, as the distance between the source and the incident point increases, the incident angle and the transmission angle will also increase. When the transmission angle increases to a right angle, the incident angle at that time is referred to as the critical angle. If the incident angle is larger than the critical angle, then the transmission angle becomes a complex, and the in-seam waves are formed by the reflection waves. In addition, if the incident angle is equal to the critical angle, the transmission angle is a right angle, and the transmission wave propagates along with the coal–rock interface with the velocity v_2. Therefore, it can be assumed that the waves will propagate at the high velocity v_2 on the rock side of the interfaces and at the low velocity v_1 on the coal side. When the high-speed waves pass through the interface, they will simultaneously produce refracted waves in the coal seam. These waves are refracted from the coal–rock interfaces to the coal seam at v_1 until they are received by receivers. However, since the distances between the source and the detectors tend to be much larger than the thicknesses of the coal seam, the propagation of the waves in the coal can be neglected. Therefore, the apparent velocity of the

refracted waves will be equal to the velocity of the waves propagating at the coal–rock interface.

When P-waves incident to a coal–rock interface, there are four types of waves generated: transmitted P-wave; reflected P-wave; transmitted SV-wave; and reflected SV-wave. The Zoeppritz equation gives the relationship between the incident angle and the amplitude changes of the different types of waves as follows (Sheriff and Geldart 1995):

$$
\begin{bmatrix}
-\sin\alpha_i & -\cos\beta_i & \sin\alpha_t & \cos\beta_t \\
\cos\alpha_i & -\sin\beta_i & \cos\alpha_t & -\sin\beta_t \\
\sin 2\alpha_i & \frac{v_{p1}}{v_{s1}}\cos 2\beta_i & \frac{\rho_2 v_{s2}^2 v_{p1}}{\rho_1 v_{s1}^2 v_{p2}}\sin 2\alpha_t & \frac{\rho_2 v_{s2} v_{p1}}{\rho_1 v_{s1}^2}\cos 2\beta_t \\
-\cos 2\beta_i & \frac{v_{s1}}{v_{p1}}\sin 2\beta_i & \frac{\rho_2 v_{p2}}{\rho_1 v_{p1}}\cos 2\beta_t & -\frac{\rho_2 v_{s2}}{\rho_1 v_{p1}}\sin 2\beta_t
\end{bmatrix}
\begin{bmatrix}
R_{pp} \\ R_{ps} \\ T_{pp} \\ T_{ps}
\end{bmatrix}
=
\begin{bmatrix}
\sin\alpha_i \\ \cos\alpha_i \\ \sin 2\alpha_i \\ \cos 2\beta_i
\end{bmatrix},
\tag{2}
$$

where for the P-wave, the incident angle and reflection angle are all α_i, the transmission angle is α_t, the velocity in the coal is v_{p1}, and the velocity in rock is v_{p2}; for the SV-wave, the reflection angle is β_i, the transmission angle is β_t, the velocity in the coal is v_{s1}, and the velocity in rock is v_{s2}; ρ_1 and ρ_2 are the densities of coal and rock, respectively. T_{pp} is the transmission coefficient of the P-wave, which represents the ratio of the amplitude of the transmitted P-wave to that of the incident P-wave. Correspondingly, R_{pp} is the reflection coefficient of the P-wave; T_{ps} is the transmitted coefficient of the SV-wave; and R_{ps} is the reflection coefficient of the SV-wave.

Therefore, in accordance with the Zoeppritz equation, this study analyzed the relationships among the transmission coefficient T_{pp}, reflection coefficient R_{pp} and incident angle α_i of the P-waves. Model with parameters is shown in Table 1. In addition, using Eq. (1), α_t, β_i and β_t are calculated from α_i and then substituted into Eq. (2) in order to obtain T_{pp} and R_{pp}. The results are shown in Fig. 1.

As can be seen in Fig. 1, when the incident angle was smaller than the critical angle, T_{pp} and R_{pp} are about 0.5. When the incident angle was similar to the critical angle, these two coefficients increased rapidly. They were observed to reach the maximum values when the incident angle was equal to the critical angle. Next, they rapidly decreased, but R_{pp} increased again to form total reflections. It can be seen that for the P-waves, only when the incident angles were equal to the critical angle where the transmitted and reflected waves are the strongest. However, if the incident angles were insufficient or exceeded the critical angles,

then the amplitudes of the transmitted and reflected waves had decayed rapidly. Therefore, the refracted P-waves were mainly generated by the P-wave incidents at the critical angles. Meanwhile, wave incidents at other angles had been attenuated. In addition, since the value of R_{pp} at the critical angle was close to 1, or even greater than 1, it could be considered that the amplitudes of the transmitted and reflected waves were not attenuated, but were enhanced under the condition of critical angles.

Only considered the incident, transmitted and reflected P-waves at the critical angle condition, the propagation
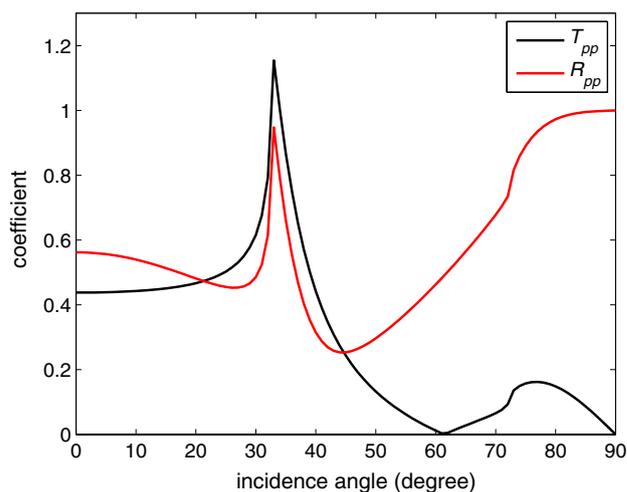
Fig. 1 Transmission coefficient T_{pp} and reflection coefficient R_{pp} for the various degrees of the incident angles

Fig. 2 Multiple transmissions and reflections of the P-wave between the coal–rock interfaces

Table 1 Parameters of the model

Layer	P-wave velocity (m/s)	S-wave velocity (m/s)	Density (g/cm³)
Roof and floor strata	3700	2100	2.7
Coal seam	2000	1100	1.3

process of a refracted P-wave in an ideal coal seam with homogeneous surrounding rock is shown in Fig. 2. As can be seen in this figure, the P-wave was excited at Point S. Then, at Point A, the wave was an incident of the coal–rock interface at the critical angle α and generated the transmitted and reflected waves. The transmitted wave propagated to Point B along the coal–rock interface at the velocity v_{p2}. During the propagation process, it was continuously refracted to the coal seam. The refracted wave at Point B propagated to the coal seam with the refraction angle α and was received by the detector located at Point R in the coal seam. The reflected wave generated at Point A also propagated at velocity v_{p1} in the coal seam with the reflection angle α. After arriving at Point C on the other coal–rock interface, this wave was again reflected with the reflection angle α and continued to propagate at v_{p1} in the coal seam. Subsequently, when it reached Point D, transmission and reflection occurred once again due to the incident angle still being the critical angle, and the transmission wave continued to propagate along the coal–rock interface at v_{p1}. Therefore, the refracted P-wave received by the detector was the result of multiple transmissions and reflections between the two coal–rock interfaces. Therefore, by assuming that the thickness of the coal seam was d, the time difference T between the two arrived P-waves can be obtained according to the geometric relationships between the rays, as detailed in Fig. 2.

$$T = 2d \frac{\sqrt{v_2^2 - v_1^2}}{v_1 v_2}. \tag{3}$$

Then, according to Eq. (3), the refracted P-wave received in the coal seam was considered to be periodic. Therefore, under the condition that the P-wave velocity of the coal seam and the surrounding rock was constant, the period of the refracted P-wave was proportional to the coal thickness.

Modeling and simulation

In order to verify the periodicity of the refracted P-waves, three numerical models of panels with different coal thicknesses were established in this study. Each model consisted of three strata areas: roof, coal seam and floor. The parameters of the roof and floor were the same, and the specific parameters are shown in Table 1. The coal thicknesses of the three models were 20 m, 10 m and 5 m, respectively, and the plans and profiles are outlined in Fig. 3.

All models have the same size, 1000 m × 300 m × 30 m $(X \times Y \times Z)$, with grid intervals of 1 m × 1 m × 0.5 m. All models contained two airways with a spacing of 200 m. The source was located in the middle of one airway, while the receivers were arranged in the other airway, with the spacing of 10 m. All of the sources and receivers were located close to the floor. The source signal was a 500 Hz Ricker wavelet, and the sampling time was 0.1 ms. A high-frequency source signal was selected to make wavelets narrow in time domain so as to clearly distinguish the periodicity of the refracted P-wave. The high-order staggered-grid finite difference method was applied for the three-dimensional simulations (He et al. 2017). The simulated Z-components of refracted P-waves are shown in Fig. 4.

It can be seen in Fig. 4 that the refracted P-waves have different periods in the simulated data of the different coal thickness models. Among those, the periods of the 20 m model and 10 m model could be read directly from the seismic records, at 18 ms and 9 ms, respectively, which was consistent with the results calculated by Eq. (3). However, for the 5 m model, since the period was less than the duration of the P-wave wavelet, the P-waves which had arrived at multiple times were observed to be overlapping, which caused the period to be unreadable in the seismic record. Therefore, although the refracted P-waves were periodic, this phenomenon had higher requirements on the thickness of coal seam and the frequency of source. However, for the

Fig. 3 Plans and profiles of the panel models

(a) 20 m **(b)** 10 m **(c)** 5 m
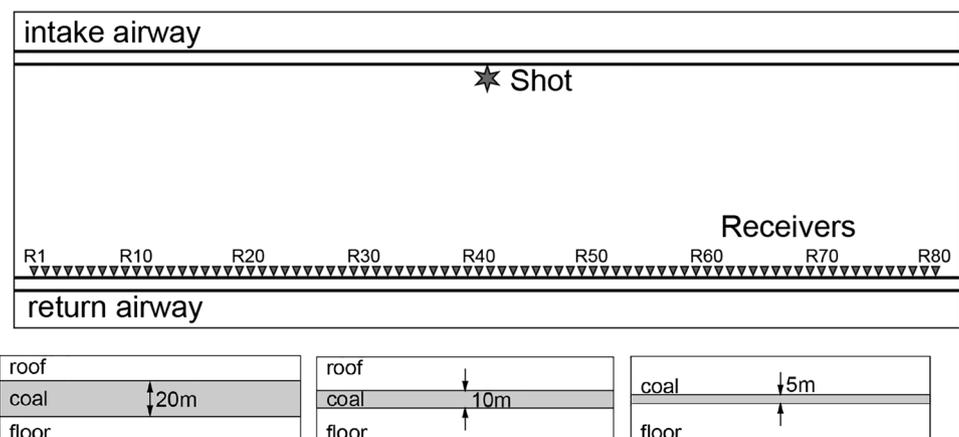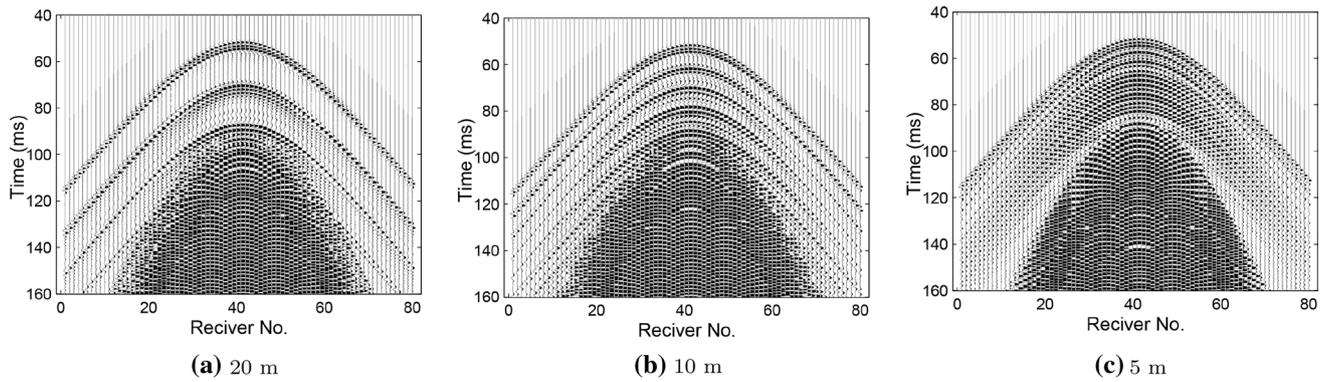
Fig. 4 Synthetic *Z*-components of refracted P-waves of models with different coal thicknesses

thick coal seams, the periods of the refracted P-waves could be obtained directly by using high-frequency sources, and the thicknesses of the coal seams could be accurately calculated. The determinations of the thicknesses of the coal seams were generally within 10 m. Therefore, it was found that in order to effectively measure the periods and obtain the thicknesses, it was necessary to process the data so as to eliminate the effects of aliasing.

It is important to place sources and receivers near either the roof or floor. For the same 20 m model, Fig. 5 shows the differences in the refracted P-waves under the conditions where all sources and receivers are located either at the bottom (Fig. 5a) or middle (Fig. 5b) of the coal seam. When located at the middle, receivers are affected by both upgoing and downgoing waves. Therefore, the period of refracted P-waves in Fig. 5 appears to be half of that in Fig. 5b. In addition, the upgoing and downgoing waves interfere with each other and take interference to refracted P-waves. This results in the noise being Fig. 5b is greater than that in Fig. 5a.

Algorithms and principles

It was found in this study that since the phases of the reflected waves and transmitted waves were constant under the condition of the critical angle, the wave form of the P-waves remained unchanged after multiple reflections and transmissions between the coal–rock interfaces. Therefore, the received refracted waves *s*(*t*) could be regarded as the convolution of the source wavelet *w*(*t*) with a sequence *a*(*t*) as follows:

$$s(t) = w(t) * a(t), \tag{4}$$

where the sequence could be expressed as:

$$a(t) = \begin{cases} 1 & t = nT \\ 0 & t \neq nT. \end{cases} \tag{5}$$

The refracted waves *s*(*t*) were recorded by the detector at the receiver point, and the source wavelets *w*(*t*) were estimated from the signals recorded by the detector near the source. As *s*(*t*) and *w*(*t*) are known, many deconvolution methods can be applied to obtain *a*(*t*), and then *T* can be calculated. However, if only one parameter *T* was estimated, a simpler method could be used.

Fig. 5 Refracted P-waves of the 20 m model with different sources and receivers location

(a) bottom **(b)** middle

In order to estimate the source wavelets, this study first calculated the main frequency f_p of the signal acquired near the source point. Then, the source wavelet was approximated using a minimum phase wavelet $w(t)$ as follows:

$$w(t) = e^{-2\pi f_p^2 t^2 \ln(k)} \sin(2\pi f_p t), \tag{6}$$

where k is a factor of the amplitude attenuation, and its value can be selected between 1.5 and 2.5 according to the actual data.

Referring the geological conditions of the panel, an interval was set to limit the variations of the coal thicknesses. For each thicknesses value d in this interval, the period T was calculated using Eq. (3). Then, a sequence $a(t)$ was constructed with Eq. (4). Subsequently, $a(t)$ was convoluted with $w(t)$ to form a synthetic signal of the refracted wave $s(t)$. The error δ^2 between the synthetic signal $s(t)$ and the actual received signal $s_1(t)$ was then calculated using the following:

$$\delta^2 = \sum_{t=0}^{N} \left(s(t) - s_1(t)\right)^2, \tag{7}$$

where N indicates the wavelet length. After all of the values in the interval had been used to form the synthetic signals, the corresponding errors were calculated. The estimated coal thickness was d, which made the errors the minimum values.

As an example, the above method was applied to process the seismic data of the 5 m coal thickness model shown in Fig. 4. The source wavelet was constructed according to Eq. (6), where $k = 1.8$, and the main frequency was 500 Hz. The constructed source wavelet is shown in Fig. 6. The interval of the coal thickness was set between 1 and 20 m. All of the errors between the actual
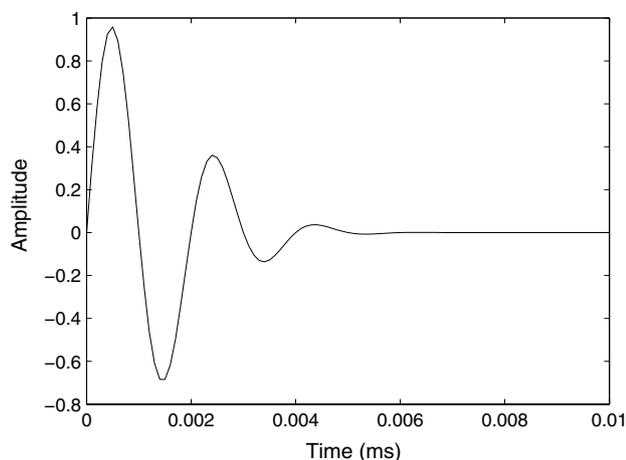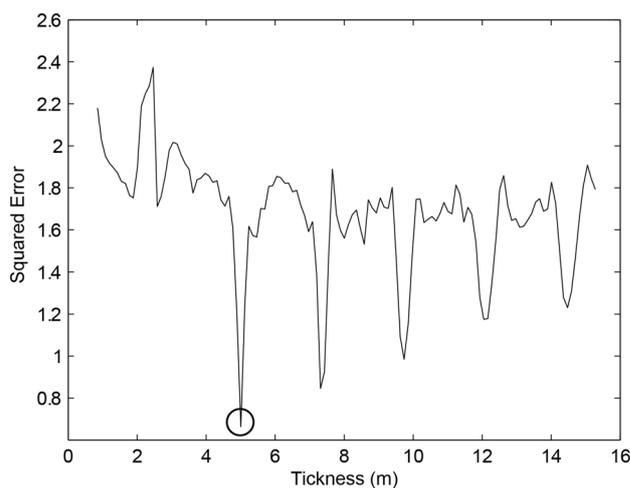
Fig. 7 Errors between the refracted P-wave signals of the 5 m model and the synthetic signals of the various coal thicknesses

refracted P-wave signals and the synthetic signals for the various thickness values are shown in Fig. 7. It can be seen that the minimum error occurred at 5 m, which was consistent with the coal thickness in that model. Figure 8 compares the synthetic signal of 5 m thickness with signal of the refracted P-wave from Fig. 4c.

Case study

The method proposed in the study was applied to a real case in order to assess its performance. The width of the No. 42224 panel in the Chaigou Mine was 122 m, and the average thickness of the coal seam was 7 m. In accordance with the airway exposure and borehole detection data, a scouting zone existed in the coal seam of the aforementioned panel. In order to verify the effectiveness of the proposed method for coal thickness detection, two positions are selected in the panel characterized with different coal thicknesses. There are two excitement shots (S1 and S2) in the intake airway. Several detectors (designated as R1 to R18) were arranged in the return airway for the purpose of receiving the seismic waves. All of the shots and detectors are located near the floor. The details of the observational system are shown in Fig. 9. Among the features, S1, R1,..., R9 were located in a normal area with approximately an 8 m coal thickness, while S2, R10,..., R18 were located in the scouting zone with approximately a 4 m coal thickness. R1 to R9 were received when S1 was excited, and R10 to R18 were received when S2 was excited. The coverage areas of the rays are shown in Fig. 9.

Shots S1 and S2 were excited within a 2 m hole with 120 g of explosive. The receivers R1 to R18 were embedded in holes with depths of 1 m and spacings of 20 m using

Fig. 6 Source wavelet with a main frequency of 500 Hz
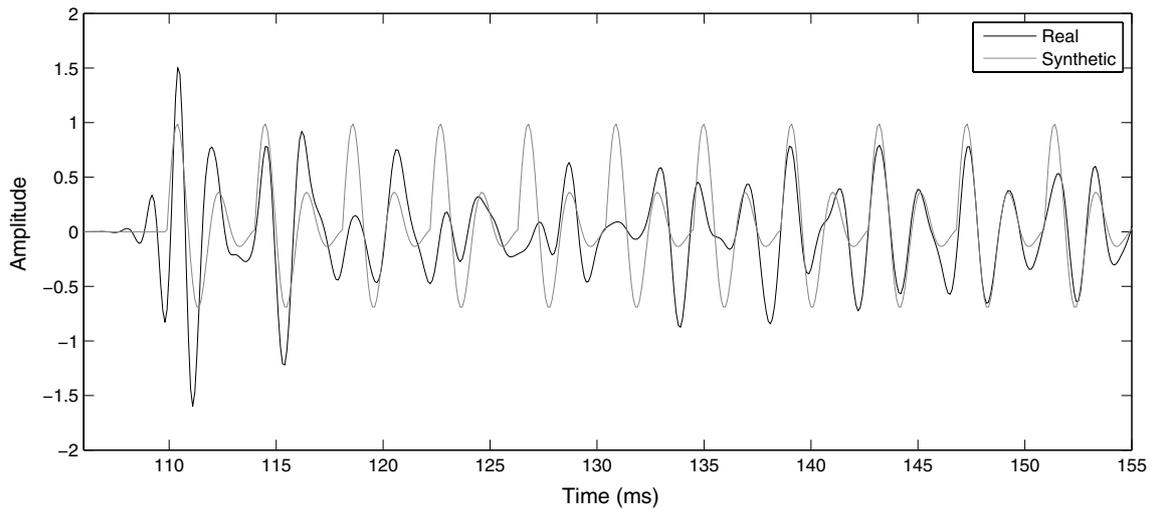
Fig. 8 Comparison of the synthetic signal of 5 m thickness with the signal of refracted P-wave from the 5 m model
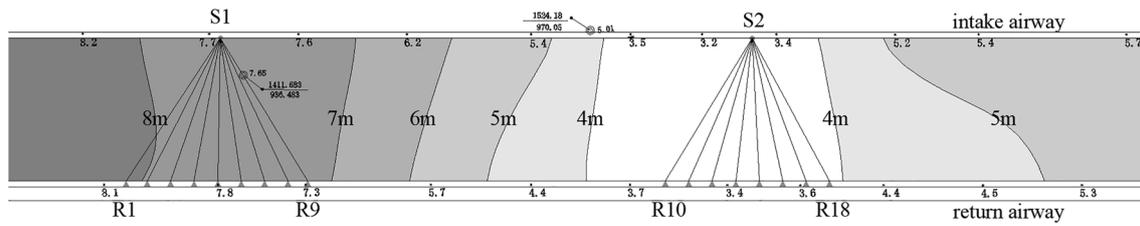
Fig. 9 Distribution of the coal thicknesses in the No. 42224 panel and the layout of the observational system
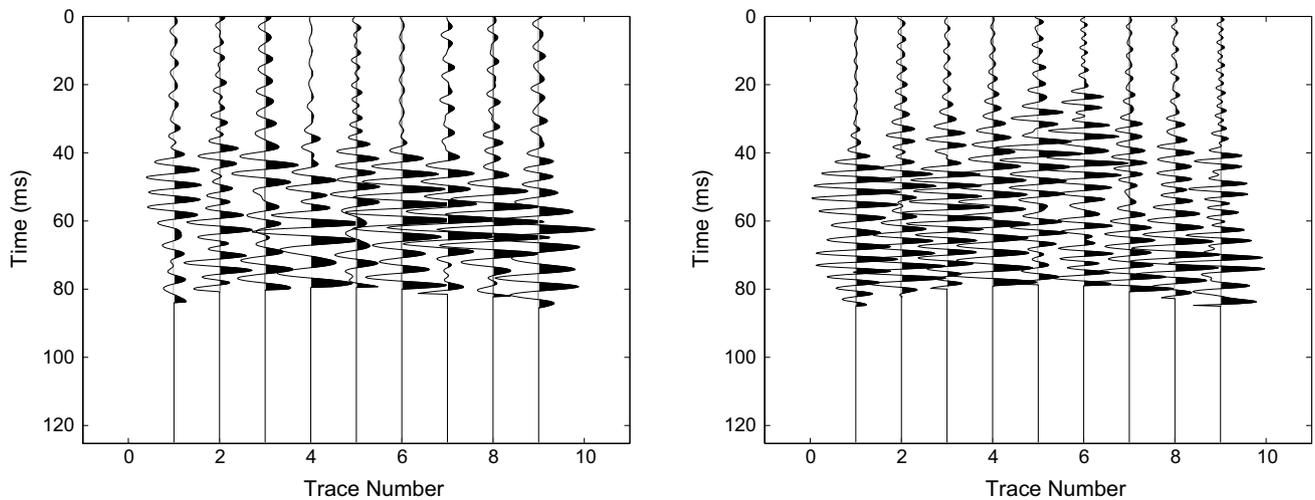
Fig. 10 The refracted P-waves of S1 (left) and S2 (right)

Z-component detectors. After pre-processing the received seismic data, this study uses 4200 m/s as the P-wave velocity of the surrounding rock for the purpose of calculating the first arrival times t_0 and to cut off the data after $t_0 + 40$ ms in order to retain the refracted P-wave only. The refracted P-waves of S1 and S2 are shown in Fig. 10. It can be seen that frequency of the S1 signals was general slightly lower than that of the S2 signals.
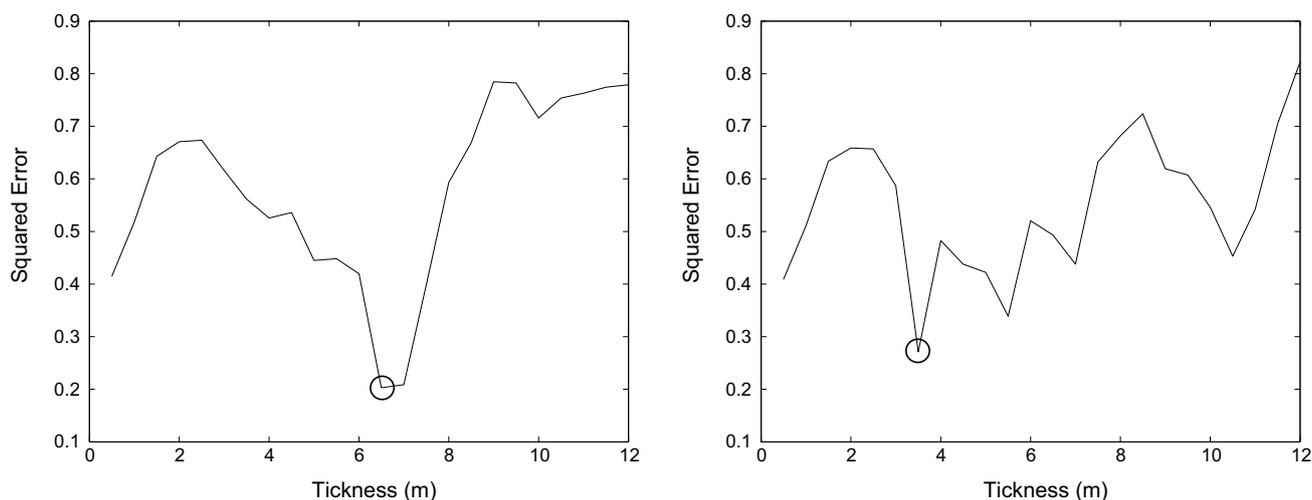
Fig. 11 Errors between the synthetic signals of the various coal thicknesses and the refracted P-waves of S1 (left) and S2 (right)

The data acquired from the detectors placed near Points S1 and S2 were analyzed, and 260 Hz was selected as the main frequency for generating the source wavelets and process the signals of S1 and S2 using the proposed method. The errors between the actual refracted P-wave signals and the synthetic signals of the various coal thicknesses obtained according to Eq. (4) are shown in Fig. 11. It can be seen from the figure that the minimum error was reached when the coal thickness was 3.8 m for S1 and 7 m for S2, which was found to be consistent with the actual geology. Figure 12 details comparison of the actual signal of the trace R7 with the synthetic signal at 3.8 m, as well as comparison of the signal of the trace R12 with that at 7 m. It can be seen that the synthesized signals of the appropriate coal thicknesses had been in good agreement with the actual P-wave signals.

Conclusions

The seismic waves which are excited and received in coal seams include both refracted P- and S-waves. The apparent velocity of the refracted P-wave has been found to be close to the P-wave velocity of the surrounding rocks, yet displays strong periodicity characteristics. These periods have been determined to be proportional to the thicknesses of the coal seams. The reason for this phenomenon is that the amplitudes of the transmitted P-waves and reflected P-waves reach the maximum values at the same time only when the incident angle is the critical angle. Therefore, P-waves can be reflected multiple times at the critical

angle between two coal–rock interfaces. In addition, each reflection produces a strong transmission wave passing along the interface. Therefore, the received refracted P-waves have strong periodicity. This study's simulations of various coal thicknesses using various models showed that the periodicity of the refracted P-waves could be directly observed when excited by focused high frequency in thick coal seams. The coal thicknesses could be simply calculated by the period. However, it was found that in thin- or medium-thick coal seams, the multiple transmitted P-waves tended to overlap, making it difficult to directly perform estimations. In order to address these issues, this study first composed the source wavelets and then composite synthetic P-waves of various coal thicknesses. It was found that the suitable estimated coal thicknesses corresponded to the minimum values of the errors between the synthetic and actual refracted P-waves.

In this study, three assumptions were made in order to simplify the processes of wave formation and propagation:

1. The lithology of roof and floor strata was considered to be the same. If the differences were significant, then the two critical angles on the coal–rock interfaces would also differ greatly.
2. The two interfaces were parallel to each other. This condition ensured that the reflected waves on one interface would occur at the critical angle on the other interface.
3. The coal thicknesses between the source point and the receiving point were constant. As a result, the coal thicknesses estimated by the proposed method could be regarded as the average thicknesses of the coal seam between those two points.
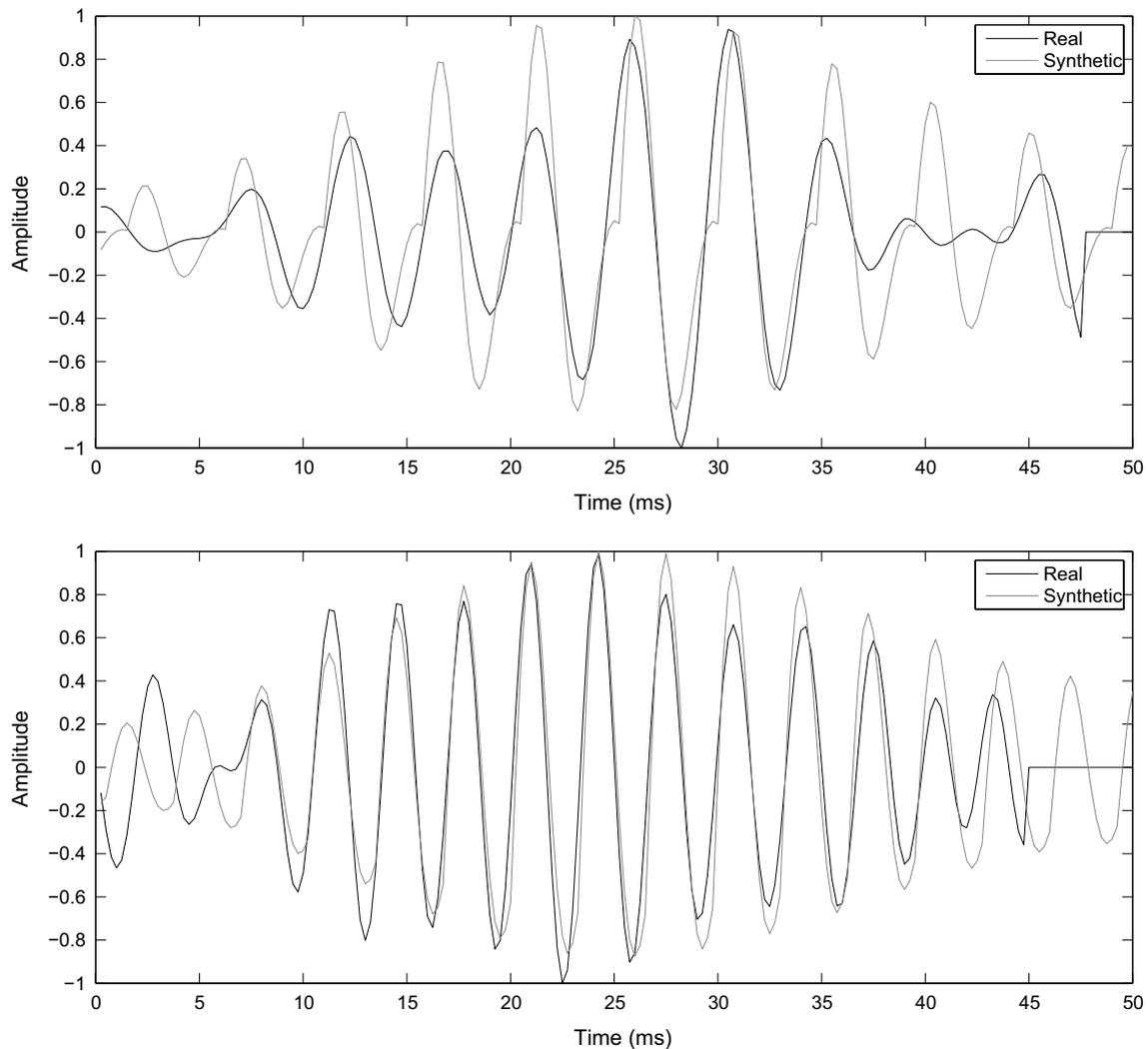
Fig. 12 Comparison of the synthetic signal with the actual signal from S1 (up) and S2 (down)

The local variations of the coal seams in regard to lithology, dip angles and thicknesses will result in local changes in the amplitudes and periodicity of the refracted P-waves. Therefore, it should be considered feasible to use the refracted P-waves to predict the thickness distributions of coal seams in mining panels in the future.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Buchanan DJ (1978) The propagation of attenuated SH channel waves. Geophys Prospect 26(1):16–28

Gritto R (2003) Subsurface void detection using seismic tomographic imaging. Lawrence Berkeley National Laboratory, Berkeley

He W, Ji G, Dong S et al (2017) Theoretical basis and application of vertical Z-component in-seam wave exploration. J Appl Geophys 138:91–101

Hu Z, Zhang P, Xu G (2018) Dispersion features of transmitted channel waves and inversion of coal seam thickness. Acta Geophys 66(5):1001–1009

Ji G, Cheng J, Hu J (2014) In-seam wave imaging using attenuation coefficient: method and application. J China Coal Soc 39(S2):471–475 (**in Chinese**)

Krey TC (1963) Channel waves as a tool of applied geophysics in coal mining. Geophysics 28(5):701–714

Liu S, Zhang J, Li C (2019) Method and test of mine seismic multi-wave and multi-component. J China Coal Soc 44(01):271–277 (**in Chinese**)

Liu S, Zhao W, Gao S (2019) Experimental study on coal seam thickness measurement of ultra-wide band ground penetrating radar. Coal Sci Technol 47(08):207–212 (**in Chinese**)

RÄDer D, Schott W, Dresen L et al (1985) Calculation of dispersion curves and amplitude-depth distributions of Love channel waves in horizontally layered media. Geophys Prospect 33:800–816

Regueiro SJ (1990) Seam waves: what are they used for? Part 2. Lead Edge 9(8):32–34

Schott W, Waclawik P (2015) On the quantitative determination of coal seam thickness by means of in-seam seismic surveys. Can Geotech J 52(10):1496–1504

Sheriff RE, Geldart LP (1995) Exploration seismology, 2nd edn. Cambridge Univ. Press, Cambridge

Wang B, Liu S, Jiang Z et al (2011) Advanced forecast of coal seam thickness variation by integrated geophysical method in the laneway. Procedia Eng 26:335–342

Wang G, Pang Y, Ren H (2020a) Intelligent coal mining pattern and technological path. J Min Strata Control Eng 2(01):5–19 (**in Chinese**)

Wang J, Liu Z, Niu H (2020b) Study on influence of collapse columns on refracted waves between coal seam and rocks and corresponding tomography method. Coal Sci Technol 48(02):214–219 (**in Chinese**)

Yancey DJ, Imhof MG, Feddock JE et al (2007) Analysis and application of coal-seam seismic waves for detecting abandoned mines. Geophysics 72(5):M7–M15

Zhang J, Liu S, Wang B et al (2019) Response of triaxial velocity and acceleration geophones to channel waves in a 1-m thick coal seam. J Appl Geophys 166(166):112–121

Zhu M, Cheng J, Cui W et al (2019) Comprehensive prediction of coal seam thickness by using in-seam seismic surveys and Bayesian kriging. Acta Geophys 67(3):825–836

RESEARCH ARTICLE - HYDROLOGY

Daily streamflow prediction using support vector machine-artificial flora (SVM-AF) hybrid model

Reza Dehghani[1] · Hassan Torabi Poudeh[2] · Hojatolah Younesi[2] · Babak Shahinejad[2]

Abstract

Precise estimation of river flow in catchment areas has a significant role in managing water resources and, particularly, making firm decisions during flood and drought crises. In recent years, different procedures have been proposed for estimating river flow, among which hybrid artificial intelligence models have garnered notable attention. This study proposes a hybrid method, so-called support vector machine–artificial flora (SVM-AF), and compares the obtained results with outcomes of wavelet support vector machine models and Bayesian support vector machine. To estimate discharge value of the Dez river basin in the southwest of Iran, the statistical daily watering data recorded by hydrometric stations located at upstream of the dam over the years 2008–2018 were investigated. Four performance criteria of coefficient of determination (R^2), root-mean-square error, mean absolute error, and Nash–Sutcliffe efficiency were employed to evaluate and compare performances of the models. Comparison of the models based on the evaluation criteria and Taylor's diagram showed that the proposed hybrid SVM-AF with the correlation coefficient $R^2 = 0.933$–0.985, root-mean-square error RMSE = 0.008–0.088 m^3/s, mean absolute error MAE = 0.004–0.040 m^3/s, and Nash-Sutcliffe coefficient NS = 0.951–0.995 had the best performance in estimating daily flow of the river. The estimation results showed that the proposed hybrid SVM-AF model outperformed other models in efficiently predicting flow and daily discharge.

Keywords Artificial flora · Prediction · Streamflow · Support vector machine

Introduction

Precise estimation of river flow is the most important factor in managing flood and preventing economic losses. Therefore, it is of necessity to secure a viable method for estimating river flow (Beven and Kirkby 1979; Duan et al. 1992, 1994; Cameron et al. 1999; Dolling and Eduardo 2002; Guven and Kisi 2011a, b; Coken et al. 2016). Prediction models of river flow are appropriate tools for managing water resources. Predicting the probability of surface run-off and having prior information of available water amount can be put into good use, e.g., (a) facilitating irrigation, production of electricity, and flood control; (b) ensuring better water allocations to industries and agriculture; (c)

controlling water pollution and expanding recreational and green areas; and (d) reaping economic profits (Jakeman and Hornberger 1993; Hsu et al. 1995; Haykin 1998; Guven and Kisi 2011a, b; Kumar and Jothiprakash 2013). Statistical models and regression analysis are the most common analytical methods that are often subject to erroneous results and are unable to model time changes with sufficient accuracy due to the linear analysis of phenomena. Selecting a model that could consider many internal and external factors to estimate river flow with sufficient accuracy is essential (Robert and Linda 2000; Pulido-Calvo and Portela 2007; Macfarlane and Ogden 2012; Lima et al. 2016). Nowadays, artificial intelligence systems are extensively used to assess nonlinear phenomena (Salas et al. 2000; Tokar and Markus 2000; Solomatine 2005; Srivastava et al. 2006). Support vector machine (SVM) model is a popular method in the domain of hydrology and enjoys an appropriate performance. To increase the accuracy and reduce the rate of errors of the mentioned model (Boser 1992; Yu et al. 2005; Bhagwat and Maity 2012), some optimization algorithms have been used in recent years (Suliman et al. 2013; Kalteh 2013; Gizaw and

✉ Hassan Torabi Poudeh
  torabi.ha@lu.ac.ir

1 Lorestan University, Khorramabad, Iran

2 Department of Water Engineering, Lorestan University, Khorramabad, Iran

Gan 2016). Lately, artificial intelligence techniques including SVMs and other methods have been widely applied to the domain of hydrology for water resources applications (Kisi and Cobaner 2009). This study employs SVM-based models as put forward by Vapnik (1998).

SVM is efficiently applicable to problems associated with regression, nonlinear classification, and time series prediction (Wang et al. 2008). It is in the same league as kernel-based learning approaches and acts as a supervised machine learning system that uses feature space, i.e., a linear high-dimensional hypothesis space. The aforementioned characteristics make SVM widely popular. As its working principle, SVM implicitly uses kernel functions and maps the data to a higher-dimensional space (Bhagwat and Maity 2012). Examples of models with SVM capability include stage-discharge modeling (Sivapragasam and Muttil 2005; Aggarwal et al. 2012), streamflow or stage modeling (Asefa et al. 2004; Yu et al. 2006), and rainfall–runoff modeling (Dibike et al. 2001). Intelligence and hybrid models are used for predicting river flow.

Huang et al. (2014) employed the SVM model for predicting the monthly river flow of China Huaxi station and found that the model enjoyed acceptable prediction accuracy. To predict the surface runoff amount in the Roodak catchment area located in the northeast of Tehran, Sedighi et al. (2016) applied the artificial neural network model and SVM based on the MODIS image data obtained during 2003–2005 and showed that the SVM could estimate the amount of surface runoff. Ghorbani et al. (2016) predicted the daily water flow of Cypress, Texas, by utilizing support vector models and artificial neural network. They also used correlation coefficient and root-mean-square error to assess the model performance and found the performance of the support vector model quite satisfactory in predicting river flow. According to their findings, the model outperformed the artificial network model in terms of accuracy. Zhu et al. (2016) predicted Jinsha river flow in China and applied the SVM model to discrete wavelet transform (DWT) and empirical mode decomposition (EMD). The results showed that EMD and DWT were efficient in predicting flow. The results of EMD–DWT comparison showed that the vector machine model equipped with DWT outperformed EMD in terms of efficiency. To predict monthly flow in Igdir river in Turkey, Ghorbani et al. (2018a, b) investigated the performance of hybrid artificial intelligence models and found that the combination model with the firefly algorithm had better performance than other combination models. Ghorbani et al. (2018a, b) studied the performances of artificial neural network models and SVM to estimate the discharge of Zarineh-rood river in Iran, and their results pointed out that the artificial neural network model had better accuracy than SVM. Adnan et al. (2019) employed SVM models to predict the monthly discharge of Swat river in Pakistan and confirmed its satisfactory performance. Alizadeh et al. (2020) probed

the ability of hybrid support-wavelet vector machine model to predict the daily river flow in the USA and confirmed its highly satisfactory prediction accuracy.

Based on the significance of the above results and given that the Dez river is one of the main catchment areas in Iran that provides much water for drinking and nearby agriculture, especially under low river water volume due to the recent drought, the importance of simulating river flow and taking constructive actions for water management is now felt more than ever. Therefore, this study aims to predict the daily flow of the Dez catchment area by using the proposed hybrid models of SVM-AF and to compare obtained results with those of other hybrid models.

Materials and methods

The studied region

Iran has six main catchments: Central Plateau, Hamoon, Sarakhs, Caspian Sea, Lake Urmia, Persian Gulf, and Sea of Oman. The Persian Gulf and the Sea of Oman are the most important catchments in Iran. Several catchments including Karune-Bozorg basin were considered in this study. The catchment area of Dez river as a third-class catchment is a sub-basin of Karoun area in the country of Iran, and in a broader sense, it is categorized as a subgroup of the Persian Gulf and Oman Sea. Sezar and Bakhtiari are the two main branches of Dez river. Sezar river is situated northeast of Dez area and comprises Marbereh, Tireh, and Sabzeh. The attachment of waterways including Azna river in Aligoodarz region constitutes Marbereh river, which flows westward in Lorestan province. Dez catchment area in southwestern Iran is located in a semiarid, mountainous region at 48° 9′ 15″ to 50° 18′ 37″ E altitude and 31° 35′ 51″ to 34° 7′ 46″ N latitude. Figure 1 shows Dez catchment and other selected stations, whose characteristics are shown in Table 1.

Support vector machine

Vapnik et al. developed SVM in the early 1990s (Vapnik 1998; Misra et al. 2009). SVM incorporates structural risk minimization (SRM) principle, which reduces any possible error relating to the learning model, minimizes the overfitting problem, and facilitates more efficient generalization (Vapnik 1998). SVM represents a productive learning system and is coupled with optimization theory to minimize possible structural errors and attain an optimum response (Vapnik 1998). In the regression model, the SVM is a function of the dependent variable Y, which is in turn a function of the independent variable X. Similar to other regression-based problems, it is supposed that the relation between independent and dependent variables and algebraic function $f(x)$ plus the allowable error (ε) is determined as follows (Hamel 2009).
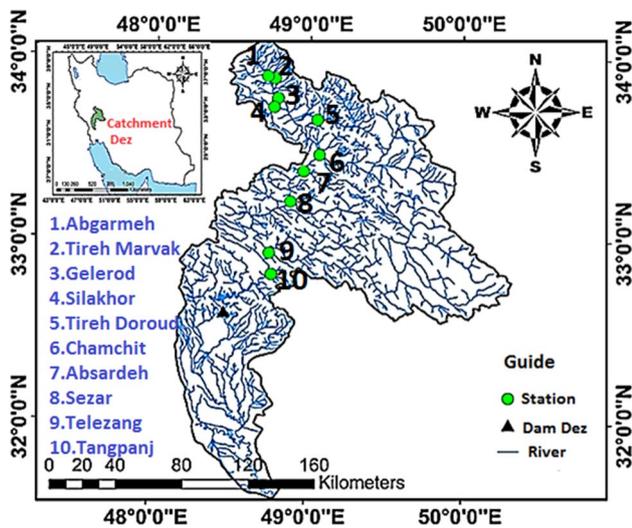
Fig. 1 The studied region

Table 1 Station characteristics

	Station name	Area (km^2)	Latitude	Longitude
1	Ab Garmeh	288	33° 54′ 7″	48° 43′ 19″
2	Tireh marvak	3400	33° 28′ 37″	49° 3′ 46″
3	Geleh-rood	69	33° 54′ 8″	48° 36′ 23″
4	Silakhor	1000	33° 46′ 43″	48° 47′ 52″
5	Tireh	3400	33° 28′ 37″	49° 3′ 46″
6	Cham Chit	345	33° 22′ 43″	47° 57′ 58″
7	Ab Sardeh	223	33° 44′ 13″	48° 45′ 17″
8	Sezar	9434	33° 54′ 7″	49° 33′ 15″
9	Taleh Zang	500	32° 49′	48° 46′ 17″
10	Tang Pangeh	555	33° 56′	48° 46′

$$f(x) = W^T \cdot \emptyset(x) + b \tag{1}$$

$$y = f(x) + \text{noise} \tag{2}$$

If W^T, b, and $\emptyset$ represent the transpose of the coefficient vector, a constant of regression function characteristics, and Kernel function, respectively, the objective function is to seek a function form for $f(x)$. The mentioned objective is achieved by training the SVM using a set of data (training dataset) (Misra et al. 2009). To calculate W and b, the error function ε (Eq. 3) in SVM must be minimized with respect to the conditions of Eqs. 3 and 4 (Hamel 2009).

$$\frac{1}{2} W^T \cdot W + C \sum_{i=1}^{N} \varepsilon_i + C \sum_{i=1}^{N} \varepsilon_i^* \tag{3}$$

$$W^T \cdot \emptyset(X_i) + b - y_i \leq \varepsilon + \varepsilon_i^*, \tag{4}$$

$$y_i - W^T \cdot \emptyset(X_i) - b \leq \varepsilon + \varepsilon_i, \varepsilon_i, \varepsilon_i^* \geq 0, \quad i = 1, 2, \ldots, N \tag{5}$$

where C is a positive integer that can be used to determine the penalty value in case of any model training error, $\emptyset$ is the kernel function, N is the number of samples, and ε_i and ε_i^* are the shortage variables. The regression function of SVM can be rewritten as follows:

$$f(x) = \sum_{i=1}^{N} \bar{\alpha}_i \emptyset(x_i)^T \cdot \emptyset(x) + b \tag{6}$$

where $\bar{\alpha}_i$ is the Lagrange coefficient. Calculating $\emptyset(x)$ in its particular space can be quite complex (Yoon et al. 2011). To solve the problem, in SVM, a kernel function should be selected as follows:

$$K(X_J, X) = \emptyset(X_i)^T \sqrt{b^2 - 4ac} \tag{7}$$

Different kernel functions can be used to fabricate different types of ε-SVM as follows: polynomial kernel, radial basis function (RBF), and linear kernel. Due to the widespread utility of these functions (Basak et al. 2007; Vapnik and Chervonenkis 1991), they were used in this study. Of note, calculations of vector machine were conducted based on coding in MATLAB, and the parameters were optimized through trial and error.

$$k(x, x_j) = (t + x_i \cdot x_j)^d \tag{8}$$

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{9}$$

$$k(x, x_j) = x_i \cdot x_j \tag{10}$$

Algorithm of artificial flora

Biological bases

Plants disperse their seeds through either autochory or allochory. Autochory-based plants disperse their seeds by themselves (self-dispersal), while allochory-based plants disperse their seeds by external forces or vector. Autochory provides a condition in which plants can independently migrate to a nearby hospitable environment, whereas allochory facilitates seed migration to far-off regions. Different methods of seed dispersal reduce the probability of extinction. Natural environment with difficult conditions and competition may reduce flora distribution. Following the migration of flora to a new environment, plant species can flourish and develop (Cheng et al. 2018). Therefore, flora migration may cause changes in and variations to the distribution area and contribute to

the development, extinction, and rebirth of flora. A plant is stationary and is not smart; however, flora seeks the ideal place for life through random seed dispersal during migration and reproduction. Plant survival depends on the surrounding environment. A plant develops and adapts to the environment under hardship. Before extinction of a specific plant in a region, flora may grow in a new environment. Seeds may grow in a new region and replicate through multi-generational replication. Flora finds an optimum region for growth, development, extinction, and rebirth (Cheng et al. 2018).

Artificial flora algorithm theory

Artificial flora algorithm comprises four main components: main plant, child plant, plant position, and distribution distance. Main plants are those plants that are ready to disperse their seeds. Child plants are the seeds of the main plants and unable to disperse grain before further growth. Plant position refers to the position or location in which it grows. Distribution distance signifies the seed dispersal distance. Artificial flora algorithm shows three main behavioral patterns: development behavior, distribution behavior, and selection behavior (Rosin and Belew 1995; Pagie and Mitchell 2002; Wiegand and Sarma 2004). Development behavior signifies the probability that plant development/maturity would facilitate its adaptation with environmental behavior (Hillis 1990; Cartlidge and Bulloc 2004; Williams and Mitchell 2005). Distribution behavior represents the movement of seeds/grains. Seeds can move through either allochory or autochory. Selection behavior indicates survival and extinction due to the environmental reasons. Figure 2 shows the flowchart of the AF algorithm application.

Wavelet transform

Wavelet transform is presented as an alternative to short-term Fourier transform, and the former is aimed at overcoming frequency resolution problems of the latter. To transform wavelet, similarly to the mechanism of short-term Fourier transformation, signal is divided into windows and each signal is transformed on each one of these windows (Vapnik 1998). The most significant difference between the two methods is that as the frequency resolution of every signal (window length) changes in parallel to frequency type, window width or frequency scale changes in proportion with the type of frequency. To put it differently, scale is used to transform wavelet instead of frequency. Therefore, wavelet transform is of timescale transform type. Accordingly, while signals can be dilated on a large scale with their details analyzed, they can be contracted on a small scale with their general features analyzed (Wang et al. 2000). A wavelet represents a small wave or a small window of the main signal whose energy is concentrated in time. Through wavelet transform, a mother signal can be decomposed into wavelets of different resolution

levels or scales. Therefore, wavelets are the translated and dilated samples of the mother wavelets, which are subject to oscillations in a finite length and critically damped. According to the mentioned property, the transient time series can be analyzed locally (Shin et al. 2005). Wavelet transform is defined in continuous and discrete forms.

Continuous wavelet transform (CWT)

CWT of the function $f(t)$ is defined as in Eqs. 11 and 12 (Vapnik 1998):

$$CWT_f^{\psi}(s, \tau) = \Psi_f^{\psi}(s, \tau) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{+\infty} f(t)\psi^* \left(\frac{t-\tau}{s}\right) dt = \langle f(t), \psi_{s,\tau}(t)\rangle$$

(11)

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-\tau}{s}\right)$$

(12)

Equation 12 is a relation to two variables of s and τ, where S and τ denote the scaling and translation parameters, respectively; * denotes complex conjugate; ψ is the window function or the mother wavelet; and $\frac{1}{\sqrt{|s|}}\psi\left(\frac{t-\tau}{s}\right)$ represents the wavelets obtained from the translation and variation of the mother wavelet scale (Wang et al. 2000). All translated and dilated versions of wavelets (daughter wavelets) are obtained from the mother wavelet, hence the term 'Mother' (Safavi and Romagnoli 1997). In other words, the Mother wavelet represents a pattern for other windows. Herein, $\langle\ldots\rangle$ denotes the vector product of two functions in the signal space.

Bayesian network

Bayesian networks are graphical models that use Bayesian inference in case of complexity and uncertainty. It is a graph that shows random variables and their dependence (Kevin and Nicholson 2010). In this graph, nodes denote random variables of continuous and discrete types and arrows that connect each pair of nodes show a correlation between variables. In fact, the implied network is a directed graph without any cycle (Heckerman 1997). A simple graph is shown in Fig. 3.

Concerning the structure of Bayesian network, if a dataset comprises n variables $\{x_1, x_2, x_3,\ldots, x_n\}$ where random variables and parent nodes are, respectively, denoted by x_i and (x_i), then the probability of every occurrence like $X = (x_1, x_2,\ldots, x_n)$ is calculated as follows:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \text{parents}(X_i))$$

(13)

Fig. 2 Flowchart of AF algorithm
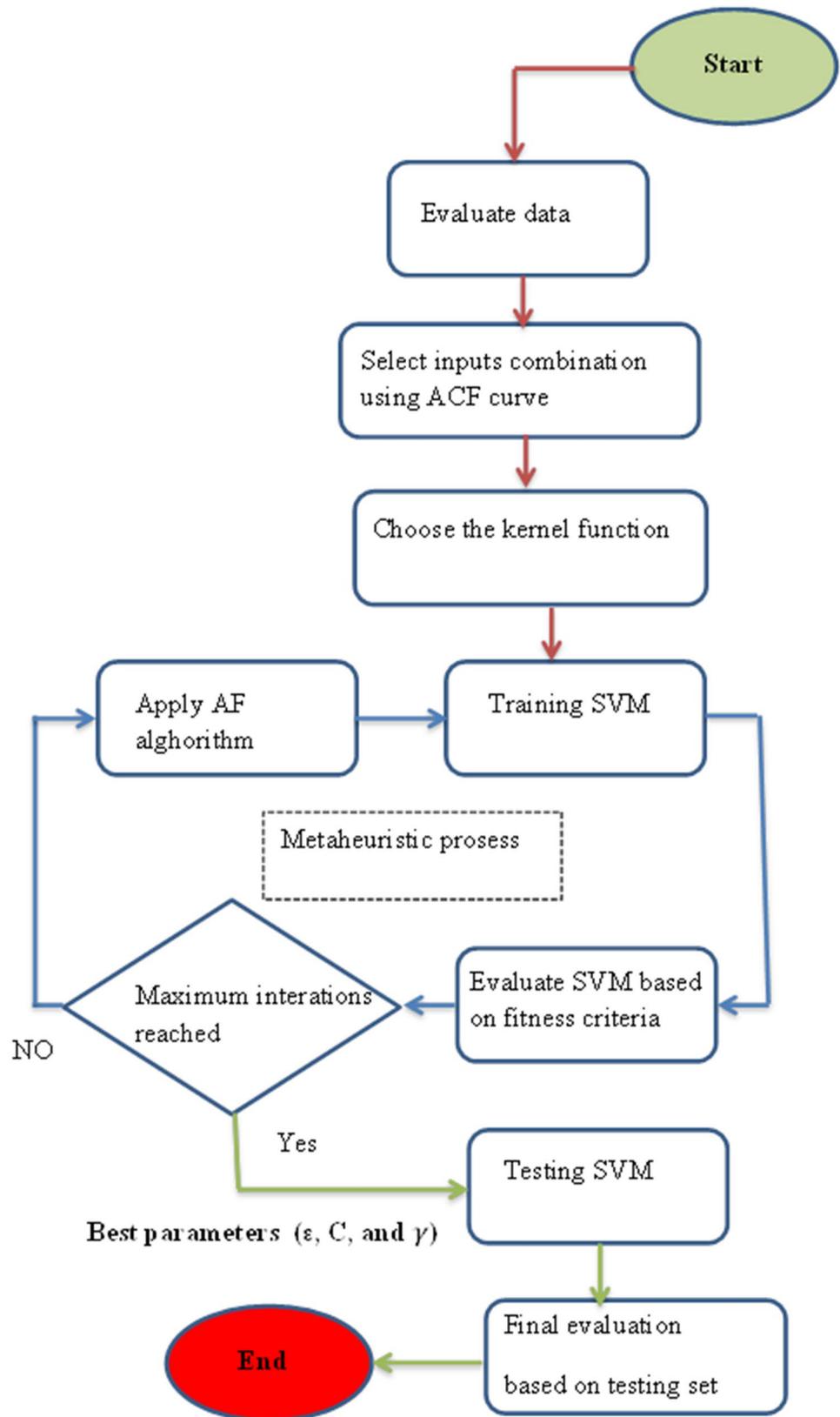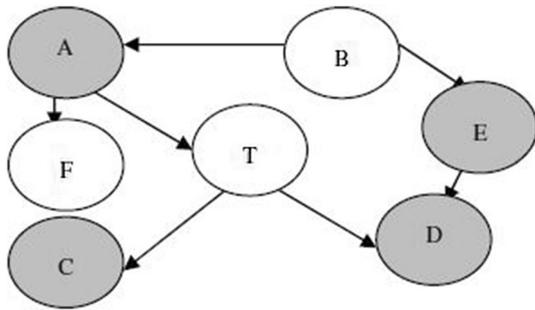
Fig. 3 A simple Bayesian network

Following the network fabrication, the values of the relations among the interconnected nodes are determined and the conditional probability value of each variable is obtained. The results are shown in a table called the conditional probability table. Each node has a conditional probability table. For each combination of parent node values, the conditional probability of each value is maintained in the table. Then, the next probability distribution of every node is calculated and, therefore, it is called probability belief (Nguyen and Widrow 1990).

Evaluation criteria

This study employed various criteria to assess the accuracy and performance of the proposed model. The criteria included the coefficient of determination (R^2), root-mean-square error (RMSE), mean absolute error (MAE), and Nash–Sutcliffe efficiency (NSE) (Zhu et al. 2007). The best values of these criteria are 1, 0, 0, and 1, respectively. In addition, for (R^2), the value (1) shows higher prediction accuracy and (0) indicates the failure of the proposed model to determine the mean variability (Nagelkerke 1991). NSE is a normalized statistic that helps determine residual variance (Nash and Sutcliffe 1970; Moriasi et al. 2007), and different ranges of NSE are between $-\infty < \text{NSE} < 1$ with an optimum value of 1. In other words, a response is optimum if the value of NSE is equal to 1. Moreover, NSE values ranging from 0 to 1 are acceptable in general, showing that the observed data are more viable than the estimated data (Duie Tien et al. 2020). NSE was also suggested by ASCE (1993), and it is highly popular and presents a broad range of information on the reported values. Moreover, its applicability to different scientific fields has surged in recent years (Sevat and Dezetter 1991; Kesgin et al. 2020).

Formulations required for calculating the considered indicators are given in Eqs. (14)–(18). In this study, in addition to the above criteria, boxplot and Taylor diagram were used. Taylor diagram and boxplots are common graphical procedures used in model performance comparison. Taylor

diagram is an appropriate tool for assessing different methods and has recently been applied to the fields of weather forecast, water science, and so on. It is represented in the form of a semicircle that demonstrates positive and negative correlations. (A quadrant shows only positive correlation.) The values of correlation coefficient are given in the form of radii, and standard deviations are concentric circles (Taylor Taylor 2001a, b; Wehner 2013). Boxplot was introduced by John Tukey (1969). It is a common figure that shows statistical values for data. In other words, it compares data in the observation form (Lo Conti et al. 2014).

$$R^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad -1 \le R \le 1 \quad (14)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \quad 0 \le \text{RMSE} \le 1 \quad (15)$$

$$\text{MAE} = \frac{1}{n}\sum |x_i - y_i| \quad 0 \le \text{MAE} \le 1 \quad (16)$$

$$\text{NS} = 1 - \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(x_i - \bar{y})^2} \quad -\infty \le \text{NS} \le 1 \quad (17)$$

where x_i and y_i show the observed and calculated values in the phase i, respectively, N is the time step, and $\bar{x}$ and $\bar{y}$ show the means of the observed and calculated values, respectively. Scatter figures and time series of observed and calculated values were also used.

Results and discussion

The values of extracurricular algorithms may accidently exceed their pre-defined domain due to the addition of velocities with random values to the problem variables. In other algorithms, given that problem variables take discrete values, optimal solutions are added to the problem domain at all iterations. Therefore, obtaining a global optimal solution to a problem becomes quite time-consuming and, accordingly, the problem should be dealt with in local optimum. Therefore, the artificial plant algorithm, which is a combination of continuous and discrete optimization with respect to large-scale problems, shortens the computational time required for achieving global optimum search and moves away from local optimum, hence making the algorithm optimally capable to solve nonlinear problems with large dimensions and high velocity. Therefore, this study applied a compound approach composed of artificial

plant algorithm and support vector machine. The artificial plant algorithm is inspired by the migration and reproduction behaviors of plants that are of three main behaviors: evolution, propagation, and selection. In the case of evolutionary behavior, the propagation distance of the child plant evolves based on the parent plant distance. Given that the propagation distance of the child plant is not dependent on the parent plant, the algorithm may exit the local optimal solution. The propagation behavior includes allochory and autochory; the former facilitates seed dispersal to far-off spaces, while the latter provides the chance to explore areas close to the original plant. These behaviors provide the algorithm with a better local search ability and enhances the convergence speed.

One of the most important stages of modeling is to select a proper combination of input variables. Thus, the cross-correlation between input and output variables was measured and input parameters were selected in order to achieve an optimum model for predicting river flow of Dez catchment, the results of which are given in Table 2. The values of the mentioned parameters, as shown in Table 2, were considered to be 0.750, according to Pearson correlation coefficient. Further to these data, $Q(t-1)$, $Q(t-2)$, $Q(t-3)$, and $Q(t-4)$ columns show

river flows at times $t-1$, $t-2$, $t-3$, and $t-4$, respectively. These columns represent input values, and $Q(t)$ (river flow at time t) is the model outcome. The nature of the dominant mechanism of a phenomenon not only increases pattern complexity and RAM (computer memory) involvement, but also attenuates the model accuracy. To model the river flow, attempt must be made to select the most efficient observation data as the training data. This study applied return flow to investigate the effects of streamflow. Since the cross-correlation between input and output data was found higher than 0.750 (as shown in Table 3), different combinations of input parameters were used for estimating the performance of an optimum model in predicting the water flow of Dez catchment. To this end, this study employed necessary data obtained from hydrometric stations of Ab Garmeh, Tireh marvak, Geleh-rood, Silakhor, Tireh, Cham Chit, Cham Chit, Ab Sardeh, Taleh Zang, and Tang Pangeh, which provided us with over 3650 records in the years 2008–2018. The total of 2920 and 730 records were utilized for training and validation purposes, respectively. Of note, this study adopted 80% and 20% of the available data randomly for training and testing, respectively (Nagy et al. 2002a, b; Kisi et al. 2006a, b). Table 2 shows the cross-correlation of input and output variables.

Results of SVM-AF algorithm

The proposed hybrid method consisted of the SVM and AF algorithm. Obtained data were simulated to reduce the

Table 2 Cross-correlation of input and output variables

Station	Output	$Q(t-1)$	$Q(t-2)$	$Q(t-3)$	$Q(t-4)$
Ab Garmeh	$Q(t)$	0.930	0.890	0.840	0.780
Tireh marvak		0.941	0.895	0.862	0.810
Geleh-rood		0.963	0.923	0.884	0.825
Silakhor		0.952	0.901	0.858	0.812
Tire		0.924	0.881	0.825	0.772
Cham Chit		0.893	0.833	0.804	0.750
Ab Sardeh		0.926	0.884	0.832	0.793
Sezar		0.954	0.798	0.862	0.821
Taleh Zang		0.943	0.895	0.868	0.815
Tang Pangeh		0.933	0.887	0.845	0.792

Table 3 Selected combinations of input parameters

Number	Input structure	Output
1	$Q(t-1)$	$Q(t)$
2	$Q(t-1)$, $Q(t-2)$	$Q(t)$
3	$Q(t-1)$, $Q(t-2)$, $Q(t-3)$	$Q(t)$
4	$Q(t-1)$, $Q(t-2)$, $Q(t-3)$, $Q(t-4)$	$Q(t)$

Table 4 Analysis of SVM-AF model applied to the selected stations

Station	Training				Testing			
	R	RMSE (m³/s)	MAE (m³/s)	NS	R	RMSE (m³/s)	MAE (m³/s)	NS
Ab Garmeh	0.915	0.060	0.025	0.936	0.942	0.037	0.018	0.958
Tiremarvak	0.930	0.071	0.028	0.947	0.968	0.033	0.021	0.987
Geleh-rood	0.958	0.036	0.015	0.942	0.970	0.018	0.012	0.982
Silakhor	0.935	0.065	0.026	0.962	0.970	0.031	0.015	0.988
Tireh	0.912	0.065	0.030	0.928	0.933	0.042	0.020	0.951
Cham Chit	0.850	0.118	0.076	0.874	0.898	0.088	0.040	0.917
Ab Sardeh	0.918	0.052	0.017	0.942	0.938	0.035	0.015	0.936
Sezar	0.936	0.038	0.028	0.966	0.96	0.021	0.011	0.981
Taleh Zang	0.946	0.029	0.020	0.971	0.972	0.015	0.008	0.989
Tang Panj	0.953	0.018	0.009	0.988	0.985	0.008	0.004	0.995

variations in the data of daily water flow of rivers in the Dez catchment. Optimum values of the SVM model properties including ε and C were determined. Different kernels were investigated, and the RBF kernel function was selected due to its high performance in estimating daily flows of rivers (Ghorbani et al. 2016; Ghorbani et al. 2018b). The value of

Fig. 4 Scatter figure of the observed and calculated values of the hybrid AF-SVM model performance at the accuracy verification phase
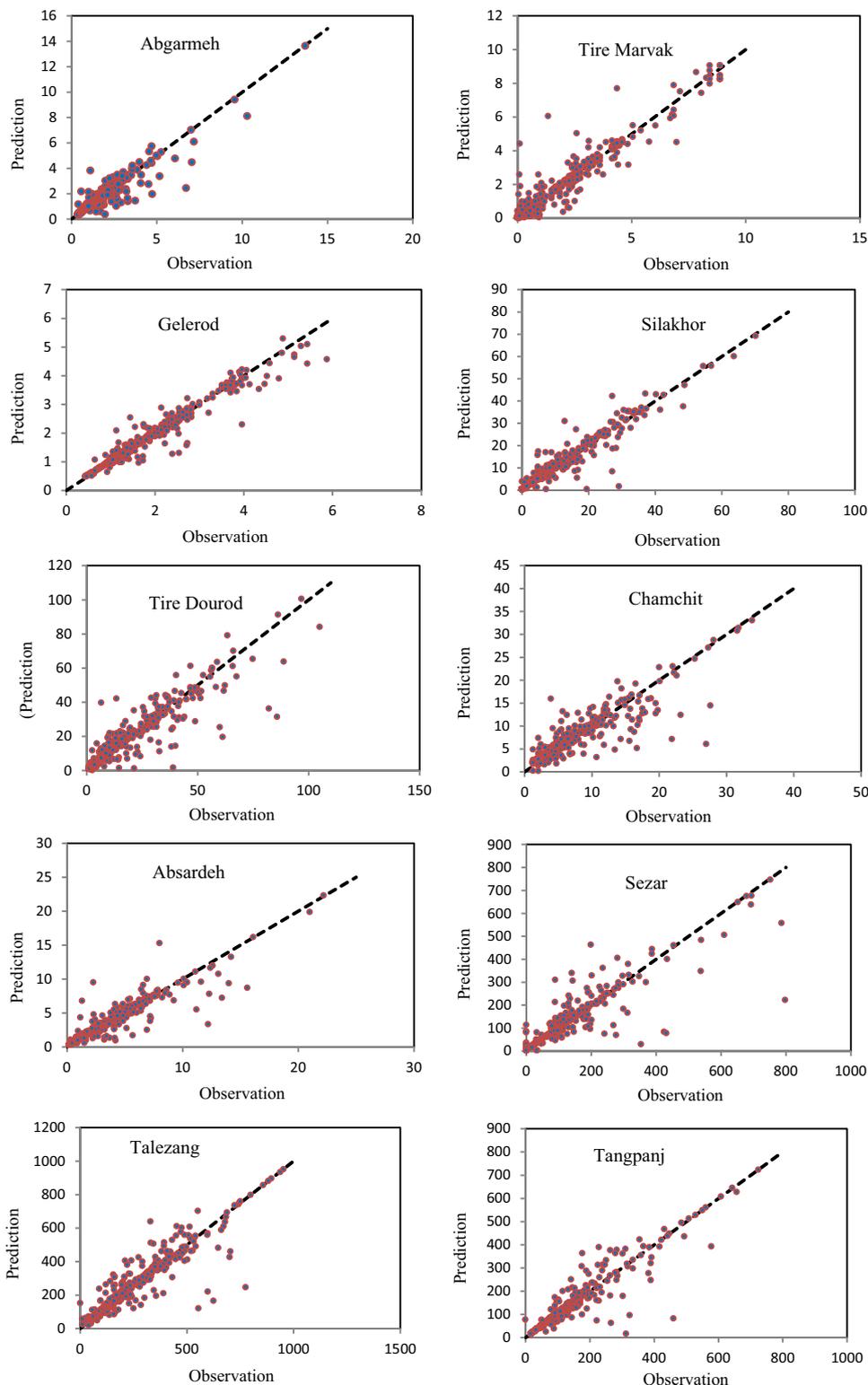
the parameter γ should be determined. Therefore, in general, to predict the daily water flows of the rivers by using the SVM, the optimum values of ε, C, and γ properties should be determined and fitted by using AF algorithm. Afterward, it would be quite easy to choose the best model with minimum error and, then, determine the optimum values of ε, C, and γ. The SVM performance results are given in Table 4. Accordingly, the proposed hybrid model applied to the final catchment stations showed minimum error and high accuracy by using base discharge of the river flow. In Fig. 4, the scatter diagram of the proposed hybrid model at the verification stage shows the fitting of the calculated values to the best fitting line $y = x$. According to the figure, the estimated and observed values of the model performance, except for some points, are positioned on the bisector and, thus, are equal based on $y = x$. In addition, the hybrid model outperforms other models in predicting minimum and maximum values such that the obtained models approximate real data.

The wavelet support vector machine results

To investigate the results of the hybrid model, wavelet transform was employed to decompose the input parameters into sub-signals, which were then added to the SVM model as input. In doing so, the hybrid model called 'wavelet support vector machine' was achieved. Various wavelet functions were compared, and it was found that the Mexican hat wavelet outperformed other functions. RBF, polynomial, and linear kernels were used to predict river flow, among which RBF showed the highest accuracy. Table 5 presents the results of the hybrid model applied to the selected stations. Accordingly, the hybrid model enjoyed higher accuracy and less error in case of Geleh-rood, Silakhor, Sezar,

and Taleh Zang stations. In Fig. 5, the scatter diagram of the hybrid wavelet-SVM at the accuracy verification phase shows the fitting of the calculated values to the best fitting line $y = x$. According to the figure, the estimated and observed values of the model performance, except for some points, are positioned on the bisector and, thus, are equal based on $y = x$. As it is clear, the hybrid model exhibited a desirable performance in predicting the lowest and highest values of the model performance, which approximated real-world data.

The results of Bayesian support vector machine

To compare the results of the hybrid wavelet-SVM model, Bayesian-SVM model was used. To predict the daily flow of rivers by using SVM, the optimum values of the aforementioned model properties (ε, C, and γ) should be calculated. To this end, two parameters of ε and C as well as parameter γ were calculated based on Bayesian probabilities and trial and error, respectively. The results of evaluating the performance of the given models are shown in Table 6. Accordingly, for Ab Garmeh station, the Bayesian-SVM model outperformed others in terms of higher accuracy and less error. Figure 6 shows the best model at the performance verification stage. The results indicate that most of the estimated and observed values, except for only one point, are positioned on the bisector due to the equality between the observed and calculated values of the model performance.

Table 5 Analysis of wavelet-SVM applied to the selected stations

Station	Training				Testing			
	R	RMSE (m³/s)	MAE (m³/s)	NS	R	RMSE (m³/s)	MAE (m³/s)	NS
Ab Garmeh	0.910	0.064	0.028	0.934	0.933	0.042	0.020	0.951
Tiremarvak	0.928	0.074	0.032	0.944	0.965	0.036	0.025	0.984
Geleh-rood	0.944	0.042	0.022	0.968	0.952	0.032	0.021	0.968
Silakhor	0.931	0.069	0.029	0.958	0.967	0.034	0.017	0.985
Tireh	0.908	0.067	0.032	0.925	0.93	0.045	0.022	0.948
Cham Chit	0.846	0.126	0.084	0.868	0.89	0.094	0.045	0.911
Ab Sardeh	0.915	0.058	0.021	0.937	0.935	0.038	0.018	0.932
Sezar	0.932	0.042	0.031	0.960	0.948	0.025	0.017	0.965
Taleh Zang	0.924	0.048	0.033	0.941	0.957	0.028	0.017	0.973
Tang Panj	0.917	0.055	0.018	0.938	0.938	0.035	0.016	0.954

Fig. 5 Scatter figure of the observed and calculated values of the hybrid wavelet-SVM model
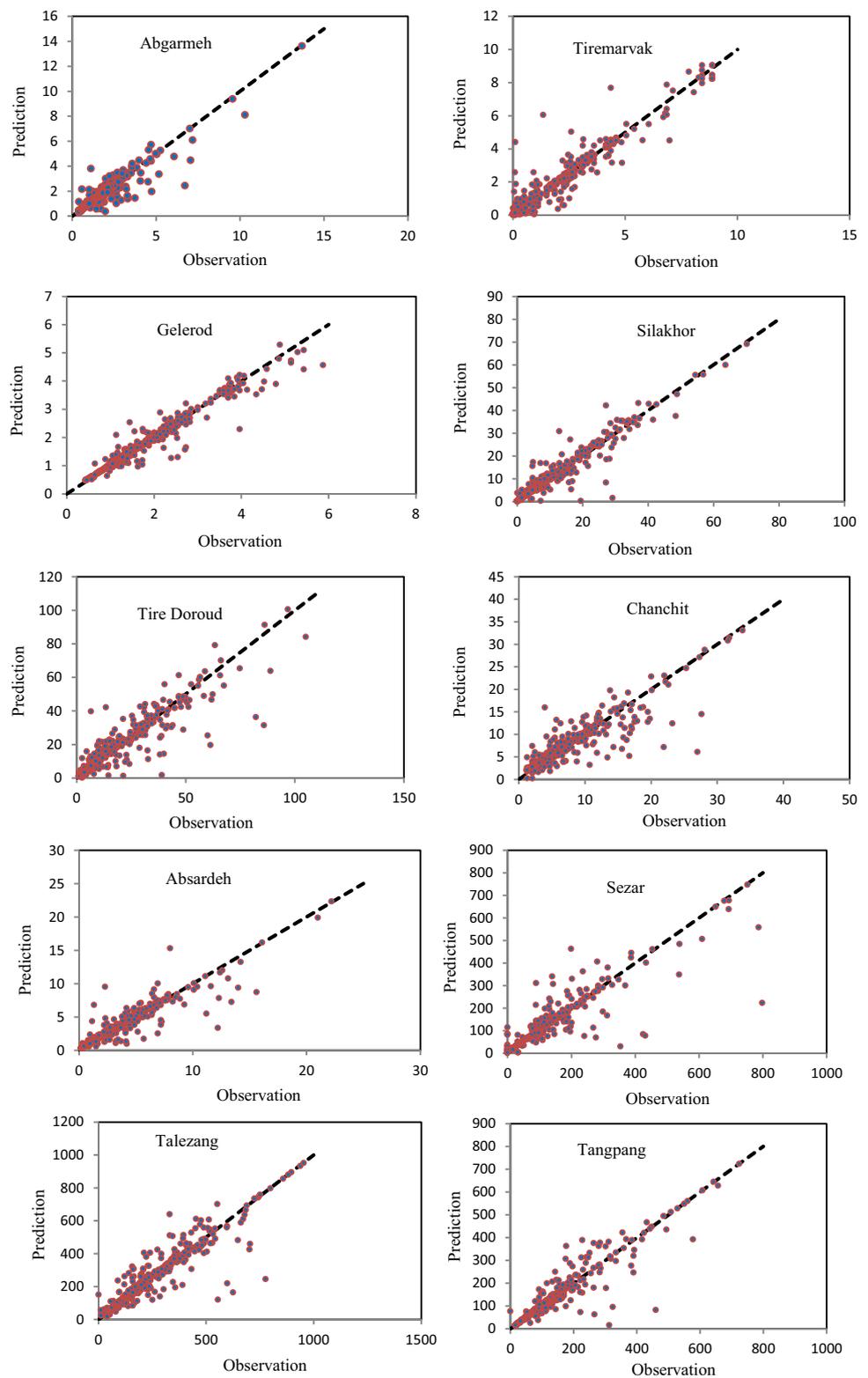
Table 6 Analysis of Bayesian-SVM applied to the selected stations

Station	Training				Testing			
	R	RMSE (m³/s)	MAE (m³/s)	NS	R	RMSE (m³/s)	MAE (m³/s)	NS
Ab Garmeh	0.892	0.112	0.069	0.912	0.918	0.082	0.043	0.936
Tiremarvak	0.925	0.074	0.031	0.948	0.952	0.052	0.027	0.967
Geleh-rood	0.942	0.065	0.025	0.961	0.948	0.035	0.010	0.957
Silakhor	0.932	0.065	0.028	0.952	0.951	0.044	0.018	0.958
Tireh	0.901	0.102	0.065	0.921	0.92	0.081	0.042	0.938
Cham Chit	0.824	0.125	0.081	0.849	0.855	0.098	0.058	0.891
Ab Sardeh	0.894	0.105	0.081	0.914	0.92	0.082	0.043	0.926
Sezar	0.922	0.081	0.032	0.031	0.938	0.057	0.041	0.944
Taleh Zang	0.918	0.088	0.041	0.931	0.932	0.065	0.041	0.048
Tang Panj	0.892	0.104	0.062	0.915	0.902	0.073	0.054	0.921

Fig. 6 Scatter figure of the calculated and observed values achieved by the Bayesian-SVM model
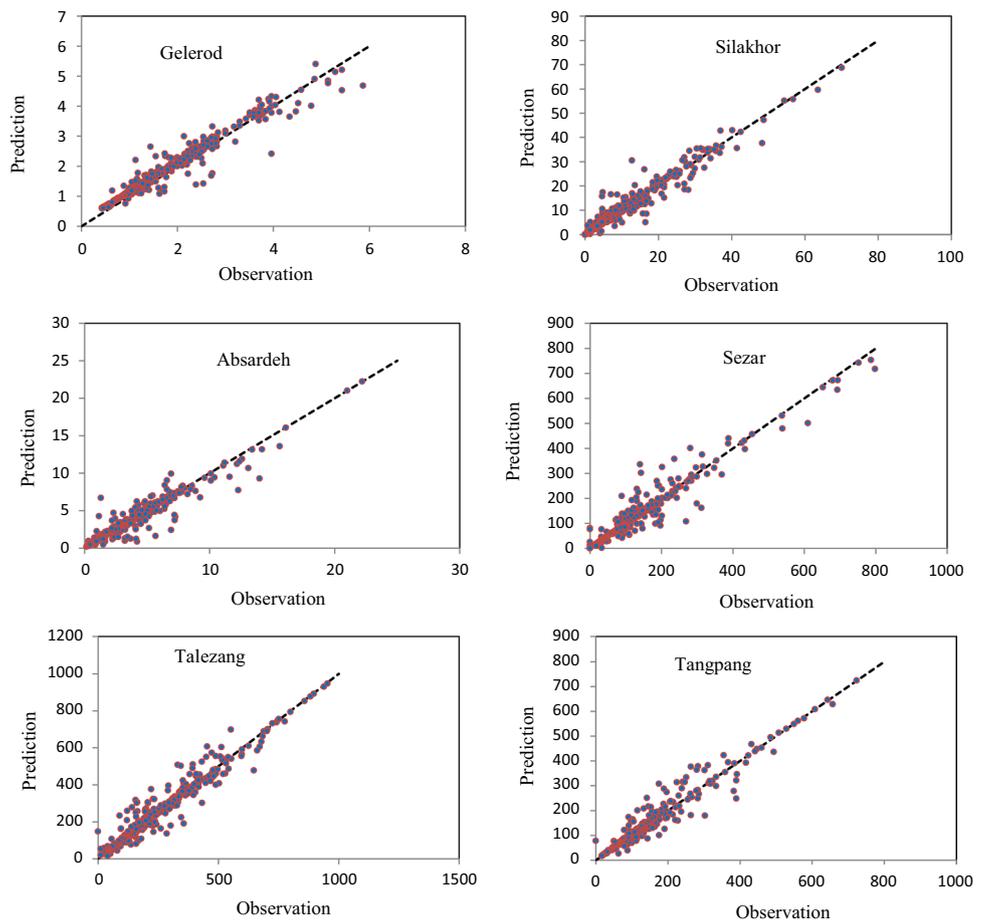
Comparison of the performances of the models

After selecting the optimal response of each hybrid AI model and comparing their responses and performances, it was concluded that all of the three models could simulate the discharge of Dez river with satisfactory accuracy. Figure 7 shows the values of CC (correlation coefficient) at the performance verification stage concerning the selected stations. As shown in the figure, at the final stations of the Dez catchment area, due to the role of the base flow of
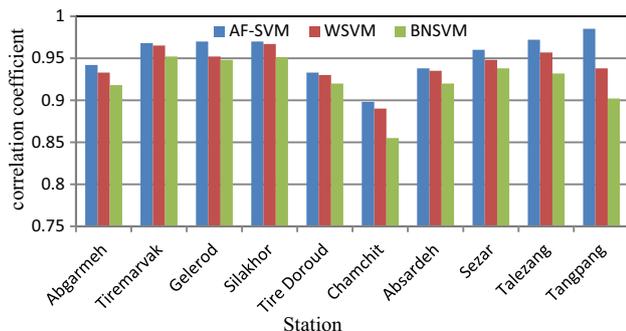
Fig. 7 Correlation coefficient of models at the accuracy verification stage

Fig. 8 Diagram of RMSE of the models studied at the verification stage

Dez river, the hybrid SVM-AF model showed viable performance and high accuracy. Figure 8 also shows RMSE values at the model verification stage. As it is clear, the SVM-AF model enjoyed less error than other models for Tang Panj and Tele Zang stations. It can be concluded that by separating the high- and low-frequency components of the signal, wavelet transform will have multi-scale properties of the signal and, thus, increase the accuracy of the model to a greater degree.

Taylor diagrams were utilized to analyze and assess the models applied to the study, as shown in Fig. 9. These diagrams enjoy the advantage of using correlation coefficient and standard deviation as two common correlation statistics (Taylor 2001a, b). When the predicted value is closer to the observed value in the matter of correlation coefficient and standard deviation, one can expect more accurate prediction results. Taylor performance diagram depicts that the AF-SVM model has the highest efficiency and performance, because the predicted standard deviation approximates the observed standard deviation of the

data and the correlation coefficient shows the highest value. According to all of the evaluation criteria, AF-SVM and BNSVM had the highest predictive performance and the lowest prediction error.

Conclusion

This study aimed to simulate river flow in Dez catchment based on the daily flow data. The applied models included hybrid SVM-AF, wavelet-SVM, and Bayesian-SVM. Predicted and observed values of the model performances were assessed based on the given evaluation criteria. In sum, the applied models showed higher efficiency than other models. They exhibited higher accuracy in predicting Dez river flow. The proposed hybrid model of SVM-AF showed better accuracy and less error than wavelet-SVM and Bayesian-SVM. The accuracy of the hybrid model was found optimum due to the optimization of SVM. Taylor
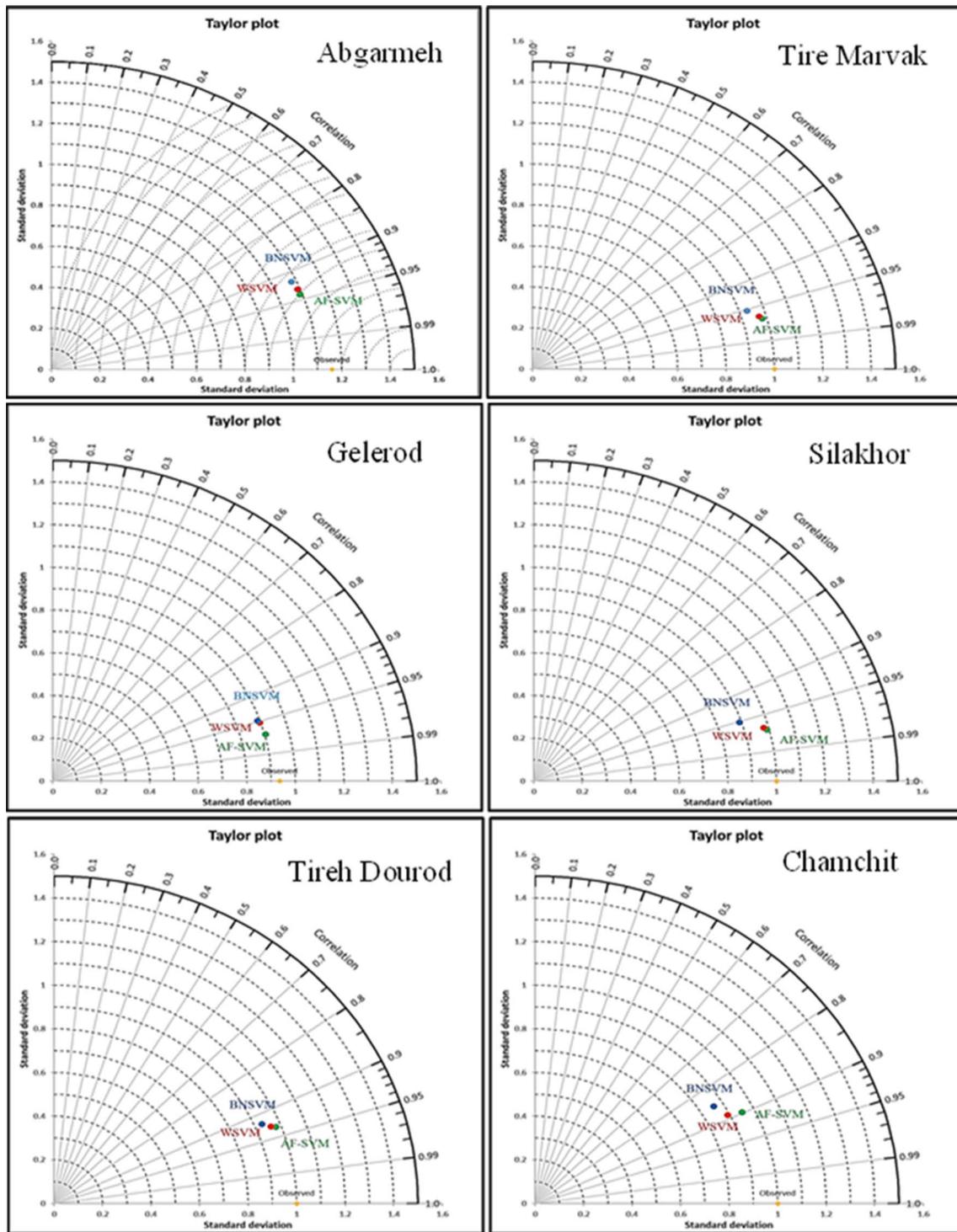
Fig. 9 Taylor diagram of the stations studied

performance diagram showed that the AF-SVM model achieved the highest efficiency and performance, because the predicted standard deviation approximated the observed

standard deviation and the correlation coefficient had the highest value. According to all of the evaluation criteria, AF-SVM and 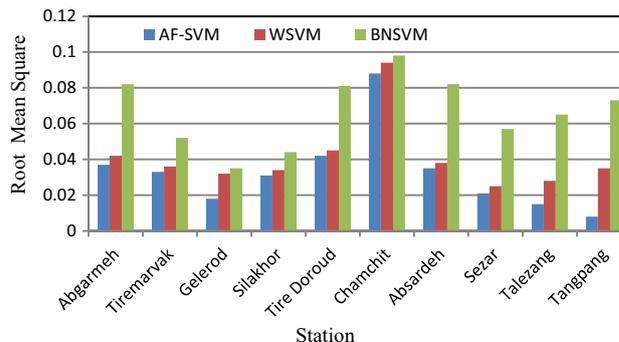BNSVM had the highest predictive performance and the lowest prediction error. In sum, the results pointed out that the proposed hybrid model could properly
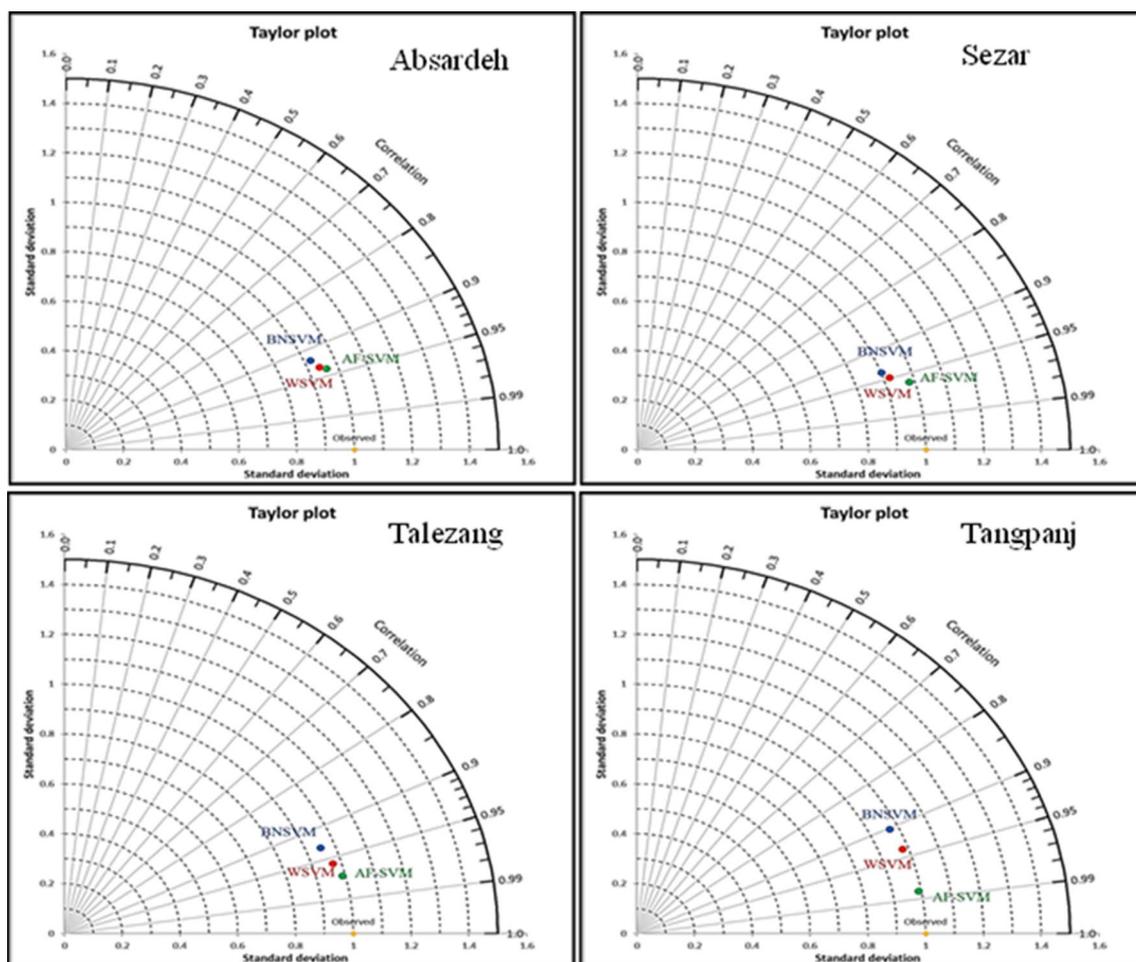
Fig. 9 (continued)

predict the daily flow of rivers. This model can be applied to the face-lifting and developing strategies that facilitate surface water management and decisions on enhancing the water flow quantity of rivers.

References

Adnan R, Liang Z, Heddam S, Kermani M, Kisi O, Li B (2019) Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. J Hydrol 19(4):432–448

Aggarwal SK, Goel A, Singh VP (2012) Stage and discharge forecasting by SVM and ANN techniques. Water Resour Manag 26:3705–3724

Alizadeh F, Gharamaleki A, Jalilzadeh M, Akhoundzadeh A (2020) Prediction of river stage-streamflow process based on a conceptual model using EEMD-WT-LSSVM Approach. Water Resour 47:41–53

ASCE (1993) Criteria for evaluation of watershed models. J Irrig Drain Eng 119(3):429–442

Asefa T, Kemblowski MW, Urroz G, McKee M, Khalil A (2004) Support vector based groundwater head observation networks design. Water Resour Res. https://doi.org/10.1029/2004WR003304

Basak D, Pal S, Patranabis DC (2007) Support vector regression. Neural Inf Process 11:203–225

Beven KJ, Kirkby MJ (1979) A physically based variable contributing area model of basin hydrology Hydrological. Sci Bull 43:43–69

Bhagwat PP, Maity R (2012) Multistep-ahead River Flow Prediction using LS-SVR at Daily Scale J. Water Resour 4:528–539

Boser JA (1992) Microcomputer needs assessment of american evaluation association members. Am J Eval 13:92–93

Cameron M, Donland H, Slobodan P (1999) Short term streamflow forecasting using artificial neural networks. J Hydrol 214(4):32–48

Cartlidge JP, Bulloc SG (2004) Combating coevolutionary disengagement by reducing parasite virulence. Evol Comput 12(2):193–222

Cheng L, Wu X, Wang Y (2018) Artificial flora (AF) optimization algorithm. Appl Sci 329(8):2–22

Coken U, Forman A, Fensoy A (2016) Streamflow forecasting using different neural network models with satellite data for a snow dominated region in Turkey. Procedia Eng 154:1185–1192

Dibike YB, Velickov S, Solomatine D, Abbott MB (2001) Model induction with support vector machines: introduction and applications. J Comput Civ Eng 15:208–216

Dolling O, Eduardo A (2002) Artificial Neural Networks for Streamflow Prediction. J Hydraul Res 40(5):547–554

Duan Q, Sorooshian S, Gupta VK (1992) Effective and efficient global optimization for conceptual rainfall–runoff models. Water Resour Res 28(4):1015–1031

Duan Q, Sorooshian S, Gupta VK (1994) Optimal use of SCEUA global optimization method for calibrating watershed models. J Hydrol 158:265–284

Duie Tien B, Khosravi K, Tiefenbacher J, Nguyen H, Kazakis N (2020) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. J Sci Total Environ 721:136612

Ghorbani MA, Khatibi R, Geol A, Fazelifard MH, Azani A (2016) Modeling river streamflow time series using support vector machine and artificial neural networks. Environ Earth Sci 75(4):675–685

Ghorbani MA, Deo R, Karimi V, Yassen ZM, Terzi O (2018a) Implementation of a hybrid MLP-FFA model for water level prediction of Lake Egirdir Turkey. Stoch Environ Res Risk Assess 32(6):1683–1697

Ghorbani MA, Khatibi R, Karimi V, Yaseen ZM, Zounemat-Kermani M (2018b) Learning from multiple models using artificial intelligence to improve model prediction accuracies: application to river flows. Water Resour Manag 32(13):4201–4215

Gizaw MS, Gan TY (2016) Regional flood frequency analysis using support vector regression under historical and future climate. J Hydrol 538:387–398

Guven A, Kisi O (2011a) Daily pan evaporation modeling using linear genetic programming technique. Irrig Sci 136(10):715–723

Guven A, Kisi O (2011b) Daily pan evaporation modeling using linear genetic programming technique. Irrig Sci 29(2):135–145

Hamel L (2009) Knowledge discovery with support vector machines. Wiley, Hoboken

Haykin A (1998) Neural networks: a comprehensive foundation. Prentice Hall, Upper Saddle River

Heckerman D (1997) Bayesian networks for data mining, data mining and knowledge. Discovery 1(1):79–119

Hillis WD (1990) Co-evolving parasites improve simulated evolution as an optimization procedure. Phys D Nonlinear Phenom 42:228–234

Hsu K, Gupta HV, Sorooshian S (1995) Artificial neural network modeling of the rainfall–runoff process. Water Resour 31(10):2517–2530

Huang S, Chang J, Huang Q, Chen Y (2014) Monthly streamflow prediction using modified emd-based support vector machine. J Hydrol 511(4):764–775

Jakeman AJ, Hornberger GM (1993) How much complexity is warranted in a rainfall-runoff model. Water Resour Res 29:2637–2649

Kalteh AM (2013) Monthly river flow forecasting using artificial neural network and support vector regression models coupled with wavelet transform. Comput Geosci 54:1–8

Kesgin E, Agaccioglu H, Dogan A (2020) Experimental and numerical investigation of drainage mechanisms at sports fields under simulated rainfall. J Hydrol 580:124251

Kevin B, Nicholson E (2010) Bayesian artificial intelligence. Comput Sci Data Anal 14(8):1328–1344

Kisi O, Cobaner M (2009) Modeling river stage–discharge relationships using different neural network computing techniques. CLEAN Soil Air Water 37:160–169

Kisi O, Karahan M, Sen Z (2006a) River suspended sediment modeling using fuzzy logic approach. Hydrol Process 20(2):4351–4362

Kisi O, Karahan M, Sen Z (2006b) River suspended sediment modeling using fuzzy logic approach. Hydrol Process 20:4351–4362

Kumar RA, Jothiprakash V (2013) Reservoir evaporation prediction using data-driven techniques. J Hydrol Eng 18(1):40–49

Lima AR, Cannon AJ, Hsieh WW (2016) Forecasting daily streamflow using online sequential extreme learning machines. J Hydrol 537:431–443

Lo Conti F, Hsu KL, Noto LV, Sorooshian S (2014) Evaluation and comparison of satellite precipitation estimates with reference to a local area in the Mediterranean Sea. Atmos Res 138:189–204

Macfarlane C, Ogden GN (2012) An improved evaporation dome for forest environments Comput. Electron Agric 89:126–129

Misra D, Oommen T, Agarwa A, Mishra SK, Thompson AM (2009) Application and analysis of support vector machine based simulation for runoff and sediment yield. Biosyst Eng 103(3):527–535

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Am Soc Agric Biol Eng 50(3):885–900

Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. Biometrika. https://doi.org/10.1093/biomet/78.3.691

Nagy H, Watanabe K, Hirano M (2002a) Prediction of sediment load concentration in rivers using artificial neural network model. J Hydraul Eng 128(3):558–559

Nagy H, Watanabe K, Hirano M (2002b) Prediction of sediment load concentration in rivers using artificial neural network model. J Hydraul Eng 128:558–559

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models: part 1. A discussion of principles. J Hydrol 10(3):282–290

Nguyen D, Widrow B (1990) Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. Proc IJCNN 3:21–26

Pagie L, Mitchell MA (2002) Comparison of evolutionary and coevolutionary search. Int J Comput Intell Appl 2:53–69

Pulido-Calvo I, Portela M (2007) Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds. J Hydrol 332(4):1–15

Robert JA, Linda S (2000) Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. Hydrol Process 14:2157–2172

Rosin CD, Belew RK (1995) Methods for competitive co-evolution: finding opponents worth beating. In: Proceedings of the international conference on genetic algorithms, Pittsburgh, pp 373–381

Safavi AA, Romagnoli JA (1997) Application of wavelet-based neural networks to modelling and optimisation of an experimental distillation column. IFAC J Eng Appl Artif Intell 10(3):301–313

Salas G, Markuus M, Tokar AS (2000) Streamflow forecasting based on artificial neural networks. Artif Neural Netw Hydrol 154(2):918–925

Sedighi F, Vafakhah M, Javadi MR (2016) Rainfall-runoff modeling using support vector machine in snow-affected watershed. Arab J Sci Eng 41(10):4065–4076

Sevat E, Dezetter A (1991) Selection of calibration objective functions in the context of rainfall-runoff modeling in a Sudanese savannah area. Hydrol Sci J 36(4):307–330

Shin S, Kyung D, Lee S, Taik K, Kim J, Hyun J (2005) An application of support vector machines in bankruptcy prediction model. Expert Syst Appl 28(4):127–135

Sivapragasam C, Muttil N (2005) Discharge rating curve extension—a new approach. Water Resour Manag 19:505–520

Solomatine DP (2005) Data-driven modeling and computational intelligence methods in hydrology. In: Encyclopedia of hydrological sciences. Wiley

Srivastava P, McVair JN, Johnson TE (2006) Comparison of process-based and artificial neural network approaches for streamflow modeling in an agricultural watershed. J Am Water Resour Assoc 2(3):545–563

Suliman A, Nazri N, Othman M, Abdul M, Ku-mahamud KR (2013) Artificial neural network and support vector machine in flood forecasting: a review. J Hydroinform 15:327–332

Taylor E (2001a) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106(7):7183–7192

Taylor KE (2001b) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 106:7183–7192

Tokar AS, Markus M (2000) Precipitation-runoff modeling using artificial neural networks and conceptual models. J Hydrol Eng 5:156–161

Tukey JW (1969) Analyzing data: sanctification or detective work. Am Psychol 24

Vapnik VN (1998) Statistical learning theory. Wiley, New York

Vapnik V, Chervonenkis A (1991) The necessary and sufficient conditions for consistency in the empirical risk minimization method. Pattern Recognit Image Anal 1(3):283–305

Wang D, Safavi AA, Romagnoli JA (2000) Wavelet-based adaptive robust M-estimator for non-linear system identification. AIChE J 46(4):1607–1615

Wang W, Men C, Lu W (2008) Online prediction model based on support vector machine. Neurocomputing 71:550–558

Wehner MF (2013) Very extreme seasonal precipitation in the NARC-CAP ensemble: model performance and projections. Clim Dyn 40(1–2):59–80

Wiegand RP, Sarma J (2004) Spatial embedding and loss of gradient in cooperative coevolutionary algorithms. In: Proceedings of the international conference on parallel problem solving from nature, Berlin, Germany, vol 43, pp 912–921

Williams N, Mitchell M (2005) Investigating the success of spatial coevolution. In: Proceedings of the 7th annual conference on genetic and evolutionary computation Washington, vol 46, pp 523–530

Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK (2011) A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J Hydrol 396(4):128–138

Yu PS, Chen ST, Chang IF (2005) Flood stage forecasting using support vector machines. Geophys Res 7:41–76

Yu PS, Chen ST, Chang IF (2006) Support vector regression for real-time flood stage forecasting. J Hydrol 328:704–716

Zhu YM, Lu XX, Zhou Y (2007) Suspended sediment flux modeling with artificial neural network: an example of the longchuan-jiang river in the upper yangtze catchment. Geomorphology 84(4):111–125

Zhu S, Zhou J, Ye L, Meng C (2016) Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River. China. Environ Earth Sci 75(6):531–544

Sea level prediction using climatic variables: a comparative study of SVM and hybrid wavelet SVM approaches

S. Sithara[1] · S. K. Pramada[1] · Santosh G. Thampi[1]

Abstract

Climate change is expected to adversely affect the coastal ecosystem in many ways. One of the major consequences of climate change in coastal areas is sea level rise. In order to manage this problem efficiently, it is essential to obtain reasonably accurate estimates of future sea level. This study focuses essentially on the identification of climatic variables influencing sea level and sea level prediction. Correlation analysis and wavelet coherence diagrams were used for identifying the influencing variables, and support vector machine (SVM) and hybrid wavelet support vector machine (WSVM) techniques were used for sea level prediction. Sea surface temperature, sea surface salinity, and mean sea level pressure were observed to be the major local climatic variables influencing sea level. Halosteric effect is found to have a major impact on the sea level. The variables identified were subsequently used as predictors in both SVM and WSVM. WSVM employs discrete wavelet transform to decompose the variables before being input to the SVM model. The performance of both the models was compared using statistical measures such as root mean square error (RMSE), correlation coefficient (r), coefficient of determination (r^2), average squared error, Nash–Sutcliffe efficiency, and percentage bias along with graphical indicators such as Taylor diagrams and regression error characteristic curves. Results indicate that the WSVM model predicted sea level with an RMSE of 0.029 m during the training and 0.040 m during the testing phases. The corresponding values for SVM are 0.043 m and 0.069 m, respectively. Also, the other statistical measures and graphical indicators suggest that WSVM technique outperforms the SVM approach in the prediction of sea level.

Keywords Climate change · Modelling · Sea level · SVM · WSVM

Introduction

Accumulation of greenhouse gases in the atmosphere is the main cause of global warming and consequent change in the earth's climate. Climate change is projected to have serious ramifications on the oceans, sea level rise (SLR) being one of the major impacts. The sea level is changing continuously owing to thermal expansion of seawater, changes in salinity, and melting of glaciers and ice sheets. There are many climatic and other variables contributing to the observed changes in sea level. Some of the short-lived anthropogenic greenhouse gases such as methane (CH_4), chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs), and hydrofluorocarbons (HFCs) contribute to sea level changes which may persist for long time (Zickfeld et al. 2017). The Intergovernmental Panel on Climate Change (IPCC) reported a rise in sea level of about 1.7 mm/year on a global scale over the period 1901–2010 (IPCC 2014). A relatively recent study on sea level rise conducted using altimeter observations reported a rise of about 3 ± 0.4 mm/year after 1993 (Nerem et al. 2018). Sea level rise may bring about coastal flooding, coastal erosion and degeneration of the coastal biome (Chang et al. 2014). Therefore, in the present scenario it is vital to study the changes in sea level and also to obtain a reasonable future estimate of sea level to facilitate implementation of realistic and efficient management strategies in the coastal zone.

Before performing modelling to obtain accurate estimates of sea level, it is essential to identify the climatic variables

✉ Santosh G. Thampi
santosh@nitc.ac.in

S. Sithara
sitharasasidharan5@gmail.com

S. K. Pramada
pramada@nitc.ac.in

[1] Department of Civil Engineering, National Institute of Technology Calicut, Kozhikode, Kerala 673601, India

that influence the sea level. Global and local variations in sea level will be different due to dynamic ocean processes, seabed movements, and redistribution of water mass (IPCC 2014). Dynamical ocean processes include steric effects, thermal expansion, ocean currents, large-scale circulation, melting of glaciers, etc. (Cui et al. 1995). Maddah (2016) developed a linear relationship between sea level and surface temperature. Also, it has been reported that wind forcing and ocean salinity structure play a substantial role in regional steric level changes (Cui et al. 1995; Durack et al. 2014; IPCC 2014). Halosteric contraction caused by salinity enhancement can negate the thermosteric contribution to sea level (Durack et al. 2014; Suzuki and Ishii 2011). Local-scale sea level variations are governed by other climatic variables such as sea surface temperature and sea level pressure (Cui et al. 1995; Heyen et al. 1996). Ishida et al. (2020) developed a model to predict sea level in coastal areas on an hourly time scale using variables such as wind, air temperature, mean sea level pressure, relative positions of sun and moon along with annual global air temperature, which is an indicator of climate change. Karamouz et al. (2013) established a relationship between sea level fluctuations and climatic variables such as sea level pressure and temperature. Naren and Maity (2017) introduced a semi-empirical approach to predict changes in sea level using both atmospheric and oceanic variables.

Machine learning algorithms have wide range of applications in the field of climate studies, especially in modelling. Among these methods, artificial neural networks (ANN) have been extensively adopted for prediction and downscaling in the field of water resources and climate change (Dorado et al. 2003; Juan et al. 2017; Yadav and Chandel 2014). Another technique called decision tree, which can be employed for both classification and regression analysis, has also been widely used for prediction (Mangai and Gulyani 2016). Apart from these techniques, there are other competent machine learning approaches; support vector machine (SVM) is one such learning algorithm. Initially, SVM technique was widely employed for classification studies (Elhag et al. 2013; Modaresi and Araghinejad 2014; Zhu and Blumberg 2002). In recent times, SVM has been widely accepted as a promising approach for modelling and prediction applications (Khaledian et al. 2020; Tripathi et al. 2006; Voyant et al. 2017). This technique has also been used to predict water quality (Yunrong and Liangzhong 2009), field hydraulic conductivity (Das et al. 2011), soil moisture, and stream flow (Gill et al. 2006; Lin et al. 2006).

A wide variety of hybrid models are also available that enhances the accuracy of prediction. Fourier transforms and wavelet transforms are the commonly adopted signal denoising techniques. Wavelet transforms have been used in the field of climate change, hydrology as well as for image processing (Koirala and Gentry 2012; Pišoft et al. 2004; Xiao

and Zhang 2011). Wavelet neural network (WNN) is the most widely implemented hybrid model, in which wavelet transforms are employed to denoise the input signals before being input to the prediction model. Krishna et al. (2011) and Nury et al. (2017) employed WNN for the prediction of river flow and temperature. These authors reported that WNN model outperformed the standard ANN and other autoregressive (AR) models. Koirala and Gentry (2012) coupled wavelet transform with SWAT to predict future water yield at the outlet of a watershed. Similar to ANN, SVM has also been coupled with wavelet transforms to improve prediction accuracy (Kalteh 2013; Mohammadi et al. 2015). Mohammadi et al. (2015) forecasted horizontal global solar radiation using a hybrid wavelet support vector machine. Kalteh (2013) employed both WNN and WSVM to predict river flow on a monthly time scale; it was observed that WSVM outperformed WNN. Detailed literature review revealed that only few studies have been reported on the application of Artificial Intelligent-based techniques for prediction of sea level. Furthermore, it was found that most of the studies reported considered only thermosteric effect on sea level, which could overestimate predicted sea levels. The objectives of the present study are to identify the local climatic variables that impact sea level the most and to predict sea level using SVM and WSVM. The highlight of this work is that a methodology is proposed for the prediction of sea level using climatic variables employing a hybrid wavelet machine learning approach.

Study location

Data pertaining to climatic variables that were likely to influence sea level were obtained for a location (9.75° N, 76.25° E) near to the Willingdon Island tidal gauge station in Ernakulam, Kerala. Satellite altimeter observations of sea level (ssh), sea surface salinity (ss), and ocean current (oc) were retrieved from Copernicus Marine Environment Monitoring Service (CMEMS) web portal for the period 2000–2014. Similarly, data pertaining to evaporation (e), sea surface temperature (sst), surface pressure (sp), and mean sea level pressure (psl) were retrieved through the web portal of the European Centre for Medium-Range Weather Forecasts (ECMWF) for the period 2000–2014.

Methods

The methodology mainly consists of three distinctive phases. The first phase was to identify the climatic variables (predictors) which impact the sea level using correlation analysis and wavelet coherence diagrams; this was followed by the second phase in which the SVM and WSVM sea level prediction

models were developed and computations were performed. The final phase involved comparison of the performance of these two models. In the first phase, correlation analysis was performed between the dependent variable (ssh) and independent climatic variables to identify the predictors of the model. Furthermore, wavelet coherence diagrams were prepared for potential climatic variables based on the results of correlation analysis. Wavelet coherence is an indicator of relationship between variables in a time–frequency frame, with values ranging from 0 (low coherence) to 1 (high coherence). Sea level was modelled with SVM and WSVM techniques using the predictors identified from the correlation analysis and the wavelet coherence diagrams.

Support vector machine (SVM)

SVM is a machine learning algorithm, which has been widely accepted as a tool for data mining, especially in modelling. This technique can be employed for performing regression and pattern recognition. The notion behind this algorithm is to separate data points by an optimal hyperplane using kernel trick (Haykin 2009). Even nonlinearly separable cases can be handled with this algorithm. The data points in the input feature space are mapped onto a higher dimensional feature space using a suitable kernel function (Gunn 1998). SVM has the advantage of convex optimization that it does not confront the problem of local minima and also allows a sparse solution to be obtained (Voyant et al. 2017). The SVM learning algorithm can solve problems with small samples; it doesn't require too many training samples to develop a model (Li and Kong 2014).

The objective function is to minimize

$$J = C \sum_{n=1}^{N} E[y(x_n) - y_{dn}] + \frac{1}{2} \|w^2\|, \tag{1}$$

where C is the inverse weight penalty parameter, E: error function, and $\frac{1}{2}\|w^2\|$: weight penalty term. $y(x_n)$ is the estimator output, x_n is the input data, N is the total number of training patterns, and y_{dn} is the target data. In order to preserve the sparseness characteristic of the SVM classifier, E is defined as:

$$E_\epsilon(z) = \begin{cases} |z| - \epsilon, & \text{if } |z| > \epsilon \\ 0 & \text{Otherwise} \end{cases}. \tag{2}$$

This is called ϵ-insensitive error function; it disregards errors lesser than ϵ. $|z| = |y_{dn} - y(x_n)|$. Two slack variables ξ_n and ξ_n' are introduced such that $\xi_n \geq 0$ and $\xi_n' \geq 0$. $\xi_n > 0$ denotes data points over the ϵ-tube and $\xi_n' > 0$ denotes those beneath the ϵ-tube.

i.e. $y_{dn} > y(x_n) + \epsilon$, if $\xi_n > 0$

$y_{dn} < y(x_n) + \epsilon$, if $\xi_n' > 0$.

The objective function can be modified as

$$\text{Minimize } J = C \sum_{n=1}^{N} (\xi_n + \xi_n') + \frac{1}{2} \|w^2\| \tag{3}$$

subject to $\xi_n \geq 0$, $\xi_n' \geq 0$, $y_{dn} \leq y(x_n) + \epsilon + \xi_n$, $y_{dn} \geq y(x_n) - \epsilon - \xi_n'$.

In order to handle the constraints, Lagrange multipliers are introduced. Support vector regression is performed in a feature space,

$$y(x) = w^T \phi(x) + w_0, \tag{4}$$

where ϕ is the feature map and w_0 is the bias.

$$w = \sum_{n=1}^{N} (\lambda_n - \lambda_n') \phi(x_n), \tag{5}$$

where λ_n and λ_n' are the Lagrangian multipliers.

Substituting (5) into (4) gives

$$y(x) = \sum_{n=1}^{N} (\lambda_n - \lambda_n') K(x, x_n) + w_0. \tag{6}$$

$K(x, x_n) = \phi^T(x)\phi(x_n)$, which is the kernel function. Support vectors are those data points that contribute to Eq. (6) (Hsieh 2009). First degree polynomial kernel was used for this study. The architecture and flow chart of SVM based model are presented in Figs. 1 and 2, respectively.

Wavelet support vector machine (WSVM)

Wavelet transform is an advanced form of Fourier transform. In the case of Fourier transforms, it is not possible to obtain both time and frequency information. But, wavelet transforms can produce both time and frequency information; also, it can handle both stationary and non-stationary data (Kalteh 2013; Santos and da Silva 2014). The wavelet transform is just a waveform that exists over a limited period with a mean value of zero (Misiti et al. 1996). There are mainly two types of wavelet transforms, viz., continuous wavelet transform and discrete wavelet transform (Santos and da Silva 2014). Discrete wavelet transform can be employed as a digital filter bank to get rid of the noise present in a signal (Misiti et al. 1996).

Discrete wavelet transform (DWT)

DWT is a piecewise function; it decomposes the original signal into a number of components. DWT can be obtained by discretizing the continuous wavelet transform (CWT)

Fig. 1 Architecture of SVM
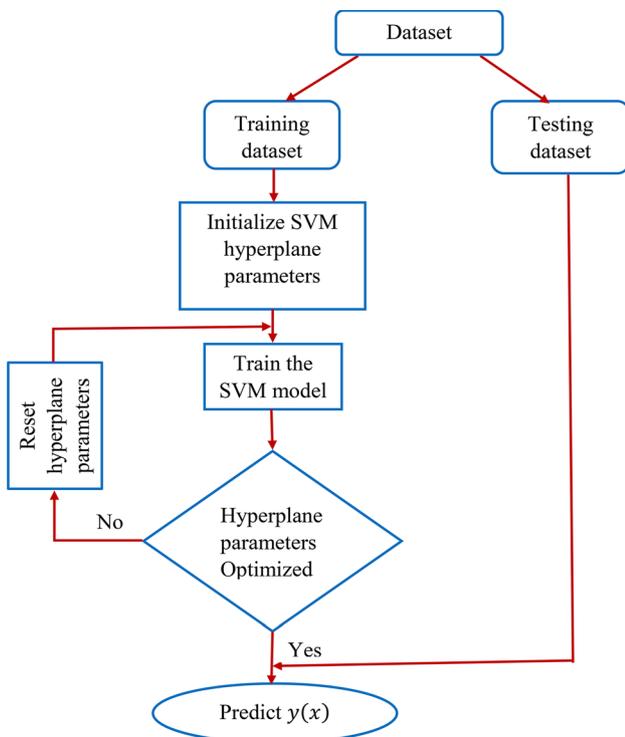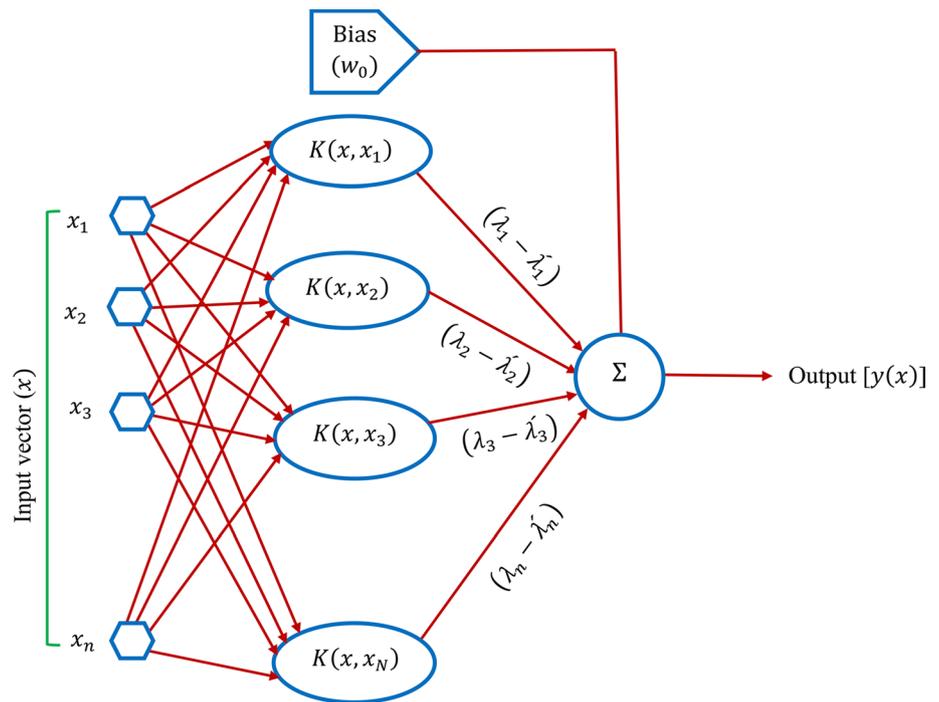
Fig. 2 Flow chart of SVM

function. Mathematically, CWT can be expressed as follows (Chan 1996):

$$\mathrm{CWT}(a, \tau) = \frac{1}{\sqrt{a}} \int S(t)\psi\left(\frac{t - \tau}{a}\right)\mathrm{d}t, \tag{7}$$

where a is the scale and τ is the shift.

DWT can be derived as given below:

$$\mathrm{DWT}(m, n) = a_0^{-\frac{m}{2}} \int S(t)\psi(a_0^{-m}t - n\tau_0)\mathrm{d}t, \tag{8}$$

where $a = a_0^m$ and $\tau = n\tau_0 a_0^m$, the variables m and n are integers.

Mother wavelet

Wavelets are produced from a fundamental wavelet, termed mother wavelet through scaling and translation. The various mother wavelets accessible are Haar, Daubechies, Biorthogonal, Coiflets, Symlets, Morlet, Mexican Hat, and Meyer. Daubechies family wavelets are denoted as dbN, where N indicates the number of vanishing moments (available from db1 to db10); theoretically, it may vary from 1 to infinity. db1 and Haar wavelet are alike. Region of support and number of vanishing moments are the two characteristic properties of the mother wavelet function. Region of support refers to the span of the wavelet and which in succession influences the feature localization abilities of the wavelet. Vanishing moment describes the capability of the wavelet to portray the information or polynomial behaviour in a given signal (Maheswaran

and Khosa 2012). For instance, db1 wavelet, with one moment, simply encodes polynomials with one coefficient or constant signal elements. db2 wavelet encodes (2 coefficients) constant and linear signal elements; db3 wavelet encodes constant, linear as well as quadratic signal elements (Maheswaran and Khosa 2012).

In this study, Daubechies wavelets were used, since they are compactly supported orthonormal wavelets with a maximal number of vanishing moments for the specified support; also, it has been employed in many studies (Maheswaran and Khosa 2012; Misiti et al. 1996). db10 is adopted in the present study because of its smoother form, which results in better depiction of time series data (Misiti et al. 1996; Nalley et al. 2012; Santos and da Silva 2014).

Decomposition process

The minimum and maximum number of decomposition levels was calculated using appropriate formulae. According to Nourani et al. (2014), the minimum level of decomposition can be computed using the equation

$$L = \text{int}[\log N_S]. \tag{9}$$

According to Lei et al. (2013), the maximum level of decomposition for a discrete wavelet transform can be computed as

$$L = \text{int}[\log_2 N_S], \tag{10}$$

where L is the decomposition level and N_S is the signal length (number of data points)

Wavelet transform was applied to the original series (time series of input variables), producing two signals, viz., approximations (A) and details (D). Approximations are the high-scale, low-frequency components of the signal. The details are the low-scale, high-frequency components. Approximation is the most important component, which reveals signal identity and detail is the nuance. This decomposition is an iterative process and so the successive approximations are decomposed in turn into approximations and details (Krishna et al. 2011). Each of these decomposed signals was input to the SVM model and the outputs were summed up to get the final predicted signal. The flow chart of WSVM is depicted in Fig. 3.

Model evaluation

The performance of both SVM and WSVM models was evaluated using statistical measures such as root mean squared error (RMSE), correlation coefficient (r), coefficient of determination (r^2), average squared error (ASE), Nash–Sutcliffe efficiency (NSE), and percentage bias (PBIAS) (Elshorbagy et al. 2000; Koirala and Gentry 2012; Moriasi et al. 2007; Tao et al. 2018). In addition to the above measures, graphical indicators such as Taylor diagrams and regression error characteristic (REC) curves were constructed to visually aid in identification of the outperforming model. Formulas used for the computation of the statistical measures are as follows:

Fig. 3 Flow chart of WSVM
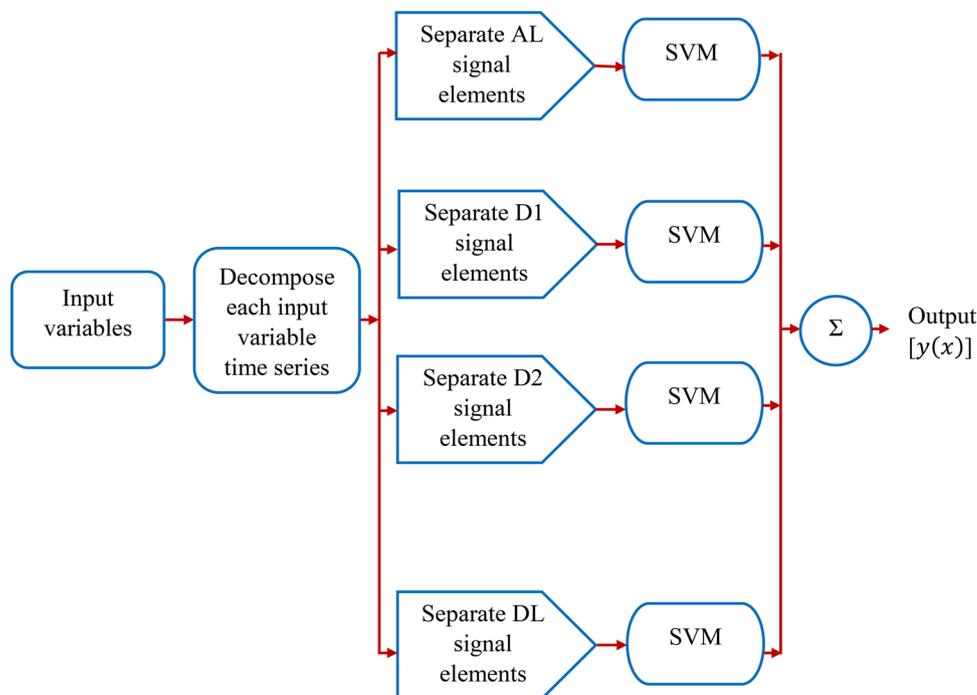
$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(Y_i^O - Y_i^P \right)^2} \tag{11}$$

$$r^2 = \left[\frac{\sum_{i=1}^{N} \left(Y_i^O - \overline{Y^O} \right)\left(Y_i^P - \overline{Y^P} \right)}{\sqrt{\sum_{i=1}^{N} \left(Y_i^O - \overline{Y^O} \right)^2} \sqrt{\sum_{i=1}^{N} \left(Y_i^P - \overline{Y^P} \right)^2}} \right]^2 \tag{12}$$

$$\text{ASE} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i^O - Y_i^P \right)^2 \tag{13}$$

$$\text{NSE} = 1 - \left[\frac{\sum_{i=1}^{N} \left(Y_i^O - Y_i^P \right)^2}{\sum_{i=1}^{N} \left(Y_i^O - \overline{Y^O} \right)^2} \right] \tag{14}$$

$$\text{PBIAS} = \left[\frac{\sum_{i=1}^{N} \left(Y_i^O - Y_i^P \right) \times 100}{\sum_{i=1}^{N} Y_i^O} \right], \tag{15}$$

where Y_i^O and Y_i^P are the observed and predicted ith value of sea level, $\overline{Y^O}$ and $\overline{Y^P}$ are the mean of observed and predicted values of sea level, and N is the total number of data points.

Results and discussion

Identification of climatic variables

Sea level is influenced by many local climatic variables; the identification of the most influencing variables (predictors) will enable more accurate predictions of sea level. The variables chosen for the correlation analysis to identify the predictors were sea level (ssh), sea surface salinity (ss), ocean current (oc), evaporation (*e*), sea surface temperature (sst), surface pressure (sp), and mean sea level pressure (psl).

Correlation analysis was performed between the sea level (predictand) and the other independent variables. The results of the analysis are presented in Table 1. From Table 1, it can be observed that variables such as mean sea level pressure, ocean current, surface pressure, sea surface salinity, and sea surface temperature possess very good correlation with sea level. Since sea surface salinity has a good correlation with sea level, it is clear that halosteric effect is significant. Furthermore, surface pressure holds strong relation with mean sea level pressure; besides that, ocean current exhibits good relation with mean sea level pressure, surface pressure, and sea surface temperature. These inter-correlated climatic variables such as ocean current and surface pressure were removed from the predictors list. Thus, climatic variables such as sea surface temperature, sea surface salinity, and mean sea level pressure were chosen as the predictors in this study and hence it can be said that both thermosteric and halosteric effects are accounted for in the prediction of sea level. Wavelet coherence diagrams constructed between the predictand and predictors are presented in Figs. 4, 5, and 6. The Wavelet coherence diagram depicts the correlation between two variables in a time–frequency frame. In all three diagrams (Figs. 4, 5, and 6), the thick black lines bounding the red coloured regions reflect 95% confidence level and the thin solid line represents the cone of influence. The faded region beyond it should be carefully treated due to potential edge effects. The colour scale ranges from blue (zero correlation) to red (perfect correlation). From Figs. 4, 5, and 6, it is evident that the red region is more predominant, indicating that there is a strong correlation between the sea level and the climatic variables considered. In addition, the prominent orientation of the arrows in the first two figures (Figs. 4 and 5) is to the right, indicating a positive correlation. Mean sea level pressure and sea surface temperature are therefore positively correlated with the sea level. In Fig. 6, the orientation of the arrows is to the left, indicating a negative correlation between sea level and sea surface salinity. It is also clear from the wavelet coherence diagram (Fig. 6) that sea level is significantly influenced by

Table 1 Results of correlation analysis

Climatic variables	Sea level	Sea surface temperature	Mean sea level pressure	Surface pressure	Evaporation	Sea surface salinity	Ocean current
Sea level	1	**0.552**	**0.622**	0.618	−0.378	**−0.743**	−0.644
Sea surface temperature	0.552	1	0.036	0.049	−0.540	−0.384	−0.565
Mean sea level pressure	0.622	0.036	1	0.999	−0.088	−0.484	−0.497
Surface pressure	0.618	0.049	0.999	1	−0.088	−0.495	−0.510
Evaporation	−0.378	−0.540	−0.088	−0.088	1	0.179	0.152
Sea surface salinity	−0.743	−0.384	−0.484	−0.495	0.179	1	0.369
Ocean current	−0.644	−0.565	−0.497	−0.510	0.152	0.369	1

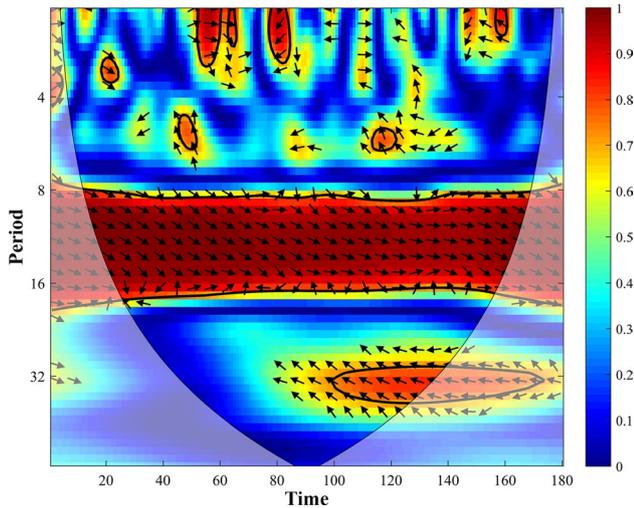Bold indicates finally selected variables (predictors) and their correlation with the sea level

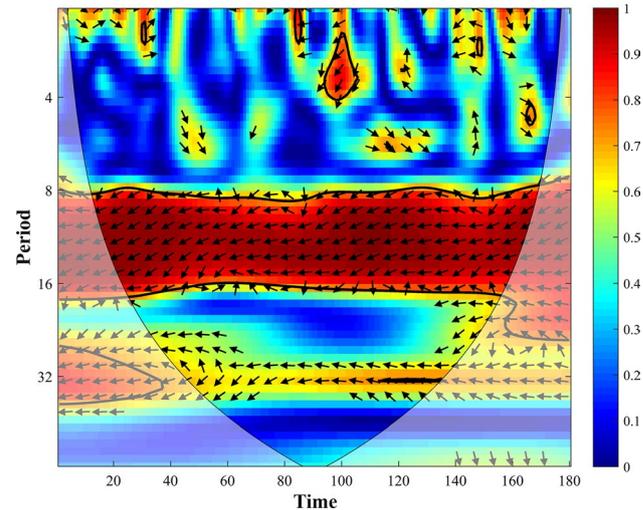**Fig. 4** Wavelet coherence between sea level and mean sea level pressure

Fig. 6 Wavelet coherence between sea level and sea surface salinity

formulation of models and sea level was the output or target variable. The entire data set was divided into two sets, such as training data set and testing data set. About 70% of the data (January 2000–June 2010) were employed for training and the remaining 30% data were used for testing.

SVM model

In the course of training, the SVM model parameters such as C and ϵ were optimized through trial and error, and the values obtained are 10 and 0.001, respectively. The trained model was then tested by predicting the values of sea level using the remaining 30% data and the results were compared with the observations of a satellite altimeter. A comparison of sea level predicted by the SVM model with satellite altimetry observations, during both training and testing phases, is presented (Fig. 7). It is clear from the figure that there is a reasonably good match between the predicted sea level and the altimetry observations during both training and testing phases. Furthermore, most of the peaks match closely. It is hence evident that the SVM based model performs reasonably well in predicting sea level.

WSVM model

In this model, the dependent variable sea level and all the independent variables such as sea surface temperature, sea surface salinity and mean sea level pressure were decomposed using DWT. The mother wavelet adopted was db10. The minimum and maximum levels of decomposition were calculated using Eqs. (9) and (10), respectively, and these were 2 and 7, respectively. Decomposition of a signal at a level greater than the optimum may disturb the characteristic properties of the data set

Fig. 5 Wavelet coherence between sea level and sea surface temperature

the halosteric effect, as red coloured regions are more in the diagram. It is, therefore, evident from these figures that there exists strong correlation between the predictand and predictors.

Sea level prediction

In this phase of the study, a relationship was established between the predictand (sea level) and the predictors (sea surface temperature, sea surface salinity, and mean sea level pressure) using the SVM and WSVM models. Predictors such as sea surface temperature, sea surface salinity, and mean sea level pressure were the input variables for the

Fig. 7 Comparison of altimeter and SVM predicted sea level

(Pramanik et al. 2011). Therefore, the uppermost level of decomposition was avoided and a decomposition level of 6 was selected. The decomposition process yields six details (D1, D2, D3, D4, D5, D6) and one approximation (A6). The original signal can be represented as $S = D1 + D2 + D3 + D4 + D5 + D6 + A6$. The details and approximations obtained for each climatic variable are depicted in Figs. 8 and 9.

Each of these decomposed components was segregated and given as input to the SVM model. A total of seven output series were obtained from the above seven segregated components. The final output (sea level) was obtained by adding all the above outputs. The results obtained during training and testing are depicted in Fig. 10. It is evident from the figure that there is a good match between the observed

and WSVM predicted sea levels during, both training and testing periods.

Comparison between SVM and WSVM models

In order to identify the best sea level prediction model, the statistical parameters of both SVM and WSVM models were compared (Table 2). The model performance evaluation criteria, based on various statistical parameters as specified in Moriasi et al. (2007), are presented in Table 3.

During the training period, the SVM model yields $r^2 > 0.5$ (acceptable), $0.75 < NSE \leq 1$ (very good) as well as PBIAS $< \pm 10$ (very good), whereas during the testing period NSE > 0.5 (satisfactory) and $\pm 10 \leq$ PBIAS $< \pm 15$ (good). Therefore, overall, the model performance can be
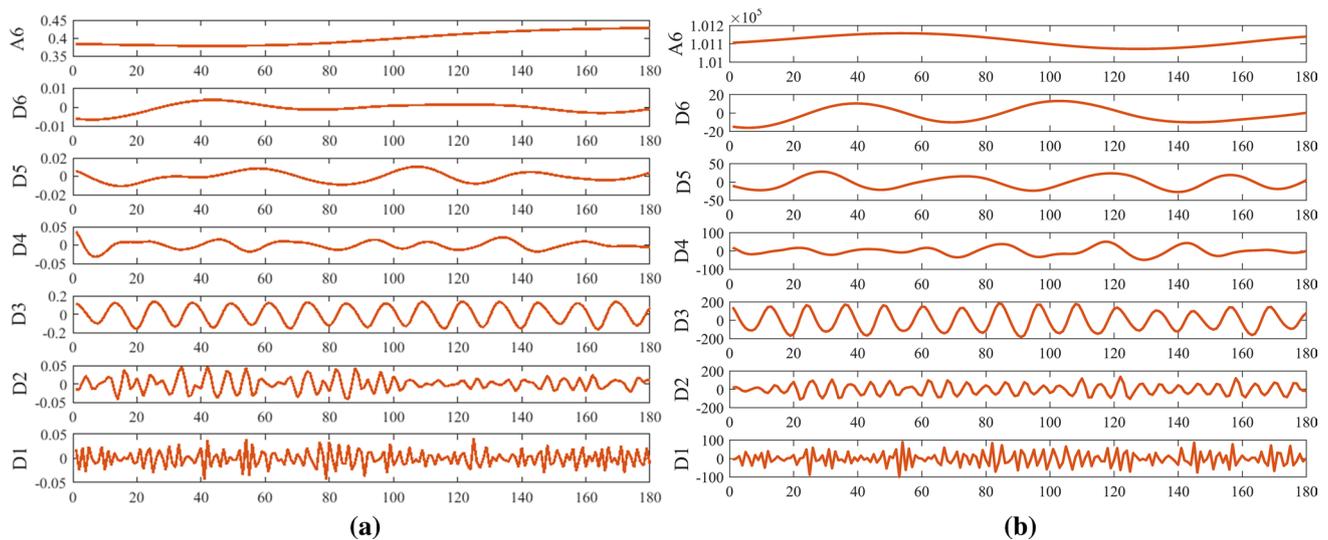
Fig. 8 DWT decomposed components of **a** sea level **b** mean sea level pressure
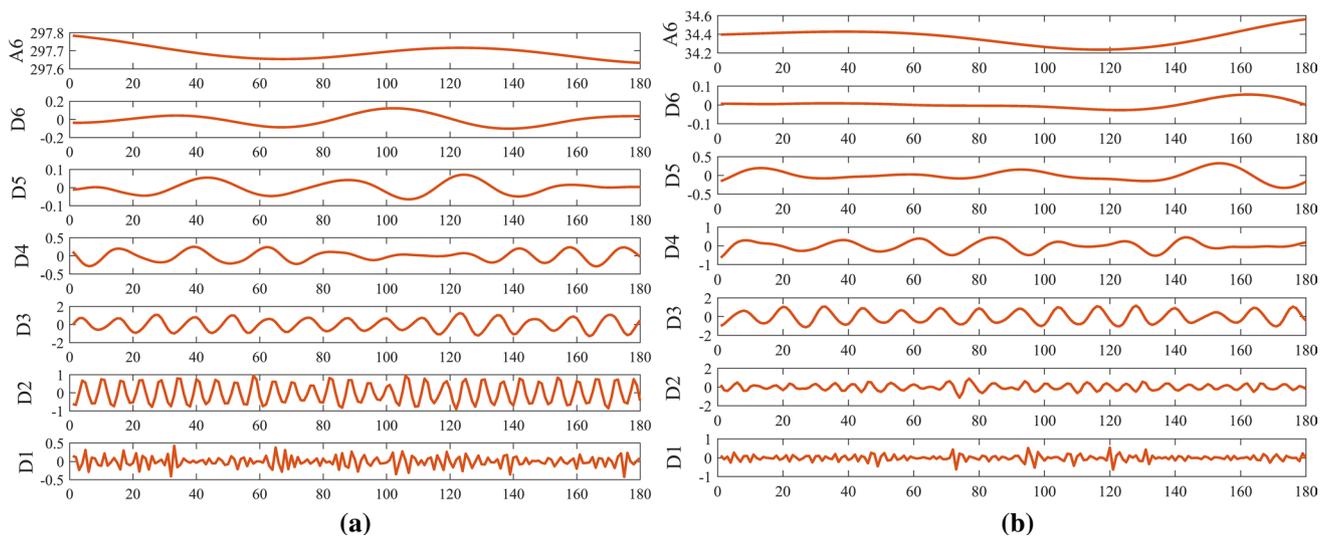
Fig. 9 DWT decomposed components of **a** sea surface temperature **b** sea surface salinity
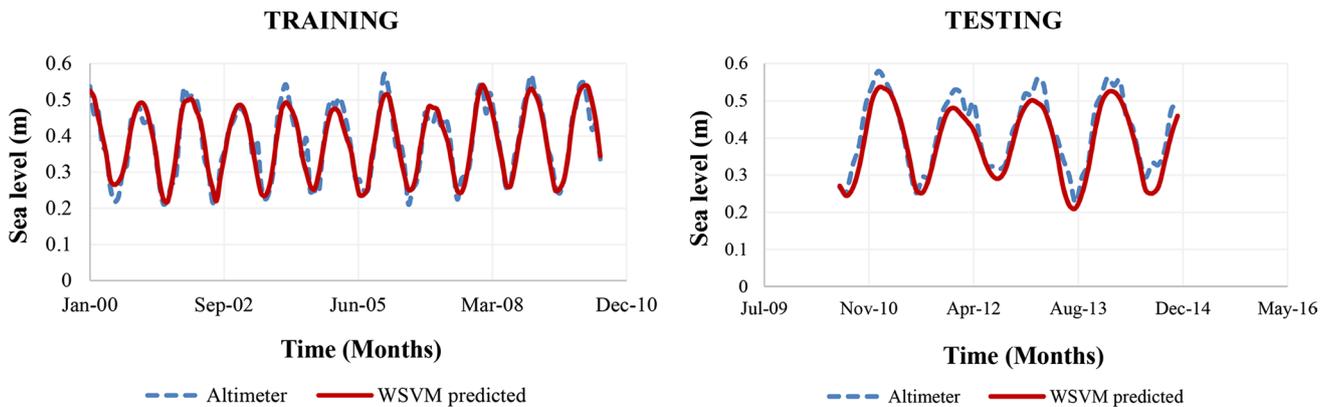
Fig. 10 Comparison of altimeter (observed) and WSVM predicted sea level

Table 2 Performance statistics

Statistical parameter	SVM		WSVM	
	Training	Testing	Training	Testing
RMSE (m)	0.043	0.069	0.029	0.040
r	0.902	0.916	0.957	0.964
r²	0.813	0.839	0.916	0.931
ASE	0.001	0.004	0.0008	0.001
NSE	0.809	0.516	0.915	0.833
PBIAS	− 0.840	13.40	0.493	7.485

Table 3 Performance evaluation criteria

Parameter	Criteria	Performance category
r^2	$r^2 > 0.5$	Acceptable
NSE	$0.75 < \text{NSE} \leq 1$	Very good
	$0.65 < \text{NSE} \leq 0.75$	Good
	$0.5 < \text{NSE} \leq 0.65$	Satisfactory
	$\text{NSE} \leq 0.5$	Unsatisfactory
PBIAS	$\text{PBIAS} < \pm 10$	Very good
	$\pm 10 \leq \text{PBIAS} < \pm 15$	Good
	$\pm 15 \leq \text{PBIAS} < \pm 25$	Satisfactory
	$\text{PBIAS} \geq \pm 25$	Unsatisfactory

rated as good. Similarly, for the WSVM model, during both training and testing periods, $r^2 > 0.5$, $0.75 < \text{NSE} \leq 1$ as well as $\text{PBIAS} < \pm 10$. Hence, the overall performance of this model can be rated as very good. In addition to this, RMSE and ASE values are relatively small for the WSVM

model compared to the SVM model. Both the models are acceptable; however, based on the performance statistics, WSVM model outperforms the SVM model.

Fig. 11 Taylor diagram for the training data set
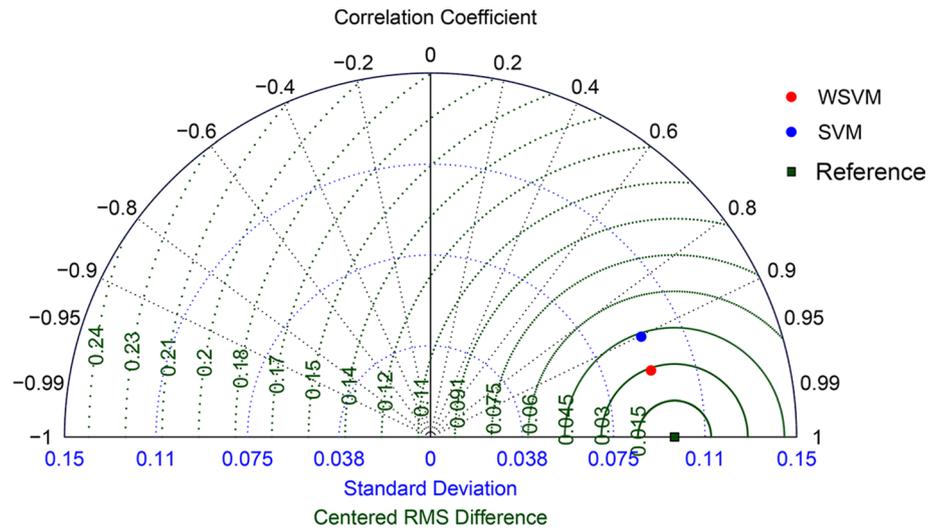
Fig. 12 Taylor diagram for the testing data set
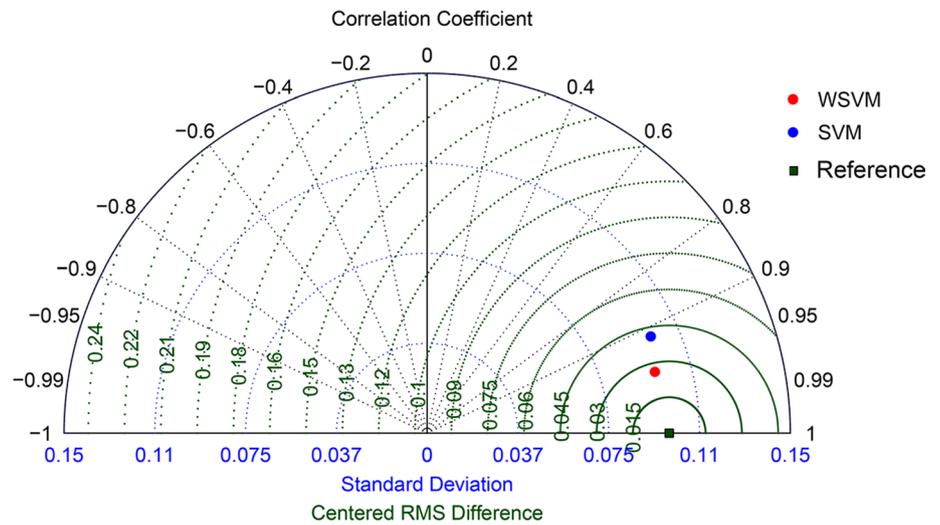
Taylor diagrams for both training and testing data sets were constructed (Figs. 11 and 12) by plotting the results of the SVM and WSVM models against the reference (observed data). Each model in the Taylor diagram was represented with three major statistical parameters, viz., correlation coefficient, standard deviation, and centred RMS difference. It is evident from the diagram that the WSVM model performs better than the SVM model as the point representing the WSVM model is close to the reference.

In addition, REC curves for the SVM and WSVM models during training and testing period were plotted (Fig. 13). The REC curve determines the cumulative distribution function of the error; the area above the curve is a metric associated with the error of the prediction model. It is apparent from the figure that the WSVM model outperforms the SVM model as the area over the curve is small for this model.
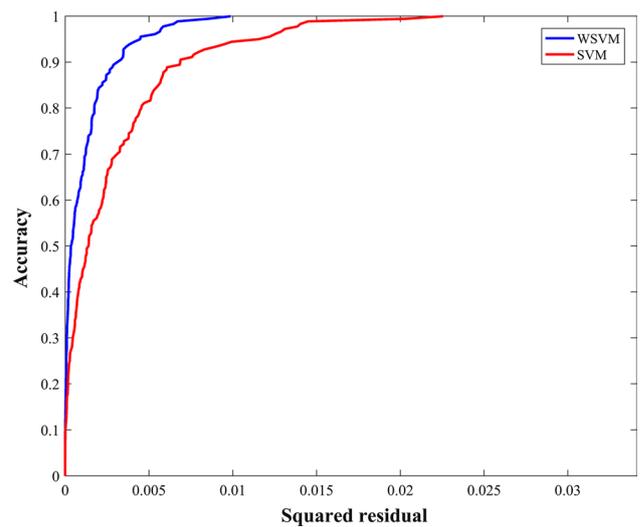
Fig. 13 REC curves

Conclusions

Climatic variables have a significant influence on the sea level. Regional sea level can be better analysed by identifying the local climatic variables which drive sea level fluctuations. From the correlation analysis and wavelet coherence diagrams, it was observed that sea surface temperature, sea surface salinity, and mean sea level pressure are the major driving forces that influence sea level. Also, it was observed that mean sea level pressure and sea surface temperature are positively correlated with sea level, whereas sea surface salinity is negatively correlated with sea level.

In this study, sea level was predicted with both the SVM and WSVM models. Satellite altimetry observations of sea level and data on climatic variables such as mean sea level pressure, sea surface temperature, and sea surface salinity, retrieved from the CMEMS and ECMWF web portals, were used for training and testing of the models. The SVM based model was developed using the variables in its original form, whereas the WSVM based model was developed using decomposed elements of the original variables. The performance of both the models was compared using statistical parameters. Furthermore, graphical indicators such as Taylor diagrams and REC curves were used for visual identification of the best performing model. It was found that the WSVM based model was superior to the SVM model in terms of its predictions.

One of the important findings from the study is that the impact of halosteric effects on sea level is very significant and it has to be taken into account in sea level prediction. In addition, machine learning techniques like SVM and WSVM can be effectively used to predict sea level. The methods suggested in this research can be used to obtain reasonably accurate estimates of sea level. The findings of this research would enable scientists to predict sea level based on climatic variables and to evaluate the impact of climate change on rising sea levels. Scientifically sound and accurate information pertaining to the sea level is very much needed for developing appropriate coastal management schemes under changing climate. Comparison of this approach with other machine learning techniques can be set as the future scope of this research.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Chan YT (1996) Wavelet basics. Kluwer Academic Publishers Group, Netherlands

Chang B, Guan J, Aral MM (2014) Scientific discourse: climate change and sea-level rise. J Hydrol Eng 20:1–14. https://doi.org/10.1061/(asce)he.1943-5584.0000860

Cui M, Storch HV, Zorita E (1995) Coastal sea level and the large-scale climate state a downscaling exercise for the Japanese Islands. Tellus A 47:132–144. https://doi.org/10.3402/tellusa.v47i1.11498

Das SK, Samui P, Sabat AK (2011) Prediction of field hydraulic conductivity of clay liners using an artificial neural network and support vector machine. Int J Geomech 12:606–611. https://doi.org/10.1061/(asce)gm.1943-5622.0000129

Dorado J, RabuñAL JR, Pazos A, Rivero D, Santos A, Puertas J (2003) Prediction and modeling of the rainfall-runoff transformation of a typical urban basin using ANN and GP. Appl Artif Intell 17:329–343. https://doi.org/10.1080/713827142

Durack PJ, Wijffels SE, Gleckler PJ (2014) Long-term sea-level change revisited: the role of salinity. Environ Res Lett 9:114017. https://doi.org/10.1088/1748-9326/9/11/114017

Elhag M, Psilovikos A, Sakellariou-Makrantonaki M (2013) Land use changes and its impacts on water resources in Nile Delta region using remote sensing techniques. Environ Dev Sustain 15:1189–1204. https://doi.org/10.1007/s10668-013-9433-5

Elshorbagy A, Simonovic SP, Panu US (2000) Performance evaluation of artificial neural networks for runoff prediction. J Hydrol Eng 5:424–427

Gill MK, Asefa T, Kemblowski MW, Mckee M (2006) Soil moisture prediction using support vector machines. J Am Water Resour 42:1033–1046

Gunn SR (1998) Support vector machines for classification and regression. Image speech and intelligent systems Technical Report

Haykin S (2009) Neural networks and learning machines, 3rd edn. Pearson Prentice Hall, Upper Saddle River

Heyen H, Zorita E, Storch HV (1996) Statistical downscaling of monthly mean North Atlantic air-pressure to sea level anomalies in the Baltic Sea. Tellus A 48:312–323. https://doi.org/10.3402/tellusa.v48i2.12062

Hsieh WW (2009) Machine learning methods in the environmental sciences: neural networks and kernels. Cambridge University Press, Cambridge

IPCC (2014) Climate Change 2014: synthesis report. Contribution of working groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change [Core Writing Team, R.K. Pachauri and L.A. Meyer (Eds.)]. IPCC, Geneva, Switzerland, p 151

Ishida K, Tsujimoto G, Ercan A, Tu T, Kiyama M, Amagasaki M (2020) Hourly-scale coastal sea level modeling in a changing climate using long short-term memory neural network. Sci Total Environ 720:137613. https://doi.org/10.1016/j.scitotenv.2020.137613

Juan C, Genxu W, Tianxu M, Xiangyang S (2017) ANN model-based simulation of the runoff variation in response to climate change on the Qinghai-Tibet Plateau, China. Adv Meteorol. https://doi.org/10.1155/2017/9451802

Kalteh AM (2013) Monthly river flow forecasting using artificial neural network and support vector regression models coupled with wavelet transform. Comput Geosci 54:1–8. https://doi.org/10.1016/j.cageo.2012.11.015

Karamouz M, Zahmatkesh Z, Nazif S (2013) Realization of sea level rise using climate variables considering climate change scenarios. In: World environmental and water resources congress 2013: showcasing the future, pp 1402–1409

Khaledian MR, Isazadeh M, Biazar SM, Pham QB (2020) Simulating Caspian Sea surface water level by artificial neural network and support vector machine models. Acta Geophys 68:553–563. https://doi.org/10.1007/s11600-020-00419-y

Koirala SR, Gentry RW (2012) SWAT and wavelet analysis for understanding the climate change impact on hydrologic response. Open J Mod Hydrol 02:41–48. https://doi.org/10.4236/ojmh.2012.22006

Krishna B, Rao YRS, Nayak PC (2011) Time series modeling of river flow using wavelet neural networks. J Water Resour Prot 03:50–59. https://doi.org/10.4236/jwarp.2011.31006

Lei L, Wang C, Liu X (2013) Discrete wavelet transform decomposition level determination exploiting sparseness measurement. Int J Electr Comput Energ Electron Commun Eng 7:1182–1185

Li XZ, Kong JM (2014) Application of GA-SVM method with parameter optimization for landslide development prediction. Nat Hazards Earth Syst Sci 14:525–533. https://doi.org/10.5194/nhess-14-525-2014

Lin J, Cheng C, Chau K (2006) Using support vector machines for long-term discharge prediction. Hydrol Sci J 51:599–612. https://doi.org/10.1623/hysj.51.4.599

Maddah HA (2016) Modeling the relation between carbon dioxide emissions and sea level rise for the determination of future (2100) sea level. Am J Environ Eng 6:52–61. https://doi.org/10.5923/j.ajee.20160602.03

Maheswaran R, Khosa R (2012) Comparative study of different wavelets for hydrologic forecasting. Comput Geosci 46:284–295. https://doi.org/10.1016/j.cageo.2011.12.015

Mangai JA, Gulyani BB (2016) Induction of model trees for predicting BOD in river water: a data mining perspective. In: Perner P (ed) Advances in data mining. Applications and theoretical aspects. ICDM 2016. Lecture notes in computer science, vol 9728. Springer, Cham. https://doi.org/10.1007/978-3-319-41561-1_1

Misiti M, Misiti Y, Oppenheim G, Poggi J-M (1996) Wavelet toolbox user's guide. The Math Works Ins., USA

Modaresi F, Araghinejad S (2014) A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water Resour Manag 28:4095–4111. https://doi.org/10.1007/s11269-014-0730-z

Mohammadi K, Shamshirband S, Tong CW, Arif M, Petković D, Sudheer C (2015) A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation. Energy Convers Manag 92:162–171. https://doi.org/10.1016/j.enconman.2014.12.050

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans ASABE 50:885–900. https://doi.org/10.13031/2013.23153

Nalley D, Adamowski J, Khalil B (2012) Using discrete wavelet transforms to analyze trends in streamflow and precipitation in Quebec and Ontario (1954–2008). J Hydrol 475:204–228. https://doi.org/10.1016/j.jhydrol.2012.09.049

Naren A, Maity R (2017) Hydroclimatic modelling of local sea level rise and its projection in future. Theor Appl Climatol 130:761–774. https://doi.org/10.1007/s00704-016-1897-4

Nerem RS, Beckley BD, Fasullo JT, Hamlington BD, Masters D, Mitchum GT (2018) Climate-change-driven accelerated sea-level rise detected in the altimeter era. Proc Natl Acad Sci 115:2022–2025. https://doi.org/10.1073/pnas.1717312115

Nourani V, Hosseini A, Adamowski J, Kisi O (2014) Applications of hybrid wavelet—artificial intelligence models in hydrology: a review. J Hydrol 514:358–377. https://doi.org/10.1016/j.jhydrol.2014.03.057

Nury AH, Hasan K, Bin Alam MJ (2017) Comparative study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh. J King Saud Univ Sci 29:47–61. https://doi.org/10.1016/j.jksus.2015.12.002

Pišoft P, Kalvová J, Brázdil R (2004) Cycles and trends in the Czech temperature series using wavelet transforms. Int J Climatol 24:1661–1670. https://doi.org/10.1002/joc.1095

Pramanik N, Panda RK, Singh A (2011) Daily river flow forecasting using wavelet ANN hybrid models. J Hydroinformatics 13:49–63. https://doi.org/10.2166/hydro.2010.040

Santos CAG, da Silva GBL (2014) Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models. Hydrol Sci J 59:312–324. https://doi.org/10.1080/02626667.2013.800944

Suzuki T, Ishii M (2011) Long—term regional sea level changes due to variations in water mass density during the period 1981–2007. Geophys Res Lett 38:L21604. https://doi.org/10.1029/2011gl049326

Tao H, Sulaiman SO, Yaseen ZM, Asadi H, Meshram SG, Ghorbani MA (2018) What is the potential of integrating phase space reconstruction with SVM-FFA data-intelligence model? application of rainfall forecasting over regional scale. Water Resour Manag 32:3935–3959. https://doi.org/10.1007/s11269-018-2028-z

Tripathi S, Srinivas VV, Nanjundiah RS (2006) Downscaling of precipitation for climate change scenarios: a support vector machine approach. J Hydrol 330:621–640. https://doi.org/10.1016/j.jhydrol.2006.04.030

Voyant C, Notton G, Kalogirou S, Nivet ML, Paoli C, Motte F, Fouilloy A (2017) Machine learning methods for solar radiation forecasting: a review. Renew Energy 105:569–582. https://doi.org/10.1016/j.renene.2016.12.095

Xiao F, Zhang Y (2011) A comparative study on thresholding methods in wavelet-based image denoising. Procedia Eng 15:3998–4003. https://doi.org/10.1016/j.proeng.2011.08.749

Yadav AK, Chandel SS (2014) Solar radiation prediction using artificial neural network techniques: a review. Renew Sustain Energy Rev 33:772–781. https://doi.org/10.1016/j.rser.2013.08.055

Yunrong X, Liangzhong J (2009) Water quality prediction using LS-SVM with particle swarm optimization. In: 2009 second international workshop on knowledge discovery and data mining. IEEE, pp 900–904

Zhu G, Blumberg DG (2002) Classification using ASTER data and SVM algorithms: the case study of Beer Sheva, Israel. Remote Sens Environ 80:233–240. https://doi.org/10.1016/s0034-4257(01)00305-4

Zickfeld K, Solomon S, Gilford DM (2017) Centuries of thermal sea-level rise due to anthropogenic emissions of short-lived greenhouse gases. Proc Natl Acad Sci 114:657–662. https://doi.org/10.1073/pnas.1612066114

Improving ANN-based streamflow estimation models for the Upper Indus Basin using satellite-derived snow cover area

Muhammad Hassan[1] · Ishtiaq Hassan[1]

Abstract

The mountainous catchments often witness contrasting regimes and the limited available meteorological network creates uncertainty in both the hydrological data and developed models. To overcome this problem, remotely sensed data could be used in addition to on-ground observations for hydrological forecasting. The fusion of these two types of data gives a better picture and helps to generate adequate hydrological forecasting models. The study aims at the improvement of ANN-based streamflow estimation models by using an integrated data-set containing, the satellite-derived snow cover area (SCA) with on-ground flow observations. For this purpose, SCA of three sub catchments of Upper Indus Basin, namely Gilgit, Astore and Bunji coupled with their respective gauge discharges is used as model inputs. The weekly stream-flow models are developed for inflows at Besham Qila located just upstream of Tarbela dam. The data-set for modeling is prepared through normalizing all variables by scaling between 0 and 1. A mathematical tool, Gamma test is applied to fuse the inputs, and a best input combination is selected on the basis of minimum gamma value. A feed forward neural network trained via two layer Broyden Fletcher Goldfarb Shanno algorithm is used for model development. The models are evaluated on the basis of set of performance indicators, namely, Nash–Sutcliffe Efficiency, Root Mean Square Error, Variance and BIAS. A comparative assessment has also been made using these indicators for models developed, through data-set containing gauge discharges, only and the data-set fused with satellite-derived SCA. In particular, the current study concluded that the efficiency of ANN-based streamflow estimation models developed for mountainous catchments could be improved by integrating the SCA with the gauge discharges.

Keywords Mountainous catchments · Upper Indus Basin · Snow cover area · Streamflow · Artificial neural network (ANN)

Introduction

The runoff originating from the Hindukush–Karakoram–Himalaya (HKH) ranges, is mainly generated through melting of snow and glaciers (Fowler and Archer 2005; Hasson et al. 2017), and contributes up to 80% of mean annual flows of Upper Indus Basin (UIB) (Immerzeel et al. 2009;

✉ Muhammad Hassan
hassan.ce@must.edu.pk

Ishtiaq Hassan
eishtiaq@cust.edu.pk

1 Capital University of Science and Technology (CUST), Islamabad, Pakistan

Arfan et al. 2019). The glacier cover over HKH ranges is considered as one of the fastest retreating cover in the world. Prasad (2009) and Liu and Chen (2000) reported that global warming instigated glacial recession and created significant changes on hydrology and water resources over HKH region. According to Wang et al. (2014), glacier mass of the region is shrinking, resulting in increased melt-water contribution to river flows downstream of the area, specifically during summer season. The reduction in snow covered areas within the UIB has also been confirmed by Mukhopadhyay and Dutta (2010) who reported a decline of about 2.15% during period 1992 to 2010. On the other hand, Tahir et al. (2016) claimed that snow cover in UIB has slightly increased in south (Western Himalaya) and in North (Central Karakorum). Archer and Fowler (2004) have also reported the same trend of increased ice mass over UIB for the last two decades. These discrepancies in previous studies have created uncertainties in climate estimation models (Arfan

et al. 2019). It has been confirmed by previous researchers that UIB has contrasting hydro-meteo-cryospheric regimes because of extreme complexity of HKH terrain (Hasson et al. 2014). Being a complex terrain, it has a limited meteorological network that is unable to cover the extent of this basin in both horizontal and vertical directions (Hasson et al. 2017).

In developing countries, most of the watersheds are ungauged or poorly gauged due to confined economical resources (Razavi and Coulibaly 2016). Thus, creating a problem of limited availability of the data for hydrological models, especially in the case of data-driven models. Similarly, the inaccurate calibration and validation may create uncertainty not only in the input data but in the hydrological models themselves (Naeem et al. 2014). The uncertainty in hydrological models could be reduced using multiple source/type information instead of using single source/type information. Azmi et al. (2010) reported that the hydrological models could only be improved through "observations" and not by increasing the complexity of the models. This is the reason that despite choosing advanced modeling options to model the response of complex mountainous catchments, it is difficult to eliminate the "uncertainty" because the available data is limited due to little coverage over the terrain (Behrangi et al. 2011). Dijk (2011) suggested to use the new breed of observations, in combination with the on-ground observations, containing remotely sensed data, airborne data and in situ sensor's data to improve the performance of hydrological models. Previously, the researchers have utilized the concept of fusing/integrating on-ground (time series data) and satellite-derived observations (spatial maps) to create better hydrological models but these studies are limited as compared to the overall importance of the subject; See and Abrahart (2001), Ercan et al. (2013), Sun et al. (2017). It is notable, that the existing data-fusion practices are more valid in the field of remote sensing and geophysics and somehow less valid in the field of hydrological modeling (Nagarajan et al. 2010). Similarly, the limits of fusion are presently not characterized and there are no specific guidelines that what amount of data should be entered into a fusion process and whether or not the use of multi-source data would be advantageous (Abrahart and See 2002). Therefore, it is essential to experiment the data-fusion approaches in the field of hydrological forecasting to establish the evidence of its importance.

Snow cover area (SCA) is considered as an important factor for many climate changes and water management challenges (Brubaker et al. 2005). It plays a vital role in estimation of streamflows for mountainous areas where the flow is mostly generated through melting of glacial masses (Sönmez et al. 2014). The balance of these masses defines the contribution of snowmelt to runoff (Rastner et al. 2019). The magnitude of snowmelt could be obtained through calculation of

changing SCA of a typical region. Remote sensing offers a wide range of options through a set of satellites to check the spatial and temporal variation of snow cover extent. Most of these satellite observations which are essentially the gridded data-sets, are easy to use and freely available (e.g., Lutz et al. 2016; Sakai et al. 2015; Tahir et al. 2016). Although, these gridded datasets have the capacity of observing multiple parameters at the same time with more Ariel coverage but in some cases the use of this data alone, may create erroneous results. For example, in the complex terrain like UIB, where the grids are often larger than the spatial variability of precipitation and the adopted interpolation schemes may add up and lead to uncertain outcomes. Similarly, Andermann (2011) observed that the satellite observations underestimate the precipitation in areas where significant snowfall occurs. So, this makes the use of gridded datasets alone, for hydrological estimation, questionable. Previous studies on the UIB evidenced that a predictive relationship exists between satellite observations and on-ground observations of the basin. As Forsythe et al. (2011) reported that remotely sensed spatial data products (MODIS SCA and LST) can provide adequate analogs for these point observations. He also suggested that the fusion of these two types of data may improve the assessment of the hydrological impact of the UIB. Similarly, Hayat (2019) has shown the UIB flow dependency upon satellite-derived SCA and simulated the runoff using Snowmelt-Runoff Model (SRM). The importance of SCA derived through satellite images is increased exponentially when a snow fed catchment is complex and poorly gauged.

Artificial Neural Networks (ANNs) are capable of capturing the complex and nonlinear relationships between inputs and outputs and they do not require detailed knowledge about catchment and underlying physical processes (Kisi and Sanikhani 2015; Ahmed et al. 2018). However, an inappropriate input data to ANN models may create inappropriate learning maps, that ultimately results in reduced efficiency of hydrological models (Zhu and Wu 2004). Similarly, the accuracy of these models depends upon a careful selection of elements that are user-defined such as model structure, data length selection, parameter optimization and data normalization techniques (Uysal and Ünal 2018). Apart from the selection of suitable network-type and its architecture, a suitable method to reduce over-fitting is required for the successful application of ANN models (Piotrowski and Napiorkowski 2013). The conventional methods to reduce over fitting in ANN are; regularization (Tikhonov 1963), early stopping (Lang et al. 1990), reducing complexity (Le Cun et al. 1990) and noise injection (Giustolisi 2004). During recent past, ANN have become quite popular in runoff modeling, specifically in areas where catchment response is mostly dependent upon rainfall. There are many studies which have utilized ANN to develop hydrologic models on

rainfall dominant catchments (Srinivasulu and Jain 2006; Gökbulak et al. 2015). But a very few have been reported for mountainous catchments due to the fact that measuring on-ground precipitation and other climate variables is quite challenging for such catchments (Uysal and Arda 2016). Although there are many past studies on the application of ANN for snow estimation such as Quiroga et al. (2013), Mishra et al. (2014), Lee et al. (2019), Broxton et al. (2014) and snowmelt- runoff modeling (Ata et al. 2016), but a very few have been reported which utilized SCA as one of the input variables for ANN-based flow estimation models (Yilmaz et al. 2011; Ata et al. 2016; Uysal and Arda 2016). Whereas the satellite-derived SCA in addition to gauge observations could be used to create a better input data-set and ultimately better streamflow estimation models.

The main motive of this study is to improve the real-time streamflow forecasting for a catchment which is complex and has a limited meteorological network by combining the two sources of data (on-ground discharge data & satellite-derived SCA). It is expected that the fusion of this multi-source data could create a better initial state with less uncertainty, which ultimately results in better streamflow forecasts for a difficult topographical catchment such as UIB. The fusion of these input variables has been carried out with the help of Gamma Test that provides an initial estimate of the Mean Square Error (MSE), prior to modeling for each set of input combination/mask. Moreover, this study has also explored the use of ANN for streamflow estimation in a mountainous catchment, where greater part of the flow is derived through melting of snow. The performance of ANN models has been optimized through the process of data normalization, input combination selection, selection of nodes in hidden layers and over fitting reduction through Gamma and M-Test. The models are developed and trained using two layer BFGS algorithm in WinGamma environment (see detail of WinGamma application at Durrant 2001) for streamflow estimation at Besham Qila, utilizing SCA of three subbasins of UIB and their respective discharges as model inputs. Afzal et al. (2014) and Khan et al. (2015) found that these three subbasins, namely, Astore, Gilgit and Bunji are mainly snow fed basins. The comparative assessment of models developed with and without satellite-derived SCA has also been made with the help of performance indicators to show the importance of data-fusion.

Study area and datasets

This study considered three subbasins of UIB located in northern part of Pakistan, namely Gilgit, Astore and Bunji. Both, Astore and Gilgit rivers are tributaries of the Indus River, while Bunji basin is directly draining into the River Indus from north to west and into River Astore from eastern side. Astore River originating from Burzil Pass runs

through Astore valley, drains the Deosai Plateau and joins River Gilgit at 35° 32′ N, 74° 42′ E. The boundary of Gilgit basin is defined by location of Alam Bridge Gilgit (74° 18′ E; 35° 55′ N), where a stream gauge is installed by Water and Power Development Authority (WAPDA). The Gilgit River joins the River Indus at Juglot near Bunji where three (03) mountain ranges, Hindu Kush, Karakorum and Himalaya (HKH) meet. The total catchment areas for Astore and Gilgit are 4040 km^2 and 12,095 km^2, respectively. The peaks in Astore catchment are even higher than 8000 m.a.s.l. and remained covered with snow, especially during winter season (Naeem et al. 2012). The upper part of Gilgit basin also has persistent snow cover (Bashir and Rasul 2003). The snowmelt response of these three basins is observed at Bisham Qilla (72° 52′ E, 34° 55′ N) located, upstream of Tarbela Reservoir, at 580 m.a.s.l. (Fig. 1).

The input variables for output (discharge at Bisham Qila) are considered as weekly flow observations (Q) at Astore, Gilgit, Bunji and Bisham Qila gauging stations and SCA of Astore, Gilgit and Bunji basins (Table 1). Streamflow data are collected from WAPDA, whereas snow cover areas for these basins have been extracted from MODIS snow cover products in ArcGIS environment. This snow cover product was previously used by researchers for extraction of SCA such as; Tekeli et al. (2016), Tahir et al. (2016), Jamal et al. (2018), Hayat et al. (2019).

The whole data is transformed using normalization (on a scale of 0 to 1) to get a unit-less data set (Eq. 1).

$$\text{Normalized } X_i = (X_i - X_{\min}) / (X_{\max} - X_{\min}) \qquad (1)$$

where X_i is the ith value of input variable "X", $X_{\min}$ is the minimum value of input variable X, $X_{\max}$ is the maximum value of input variable X.

Determination of best input combination using gamma test and selection of data length through M-test

In order to develop smooth and reliable models for a given set of inputs, it is often necessary to find the unique combination of inputs that provide minimum noise and variance while predicting the desired output. WinGamma assumes those models smooth in which outputs can be determined smoothly from the inputs with only limiting factor of noise in the data. In the present study, the combination selection procedure for input variables is carried out using a mathematical tool, Gamma Test. The Gamma Test (GT) provides us the value of Gamma Statistics or best Mean Square Error (MSE) which is the measure of variance of noise on our desired output (Remesan et al. 2009). On the basis of an initial data set $\{(x_i, y_i), 1 \le i \le M\}$, an algorithm can be developed between two variables x and y to
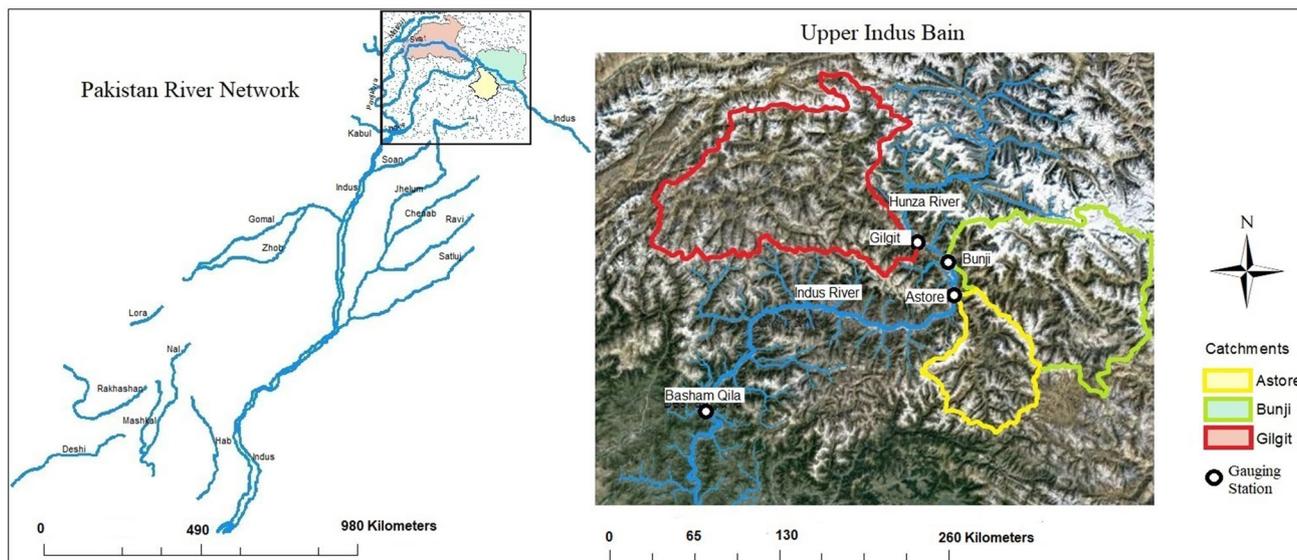
Fig. 1 Study area (UIB): catchments delineation using digital elevation model (DEM)

Table 1 Dataset used for model development

Sr. no.	Station	River	Elevation (m.a.s.l.)	Location	Data sets	Data length (2003–2010) (weeks)
1	Astore	Astore	2394	74° 42′ E, 35° 32′ N	SCA, Q	368
2	Gilgit	Gilgit	1460	74° 18′ E, 35° 55′ N	SCA, Q	368
3	Bunji	Indus	1372	74° 36′ E, 35° 42′ N	SCA, Q	368
4	Bisham Qilla	Indus	480	72° 52′ E, 34° 55′ N	Q	368

capture a relationship between them. The process involves decomposition of these variables into smooth and noisy parts while having an assumption that y is a function of x. If f is a smooth function between x and y and r is the part of noise which cannot be considered for by any even model, then their relationship can be shown by Eq. 2.

$$y = f(x) + r \qquad (2)$$

If the mean of this random variable "r" is zero then a constant bias can be engaged into this unknown function f. Despite the fact that f is unknown, this tool enables us to calculate the value of noise on an output on a certain condition which states that "As the number of data samples increase, the gamma value becomes equal to an asymptotic value which represents the variance of a noise on an output" (Stefánsson et al. 1997). In order to give stability to this asymptotic value, the gamma statistics should be determined for increasing number of data points (M). This process, performed through an algorithmic program to check the convergence of an infinite number of functions, is called the M-test. This test helps in the selection

of the most suitable data length for model training which ultimately helps in providing the best goodness of fit in the ANN models (Hassan et al. 2014; Hassan et al. 2015; Shamim et al. 2016).

In addition to achieve the best input combination, the Gamma value could also be used as a stopping criteria to reduce over fitting problems in ANN modeling. As compared to the conventional approaches used for the generalization, the Gamma test is superior in the context that the noise present in the data is already known and could be used to access the model performance, prior to the model building (Durrant 2001). Thus, reducing the need of separate validation data-set which is usually required for conventional early stopping methods. Elshorbagy et al. (2010), used Gamma Test as an assistive tool to select the appropriate modeling option for hydrological predictions among the various nonlinear modeling techniques.

In present study, GT is applied in WinGamma environment and a model identification tool is used to make number of combinations for a given set of input variables. Model Identification tool provides help in performing GT through a set of algorithms for a given set of inputs. The data length

for model training has been optimized using M-test. For this purpose, the gamma value is calculated for increasing number of inputs and the value for which standard error becomes stable against unique data points (inputs), is considered as the data length for model training.

Artificial neural networking (ANN)

The idea of ANN is inspired by the biological neuron system which has millions of neurons that are interconnected with each other. These neurons carry signals to our brain that acts as a processing unit and gives feedback on upcoming signals. ANN has a similar network of interconnected nodes that work in a same fashion. Each node acts as an artificial neuron and carries some input signals (Shamim et al. 2016). An arrow denotes a connection from the output of one node to the input of the other (Fig. 2). These connecting links receive the input signals and multiply them with the corresponding weights. Signals are then transferred, based upon the type of transfer function opted for a particular type of ANN model. A typical ANN model must have a minimum of three layers, including an input layer, intermediate layer and an output layer. Intermediate layer which is also called as a hidden layer could be one or more depending upon the structure of an ANN model (Remesan et al. 2008). A conceptual framework of ANN with two hidden layers is shown in Fig. 2.

ANN type models could be trained through two types of learning approaches which are supervised and unsupervised learnings. The unsupervised learning, also known as self-guided approach, works in the absence of output and is used to find the buried patterns in the data. On the other hand, supervised learning approach is used when a specific output/target is present. Both inputs and outputs are fed into the network and weights are adjusted to reduce the difference between inputs and outputs. Selection of the training algorithm is very important while developing an ANN type
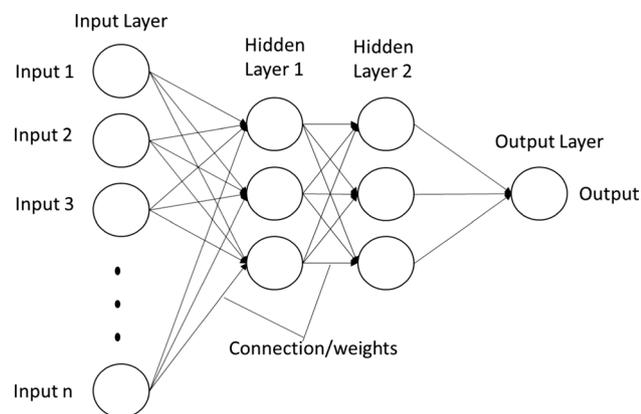
model because each training algorithm has its own specific properties. Back propagation algorithms are quite popular but are slow as their stable learning requires small learning rates. However, Levenberg–Marquardt (LM), conjugate gradient (CG) and quasi-Newton have pretty fast processing as they use the standard optimization techniques (Hassan et al. 2015b). In present study, a feed forward neural network with two hidden layers trained via Broyden Fletcher Goldfarb Shanno (BFGS) algorithm is used for model development process. Normally, one layer is used to represent the functions which are linearly separable. For complex problems, multi layers are preferred as empirically, deep learning seems to result in better generalization for a wide variety of tasks (Goodfellow et al. 2016). As the behavior of natural streams is very complex and the mountainous catchments (like Upper Indus Basin) often observe contrasting regimes. For the better understanding of this complex phenomenon, a MLP neural network is used with 2-hidden layers. The reason behind using two hidden layers in a neural network is due to their ability of solving nonlinear problems as reported by Minsky and Papert (1969) and Jones (2004). Previously the 2-layer BFGS is successfully used to train ANN models in the hydrological estimation e.g., (Ahmed et al. 2018; Shamim et al. 2016; Hassan et al. 2014, 2015). The optimum number of nodes in hidden layers of a multilayer ANN model may vary from problem to problem due to the difference in complexity level, number of inputs and outputs. The less number of nodes may create under fitting problem while too many neurons may cause over-fitting and takes more time to train a model. It is suggested by Brownlee (2018) that the best practice of finding the optimum number of nodes is to experiment them for the given set of data. In this study, different number of nodes in both the hidden layers is tried and value of Mean Square Error (MSE) and Correlation Coefficient (R^2) is computed for every set of nodes in training and testing phases (Table 2). It is quite clear from Table 2 that there is no significant difference in the values of MSE and R^2 values for different combination of nodes in hidden layers. However, a minute improvement in the R^2 values is noted in testing phase with same number of nodes in both the hidden layers.

The models are further evaluated with the help of statistical indices including Nash–Sutcliffe coefficient (NSE), Variance (VAR) and BIAS.

Results and discussion

Gamma test

The Gamma test is performed on a set of six (6) input variables (SCA of Astore, Gilgit and Bunji and Q of Astore, Gilgit and Bunji) for output which is Q at Besham Qila. A

Fig. 2 Framework for ANN with 2 hidden layers

Table 2 Selection of nodes in hidden layers on the basis of MSE and R^2

Sr. no.	Nodes in 1st layer	Nodes in 2nd layer	Combination (111111), with SCA Target MSE 0.000536				Combination (010101), without SCA Target MSE 0.001118			
			MSE (train) achieved	MSE (test) achieved	R^2 (train) %	R^2 (test) %	MSE (train) achieved	MSE (test) achieved	R^2 (train) %	R^2 (test) %
1	1	1	0.000612	0.00325	98.9	92.8	0.00104	0.00379	98.1	92.0
2	1	2	0.000534	0.00341	99.0	92.4	0.00101	0.00418	98.1	90.9
3	1	5	0.000526	0.00342	99.0	93.5	0.00108	0.00290	98.0	93.5
4	1	6	0.000533	0.00332	99.0	92.7	0.00108	0.00254	98.9	94.2
5	1	7	0.000535	0.00342	99.0	92.5	0.00109	0.00382	98.1	91.4
6	2	1	0.000532	0.00268	99.0	93.8	0.00101	0.00372	98.1	91.9
7	2	2	0.000531	0.00266	99.0	93.9	0.00099	0.00293	98.2	93.1
8	3	3	0.000534	0.00292	99.0	93.4	0.00109	0.00369	97.9	92.0
9	3	5	0.000530	0.00337	99.5	92.4	0.00092	0.00426	98.3	90.2
10	3	6	0.000525	0.00324	99.0	93.2	0.00108	0.00468	98.0	89.2
11	4	4	0.000529	0.00267	99.0	93.8	0.00078	0.00349	98.5	92.3
12	4	7	0.000531	0.00343	99.0	92.2	0.00109	0.00293	98.0	93.4
13	5	1	0.000527	0.00271	99.0	93.7	0.00112	0.00319	97.9	92.7
14	5	3	0.000536	0.00265	99.0	93.8	0.00072	0.00311	98.6	93.0
15	5	5	0.000536	0.00282	99.1	93.5	0.00103	0.00457	98.1	90.1
16	5	7	0.000529	0.00305	99.0	93.2	0.00082	0.00395	98.4	90.8
17	6	1	0.000522	0.00339	99.0	92.3	0.00103	0.00419	98.1	90.8
18	6	3	0.000530	0.00281	99.0	93.6	0.00086	0.00284	98.4	93.4
19	6	6	0.000525	0.00276	99.0	93.9	0.00077	0.00339	98.5	92.2
20	7	1	0.000512	0.00347	99.1	92.2	0.00111	0.00411	97.9	90.3
21	7	2	0.000517	0.00289	99.1	93.3	0.00103	0.00275	98.1	93.3
22	7	7	0.000534	0.00305	99.0	92.9	0.00107	0.00299	98.0	93.5

total of 63 realistic combinations are made using a model identification tool and Gamma value is computed for each (Fig. 3), which is later on considered as the targeted MSE value for model training process (Table 2).

It is found that input mask (111111) gives minimum gamma value/targeted MSE that is 0.000536. "1" means that a particular input is included and "0" means a particular input is excluded. All 1's in this combination showed that gamma value is minimum with all inputs included. So, the combination containing all input variables with minimum targeted MSE is further used for model development process. The combination that includes only on-ground observations (010101) is also used to develop models in order to compare the performance of estimation models without considering the satellite-derived SCA as a possible predictor. In order to find the suitable data length for model training, which optimize the model performance in terms of goodness of fit, the M-test has been performed on increasing number of inputs for both the input combinations 010101 (Fig. 4a) and 111111 (Fig. 4b).
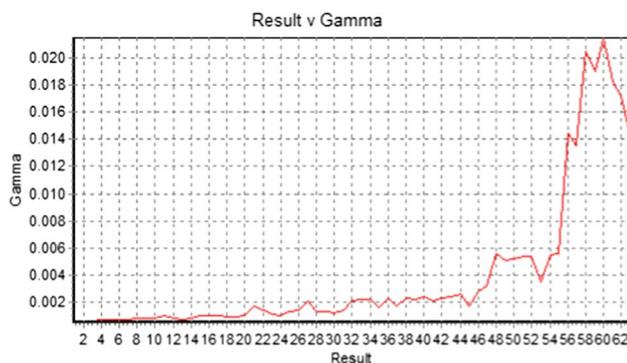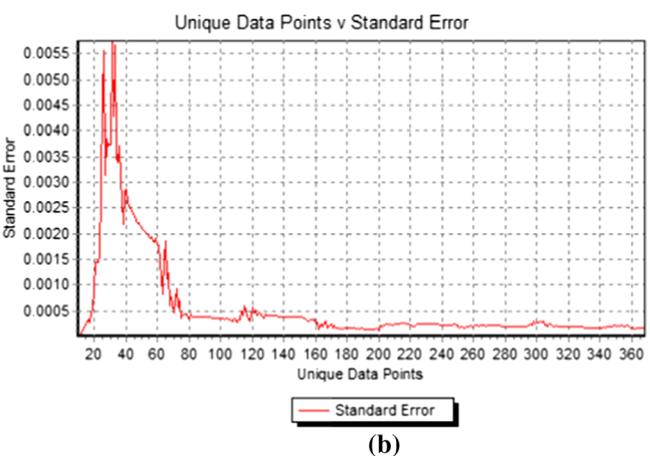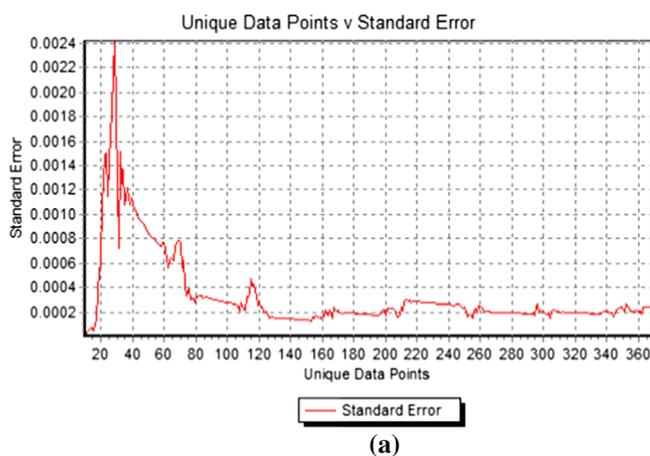
It is clear from Fig. 4a, b, that the standard error line becomes almost stable after 165 unique data points. So, the data length for model training is considered as 45% (165 weeks) and the testing/validation as 55% (203 weeks).

ANN modeling results

The number of hidden layers for all ANN models is fixed as two (2) and a variety of combinations of nodes in these layers are tried (Table 2). Since, the change in number of nodes in each layer doesn't have any significant impact upon the MSE and the correlation coefficient values for the developed models, and therefore, the models are developed for all the selected combination of nodes as presented in Table 2, tested and evaluated using performance indicators; NSE, RMSE and VAR (Table 3). Each model with a specific network architecture is trained against the output discharge at Besham Qila for both the combinations, the one (111111) obtained through the Gamma test and the other (010101) comprised of only gauge-discharges (Table 3).

It is clear from Table 3 that the models developed with integrated data-set (111111), performed better with the average values of NSE = 99.5/97.5 (training/testing), BIAS = − 0.01/− 6.6, RMSE = 251.4/532.3 and VAR = 63,218.0/286,917.1, as compared to the models developed without SCA in the input variables (010101) with average values of NSE = 99.1/97.1 (training/testing), BIAS = 14.6/− 26.1, RMSE = 327.6/531.4 and VAR = 106,390.6/284,363.4. Again, the changing number of nodes in the hidden layers does not produce any significant difference in the performance of models with almost similar values of performance indicators, especially for NSE and RMSE in the case of training and testing of models. However, the moderate difference in values for BIAS and VAR is observed during both the training and testing phases of the developed models. BIAS is a systematic error that represents

Fig. 3 Gamma values versus input combinations/masks

Fig. 4 **a** M-test for combination 010101, **b** M-test for combination 111111

Table 3 ANN modeling results for both set of input-combinations

No.	1st layer Nodes	2nd layer Nodes	Combination (111111), with SCA $\left(\frac{Training}{Testing}\right)$ NSE (%)	BIAS	RMSE	VAR	Combination (010101), without SCA $\left(\frac{Training}{Testing}\right)$ NSE (%)	BIAS	RMSE	VAR
1	1	1	99.5	0.6	255.4	65238.9	99.1	11.9	326.8	106688.9
			97.6	−36	695.3	482147.5	97.1	−61.7	560.1	309901.8
2	1	2	99.5	−4.27	253.1	64036.1	99.1	−4.2	341.8	116782.4
			97.6	−47.6	512.9	260837.4	98.1	−20.8	482.2	232050.1
3	1	5	99.8	−1.30	255.3	65176.9	98.9	0.9	369.2	136321.4
			97.7	−50.6	497.3	244770.6	97.4	33.4	521.4	270794.2
4	1	6	99.7	−0.25	212.9	45351.0	99.0	3.2	363.5	132159.3
			97.3	−55.9	742.5	548130.0	97.6	−20.6	505.4	255111.8
5	1	7	99.5	3.73	255.6	65333.1	99.2	62.5	346.3	116029.7
			97.5	−30.95	517.0	266336.0	94.4	−19.9	608.4	369781.7
6	2	1	99.5	−5.2	256.0	65534.5	99.0	−1.3	356.5	127111.3
			97.9	12.8	475.2	225674.4	97.5	−56.1	520.6	267842.0
7	2	2	99.5	1.6	252.4	63709.6	99.3	4.1	298.3	88941.2
			97.8	15.0	486.0	236002.2	96.9	−64.4	572.0	322997.1
8	3	3	99.5	−10.1	250.4	65588.1	98.9	4.0	369.1	136214.0
			97.7	18.8	500.9	250630.4	96.8	−47.8	345.4	116985.3
9	3	5	99.5	13.0	248.1	61385.9	99.5	7.6	266.5	70982.9
			97.6	5.8	511.9	262074.0	97.8	−32.8	493.4	242343.7
10	3	6	99.6	17.3	243.2	58825.1	98.9	2.5	370.0	136894.6
			96.7	10.7	500.9	250770.3	96.8	−24.3	589.8	347303.9
11	4	4	99.5	−0.9	251.1	63062.9	99.4	19.6	294.0	86056.1
			97.8	9.9	480.9	231196.7	97.5	11.2	524.2	275212.2
12	4	7	99.5	5.4	255.7	65344.6	99.4	−14.1	272.8	74220.3
			97.3	−32.4	536.6	286866.0	96.5	−33.2	512.2	261277.3
13	5	1	99.5	−15.8	254.9	64746.6	99.0	−4.9	366.2	134055.3
			97.9	9.7	474.9	225483.8	96.8	−65.2	587.1	340488.5
14	5	3	99.5	−6.2	255.3	65145.9	99.0	21.3	455.9	130288.8
			97.7	16.9	497.1	246841.9	96.6	−70.9	611.1	368447.3
15	5	5	99.5	−11.6	253.2	63994.4	99.5	3.8	262.2	68708.7
			97.2	12.7	550.2	302642.8	97.6	−22.9	505.7	255202.3
16	5	7	99.5	30.3	253.6	63384.1	99.0	60.8	363.8	128674.9
			97.7	52.6	494.0	241284.2	96.7	−27.1	595.1	353378.3
17	6	1	99.5	−14.1	251.4	63008.9	99.0	63.0	367.7	131252.5
			97.2	−8.6	550.4	302936.3	96.9	−1.9	577.7	336998.6
18	6	3	99.5	3.9	256.1	65569.7	99.3	29.4	290.0	83281.2
			97.4	−14.6	530.0	280715.3	97.4	−24.0	529.6	279876.4
19	6	6	99.5	22.3	255.2	64646.9	99.4	2.0	266.9	71221.6
			97.6	14.7	506.7	256555.3	97.5	3.2	520.9	271285.9
20	7	1	99.5	−25.0	250.3	62046.4	99.0	28.2	364.6	132152.4
			96.9	−13.4	572.4	327851.5	97.5	−15.4	510.7	260532.5
21	7	2	99.5	19.6	256.3	65297.5	99.6	9.7	236.0	55593.5
			96.9	0.24	583.5	340469.2	97.8	−10.9	485.9	235948.3
22	7	7	99.5	−23.3	254.8	64370.3	99.4	11.6	277.7	76962.2
			97.8	−34.5	493.1	241960.3	97.4	−1.7	531.2	282236.4
	Average errors		99.5	−0.01	251.4	63218.0	99.1	14.6	327.6	106390.6
			97.5	−6.6	532.3	286917.1	97.1	−26.1	531.4	284363.4

the difference in values of predicted and actual mean. Positive BIAS means that the mean of predicted discharge is more than the mean of actual or observed discharge. In the training phase of models developed with 010101 (input combination), the value of BIAS in most of the cases is positive, depicting that the trained models are overpredicting; whereas, the converse trend of negative BIAS has been observed in most of the cases of testing/validation phase, showing that the models are predicting less than the actual values. To make the ANN models predict more or less, often depends upon the careful selection of network-architecture as the complex architecture tends to predict more due to over-fitting, while the lighter network tends to predict less due to under-fitting. This is why, this study considered a number of node-combination options while training ANN models through BFGS algorithm to find the best possible option with minimum uncertainty. For combination 111111, the minimum value of BIAS in the training phase is 0.6 for Model No.1 with a node combination of 1–1. But the same model showed high values of BIAS (−36.0) and

RMSE (695.3) in the testing phase. The model No. 7 with a node combination of 2–2 and the model No. 11 with a node combination 4–4 performed well with low values of BIAS (1.6/15) & (−0.9/9.9), RMSE (252.4/486.0) & (251.1/480.9) and VAR (63,709.6/236,002.2) & (63,062.9/231,196.7), respectively, as compared to other models. Figure. 5a, b shows the training and testing models developed with a node combination of 4–4.

Although, some of the models developed without SCA also performed reasonably well, e.g., the model No. 21 with a node combination of 7-2 and the model No. 19 with a node combination of 6–6 with BIAS (9.7/− 10.9) & (2.0/3.2), RMSE (236.0/485.9) & (266.9/520.9) and VAR (55,593.5/235,948.3) & (71,221.6/271,285.9), but overall trend of results suggests that the models developed with integrated data have the better tendency to perform well with significant high values of NSE. The initial data-fusion through Gamma test has also provided the same that the inputs with a mask 111111 (SCA of Gilgit, Q at Gilgit, SCA of Bunji, Q at Bunji, SCA at Astore, Q at Astore) will
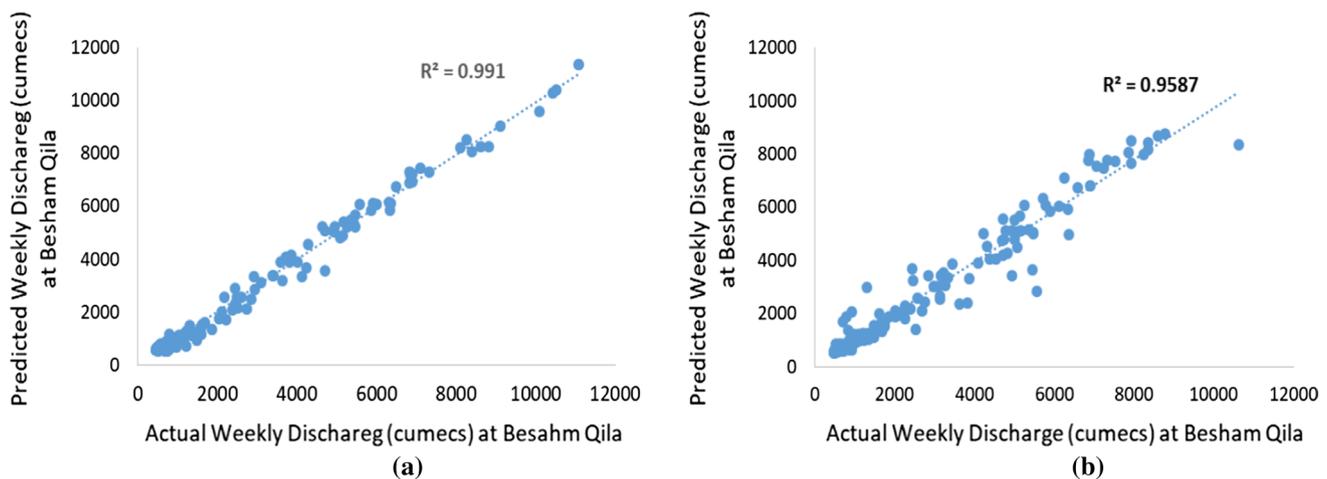
Fig. 5 **a** Training model (node 4–4), **b** testing model (node 4–4)

produce better models with least MSE. The results presented in Table 3 indicate that the discharge estimation models for Besham Qila, trained via two layer BFGS algorithm, performed better with the multi-source information as compared to the single source information. The results clearly showed the dependency of streamflow of Upper Indus Basin on the upland condition of snow cover area (SCA) and evidenced the importance of satellite-derived SCA for a complex terrain of UIB. This study evidenced that the fused data-set comprising of SCA and gauge observations represents the watershed response better than the data-set only consisting of on-ground observations. The present work also demonstrated the importance of conjunctive use of the Gamma Test and Artificial Neural Networking (ANN) approaches and the ability of ANN to perform well in the development of runoff models for mountainous catchments.

Summary and conclusions

The study is carried out to improve the real-time streamflow estimation for a complex terrain of UIB where on-ground observations are limited. Since, the most part of the watershed derives their flow from snow melt, so the satellite-derived SCA of the region could be used as a crucial input variable. In this paper, a case study of UIB is considered to improve the streamflow estimation models at Besham Qila through a fused data set, comprising of on-ground flow observations and satellite-derived SCA of three (03) subbasins of UIB (Astore, Gilgit and Bunji). The fusion process is carried out with the help of a novel mathematical tool, Gamma test, which provided the best combination 111111 (SCA of Gilgit, Q at Gilgit, SCA

of Bunji, Q at Bunji, SCA at Astore, Q at Astore), with least value of MSE (0.000536). The feed-forward ANN models are trained via two layer BFGS algorithm with a variety of node combinations. The data length for training is optimized with the help of M-test in WinGamma environment and the least value of MSE, as determined by Gamma test, is utilized as an early stopping criteria to avoid over-fitting in ANN models. In this case, the best data length for training and testing of models comes out as 45% (165 weeks) and 55% (203 weeks), respectively. The streamflow estimation models are also developed for input combination 010101, which only contains on-ground flow observations. The both type of models are compared on the basis of NSE, BIAS, RMSE and VAR. The results indicate that the models developed with integrated dataset (combination: 111111) performed better with significant high values of NSE and low values of other statistical errors; including BIAS, RMSE and VAR. The findings of this study suggest that satellite-derived SCA could be used as a possible predictor for streamflow estimation in the region. Moreover, it is suggested that the uncertainty in the hydrological estimation models could be reduced by knowing more about the watershed. So, the multi-source/type information like climate and meteorological observations (e.g., temperature, solar radiations and rainfall, etc.), more satellite observations (e.g., gridded precipitation and LST) and multi-sensors data (different satellite products and airborne data) could be used and fused with the traditional observations to improve the real-time flow forecasts.

Compliance with ethical standards

References

Abrahart RJ, See L (2002) Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. Hydrol Earth Syst Sci 6(4):655–670

Afzal MT, Arslan M, Zafar S, Waqar MM (2014) Satellite derived snow cover status and trends in the Indus Basin. J Sp Technol 4:26–31

Ahmed F, Hassan M, Nisar H (2018) Developing nonlinear models for sediment load estimation in an irrigation canal. Acta Geophys. https://doi.org/10.1007/s11600-018-0221-3

Andermann C (2011) Evaluation of precipitation data sets along the Himalayan front. Bull Am Meteorol Soc. https://doi.org/10.1029/2011GC003513

Archer DR, Fowler HJ (2004) Spatial and temporal variations in precipitation in the Upper Indus Basin, global teleconnections and hydrological implications. Hydrol Earth Syst Sci 8:47–61

Arfan M, Lund J, Hassan D, Saleem M, Ahmad A et al (2019) Assessment of spatial and temporal flow variability of the Indus River. Resources 8:103

Ata M, Tekeli AE, Dönmez S, Fouli H (2016) Use of interactive multisensor snow and ice mapping system snow cover maps (IMS) and artificial neural networks for simulating river discharges in Eastern Turkey. Arab J Geosci. https://doi.org/10.1007/s12517-015-2074-2

Azmi M, Araghinejad S, Kholghi M (2010) Multi model data fusion for hydrological forecasting using K-nearest neighbour method. Iran J Sci Technol 34:81–92

Bashir F, Rasul G (2003) Estimation of water discharge from gilgit basin using remote sensing. GIS Runoff Model Introd 6:97–113

Behrangi A, Khakbaz B, Chun T, Aghakouchak A, Hsu K (2011) Hydrologic evaluation of satellite precipitation products over a mid-size basin. J Hydrol 397:225–237. https://doi.org/10.1016/j.jhydrol.2010.11.043

Brownlee J (2018) Deep learning performance. Machine Learning Mastery

Broxton PD, Zeng X, Sulla-Menashe D, Troch PA (2014) A global land cover climatology using MODIS data. J Appl Meteorol Climatol. https://doi.org/10.1175/JAMC-D-13-0270.1

Brubaker KL, Pinker RT, Deviatova E et al (2005) Evaluation and comparison of MODIS and IMS snow-cover estimates for the continental united states using station data. J Hydrometeorol 6:1002–1017

Durrant PJ (2001) A non-linear data analysis and modeling tool with applications to flood prediction. Ph.D. Thesis, Department of Computer Science, Cardiff University, University of Wales, UK

Elshorbagy A, Corzo G, Srinivasulu S, Solomatine DP (2010) Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology: Part 2. Hydrol Earth Syst Sci 14:1943–1961. https://doi.org/10.5194/hess-14-1943-2010

Ercan MB, Asce M, Goodall JL, Asce AM (2013) Estimating watershed-scale precipitation by combining gauge- and radar-derived observations, pp 983–994. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000687

Forsythe N, Kilsby CG, Fowler HJ, Archer DR, Forsythe N, Kilsby CG, Fowler HJ, Archer DR (2011) Variability and potential change using MODIS satellite data products assessment of runoff sensitivity in the Upper Indus Basin to interannual climate variability and potential change using MODIS satellite data products. BioOne 32:16–29

Fowler HJ, Archer DR (2005) Hydro-climatological variability in the Upper Indus Basin and implications for water resources. Reg Hydrol Impacts Clim Change Impact Assess Decis Mak 295:131–138

Giustolisi O (2004) Sparse solution in training artificial neural networks. NeuroComputing 56C:284–304

Gökbulak F, Kamil Ş, Serengil Y (2015) Comparison of rainfall-runoff relationship modeling using different methods in a forested watershed. Water Resour Manag. https://doi.org/10.1007/s11269-015-1056-1

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge

Hassan M, Shamim MA, Hashmi HN, Ashiq SZ, Ahmed I, Pasha GA, Naeem UA (2014) Predicting streamflows to a multipurpose reservoir using artificial neural networks and regression techniques. Earth Sci Inform. https://doi.org/10.1007/s12145-014-0161-7

Hassan M, Shamim MA, Sikandar A, Mehmood I, Ahmed I, Ashiq SZ, Khitab A (2015) Development of sediment load estimation models by using artificial neural networking techniques. Environ Monit Assess. https://doi.org/10.1007/s10661-015-4866-y

Hasson ul S, Böhner J, Lucarini V et al (2017) Prevailing climatic trends and runoff response from Hindukush–Karakoram–Himalaya, upper Indus Basin. Earth Syst Dyn 8:337–355. https://doi.org/10.5194/esd-8-337-2017

Hasson S, Lucarini V, Khan, Petitta M, Bolch T, Gioli G (2014) Early 21st century snow cover state over the western river basins of the Indus River system. Hydrol Earth Syst Sci. https://doi.org/10.5194/hess-18-4077-2014

Hayat H, Akbar TA, Tahir AA, Hassan QK, Dewan A, Irshad M (2019) Simulating current and future river-flows in the snowmelt-runoff model and RCP scenarios. Water 11:1–19

Immerzeel WW, Droogers P, De Jong SM, Bierkens MFP et al (2009) Remote Sensing of Environment Large-scale monitoring of snow cover and runoff simulation in Himalayan river basins using remote sensing. Remote Sens Environ 113:40–49. https://doi.org/10.1016/j.rse.2008.08.010

Jamal K, Ahmad S, Li X, Rizwan M, Li H, Feng J (2018) Climate change and runoff contribution by hydrological zones of cryosphere catchment of Indus River, Pakistan. Hydrol Earth Syst Sci Discuss. https://doi.org/10.5194/hess-2018-548

Jones AJ (2004) New tools in non-linear modelling and prediction. Comput Manage Sci 1(2):109–149. https://doi.org/10.1007/s10287-003-0006-1

Khan A, Naz BS, Bowling LC (2015) Separating snow, clean and debris covered ice in the Upper Indus Basin, Hindukush–Karakoram–Himalayas, using Landsat images between 1998 and 2002. J Hydrol 521:46–64. https://doi.org/10.1016/j.jhydrol.2014.11.048

Kisi O, Sanikhani H (2015) Prediction of long-term monthly precipitation using several. Int J Climatol 35:4139–4150. https://doi.org/10.1002/joc.4273

Lang K, Waibel A, Hinton G (1990) A time-delay neural network architecture for isolated word recognition. Neural Netw 3:23–43

Le Cun Y, Denker J S, Solla S A (1990) Optimal brain damage. In: Touretzky DS (ed) Advances in neural information processing systems, vol 2, Morgan Kaufmann, San Mateo, California, pp 598–605

Lee G, Kim D, Kwon H, Choi E (2019) Estimation of maximum daily fresh snow accumulation using an artificial neural network model. Adv Meteorol. https://doi.org/10.1155/2019/2709351

Liu X, Chen B (2000) Climatic warming in the Tibetan Plateau during recent decades. Int J Climatol 1742:1729–1742

Lutz AF, Immerzeel WW, Kraaijenbrink PDA, Shrestha AB, Bierkens MFP (2016) Climate change impacts on the Upper Indus hydrology: sources, shifts and extremes. PLoS ONE. https://doi.org/10.1371/journal.pone.0165630

Minsky M, Papert S (1969) Perceptrons. MIT Press, Cambridge, MA

Mishra B, Tripathi NK, Babel MS (2014) An artificial neural network-based snow cover predictive modeling in the higher Himalayas An Artificial Neural Network-Based Snow Cover Predictive

Modeling in the Higher Himalayas. J Mt Sci 11:825–837. https://doi.org/10.1007/s11629-014-2985-5

Mukhopadhyay B, Dutta A (2010) A stream water availability model of Upper Indus Basin based on a topologic model and global climatic datasets. Water Resour Manag. https://doi.org/10.1007/s11269-010-9666-0

Naeem UA, Hashmi HN, Shamim MA, Ejaz N (2012) Flow variation in astore river under assumed glaciated extents due to climate change. Pak J Eng Appl Sci 11:73–81

Naeem U, Mughal H-U-R, Ghumman AR, Shamim MA (2014) Ranking sensitive calibrating parameters of UBC watershed model ranking sensitive calibrating parameters of UBC watershed model. KSCE J Civ Eng. https://doi.org/10.1007/s12205-013-0515-6

Nagarajan K, Krekeler C, Member S, Slatton KC, Member S, Graham WD (2010) A scalable approach to fusing spatiotemporal data to estimate streamflow via a bayesian network. IEEE Trans Geosci Remote Sens 48:3720–3732

Piotrowski AP, Napiorkowski JJ (2013) A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. J Hydrol 476:97–111. https://doi.org/10.1016/j.jhydrol.2012.10.019

Prasad AK, Yang KS, Kafatos M (2009) Melting of major Glaciers in the western Himalayas: evidence of climatic changes from long term MSU derived tropospheric temperature trend (1979–2008). Annales Geophysicae 27:4505–4519

Quiroga VM, Mano A, Kure S, Udo K (2013) Snow glacier melt estimation in tropical Andean glaciers using artificial neural networks. Hydrol Earth Syst Sci. https://doi.org/10.5194/hess-17-1265-2013

Rastner P, Prinz R, Notarnicola C, Nicholson L, Sailer R, Schwaizer G, Paul F (2019) On the automated mapping of snow cover on glaciers and calculation of snow line altitudes from multi-temporal landsat data. Remote Sens 12:1–24

Razavi T, Coulibaly P (2016) Improving streamflow estimation in ungauged basins using multi-modelling approach. Hydrol Sci J. https://doi.org/10.1080/02626667.2016.1154558

Remesan R, Shamim MA, Han D (2008) Model data selection using gamma test for daily solar radiation estimation. Hydrol Process 2274:2267–2274. https://doi.org/10.1002/hyp

Remesan R, Ali M, Han D, Mathew J (2009) Runoff prediction using an integrated hybrid modelling scheme. J Hydrol 372:48–60. https://doi.org/10.1016/j.jhydrol.2009.03.034

Sakai A, Nuimura T, Fujita K, Takenaka S, Nagai H, Lamsal D (2015) Climate regime of Asian glaciers revealed by GAMDAM glacier inventory. Cryosph. https://doi.org/10.5194/tc-9-865-2015

See L, Abrahart RJ (2001) Multi-model data fusion for hydrological forecasting. Comput Geosci 27:987–994

Shamim MA, Hassan M, Ahmad S, Zeeshan M (2016) A comparison of artificial neural networks (ANN) and local linear regression (LLR) techniques for predicting monthly reservoir levels. KSCE J Civ Eng 20:971–977

Sönmez I, Emre AT, Erdi E (2014) Snow cover trend analysis using Interactive Multisensor Snow and Ice Mapping System data over Turkey. Int J Climatol 2361:2349–2361. https://doi.org/10.1002/joc.3843

Srinivasulu S, Jain A (2006) A comparative analysis of training methods for artificial neural network rainfall: runoff models. Appl Soft Comput 6:295–306. https://doi.org/10.1016/j.asoc.2005.02.002

Stefánsson A, Končar N, Jones AJ (1997) A note on the Gamma test. Neural Comput Appl 5(3):131–133

Sun L, Anderson MC, Gao F, Hain C, Alfieri JG, Sharifi A, Mccarty GW, Yang Y, Yang Y, Kustas WP, Mckee L (2017) Investigating water use over the Choptank River Watershed using a multi-satellite data fusion approach. Water Resour Res. https://doi.org/10.1002/2017WR020700

Tahir AA, Adamowski JF, Chevallier P, Haq AU, Terzago S (2016) Comparative assessment of spatiotemporal snow cover changes and hydrological behavior of the Gilgit, Astore and Hunza River basins (Hindukush–Karakoram–Himalaya region. Meteorol Atmos Phys, Pakistan). https://doi.org/10.1007/s00703-016-0440-6

Tekeli AE, Sönmez I, Erdi E (2016) Snow-covered area determination based on satellite-derived probabilistic snow cover maps. Arab J Geosci. https://doi.org/10.1007/s12517-015-2149-0

Tikhonov AN (1963) Solution of incorrectly formulated problems and the regularization method. Doklady Akademii Nauk SSSR 151:501–504

Uysal G, Arda A (2016) Streamflow forecasting using different neural network models with satellite data for a snow dominated region in Turkey. Procedia Eng 154:1185–1192. https://doi.org/10.1016/j.proeng.2016.07.526

Uysal G, Ünal A (2018) Monthly streamflow estimation using wavelet-artificial neural network model: a case study on Çamlıdere dam basin, Turkey. Procedia Comput Sci. https://doi.org/10.1016/j.procs.2017.11.234

Van Dijk AIJM (2011) Model-data fusion: using observations to understand and reduce uncertainty in hydrological models. 19th international congress on modelling and simulation, Perth, Australia, 12–16 December 2011. http://mssanz.org.au/modsim2011/index.htm

Wang W, Xiang Y, Gao Y, Lu A, Yao T (2014) Rapid expansion of glacial lakes caused by climate and glacier retreat in the Central Himalayas. Hydrol Process. https://doi.org/10.1002/hyp.10199

Yilmaz AG, Imteaz MA, Jenkins G (2011) Catchment flow estimation using Artificial Neural Networks in the mountainous Euphrates Basin. J Hydrol 410:134–140. https://doi.org/10.1016/j.jhydrol.2011.09.031

Zhu X, Wu X (2004) Class noise versus attribute noise : a quantitative study of their impacts. Artif intell Rev 22:177–210

Utilization of coconut shell residual in green roof: hydraulic and thermal properties of expansive soil amended with biochar and fibre including theoretical model

Hao Wang[1,2,3] · Ankit Garg[1,2,3] · Xiaoyong Zhang[1,2,3] · Yangyang Xiao[1,2,3] · Guoxiong Mei[1,2,3]

Abstract

The study intends to explore hydraulic and thermal properties of expansive soils treated with fibre, biochar and biochar–fibre mix. Both fibre and biochar are derived from coconut shell, which is highly common in coastal regions around the world. Besides, benefits, limitations and engineering feasibility of these geomaterials in green roofs are explored. Theoretical framework for thermal–hydraulic analysis is proposed based on mass conservation and the first law of thermodynamics. Heat capacity, thermal conductivity, water retention curve, crack intensity factor (CIF) and saturated and unsaturated hydraulic conductivities of four kinds of soils are evaluated and compared. Characterizations of geomaterials are also investigated via thermal mass loss, micro-structure, surface area and functional groups identification. Both biochar and fibre admixtures contribute to improvement in soil heat capacity and saturated and unsaturated hydraulic conductivities. Biochar enhances saturated and residual water contents of expansive soil by 10% and 8%, respectively. Also, biochar decreases soil thermal conductivity and CIF by 31% and 5%, respectively, while fibre decreases soil-saturated and residual water contents by 15% and 29%, respectively, and reduces soil thermal conductivity and CIF by 21% and 50%, respectively. Soil–biochar–fibre composite is also recommended due to low air-entry value, acceptable water-holding capacity and limited crack propagation. The study fills the knowledge gap of how soil thermal–hydraulic properties are affected due to biochar and/or fibre admixture. It is recommended to pay more attention on production and utilization of biochar derived from coconut shell currently utilized for fibre extraction.

Keywords Biochar · Green roof · Coconut shell · Expansive soil · Fibre · Sponge city

Introduction

Due to accelerated urbanization and rising demand for land worldwide, reduction of greening areas in cities has become a non-negligible issue (Nguyen et al. 2019). There is more than 55% of global population living in urban areas nowadays, and the figure would exceed 68% by the middle of the

twenty-first century according to Department of Economic and Social Affairs of United Nations. Severe burdens have been imposed to the environment as a consequence of urbanization process, for instance, heat island effect (Aflaki et al. 2017; Chun and Guldmann 2018) and urban flood disasters (Chan et al. 2018; Rahman et al. 2019). Creating more green areas in cities by vegetated infrastructures (e.g. green roof and vertical garden) has proved to be an effective measure to tackle these environmental problems (Zhang et al. 2019). As compared to conventional roofs, green roofs delay rainwater runoff peaks and store water temporarily in soil substrate, thus achieving rainwater management (Walsh et al. 2016). Furthermore, green roofs are conducive to thermal insulation due to plant evaporation and relatively low thermal conductivity of soil. It prevents uncomfortable fluctuations of indoor temperature and reduces energy consumption made by air condition (Besir and Cuce 2018).

✉ Guoxiong Mei
meiguox@163.com

1　College of Civil Engineering and Architecture, Guangxi University, Nanning 530004, Guangxi, China

2　Key Laboratory of Disaster Prevention and Structural Safety of Ministry of Education, Guangxi University, Nanning 530004, Guangxi, China

3　Key Laboratory of Disaster Prevention and Mitigation and Engineering Safety of Guangxi, Nanning 530004, Guangxi, China

Local available soils and indigenous plants are commonly given priority in order to minimize costs (Besir and Cuce 2018; Nektarios et al. 2011) and avoid any potential threat to local ecosystem (e.g. eggs of invasive pest in soil). Also, applications of non-agricultural soils in vegetated infrastructures are critical in view of shortage of arable land currently. As a worldwide distributed soil (especially in Southeast Asia), expansive clay is unavoidable to be used as substrate soil in green roofs. Although swelling deformation of expansive soil is negligible under relatively low dry density (i.e. loose planting soil), undue acidity and insufficient fertility still hinder practical application of expansive soil in vegetated infrastructures (Kaufhold et al. 2015; Villar and Lloret 2008). However, addition of conventional fertilizer directly might be not economically acceptable because of nutrient loss. Controlled release fertilizers (CRFs) are required to ensure long-term service of green roofs. Furthermore, engineering properties of expansive soils are also not acceptable in green roofs, e.g. low hydraulic conductivity, relatively low water-holding capacity and desiccation cracks. Before large-scale utilization of expansive soil in green roofs and other vegetated infrastructures, suitable soil amendments are demanded. In addition to improving engineering behaviours mentioned above, desired soil amendments need to be durable, low cost and environmentally acceptable.

Biochar, the carbon-rich pyrolysis product of biomass under anoxic condition (Bordoloi et al. 2019a; Kumar et al. 2019; Ni et al. 2020) might be one of the promising options of expansive soil admixture in green roofs. Biochar has garnered attention nowadays as a sustainable soil admixture for environmental remediation and improving long-term agricultural productivity (Omondi et al. 2016). Both physical and chemical properties of biochars are mainly determined by biomass feedstock and pyrolysis conditions, which further influence soil amendment effects (Kumar et al. 2020; Omondi et al. 2016). There are noticeable discrepancies between plant-derived biochars (PDBs) and animal-derived biochars (ADBs), and their effects on soil properties (Bora et al. 2020). Because of unidirectionality of the food chain, toxic substances such as heavy metals would be accumulated in animal bodies. Therefore, ADBs usually contain more toxic substances, which might exert adverse influences on growth of expansive ornamental plants.

In the context of commercialization, biochar produced by crop straw is one of the most commercialized PDBs currently, due to its abundant raw materials and protecting the atmosphere from toxic gases produced by straw combustion. However, coconut shell biochar (CSB) possesses more advantages as compared to straw one, e.g. higher surface area, less volatiles and more fixed carbon (Windeatt et al. 2014). Southeast Asia and South Asia are main production

areas for coconuts, where expansive soils are also widely distributed. Production of coconut water beverage has finished the transportation and collection of coconut shells, ensuring extremely low cost of biochar production. Hence, the major cost of biochar derived from coconut shell is energy consumption of pyrolysis which is mainly related to pyrolysis temperature and time. On the other hand, most ornamental plants survive well in weak acid or neutral soil environment. The pH value of straw biochar (more than 11.5) is much higher than that of coconut shell one (about 8.5), which suggests that application amount of straw biochar should be limited in vegetated soils (Pituya et al. 2017; Windeatt et al. 2014). Without such limitation, application of CSB could be determined by engineering-oriented properties, for instance, thermal insulation, hydraulic properties and desiccation crack.

As another valued-added product of coconut shell residuals, coconut shell fibre (CSF) has been widely researched in soil reinforcement (Danso et al. 2015) and growing media for plants (Lopez-Rodriguez et al. 2016). Studies on application of CSF in vertical garden are reported in recent years (Lopez-Rodriguez et al. 2016), so CSF also seems a potential soil admixture in green roofs. Careful comparison between CSB and CSF is a task of top priority before large-scale production and application of CSB. To the best of authors' knowledge, however, investigations on comparison between fibre and homologous biochar are rarely reported till date. Actually, CSB might maximize the potential of coconut shell residuals economically, by being complementary product of CSF.

Studies have been conducted recently to examine effects of biochar derived from water hyacinth, poultry litter, saw dust and algae (Bordoloi et al. 2018a; Kumar et al. 2020; Ni et al. 2020) on soil hydraulic properties such as soil–water characteristic curve (SWCC), hydraulic conductivity and desiccation crack. Nevertheless, most of studies were conducted on non-expansive soil without any inclusion of fibre. It is reported that soil strength will decrease with biochar amendment (Bora et al. 2020), which may be detrimental to stability. Besides, there are rare studies that investigate effects of biochar and fibre on soil thermal properties. Thermal properties (thermal conductivity and heat capacity) are equally important to analyse and understand any possible heat exchange of green infrastructures in urban environment (Aflaki et al. 2017; Chun and Guldmann 2018).

A comprehensive study on thermal and hydraulic properties of loose non-agricultural soils amended with CSB and/or CSF was carried out. This study aims to (1) explore hydraulic and thermal properties of expansive soils amended with CSB and/or CSF and further (2) investigate their benefits, limitations and engineering feasibility in green roofs.

Fundamental mechanism of soil–water–air–biochar–fibre interaction is revealed at both microscale and macroscale. Thermo-hydraulic coupled model has been formulated for interpretation of expansive soil–water–air–biochar–fibre interaction theoretically. A series of experimental tests were conducted to examine thermal properties (thermal insulation and heat storage), hydraulic properties (water holding and permeability) and crack development in soil amended with biochar and fibres at different percentage contents (0,

is discussed for phase transition of water in the study. On the premise of assumptions above, the theory framework is proposed based on mass conservation of water component and air component, and the first law of thermodynamics.

Mass conservation for water component

Equation of mass conservation for water component in a representative volume element (RVE) is given as:

$$\underbrace{\frac{\partial(nS_r^l \rho_w^l)}{\partial t}}_{\text{Storage}} + \underbrace{\text{div}(nS_r^l \rho_w^l \boldsymbol{v}_w^l)}_{\text{Advection}} + 0 + \underbrace{\frac{\partial(nS_r^g \rho_w^g)}{\partial t}}_{\text{Storage}} + \underbrace{\text{div}(nS_r^g \rho_w^g \boldsymbol{v}_w^g)}_{\text{Advection}} + \underbrace{div(\boldsymbol{f}_w^g)}_{\text{Non-advection}} = 0 \tag{1}$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\text{Liquid water}} \qquad \underbrace{\qquad\qquad\qquad\qquad\qquad}_{\text{Water vapour}}$$

2.5% and 5%). Besides, micro-analyses were performed for elucidation of macro-experimental results. The study fills the knowledge gap in how to utilize coconut shell residuals in green roofs and what benefits these residues could provide technically, economically and environmentally.

Theory, materials and methods

Hydraulic–thermal model

Before discussing engineering-oriented properties of soils aiming at green roofs, there remains theoretical question such as how water, air, biochar and fibre affect soil hydraulic and thermal behaviours in nature. A thermo-hydraulic model is necessary for interpretation and modelling of seepage analysis under rainfall and drying, and thermal insulation under cooling and heating in green roofs or other vegetated infrastructures.

In the research, clayed soil is assumed to be a porous material containing three different phases, i.e. solid (s), gas (g) and liquid (l). Special physical and chemical properties of interfaces between each two phases are ignored. And any phase transitions of solid components (i.e. sublimation and melt) are neglected. Further, gas phase is composed of dry air and vaporous water. Liquid phase consists of liquid water and dissolved air in water. Vaporization and condensation of water occur simultaneously in a non-closed system, and vaporization is dominant in most air humidity range. According to vapour equilibrium method, corresponding suction of 85.1% relative humidity (created by over-saturated KCl solution) is 21.82 MPa at 20 °C (Ye et al. 2019). In such high suction (i.e. soil–water potential), moisture absorption could only occur in fine-grained soils with moisture content near shrinkage limit. Consequently, only vaporization

where symbol n stands for the porosity. Symbol S_r^k denotes the degree of saturation of phase $k(k = l, g)$. Pore volume is filled with either gas or liquid; therefore, $S_r^l + S_r^g = 1$. Symbol ρ_i^k represents the density of component $i(i = s, w, a)$ in phase $k(k = s, l, g)$. Velocity of component $i(i = s, w, a)$ in phase $k(k = s, l, g)$ is expressed by $\boldsymbol{v}_i^k$. Further, it is assumed that different components in the same phase share the same velocity, i.e. $\boldsymbol{v}^k = \boldsymbol{v}_s^k = \boldsymbol{v}_w^k = \boldsymbol{v}_a^k$. It should be noted that the water created by chemical reaction is neglected, i.e. the sources of water are only rainwater and irrigation water. Hence, the right side of the equal sign is 0. And $\boldsymbol{f}_i^k$ is the non-advection flux vector of component $i(i = s, w, a)$ in phase $k(k = s, l, g)$. Non-advection of water in liquid phase is negligible as compared to other terms in Eq. (1). Non-advection flux vector of water in gas phase can be expressed through the following equation:

$$\boldsymbol{f}_w^g = -D_{\text{atm}}^{w,a} \tau_s n S_r^g \frac{M_w^g \rho_w^g g \boldsymbol{v}_w^g}{RT} \tag{2}$$

where g and R are acceleration of gravity and gas constant of gaseous water, respectively. And $D_{\text{atm}}^{w,a}$ is the molecular diffusion coefficient of water vapour in the air and given by the following equation:

$$D_{\text{atm}}^{w,a} = \frac{T^{7/4} \left(1/M_w^g + 1/M_a^g\right)^{1/2}}{p\left[\left(\sum V\right)_w^{1/3} + \left(\sum V\right)_a^{1/3}\right]} \times 10^{-7} \tag{3}$$

where T and p denote the absolute temperature (i.e. thermodynamic temperature) and the pressure, respectively. Symbol M_i^k is the molecular weight of component $i(i = s, w, a)$ in phase $k(k = s, l, g)$. And $\left(\sum V\right)_i$ represents the molecular volume of component $i(i = s, w, a)$.

Tortuosity of soil is described by τ_s in Eq. (2), which describes the flow channel in porous material, and

$$\tau_s = \frac{L_c}{L_s} \tag{4}$$

where L_c represents the length of curved channel of the flow and L_s denotes the apparent length of straight channel hypothetically. The tortuosity is usually measured indirectly instead of being calculated through Eq. (4).

Further, Eq. (1) is rewritten as:

$$\underbrace{\underbrace{\frac{\partial\left[n\left(1-S_r^l\right)\rho_a^g\right]}{\partial t}}_{\text{Storage}} + \underbrace{\text{div}\left[n\left(1-S_r^l\right)\rho_a^g \boldsymbol{v}^g\right]}_{\text{Advection}}}_{\text{Dry air}} + 0 + \underbrace{\underbrace{\frac{\partial(nS_r^l H\rho_a^l)}{\partial t}}_{\text{Storage}} + \underbrace{\text{div}(nS_r^l H\rho_a^l \boldsymbol{v}^l)}_{\text{Advection}} + \underbrace{\text{div}(\boldsymbol{f}_a^l)}_{\text{Non-advection}}}_{\text{Air dissolved in water}} = 0 \tag{10}$$

$$\underbrace{\underbrace{\frac{\partial(nS_r^l \rho_w^l)}{\partial t}}_{\text{Storage}} + \underbrace{\text{div}(nS_r^l \rho_w^l \boldsymbol{v}^l)}_{\text{Advection}}}_{\text{Liquid water}} + 0 + \underbrace{\underbrace{\frac{\partial\left[n\left(1-S_r^l\right)\rho_w^g\right]}{\partial t}}_{\text{Storage}} + \underbrace{\text{div}\left[n\left(1-S_r^l\right)\rho_w^g \boldsymbol{v}^g\right]}_{\text{Advection}} + \underbrace{\text{div}(\boldsymbol{f}_w^g)}_{\text{Non-advection}}}_{\text{Water vapour}} = 0 \tag{5}$$

where

$$f_w^g = -\frac{T^{7/4}\left(1/M_w^g + 1/M_a^g\right)^{1/2}\tau_s n(1-S_r^l)M_w^g \rho_w^g g \boldsymbol{v}^g}{pRT\left[(\sum V)_w^{1/3} + (\sum V)_a^{1/3}\right]} \times 10^{-7}. \tag{6}$$

where $D_{\text{atm}}^{a,w}$ represents the molecular diffusion coefficient of dry air in liquid water, and given by the following equation:

$$D_{\text{atm}}^{a,w} = \frac{T^{7/4}\left(1/M_a^l + 1/M_w^l\right)^{1/2}}{p\left[(\sum V)_a^{1/3} + (\sum V)_w^{1/3}\right]} \times 10^{-7} \tag{9}$$

where definitions of parameters are same as those in Eq. (3).

Further, Eqs. (7) and (8) are derived to Eqs. (10) and (11), respectively.

where

$$f_a^l = -\frac{T^{7/4}\left(1/M_a^l + 1/M_w^l\right)^{1/2}\tau_s nS_r^l M_a^l \rho_a^l g \boldsymbol{v}^l}{pRT\left[(\sum V)_a^{1/3} + (\sum V)_w^{1/3}\right]} \times 10^{-7}. \tag{11}$$

Mass conservation for air component

Similar to the expression of Eq. (1), the mass conservation for air component in the RVE is given by:

$$\underbrace{\underbrace{\frac{\partial(nS_r^g \rho_a^g)}{\partial t}}_{\text{Storage}} + \underbrace{\text{div}(nS_r^g \rho_a^g \boldsymbol{v}^g)}_{\text{Advection}}}_{\text{Dry air}} + 0 + \underbrace{\underbrace{\frac{\partial(nS_r^l H\rho_a^l)}{\partial t}}_{\text{Storage}} + \underbrace{\text{div}(nS_r^l H\rho_a^l \boldsymbol{v}_a^l)}_{\text{Advection}} + \underbrace{\text{div}(\boldsymbol{f}_a^l)}_{\text{Non-advection}}}_{\text{Air dissolved in water}} = 0 \tag{7}$$

where H represents the Henry's coefficient of solubility that describes volume of air dissolved in liquid phase. At 25 °C and atmospheric pressure, the Henry's coefficient of air in water is 7.3×10^{-6} kPa. Non-advection of air in gas phase can be ignored as compared to other terms in Eq. (7). And $\boldsymbol{f}_a^l$ is the non-advection flux vector of component air in phase liquid, which can be expressed as:

$$f_a^l = -D_{\text{atm}}^{a,w}\tau_s nS_r^l \frac{M_a^l \rho_a^l g \boldsymbol{v}^l}{RT} \tag{8}$$

Equation of heat balance

The heat exchange between soil particles is mainly realized through conduction, while that between fluids (water and air) is through convection. Heat exchange between soil particle and fluids is combined effect of conduction and convection. Further, all heat exchanges mentioned contribute to variation of soil enthalpy, as shown in Eq. (12).

$$\begin{aligned}\Delta H = &\left[c_s^s(1-n)\rho_s^s + c_a^g nS_r^g \rho_a^g + c_a^l HnS_r^l \rho_a^l\right.\\ &\left.+ c_w^g nS_r^g \rho_w^g + c_w^l nS_r^l \rho_w^l\right](T - T_0)\end{aligned} \tag{12}$$

where T and T_0 stand for the current temperature and initial temperature, respectively. Symbol c_i^k is the mass specific heat

capacity of component $i(i = s, w, a)$ in phase $k(k = s, l, g)$ at constant pressure, and

$$c_i^k = \frac{C_{i,p}^k}{m_i^k} \tag{13}$$

where $C_{i,p}^k$ is the heat capacity at constant pressure of component $i(i = s, w, a)$ in phase $k(k = s, l, g)$. In comparison with heat capacity at constant volume, heat capacity at constant pressure is relatively easy to gauge.

Phase transition should be also taken into consideration, and the phase transitions between liquid water and water vapour (i.e. vaporization and condensation) are critical for green roofs. Although vaporization and condensation occur simultaneously, vaporization plays a predominant role in soil. Heat of vaporization (i.e. latent heat) of water is related to temperature. When discussing an extremely short unit time, latent heat could be treated as a constant value. Therefore, heat loss due to vaporization could be expressed as Eq. (14).

$$Q_v = L_T m_w^l R_w^{l \to g} \tag{14}$$

where L_T stands for the latent heat at a particular temperature T and m_w^l denotes the mass of liquid water. Ratio of vaporized water to the total at given time is represented by $R_w^{l \to g}$.

According to the first law of thermodynamics, the heat conservation equation under local heat balance is given as:

$$\Delta H + Q_v = Q_E \tag{15}$$

where Q_E is the heat from external source.

Employing formulas (12) and (14), Eq. (15) is rewritten as:

$$\begin{aligned}&\left[c_s^s(1-n)\rho_s^s + c_a^g n S_r^g \rho_a^g + c_a^l H n S_r^l \rho_a^l \right.\\&\left. + c_w^g n S_r^g \rho_w^g + c_w^l n S_r^l \rho_w^l \right](T - T_0) + L_T m_w^l R_w^{l \to g} = Q_E.\end{aligned} \tag{16}$$

After proposal of theory framework, the next work is to measure unknown parameters used in analytical derivation. Hydraulic conductivity is used to replace water velocity which is not a geomaterial parameter. Special heat capacity and heat conductivity are required in thermal field. Other geomaterial properties (such as porosity, compaction and degree of saturation) are also discussed in the research.

Experimental programme for investigating effects of biochar and fibre on expansive soil

Expansive soil, fibre and biochar properties

Expansive soil used in the current study was collected from Nanning (108° 22′ E, 22° 48′ N) that is located in Southern China. Both CSB and CSF were derived from coconuts

Table 1 Basic properties of bare soil, biochar and fibre

	BS	CSB	CSF
pH value	4.6–5.1	8–8.5	N/A
Carbon content (w/w; %)	11.9	90.4	45.9
Liquid limit (%)	58	N/A	N/A
Plastic limit (%)	30	N/A	N/A
Liquidity index	28	N/A	N/A
Shrinkage limit (%)	17	N/A	N/A
Free swelling ratio (%)	50	N/A	N/A

from Hainan Island in Southern China. Procedures for fibre extraction were based on Li et al. (2019). Biochar was produced by slow pyrolysis at 600 °C for a period of 2 h. Both expansive soil and CSB were crushed and then passed through 2-mm sieve before being tested. In order to avoid any influence due to fibre aspect ratio, only fibres with 1.5–2.5 cm length were used (Bordoloi et al. 2018b). The basic physical–chemical properties of tested materials are characterized in Table 1. The test series are divided into four sets in the study: (1) bare soil (BS), (2) soil + 5% biochar (biochar-amended soil, BAS), (3) soil + 5% fibre (fibre-amended soil, FAS) and (4) soil + 2.5% biochar + 2.5% fibre (composite soil, CS). All percentages mentioned are mass ratios of additives to bare soil. The selected percentages are commonly used for suppressing desiccation crack or improving soil strength in soil amended with biochars and fibres (Danso et al. 2015; Ganesan et al. 2020).

Hydraulic properties and cracking of expansive soil amended with biochar and fibres

Soil–water retention curves (i.e. relationship of volumetric water content and suction, also known as soil–water characteristic curves, SWCCs) were measured in cylindrical vessels with 22 cm diameter. Two soil suction sensors (TEROS-21; Meter Group Inc., USA) and two volumetric water content sensors (EC-5; Meter Group Inc., USA) were installed at diametrically opposite ends in each sample. A gap of more than 4 cm ensures preventing any interference between sensors if exist (Bordoloi et al. 2019a; Garg et al. 2020). According to studies (Bogena et al. 2007; Evett et al. 2006; Hook et al. 2004; Sreedeep and Singh 2005, Sreedeep and Singh 2011), soil properties (dry density and particle size), temperature and applied supply voltage affect accuracy and credibility of sensors. Therefore, ProCheck (Meter Group Inc., USA) with stable supply voltage was used for sensor reading measurements. According to previous studies (Bogena et al. 2007; Evett et al. 2006), influence of temperature on sensor reading was relatively acceptable in the range from 18 to 28 °C. Therefore, all samples were placed at

air-conditioned room with 24 ± 2 °C temperature. Dry density and depth of soils were 1.0 g/cm³ and 12 cm, respectively. Deionized water was used throughout the experimentation tenure for minimizing any influences of salts in water if exist on suction measurements.

Considering requirement of vegetation for green roofs, saturated hydraulic conductivity was gauged at relatively loose conditions (i.e. 0.9, 1.0 and 1.1 g/cm³ dry density) via falling head permeability test. Vacuum saturation was conducted using deionized water, and temperature difference between water used in preparation and test processes was less than 2 °C. A mini-disk infiltrometer (Meter Group Inc., USA) was utilized to measure near-saturated hydraulic conductivity. As per review by Bordoloi et al. (2019b), infiltration measurements can be deduced using a mini-disk infiltrometer under laboratory conditions. Using measurements of SWCC and saturated hydraulic conductivity, unsaturated hydraulic conductivity was also deduced by empirical formula as follows (Brooks and Corey 1964):

$$K = K_{sat} \left(\frac{S_r - S_{res}}{S_{sat} - S_{res}} \right)^3 \qquad (17)$$

where K_{sat} stands for the saturated hydraulic conductivity. Symbols S_{res} and S_{sat} denote the degree of saturation at residual state and saturated one (i.e. 100%), respectively.

When it comes to hydraulic and thermal properties, effects of cracks should be also taken into account. Rainfall infiltration and evaporation result in drying–wetting cycles in soils, leading to initiation of desiccation cracks and degradation of soil shear strength (Xie et al. 2020). On the one hand, hydraulic conductivity would directly increase by several orders of magnitude due to the presence of cracks. On the other hand, crack network connects originally isolated and closed internal pores, which contributes to improving soil permeability indirectly. Also, soil surface area exposed to atmosphere is escalated due to desiccation cracks, leading to more intense evaporation. So as to investigate crack propagation quantificationally, crack intensity factor (CIF, i.e. ratio of cracked soil area to total area) was calculated. Each soil sample with 1.0 g/cm³ dry density was placed in a

Fig. 1 Crack images processing
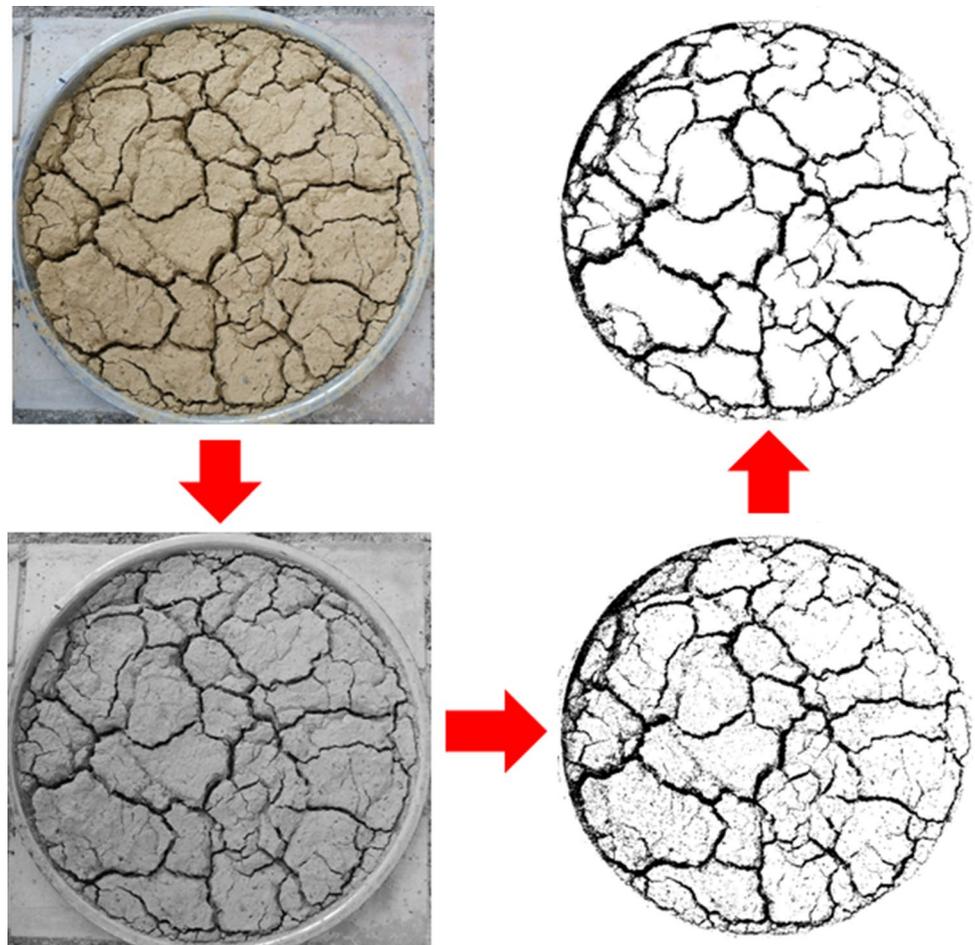
salver with 30 cm diameter and 2 cm height. All soil samples including composites were placed at outdoor environment, subjected to natural drying–wetting cycles. Meteorological parameters upon samples, e.g. temperature, solar radiation and atmospheric relative humidity, were measured using ATMOS 41 meteorological station. The average atmospheric temperature, solar radiation (from 8 a.m. to 8 p.m.) and relative humidity were 25.8 ± 7.2 °C, 958 ± 308 kJ/(m² h) and $77.5 \pm 11\%$, respectively. Fully air-dry was defined as the condition when sample mass reduction was less than 1% in 48 h. Cracks were photographed at each dry event using a digital camera (EOS80D; Canon Inc., Japan) at the same height and regular interval, and then, crack images obtained were converted to greyscale ones (Fig. 1). Intact soil appears light than cracked areas in greyscale images. Such colour contrast under same photographic conditions enables effective crack identification. After binarization and de-noising, CIF was calculated by counting pixel numbers of cracked and intact areas. It is also recommended to trace crack propagation with respect to time using time-lapse photography.

Thermal properties of expansive soils amended with biochar and fibres

Thermal properties (thermal conductivity, thermo-gravimetric behaviour and specific heat capacity) were measured for soil samples mixed with biochar and fibres. These properties are essential for interpretation of clayed soil–biochar–fibre interaction aiming at green roofs.

Thermal conductivity was measured via a TPS-2500S apparatus (Hot Disk Inc., Sweden) based on transient plane thermal source method (Gustafsson 1991). Only solid components in soil or composites are discussed in the study, because (1) thermal conductivities of air and water are known and (2) soil moisture content and void ratio cannot be measured to perfection. In order to reflect thermal conductivity of solid component, samples were highly compacted to 1.55 ± 0.02 g/cm³ dry density and dried at 110 °C for 48 h. Before testing, samples were wrapped with plastic film to prevent absorbing moisture from atmosphere. For each sample, repeated tests were carried out at different powers (i.e. 50, 100 and 200 mW).

Thermo-gravimetric analyses were performed using a thermo-gravimetric apparatus (Q5000IR; TA Instruments Inc., USA). Samples (5 mg) were placed in a crucible and heated from ambient temperature (30 °C) to 710 °C under a flow of dry air at 10 °C/min. In accordance with published study (De-la-Rosa et al. 2008), organic matters were divided into three groups, i.e. labile (mass loss observed between 200 and 380 °C), recalcitrant (between 380 and 475 °C) and refractory components (between 475 and 650 °C). Three temperature ranges mentioned were named as Exo 1, Exo 2 and Exo 3, respectively.

Heat capacity of soil sample could be calculated approximately based on proportions of soil particles, air, water, CSB and CSF (Abu-Hamdeh 2003). Since specific heat capacities of air and water are known, only soil particles, CSB and CSF were required in the study. For these solid materials, specific heat capacity was measured by a water calorimeter (Abu-Hamdeh 2003; Lyon 2015) based on the following theory. In a closed adiabatic system without any heat exchange with external environment, several components with different initial temperatures would obtain a given temperature in equilibrium. Such heat exchange process follows rules as expressed in Eq. (18):

$$m_w^l c_w^l (T_1 - T_{0,w}) + m_m^s c_m^s (T_1 - T_{0,m}) = 0 \qquad (18)$$

where m_w^l and c_w^l stand for the mass and mass specific heat capacity of liquid water. Symbols m_m^s and c_m^s denote mass of measured solid and its mass specific heat capacity. Initial temperature of component k ($k = w, m$) and final temperature of mixture were expressed as $T_{0,k}$ and T_1, respectively.

Since specific heat capacity of water is known, Eq. (18) could be rewritten as Eq. (19) to obtain specific capacity of measured solid.

$$c_m^s = \frac{m_w^l c_w^l (T_1 - T_{0,w})}{m_m^s (T_{0,m} - T_1)}. \qquad (19)$$

In order to minimize influence of air, (1) measured soil and biochar were pulverized and passed through 0.5-mm sieve to reduce closed intra-pore volume, (2) calorimeter container was full of pre-boiled water and measured solid and (3) samples are highly compacted to minimize void space. Repeatability was ensured by re-test with different initial water temperatures.

Micro-analysis of expansive soil, biochar and fibres

So as to visualize surface morphology of sample, scanning electron microscopy (SEM; EVO18, Carl Zeiss, Germany) analysis was performed. Also, Brunauer–Emmett–Teller (BET) specific surface areas of samples were measured using an ASAP-2020 apparatus (Micromeritics Inc., USA) based on nitrogen adsorption and desorption method to further explore material surface characterizations quantificationally. Two replicates (pre-dried by an oven) were provided for each tested sample.

Fourier transform infrared (FTIR) spectroscopy analyses were performed using a Nicolet iS10 spectrometer (Thermo Fisher Scientific Inc., USA) on triplicate samples for each spectrum. Oven-dried sample (1 mg) was diluted with KBr powder (150 mg) uniformly, and then, the mixture was ground using an agate mortar and compressed by a tableting machine. Conservatively, limitations of KBr technique are

listed: (1) ion exchange and displacement induced by applied load during grinding process would result in absorption peak moving towards the right by 10 cm^{-1} approximately and (2) pre-dried KBr powder absorbs moisture from surrounding atmosphere during grinding process. To reduce influences of limitations mentioned, necessary measures were adopted: (1) powders were ground with moderate grinding force under a drying lamp as soon as possible and (2) freshly prepared powder mixture was oven-dried again before being compressed. Spectra obtained by spectrometer were converted to absorbance ones followed by correctness (i.e. smoothing and baseline correction). Although ratio of sample to KBr powder was controlled at 1:150, uniformity cannot be guaranteed absolutely and only part of diluted sample was used for making tablet sample. Further, only a point within the tablet was irradiated via infrared light. Therefore, absolute peak values were not comparable among different spectra. Standardized spectra obtained via ordinate normalization, where all peaks were at the interval of 0–1, are more acceptable for comparison among different samples.

Results and discussion

Thermo-gravimetric analyses

Mass loss curves of expansive soil, CSB and CSF are shown in Fig. 2a–c, respectively. Unlike cellulose-rich straw, hemicellulose (near 50%) and lignin (30%) are predominant in coconut shell residual (Windeatt et al. 2014). Labile group is mainly composed of carbohydrates and long-chain aliphatic carbon, including cellulose, while lignin and hemicellulose mainly decompose in Exo 2 range (dominated by mass losses of aromatic-rich organic matters). This suggests that higher pyrolysis temperature is required for CSB as compared to biochar derived from cellulose-rich straw. However, costs of CSB during collection and transportation of raw material (i.e. coconut shell residues) are almost nil due to production of coconut water beverage, allowing low total costs (the sum of collection, transportation, pyrolysis and package costs).

For all samples, low mass losses (about 4%) are observed at 50–100 °C that could correspond to evaporation of liquid water within samples. Evaporations of bond water and adsorbed water in samples lead to part of mass loss from 100 to 400 °C (Wang et al. 2019). Soil organic matters mainly decompose into Exo 3 and Exo 2, while mass losses in Exo 1 and beyond 650 °C are negligible. Gradient of biochar mass loss curve is relatively invariable from 200 to 600 °C and then becomes obviously sharp after 600 °C (i.e. pyrolysis temperature of tested biochar). Biochars are produced in high temperature but anoxic conditions, so mass losses are still possible in dry air furnace atmosphere below the
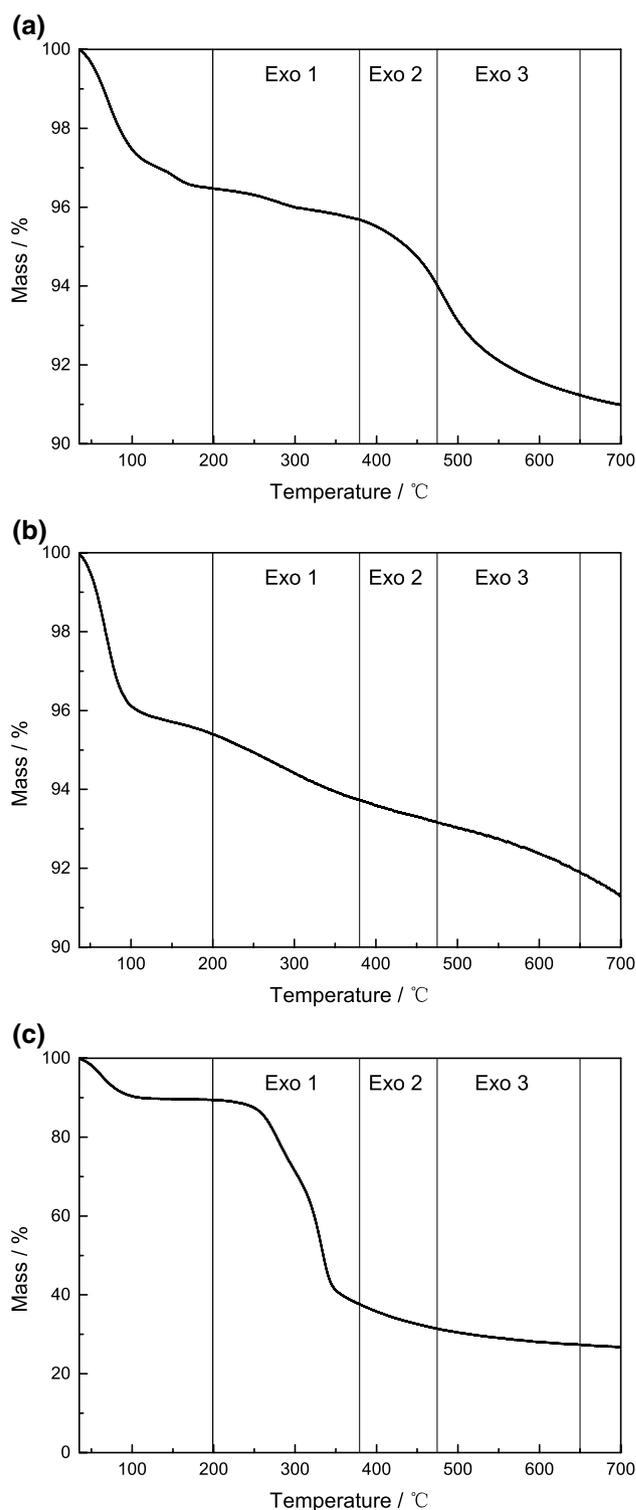

Fig. 2 **a** Mass loss curve of bare soil. **b** Mass loss curve of biochar. **c** Mass loss curve of fibre

pyrolysis temperature. Nevertheless, weight drop was also detected between 500 and 700 °C in wood biochar produced at 700 °C in TG test (with Ar stripping gas) according to

Wang et al. (2019). As per inert gas (without oxygen) used in their study, such mass loss in TG test cannot be ascribed to oxytropic bio-polymer which did not decompose under anoxic pyrolysis. On the other hand, biochar pyrolysis condition in their study maintains at 700 °C for 15 h which could ensure complete thermal decompose in comparison with heating conditions in TG test. Hence, it would be controversial to assert that all mass losses below biochar pyrolysis temperature are attributed to organic matters which only decompose in aerobic atmosphere. Such mass loss below the biochar pyrolysis temperature in inert gas might be elucidated by ageing of biochar. Although biochars are commonly described as chemically stable (Omondi et al. 2016), more investigations on biochar ageing are necessary considering

Fig. 3 FTIR spectra of bare soil, biochar and fibre

long-term services of green roofs, landfill cover system and agricultural soils.

Also, superior thermal stability enables CSB an optimal soil amendment in extreme environment, for instance, bentonite buffer in radioactive waste treatment. It is also found that total mass loss percentages of soil and biochar are all less than 10%, while coconut shell fibre loses 70% mass in TG test. In addition to mass loss due to water evaporation, almost all mass losses of fibre are observed in Exo 1. After 450 °C mass loss of fibre is negligible, 450 °C could be considered as the minimum pyrolysis temperature of biochar in manufacture. The major cost of biochar derived from coconut shell is energy consumption of pyrolysis which is mainly related to pyrolysis temperature and time. It necessities more investigations on how to balance pyrolysis cost, biochar yield and its engineering behaviours, i.e. to find the optimal pyrolysis conditions (temperature and pyrolysis time).
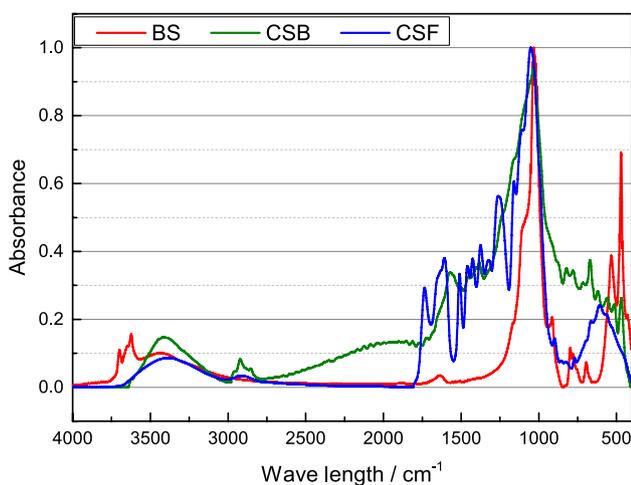
FTIR spectroscopy analyses

A FTIR spectrometer was used to detect functional groups in CSB, CSF and expansive soil as shown in Fig. 3. Functional groups with correction peak height greater than 0.1 are identified and presented in Table 2. For all samples, broad absorption peaks at 3400 cm^{-1} are corresponding to O–H stretching in water. Double peaks in the range from 3700 to 3600 cm^{-1} are attributed to mixture of illite and montmorillonite, while peak at 910 cm^{-1} represents kaolinite (Chen et al. 2017). Expansive soil intenerates and swells significantly after absorbing water and shrinks upon drying, which is mainly because of instable properties of montmorillonite and illite. Relatively low free swelling ratio of expansive

Table 2 FTIR functional group identification

| Position (cm^{-1})	Corrected peak height			Assignment	Representative characterization
	CSB	CSF	Soil		
3700	N/A	N/A	0.061	O–H stretching of phyllosilicates	Montmorillonite and illite
3620	N/A	N/A	0.108	O–H stretching of phyllosilicates	Montmorillonite and illite
3400	0.142	0.078	0.079	O–H stretching	Water
1735	N/A	0.153	N/A	C=O stretching	Ester group
1608	N/A	0.259	N/A	Vibration of aromatic ring	Aromatic compounds
1510	N/A	0.196	N/A	Vibration of aromatic ring	Aromatic compounds
1376	N/A	0.102	N/A	Vibration of aromatic ring	Aromatic compounds
1265	N/A	0.24	N/A	Aromatic CF stretching	Aromatic compounds
1050–1045	0.63	0.401	N/A	C–O stretching	Aliphatic anhydride
1030	N/A	N/A	0.121	Stretching of Si–O	Quartz
910	N/A	N/A	0.15	–OH	Kaolinite
800	N/A	N/A	0.09	Si–O–Si	Quartz
615–605	N/A	0.191	N/A	SO_4^{2-}	Sulphate
535–530	N/A	N/A	0.298	Fe_2O_3	Haematite
470	0.144	N/A	0.369	Fe_2O_3	Haematite
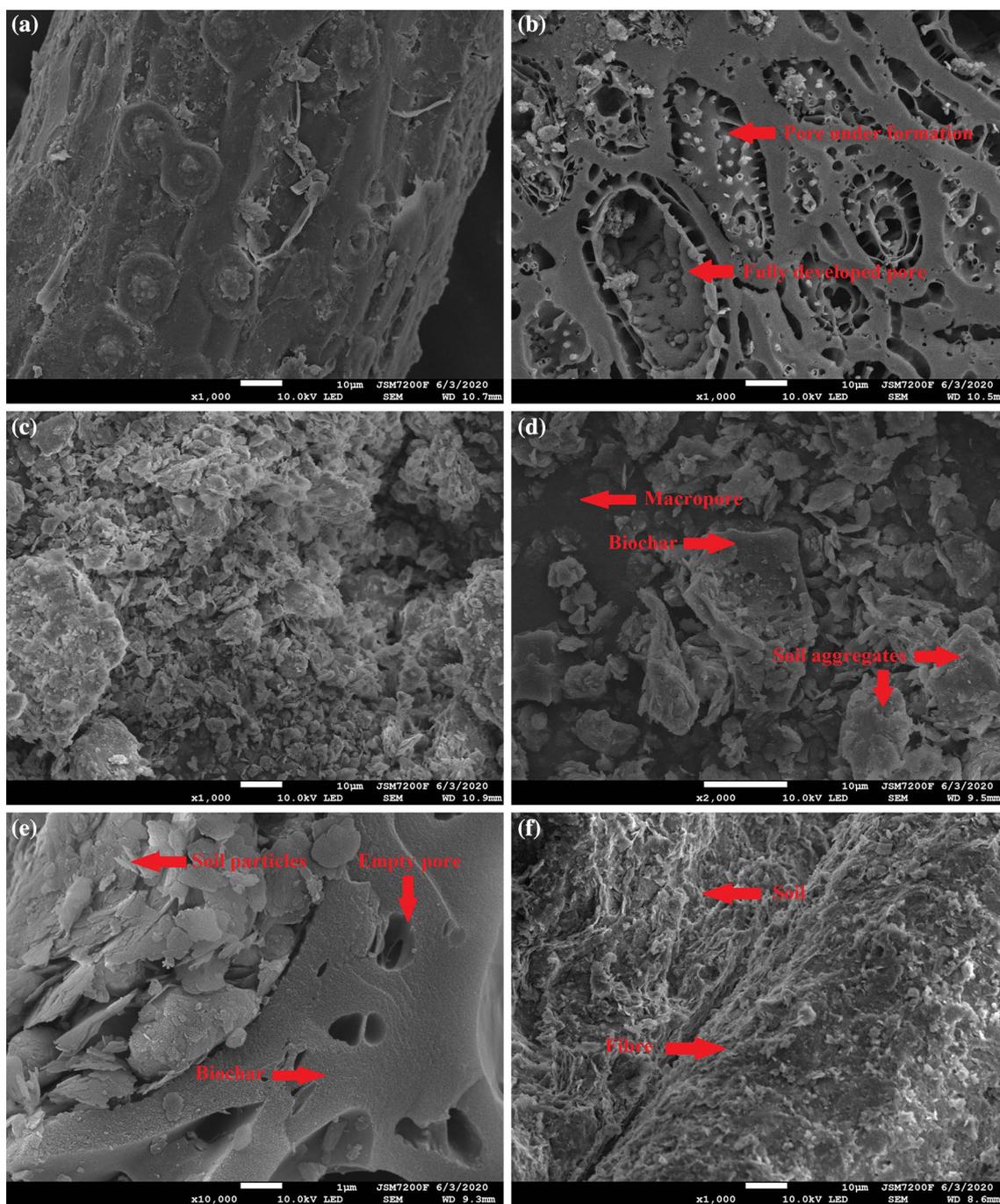
Fig. 4 **a** SEM image of fibre at × 1000 magnification. **b** SEM image of biochar at × 1000 magnification. **c** SEM image of bare soil at × 1000 magnification. **d** SEM image of biochar-amended soil at × 2000 magnification. **e** SEM image of biochar-amended soil at × 10,000 magnification. **f** SEM image of fibre-amended soil at × 1000 magnification

soil tested (Table 1) could be attributed to higher content of kaolinite than illite and montmorillonite. For soil sample, Si–O bond in quartz results in weak peaks at 1030 and 800 cm^{-1} (Reig et al. 2002). Three absorption peaks among 1625–1365 cm^{-1} are ascribed to vibration of aromatic rings (Liu et al. 2019), which are only detected in fibre sample. It is reasonably speculated that these aromatic compounds have been completely decomposed during pyrolysis.

Table 3 Specific surface areas of samples

Samples	Bare soil	Biochar	BAS	FAS	CS
Specific surface area (m^2/g)	15.026	374.359	122.979	30.712	56.542

SEM images and specific surface area

Imporous coconut shell residual is transformed to porous biochar after pyrolysis as observed via SEM at 1000× magnification (Fig. 4a, b). It is also noticeable that a few of pores are still under formation as marked in Fig. 4b, which helps to explain mass loss of biochar beyond 600 °C (i.e. biochar pyrolysis temperature) in TG test. As compared to bare soil and biochar-amended soil (Fig. 4c, d), fibres are tightly bonded by surrounding soils in fibre-amended soil (Fig. 4f). Figure 4e indicates that macropores within biochar surface are filled with soil particles, while pores with relatively small diameter (i.e. micropores and part mesopores) are still empty. Also, micropores within biochar surface enhance the capillary action of water holding.

BET results reveal that both biochar and fibre contribute to improvement in specific surface area of expansive soil (Table 3). Kaolinite-rich expansive clay has relatively lower surface area, as compared to montmorillonite-rich ones. Another mineral with small surface area, quartz, is also detected in tested soil by FTIR. Eight plant-derived biochars produced at 600 °C for 1 h were compared in the literature (Windeatt et al. 2014); coconut shell biochar showed greatest surface area (222.5 m^2/g) among them. Yet surface area reported in the literature is smaller than that measured in this research by 40.6%, because of longer pyrolysis time (i.e. 2 h) in this approach. It is also noticeable that specific surface area of BAS (122.979 m^2/g) is

much greater than the mass-weighted average (32.1 m^2/g) of bare soil and 5% biochar. This indicates that more complex physical–chemical amendments have occurred in composite of biochar and soil, instead of simple mixture.

Soil–water characteristic curves

The soil–water retention curves which represent the variation of volumetric water content with respect to suction (i.e. the difference between soil pore water pressure and the atmospheric pressure) in the range from 10 to 100,000 kPa are shown in Fig. 5. It turns out that both saturated and residual water contents are remarkably improved due to introduction of biochar, and air-entry value is also increased slightly. On the contrary, inclusion of fibre leads to decreased air-entry value, saturated and residual water contents. The results are different from water retention obtained in silty sandy soil mixed with lignocellulose fibres from water hyacinth (Ni et al. 2019). Ni et al. (2019) asserted that highly porous fibres of water hyacinth were conducive for water sorption and hence, soil–water retention. In the current study, samples were compared at different degrees of compaction. At given dry density, void space of FAS is lesser than BS due to relatively low specific gravity of FAS. If samples are compared at given degree of compaction as their study, results might be consistent with each other. On the other hand, surface of coconut shell fibre is not rough (Fig. 4a), unlike porous water hyacinth fibres.

When it comes to applications in green roofs, relatively low air-entry value means that rainwater stored in soil after rainfall could be drained relatively easily, thereby providing sufficient water storage space for next rainfall in rainy seasons. Soil air-entry value is reduced remarkably by inclusion of fibre, but improved slightly via admixture of biochar. Moreover, green roofs delay runoff peak via storing
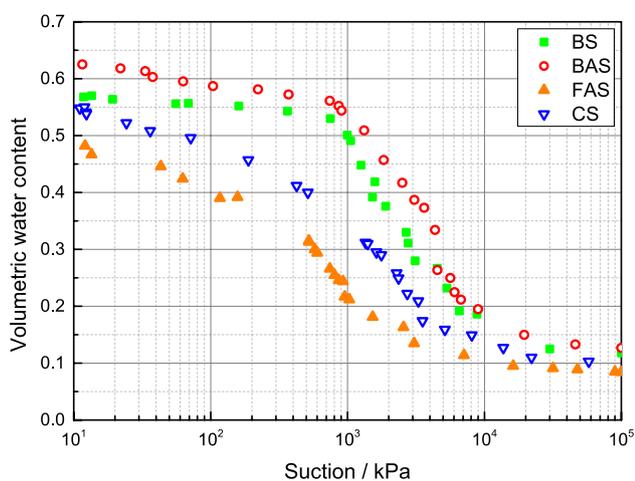
Fig. 5 Soil–water characteristic curves
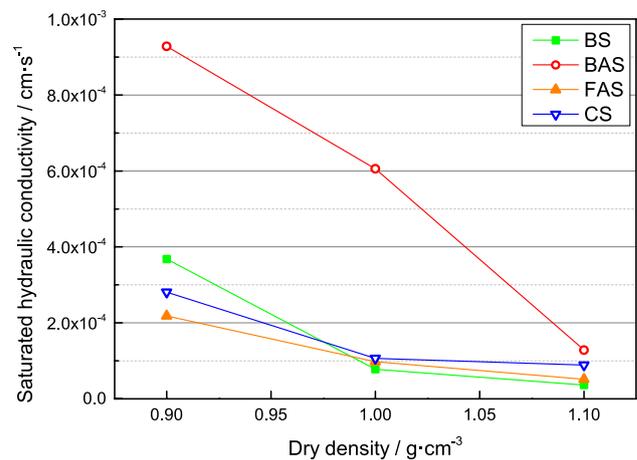
Fig. 6 Variation of saturated hydraulic conductivity with soil dry density

rainwater in soil temporarily. Therefore, difference between saturated water content and water content corresponding to plant permanent wilting point could be used to assess rainwater regulation capacity of green roofs. Suctions corresponding to residual water contents in clayed soils are much larger than permanent wilting points of most plants, so saturated water content becomes critical in the current study. Biochar helps to improve flood control capacity of green roofs as a result of improved soil maximal water-holding capacity. CS might work well in green roofs subjected to frequent but relatively light rains due to reduced air-entry value and acceptable saturated water content.

Saturated hydraulic conductivity

As shown in Fig. 6, K_{sat} of expansive clay is dramatically improved via biochar incorporation at both loose and compacted conditions, even similar to silt soil. This suggests that basic physical properties of clayed soil have been altered after introduction of biochar. It is reported by Omondi et al. (2016) using meta-analysis based on the published studies that biochar enhances K_{sat} in coarse (by 36.5%)-, medium (by 27.3%)- and fine (by 17.8%)-textured soils. However, there are few studies (Jeffery et al. 2015; Laird et al. 2010) asserting that no influence of biochar on soil K_{sat} is detected. Fibre-amended soil shows similar permeability as bare soil, especially at 1.0 and 1.1 g/cm³. As per model developed by Ni et al. (2019), it can be concluded that permeability of sandy soil was found to reduce due to inclusion of fibres. Their model had made an assumption that the inclusion of fibre will lead to filling of pores of soil, while ignoring any preferential flow effects induced by fibres. It comes to the question of comparability under different conditions again. When samples are compared at given dry density as the current research did, more solid volume (i.e. less void ratio) is found in amended soil due to low specific gravities of biochar and fibre. However, fibres might increase soil void ratio at given degree of compaction considering their spatial structures. Pores-filling effect of fibres is dominant in the former case, while preferential flow effect might be critical in the latter one.

Moreover, K_{sat} of soil reduces by near 40% after fibre amendment at 0.9 g/cm³ dry density. Also, K_{sat} of BS decreases dramatically when dry density increases from 0.9 to 1.0 g/cm³ (i.e. compaction). This phenomenon might be explained through theory of porous medium percolation: isolated pores in soil enlarge with increasing porosity, and these pores would merge quickly when porosity reaches certain value. Hydraulic conductivity would increase by several orders of magnitude after percolation threshold. This means that soil permeability is susceptive to porosity when near percolation threshold. And it could be reasonably speculated that porosity corresponding to 0.9 g/cm³ dry density is near
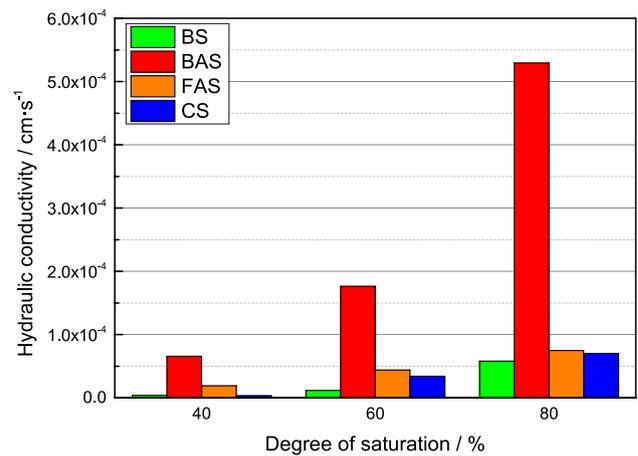

Fig. 7 Correlation between hydraulic conductivity and degree of saturation

percolation threshold of soil tested. Next, hydraulic property of CS is comparable to FAS at loose condition, but close to BAS at compacted condition.

Unsaturated hydraulic conductivity

Surface evaporation leads to a reduction in soil moisture content companied with increasing suction. This means that soil has turned from a two-phase porous medium to a three-phase one. Under unsaturated condition, relative distribution of air-filled and water-filled pores in soil affects soil hydraulic properties. As observed in Fig. 7, unsaturated hydraulic conductivities of BAS are considerably higher than others. When subjected to short but heavy rainfall following long-term draught, superior unsaturated hydraulic conductivity of BAS allows faster recharge of water. Also, it is recommended to further enhance shallow layer permeability via introducing coarse aggregate (e.g. ceramsite) or embedding open-pore plastic pipes. BS, FAS and CS share comparable hydraulic conductivities at near-saturated condition (i.e. 80% degree of saturation). Yet, hydraulic conductivities of BS and CS decline sharply at 60% and 40% degree of saturation, respectively. However, biochar was reported to reduce unsaturated hydraulic conductivity of both bare and vegetated silty sandy soil compacted in 1-D column (Ni et al. 2020). In their study, K_{sat} of bare soil without vegetation increased by an order of magnitude after 730 days. It is to say that large deformation (e.g. swelling) or crack escalation of bare soil might occur during their experiment. In the current study, hydraulic conductivities were measured upon intact soils without cracks and swelling. Another major difference is the methodology for measuring soil permeability. Their study utilizes instantaneous profile method (Ng and Leung 2012) for deducing hydraulic conductivity function. Such method is transient and indirectly measures hydraulic conductivity
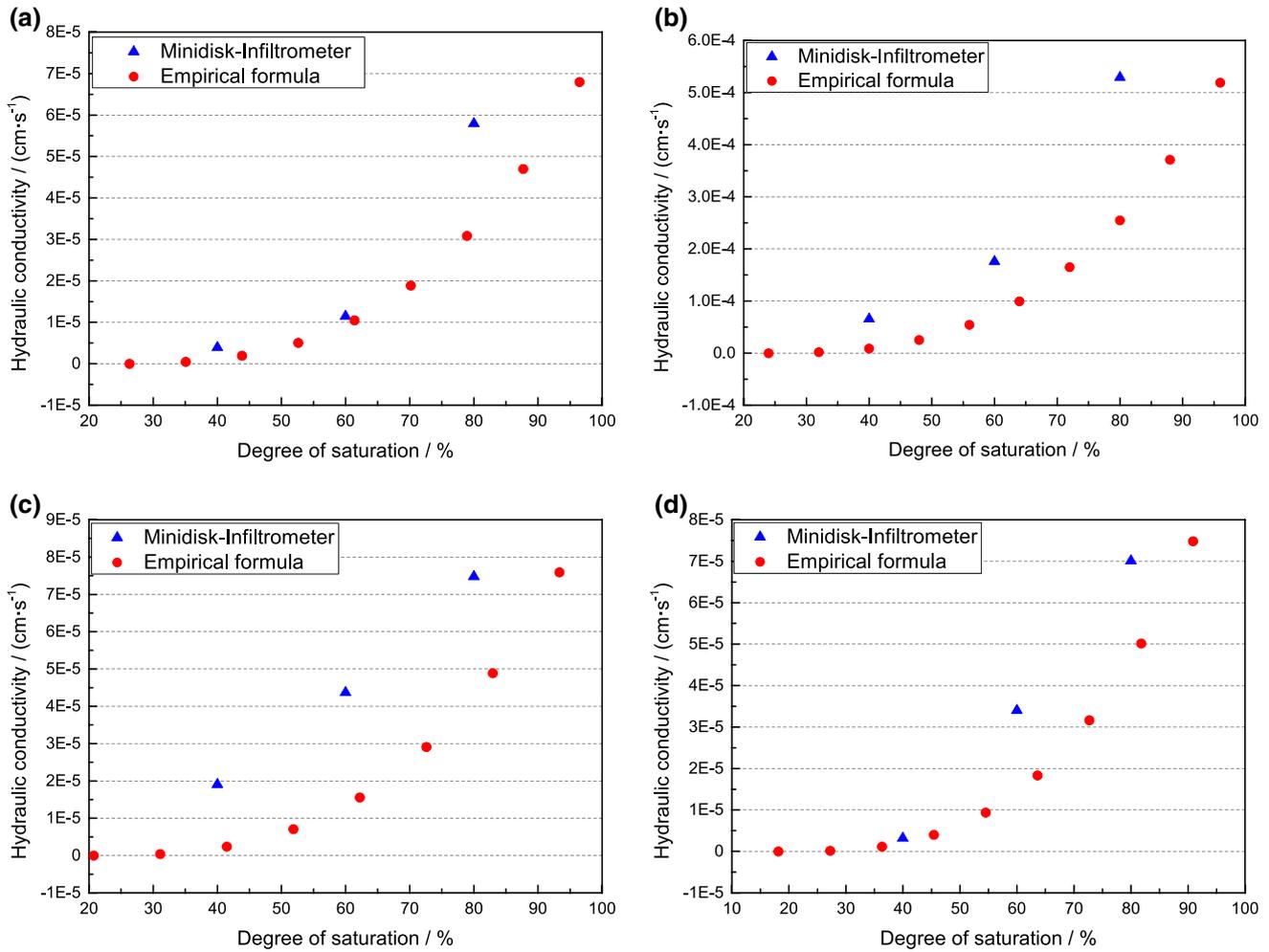
Fig. 8 **a** Comparison of unsaturated hydraulic conductivities between mini-disk infiltrometer and empirical formula for (**a**) BS. **b** Comparison of unsaturated hydraulic conductivities between mini-disk infiltrometer and empirical formula for (**b**) BAS. **c** Comparison of unsaturated hydraulic conductivities between mini-disk infiltrometer and empirical formula for (**c**) FAS. **d** Comparison of unsaturated hydraulic conductivities between mini-disk infiltrometer and empirical formula for (**d**) CS

function from simultaneous measurements of soil volumetric water content and suction. Methodology adopted in the current research is measurement using a mini-disk infiltrometer and falling head permeability tests directly.

Further, comparisons between hydraulic conductivities gauged via a mini-disk infiltrometer and calculated by empirical formula (Eq. 17) are shown in Fig. 8. It is noticeable that all data measured by infiltrometer are higher than calculated ones at given suctions. This might be explained by infiltrometer limitations, i.e. the measurement itself increases soil–water content, especially in surrounding soil; therefore, infiltrometer-measured hydraulic conductivity is higher than true value. Thus, greater difference between two methods is observed in samples with a higher degree of saturation (i.e. with higher hydraulic conductivity). More water infiltrates from a mini-disk infiltrometer into soil during measurement for near-saturated samples. It is also found that
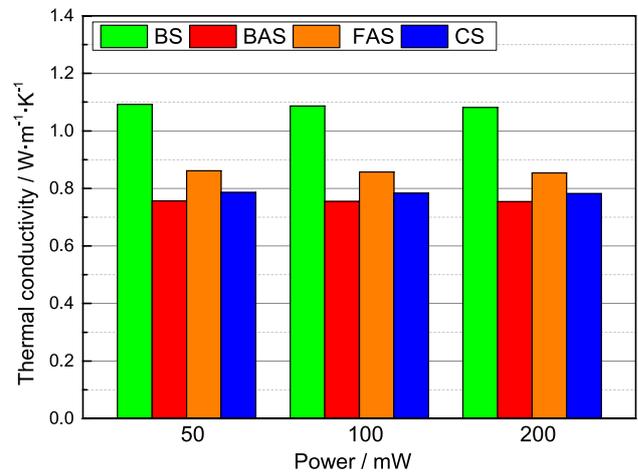

Fig. 9 Thermal conductivities of BS, BAS, FAS and CS

acceptable fittings are shared between infiltrometer measurement and empirical formula, especially for BS, BAS and CS. This suggests that empirical formula can be utilized for design of clay layer in green roofs and provide relatively conservative results.

Thermal conductivity

Thermal conductivity is found to reduce on account of inclusion of both biochar and fibre, as observed in Fig. 9. In accordance with the published studies (Usowicz et al. 2016; Zhang et al. 2013), biochar contributes to decreasing soil thermal conductivity and hence improving thermal insulation. In the current study, highly compacted tested samples (with 1.55 ± 0.02 g/cm^3 dry density) mainly describe thermal properties of solid phase in soil. When it comes to loose planting soil (with less than 1.1 g/cm^3 dry density), thermal conductivities might further decrease because thermal conductivities of both dry air (0.026 W/m K at 30 °C) and liquid water (0.62 W/m K at 30 °C) are less than tested solid samples. The relationship of thermal conductivity with respect to biochar-induced changes in soil properties could vary depending on relative distribution of air-filled and water-filled pores in soil. Briefly speaking, thermal insulation capacity of dry soil is superior to wet one. Hence, it is recommended to weigh decreasing thermal conductivity (i.e. keeping soil dry) against heat loss by soil evaporation and plant transpiration (i.e. keeping soil wet) in summer.

Soil thermal conductivities reduce after amendment and, however, are still greater than traditional thermal insulation materials. According to guidelines in China (JGJ289-2012), USA (ASTM C533-09; ASTM C610-11; ASTM C612-10) and Europe (EN13162-2008; EN13166-2008), materials with thermal conductivity less than 0.1–0.2 W/(m K) are defined as thermal insulation materials. Nevertheless, soil

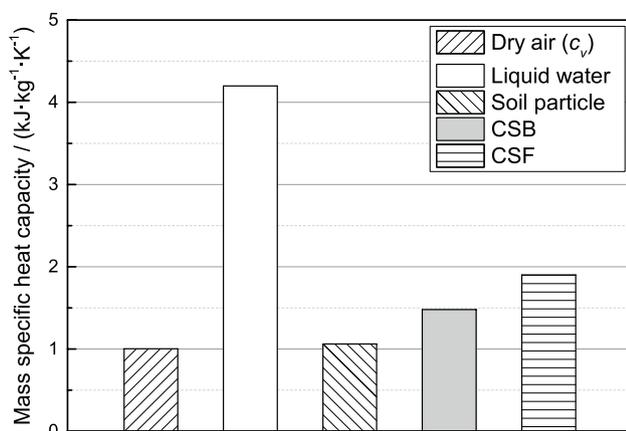depth is usually deeper than 15 cm in green roofs, while thickness of traditional thermal insulation layer is commonly less than 0.5 cm. Green roof might be a promising option to replace traditional insulation materials. It is reported that concrete surface below green roofs without any insulation layer is cooler by near 12 °C in summer and hotter by 4 °C in winter as compared to conventional roofs (Bevilacqua et al. 2016). Also, Permpituck and Namprakai (2012) asserted that heat transfer and energy consumption reduce by 59% and 31%, respectively, in green roof with 10 cm soil depth in comparison with bare roof. Careful studies are required before replacing insulation materials by green roofs at large scale. However, green roofs could work together with traditional thermal insulation layer at least and also protect the layer from natural drying–wetting cycles and thus delay ageing of the layer.

Specific heat capacity

When it comes to heat storage in winter, heat capacity should be also taken into consideration. Heat capacity of composite material could be treated as sum of heat capacities of components approximately. As shown in Fig. 10, dry air and soil particles share similar specific heat capacities. More heat could be stored in fibre and biochar in comparison with bare soil. This suggests that both biochar and fibre are conducive to heat storage in winter. However, direct influences of fibre and biochar are negligible as compared to moisture content due to superior heat capacity of liquid water. Refer to SWCCs (Fig. 5), biochar enhances water-holding capacity and hence improves heat storage capacity, while fibre is the opposite. Only considering heat capacity, it seems that frequent irrigations are required in winter to keep soil wet aiming at heat storage.
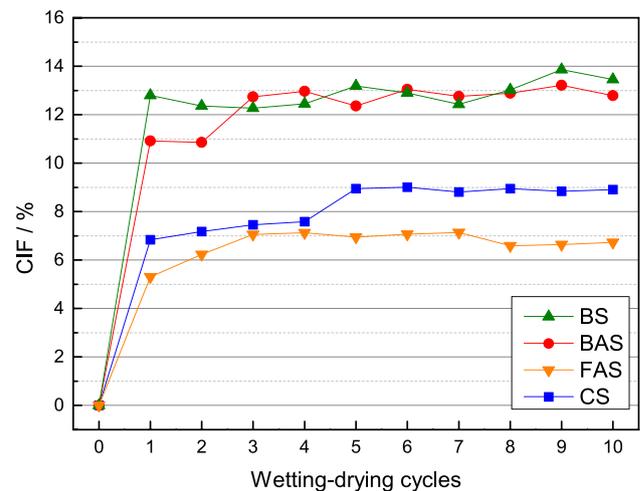

Fig. 10 Mass specific heat capacities of different components

Fig. 11 Variation of CIF with respect to wetting–drying cycles

Crack development

Desiccation cracks would form on soil surface when force induced by air–water interface menisci surpasses the tensile strength of soil during evaporation process. Damage-healing cycles are accompanied with crack propagation within soils under cyclic wetting–drying. During drying period, cracks along with soil contraction (i.e. shrinking) occur which are inherent to the nature of clay. Effects of biochar on mitigating soil crack development seem to be limited or even negligible (Fig. 11). While 5% straw biochar with less than 0.5 mm grain diameter decreased CIF of compacted expansive soil by near 30% after five wetting–drying cycles using 105 °C oven according to Wang et al. (2020), in the current study, CIF of FAS sample reaches to peak value (7%) after three wetting–drying cycles and then remains stable in the following cycles. At beginning, CS shares comparable CIF with FAS, while an increase could be observed at the fifth cycle for CIF of CS. Effects of biochar on crack are assumed to be neglected; 2.5% and 5% fibre amendments decrease soil CIF by 35% and 50%, respectively, after ten wetting–drying cycles, i.e. amendment efficiency decreases with increasing fibre content.

Effects of biochar on fibre degradation

Next, it comes to effects of biochar on fibre in expansive soil, especially degradation of fibre. Degradation of fibre buried in soil depends mainly on burial duration, properties of fibres (e.g. lignin content) and environmental conditions (e.g. soil pH, moisture, temperature and microbes). It is reasonably speculated that biochar could influence fibre degradation indirectly via affecting soil pH and microorganisms. Cellulose-rich fibres are mainly bio-decomposed by fungi, instead of bacteria. According to Rahman et al. (2017), increased fungal community was observed in expansive soil amended with straw biochar. On the other hand, degradation rate of coconut fibre is accelerated marginally in alkaline conditions (pH 9.0) as compared to acidic (pH 5.0) and neutral (pH 7.0) ones (Sumi et al. 2018). It seems that biochar could accelerate decomposition of fibre mainly by enhanced biodegradation. Unlike compacted soil used in landfill cover where anaerobic biochemical reactions play dominant role, loose soils in green roofs allow more oxygen for fungi and bacteria. Actually, coconut fibre is renowned for its greater ability to resist biodegradation as compared to other fibres (Sumi et al. 2018). Measures to extend service life of coir geotextiles, such as surface modification, could be also used in green roofs.

Conclusions

Aiming at green roofs, hydraulic and thermal properties of four kinds of geomaterials (i.e. bare soil, biochar-amended soil, fibre-amended soil and composite soil) are explored theoretically and experimentally. And the following conclusions can be drawn:

K_{sat} of clayed soil is dramatically improved due to introduction of biochar at both compacted and loose conditions. Fibres are conducive to enhancing permeability of slightly compacted clay, while decreasing soil hydraulic conductivity at 0.9 g/cm^3 dry density.

More water could be stored within BAS in entire suction range due to introduction of microporous biochar. CS could drain stored rainwater relatively fast after rainfall because of reduced air-entry value, ensuring sufficient rainwater storage spaces preparing for next rainfall.

Coconut shell biochar improves expansive soil-saturated and residual water contents by 10% and 8%, respectively, and decreases soil thermal conductivity and CIF by 31% and 5%, respectively, while coconut shell fibre decreases soil-saturated and residual water contents by 15% and 29%, respectively, and reduces soil thermal conductivity and CIF by 21% and 50%, respectively. Also, chemical stability of biochar allows long-term services in green roofs, unlike degradable fibres. Therefore, coconut shell biochar seems to be with more commercial values than homologous fibres aiming at green roof soil amendment.

The study fills the knowledge gap of how soil thermal–hydraulic properties are altered due to biochar and/or fibre incorporation. It is recommended to pay more attention on production and utilization of biochar derived from coconut shell currently utilized for fibre extraction. Further studies are required to explore effects of soil suction on hydraulic–thermal properties and interaction between biochar and fibre (especially ageing and degradation) and assess amended soils in real green roofs based on life cycle.

Compliance with ethical standards

Conflict of interest All authors declare that there is no conflict of interest.

References

Abu-Hamdeh NH (2003) Thermal properties of soils as affected by density and water content. Biosyst Eng 86:97–102

Aflaki A, Mirnezhad M, Ghaffarianhoseini A et al (2017) Urban heat island mitigation strategies: a state-of-the-art review on Kuala Lumpur, Singapore and Hong Kong. Cities 62:131–145

Besir AB, Cuce E (2018) Green roofs and facades: a comprehensive review. Renew Sustain Energy Rev 82(1):915–939

Bevilacqua P, Mazzeo D, Bruno R et al (2016) Experimental investigation of the thermal performances of an extensive green roof in the Mediterranean area. Energy Build 122:63–79

Bogena HR, Huisman JA, Oberdorster C et al (2007) Evaluation of a low-cost soil water content sensor for wireless network applications. J Hydrol 344:32–42

Bora MJ, Bordoloi S, Kumar H et al (2020) Influence of biochar from animal and plant origin on the compressive strength characteristics of degraded landfill surface soils. Int J Damage Mech. https://doi.org/10.1177/1056789520925524

Bordoloi S, Garg A, Sreedeep S et al (2018a) Investigation of cracking and water availability of soil–biochar composite synthesized from invasive weed water hyacinth. Bioresour Technol 263:665–677

Bordoloi S, Kashyap V, Garg A et al (2018b) Measurement of mechanical characteristics of fiber from a novel invasive weed: a comprehensive comparison with fibers from agricultural crops. Measurement 113:62–70

Bordoloi S, Gopal P, Boddu R et al (2019a) Soil-biochar-water interactions: role of biochar from *Eichhornia crassipes* in influencing crack propagation and suction in unsaturated soils. J Clean Prod 210:847–859

Bordoloi S, Yamsani SK, Garg A et al (2019b) Critical assessment of infiltration measurements for soils with varying fine content using a mini disk infiltrometer. J Test Eval 47(2):868–888

Brooks RH, Corey AT (1964) Hydraulic properties of porous media. Colorado State University, Fort Collins

Chan FKS, Griffiths JA, Higgitt D et al (2018) "Sponge City" in China—a breakthrough of planning and flood risk management in the urban context. Land Use Policy 76:772–778

Chen H, Koopal LK, Xiong J et al (2017) Mechanisms of soil humic acid adsorption onto montmorillonite and kaolinite. J Colloid Interfaces Sci 504:457–467

Chun B, Guldmann J (2018) Impact of greening on the urban heat island: seasonal variations and mitigation strategies. Comput Environ Urban Syst 71:165–176

Danso H, Martinson DB, Ali M et al (2015) Effect of fibre aspect ratio on mechanical properties of soil building blocks. Constr Build Mater 83:314–319

De-la-Rosa JM, Knicker H, Lopez-Capel E et al (2008) Direct detection of black carbon in soils by Py-GC/MS, carbon-13 NMR spectroscopy and thermogravimetric techniques. Soil Sci Soc Am J 72:258–267

Evett SR, Tolk JA, Howell TA (2006) Soil profile water content determination: sensor accuracy, axial response, calibration, temperature dependence, and precision. Vadose Zone J 5:894–907

Garg A, Huang H, Kushvaha V et al (2020) Mechanism of biochar soil pore–gas–water interaction: gas properties of biochar-amended sandy soil at different degrees of compaction using KNN modeling. Acta Geophys 68:207–217

Gustafsson SE (1991) Transient plane source techniques for thermal conductivity and thermal diffusivity measurements of solid materials. Rev Sci Instrum 62(3):797–804

Hook WR, Ferre TPA, Livingston NJ (2004) The effects of salinity on the accuracy and uncertainty of water content measurement. Soil Sci Soc Am J 68:47–56

Jeffery S, Meinders MBJ, Stoof CR et al (2015) Biochar application does not improve the soil hydrological function of a sandy soil. Geoderma 251:47–54

Kaufhold S, Baille W, Schanz T et al (2015) About differences of swelling pressure—dry density relations of compacted bentonites. Appl Clay Sci 107:52–61

Kumar H, Ganesan SP, Bordoloi S et al (2019) Erodibility assessment of compacted biochar amended soil for geo-environmental applications. Sci Total Environ 672:698–707

Kumar H, Cai WL, Lai JL et al (2020) Influence of in-house produced biochars on cracks and retained water during drying-wetting cycles: comparison between conventional plant, animal, and nano-biochars. J Soil Sedim 20:1983–1996

Laird DA, Fleming P, Davis DD et al (2010) Impact of biochar amendments on the quality of a typical Midwestern agricultural soil. Geoderma 158:443–449

Li J, Wang SL, Zheng L et al (2019) Sorption of lead in soil amended with coconut fiber biochar: geochemical and spectroscopic investigations. Geoderma 350:52–60

Liu XY, Yang G, Lipik V (2019) Permanent water repellent chemical modification of cotton fabric with reagents containing aromatic rings. Fiber Polym 20:51–56

Lopez-Rodriguez G, Perez-Esteban J, Ruiz-Fernandez J et al (2016) Behavior and evolution of sustainable organic substrates in a vertical garden. Ecol Eng 93:129–134

Lyon RE (2015) Thermal dynamics of bomb calorimeters. Rev Sci Instrum 86:125103

Nektarios PA, Amountzias I, Kokkinou I et al (2011) Green roof substrate type and depth affect the growth of the native species Dianthus fruticosus under reduced irrigation regimens. HortScience 46(8):1208–1216

Ng CWW, Leung AK (2012) Measurements of drying and wetting permeability functions using a new stress-controllable soil column. J Geotech Geoenviron 138(1):58–68

Nguyen TT, Ngo HH, Guo W et al (2019) Implementation of a specific urban water management—Sponge City. Sci Total Environ 652:147–162

Ni JJ, Bordoloi S, Garg A et al (2019) Simple model on water retention and permeability in soil mixed with lignocellulose fibres. KSCE J Civ Eng 23(1):138–146

Ni JJ, Bordoloi S, Shao W et al (2020) Two-year evaluation of hydraulic properties of biochar-amended vegetated soil for application in landfill cover system. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2019.136486

Omondi MO, Xia X, Nahayo A et al (2016) Quantification of biochar effects on soil hydrological properties using meta-analysis of literature data. Geoderma 274:28–34

Permpituck S, Namprakai P (2012) The energy consumption performance of roof lawn gardens in Thailand. Renew Energy 40:98–103

Pituya P, Sriburi T, Wijitkosum S (2017) Properties of biochar prepared from acacia wood and coconut shell for soil amendment. Eng J Thailand 21:63–75

Rahman MT, Zhu QH, Zhang ZB et al (2017) The roles of organic amendments and microbial community in the improvement of soil structure of a Vertisol. Appl Soil Ecol 111:84–93

Rahman MA, Moser-Reischl A, Anderson M et al (2019) Comparing the infiltration potentials of soils beneath the canopies of two contrasting urban tree species. Urban For Urban Green 38:22–32

Reig FB, Adelantado J, Moreno M (2002) FTIR quantitative analysis of calcium carbonate (calcite) and silica (quartz) mixtures using the constant ratio method. Application to geological samples. Talanta 58(4):811–821

Sreedeep S, Singh DN (2005) A study to investigate the influence of soil properties on suction. J Test Eval 33:61–66

Sreedeep S, Singh DN (2011) Critical review of the methodologies employed for soil suction measurement. Int J Geomech 11:99–104

Sumi S, Unnikrishnan N, Mathew L (2018) Durability studies of surface-modified coir geotextiles. Geotext Geomembr 46:699–706

Usowicz B, Lipiec J, Lukowski M et al (2016) The effect of biochar application on thermal properties and albedo of loess soil under grassland and fallow. Soil Till Res 164:45–51

Villar MV, Lloret A (2008) Influence of dry density and water content on the swelling of a compacted bentonite. Appl Clay Sci 39:38–49

Walsh CJ, Booth DB, Burns MJ et al (2016) Principles for urban stormwater management to protect stream ecosystems. Freshw Sci 35(1):398–411

Wang L, Chen L, Tsang DCW (2019) The roles of biochar as green admixture for sediment-based construction products. Cem Concr Comp 104:103348

Wang H, Garg A, Huang S et al (2020) Mechanism of compacted biochar-amended expansive clay subjected to drying-wetting cycles: simultaneous investigation of hydraulic and mechanical properties. Acta Geophys 68:737–749

Windeatt JH, Ross AB, Williams PT et al (2014) Characteristics of biochars from crop residues: potential for carbon sequestration and soil amendment. J Environ Manag 146:189–197

Xie CR, Ni PP, Xu MJ et al (2020) Combined measure of geometry optimization and vegetation for expansive soil slopes. Comput Geotech. https://doi.org/10.1016/j.compgeo.2020.103588

Ye YX, Zou WL, Han Z et al (2019) Predicting the entire soil-water characteristic curve using measurements within low suction range. J Mt Sci 16:1198–1214

Zhang QZ, Wang YD, Wu YF et al (2013) Effects of biochar amendment on soil thermal conductivity, reflectance, and temperature. Soil Sci Soc Am J 77(5):1478–1487

Zhang L, Fukuda H, Liu Z (2019) Households' willingness to pay for green roof for mitigating heat island effects in Beijing (China). Build Environ 150:13–20

An innovative application of borehole acoustic image and amplitude logs for geotechnical site investigation

Hsien-Chang Kao[1,2] · Po-Yi Chou[2] · Hung-Chieh Lo[2]

Abstract

Down-hole acoustic televiewer (ATV logging) is a geophysical well-logging device known to be able to reveal underground in-depth image. However, reviewing the current developments in the area of geotechnical engineering, ATV logging data were predominantly used for the purpose of visualizing and calculating the specific structural features in an open borehole, and nothing more. Noticeably, to the majority of engineering applications, it appears that ATV logging played a minor role compared to other geophysical borehole logging tools. Our study does not take it for granted. A systematic study is presented of the significance and consequent applications of ATV logging. The significance of correlation between acoustic amplitude and different geotechnical parameters is discussed and found to be consistent with previous findings. The result tells that ATV logging is actually able to provide timely and ground-truth information during site characterization, if properly used. Additionally, in our study, a simple rock mass evaluation index, Aco-RQD, is proposed, which is applicable to all aspects of geotechnical engineering as long as the continuous ATV logging data are available.

Keywords Acoustic image · Conductive fractures · Straddle-packer test · Geotechnical investigation

Introduction

Receiving adequate and in-depth knowledge on the sub-surface conditions at site is essential to many geological engineering practices. It means a systematic examination conducted by the professionals to obtain ground-truth data with down-hole exploration. Presently, a variety of advanced and exclusive down-hole techniques have been developed for practical use, and the choice can be made from a combination of electrical, nuclear, electromagnetic, seismic, thermal, fluid velocity, acoustic and optical image logging, and so on. Different principles and configurations associated with

specific objectives of investigation can be adopted. In order to cope with more extreme working environment, the development of more representative and reliability-based methods for down-hole characterization is needed.

Down-hole image logging by acoustic televiewer (ATV), as its name implies, is based on analyzing the amplitude of the acoustic reflectivity from the water–solid boundary at the face of borehole (Deltombe and Schepers 2004) and generating an unwrapped real-time images from the reflected acoustic signal of wellbore. It is thus taken for granted that this logging instrument has to be deployed in liquid-saturated open holes, and the amplitude of acoustic reflectivity is closely related to the hardness and structure of the object. As well known, when ATV is deployed in the borehole, the occurrences of breakouts, fractured zones, and weak planes are able to be displayed, which facilitates a number of applications in geotechnical site investigation, ranging from the evaluation of rock mass porosity, the identification of induced fractures, the measurement of strike and dip in the borehole, and to the use in the construction of a fracture network model (Thill and D'Andrea 1976; Tezuka and Niitsuma 2000; Wang et al. 2001; Okko et al. 2003; Williams and Johnson 2004; Cha et al. 2006; Hamm et al. 2007; Nastev et al. 2008;

✉ Po-Yi Chou
  poyi.chou@sinotech.org.tw

  Hsien-Chang Kao
  hckao@sinotech.org.tw

  Hung-Chieh Lo
  jaylo@sinotech.org.tw

1   Department of Civil Engineering, National Taiwan University, 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan

2   Geotechnical Engineering Research Center, Sinotech Engineering Consultants, Inc., 280, Xinhu 2nd Rd., Neihu Dist., Taipei 11494, Taiwan

Comte et al. 2012; Zhao et al. 2013; Zhang et al. 2017; Liu and Leung 2018; Jo et al. 2019; Lin et al. 2020).

However, to the majority of current studies, ATV data was only taken to be used for visualizing borehole structural features, and played a minor role in complement to the field mapping or conventional core logging. Most of the present applications that employing the use of acoustic televiewer only gave (semi)qualitative assessments on the integrity of geological formation (Gwynn et al. 2013; Peyras et al. 2015). As indicated by Nastev et al. (2008) that in some specific conditions those high-angle fractures were hard to be clearly observed by ATV logging, and the effectiveness was likely to be restricted along irregular cavity surfaces (Carnevale and Domaracki 2018). In other words, based on our observations in practice, ATV logging data have been considered less important than other types of borehole logging techniques. For instance in practical use in hydrogeological research, ATV logging was mainly applied for identifying how likely a specific depth interval is to gain access to groundwater, noticeably absent from the discussion was its potential for site characterization and hydraulic parameter estimation. Although a number of previous studies have taken formation porosity (Paillet and Crowder 1996; Kennel and Parker 2019) or fracture aperture (Ku et al. 2009; Chou et al. 2012; Lo et al. 2012) that calculated from ATV logging into account and used for hydraulic conductivity prediction, their applicability is exclusively restricted to the assessment for discrete intervals in a borehole. It appeared to be challenging to further increase the added value of ATV logging with regard to the quantitative characterization in a field scale.

With an aim to demonstrate the improvement and make the best of conventional ATV logging technique for geotechnical site investigation, the current study takes an exploratory borehole NH-01 drilled at the upstream catchment of Nanhua Reservoir (Southern Taiwan) as an example. This borehole investigation program is undertaken by Taiwan Water Resources Planning Institute. Three types of experiments including ATV logging, straddle packer hydraulic testing (SPH), and borehole deformation (BD) testing are conducted in the field, and a number of rock core samples are collected for triaxial compression (TC) testing, shear strength (SS) testing, and elastic modulus (EM) testing in the laboratory. By integrating all measured parameters for systematic evaluation, a strategic approach based on ATV data is developed to interpret the physical behavior of local geological formation. The result provides a new way of thinking about the use of ATV logging in geotechnical site investigation, as both qualitative and quantitative subsurface information can be obtained, the application of this technique could provide a more efficient assessment

means for reservoir safety and stability than current other methods.

Brief description of the study area

The study was carried out at Nanhua district, a rural township locating near the border of Tainan and Kaohsiung city, Southern Taiwan. Its geographical location is situated close to Houku River and Dapu River in Mt. Sanjiaonan, a Plio—Pleistocene sedimentary field in the western foothills belt. The surrounding outcrops are mainly found to be silty shale and sandstone. As shown in Fig. 1, the exploratory borehole NH-01 was drilled in Nanhua Reservoir catchment (23° 05′ N and 120° 32′ E, with an elevation of 260 m) with an aim to provide geologic data on the host rocks for a safety assessment of dam construction, since the aging Nanhua Reservoir located at the downstream has suffered from excess sedimentation. As reported, the capacity of reservoir has reduced to less than half due to periodic erosion and siltation, which causes serious impacts on the operation and water supply efficiency of the reservoir. Accordingly, numerous previous studies have been carried out on the sediment transport caused by landslide and rainfall around this area, and the associated impact on land use, water availability, water quality, and habitat loss, while relatively few were reported about the hydrogeological characterization (Chang and Chuang 2001; Shieh et al. 2010; Chen et al. 2014; Wang et al. 2018).

Methodology

To facilitate the assessment of whether the local geological condition is suitable for dam construction, being as a part of site investigation program, our study was carried out as the following order. As the flowchart of working illustrated in Fig. 2, each stage of assessment was conducted in an orderly manner. Borehole drilling, rock core recording, lateral load testing, ATV logging, and packer testing were carried out in the field, while triaxial compression testing, shear strength testing, and elastic modulus testing of collected rock core specimens were implemented in the laboratory.

Test boring

The test boring (NQ size, drilling coring size of 47.6 mm) was first drilled to a depth of 80 m (262.5 ft). The drill cuttings were collected for preliminary identification of lithostratigraphic units, geological attributes, and later for laboratory analysis. According to the field geologists' report, the drilling has covered the weathered clastic crust
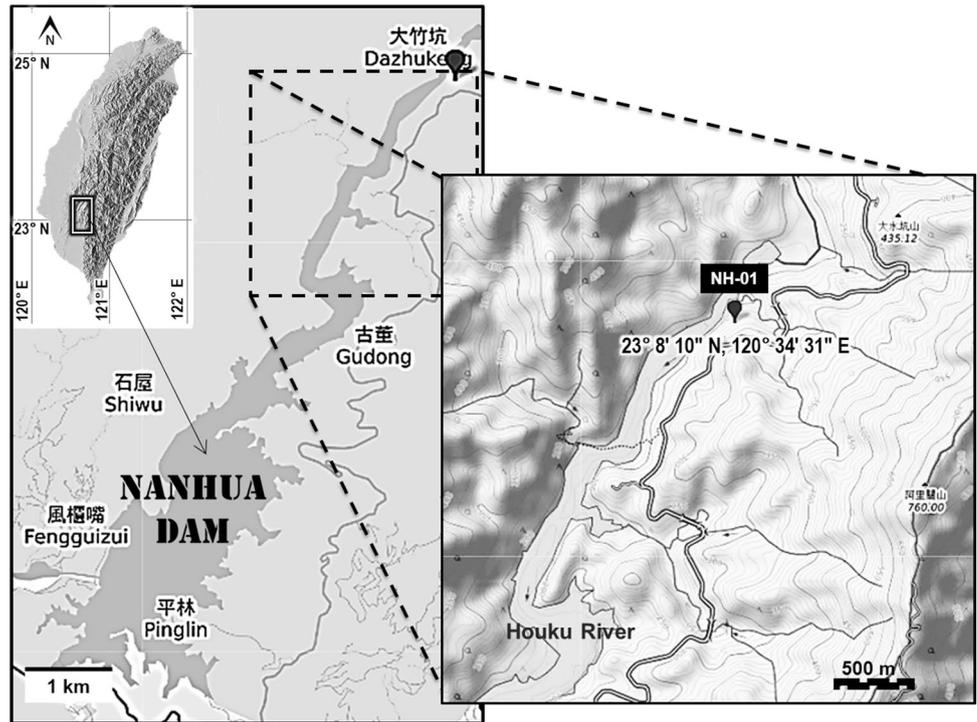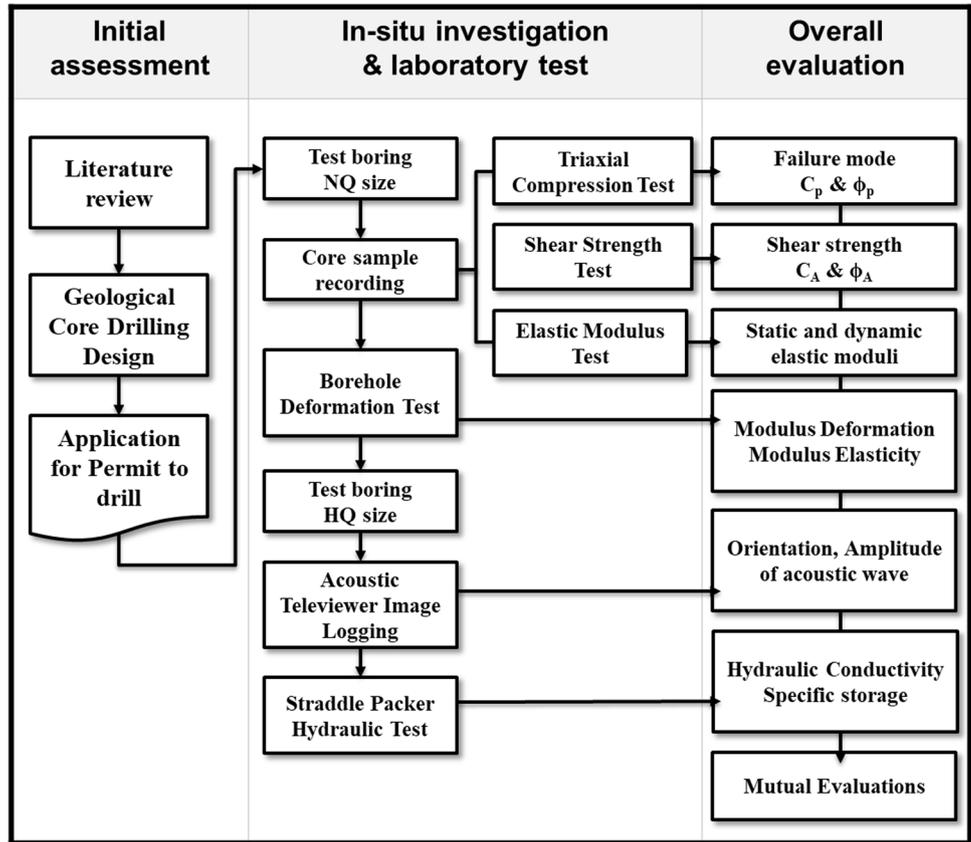
Fig. 1 Profile of study area

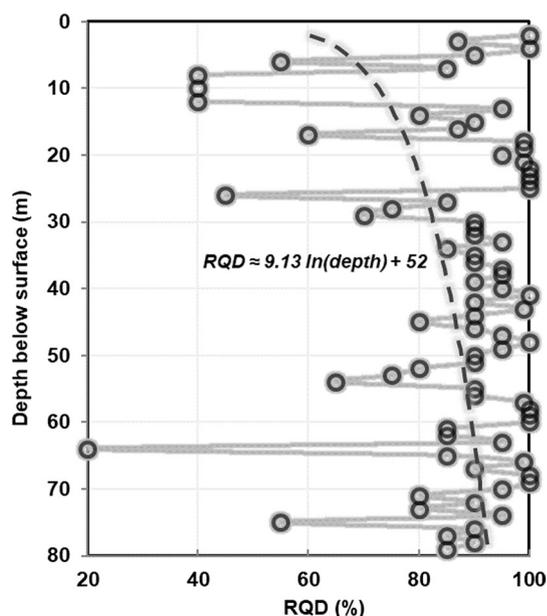**Fig. 2** Flowchart of investigation procedure of this study

Fig. 3 Plot of RQD as a function of depth

the top 2 m, the dominated rock type was composed of shale with few intercalated sandstone layers, and gouge layers were identified at the depth from 8 to 12 and 64 m. The rock quality designation index (RQD, the percentage of intact core pieces longer than 100 mm) ranged from 20 to 100% of the run, with an average of 86%. As illustrated in Fig. 3, it can be seen that the distribution of rock quality in this study area showed a gradual positive exponential increase along the depth, the majority ranged from fair (RQD = 50–75%) to excellent (RQD = 90–100%), while the value over the gouge layers was poor (RQD = 25–50%).

Borehole deformation (BD) test: 6 m, 39 m, and 60 m

This test is designed to measure the radial displacements of borehole wall, right after drilling, and to determine the existing stresses in the rock. For understanding the deformation characteristics of the ground (also called the lateral load test, LLT), this test is conducted to study the cycled loading–unloading stress–strain relationship of rock under different confining pressures, and to provide a straightforward estimation of the corresponding deformation and elastic modulus in the field following the American Society for Testing and Materials (ASTM) (2020) procedure D4719. The instrument used in this study is the TRI-MOD-S Rock Pressuremeter from Roctest (73 mm in diameter). Three depths were selected for deformation test during the drilling operation, the

upper (6 m), middle (39 m), and the lower (60 m) section. After the borehole deformation test was completed, the test boring was then further enlarged with HQ-size core bits (63.5 mm) for the subsequent testing steps.

Acoustic televiewer logging (ATV): 2–80 m

Acoustic televiewer logging was performed, as previously mentioned, to provide a visual indication of the borehole-wall images, whereby the geometric attributes of the fracture pattern can be quantified. This study used the high resolution ATV probe (42 mm in diameter, 1.99 m in length) provided by Robertson Geo for borehole image logging associated with a focused ultrasound beam. To convert the amplitude signal into photographic images, the whole assembly consists of a fixed acoustic transducer, a rotating acoustic mirror and the actuating device. Aiming at providing a high resolution oriented image of the borehole wall, during the test, depth increment was set to 0.002 m and the sampling rate was set as 180/rev. Figure 4 shows the obtained 360° acoustic image and amplitudes of the borehole. As can be seen, the bright areas generally correspond to zones of intact rock associated with good acoustic transmission (high amplitude acoustic responses), whereas the darker areas are the softer formation, strata of low strength or fractured zone associated with attenuating acoustic response, that is the zones at depths of 5–15 m and 60–70 m, at which the variations of rock quality designation index (RQD) are also much significant.

Straddle packer hydraulic (SPH) test: 16 m, 39 m, 52 m, and 73 m

After the lithostratigraphic and geometric characteristics of the borehole have been identified, the straddle packer hydraulic testing is then performed. It is a classical controlled field experiment by which groundwater is pressurized at a given level that inducing a change of flow rate, and as such it allows in situ depth-discrete permeability to be determined. As designed, a pair of inflatable packers was deployed to the depths of interest at which fractures or breakouts are shown, and the test section (i.e,. the targeted zone) was packed off and isolated from the rest of the borehole after gas injection. In principle, this test allows at any specific depths in the borehole to be observed and assessed, as long as the borehole remains open and stable. The inflatable packers used in this study are the ϕ72 Type TZ from Geopro (35 mm inner diameter, 160 mm maximum expansion diameter), and the packed off interval is 1.5 m. The variation of hydraulic head and injected flow rate was archived in the data-logger developed by Sinotech Engineering Consultants Inc. following the ASTM procedure D4630 and D4631.
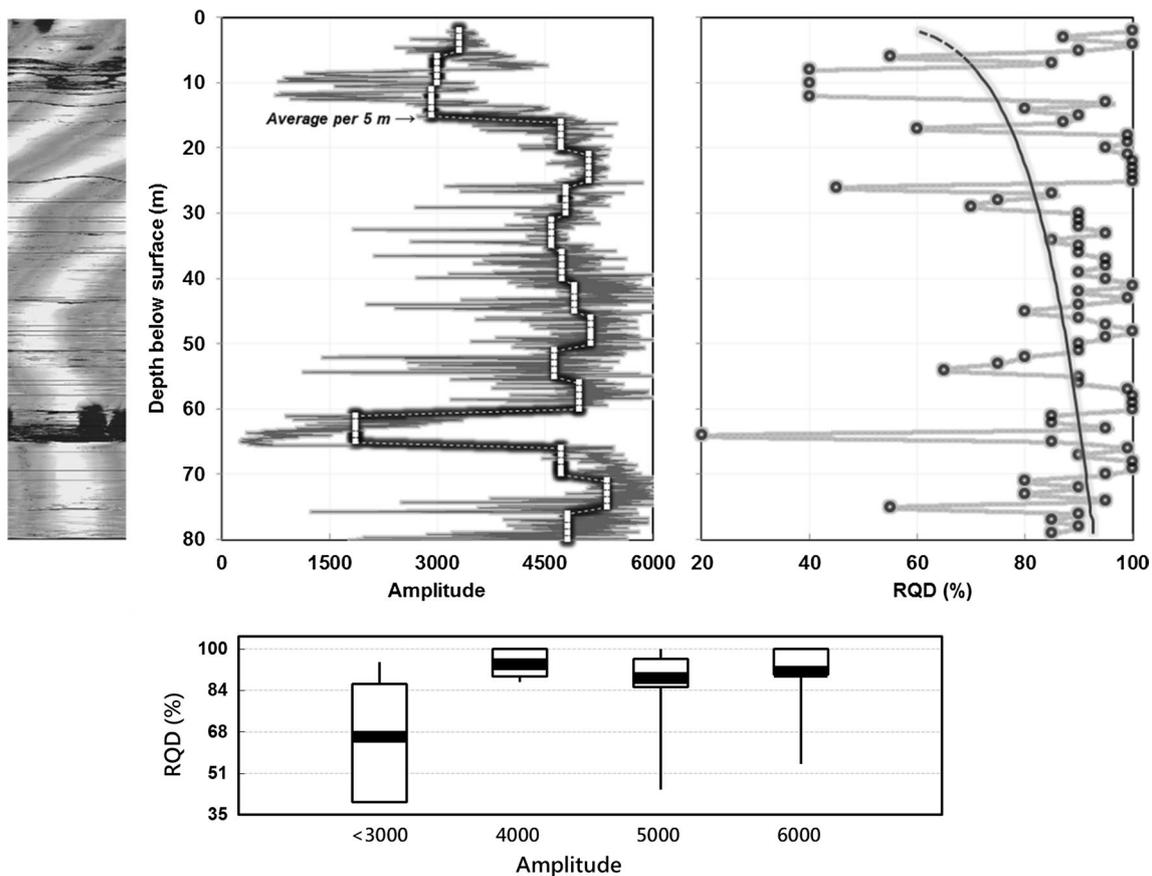
Fig. 4 Acoustic amplitude logging results and a comparison with RQD distribution

Triaxial compression (TC) test: 3–5 m, 18–20 m, 22–24 m, 46–48 m, 72–74 m

In order to understand in detail the mechanical failure behavior of rock mass under confining pressure, the rock core samples collected from five different sections in the borehole were prepared for the indoor TC test. The purpose of this test is to obtain the strength of shale rock core samples as a function of confining pressure, and to estimate the peak cohesion (C_p) and friction angle (ϕ_p). This study used the strain-controlled triaxial compression apparatus (Autotriax) manufactured by Wykeham Farrance. The strength envelope produced under the triaxial state of confining pressure was following the ASTM procedure D2664-95 and the suggested method of ISRM (1978) (Dohnalik et al. 2015).

Shear strength (SS) test along discontinuities: 15 m, 37 m, 44 m, 54 m, and 67 m

In order to understand the bearing capacity of rock against sliding, five rock core specimens collected at different depths were prepared for the SS test. This test is known to be of great importance in rock engineering practices, as the purpose is

to obtain the shear resistance as a function of normal stress present in strata, and to estimate the corresponding shear-strength parameters, i.e. apparent cohesion (C_A) and friction angle (ϕ_A). During the test, the rock core specimen was confined in a metal shear box and then placed in a shearing test machine. Under the constant normal loading condition, increasing horizontal loads were applied to the specimen. In our study, a multi-stage shear tests under three different normal loads were applied. The variation of shear stress corresponding to the applied specific initial apparent normal stress was recorded, and the stress–strain relationship (the failure curve) under different normal stress was established following the ASTM (2014) procedure D7012 and the suggested method by ISRM (1981).

Elastic modulus (EM) test: 3–4 m, 18–19 m, 22–23 m, 46–47 m, 72–73 m

In order to understand the deformability of rock mass under uniaxial compression, in this study, ten rock core specimens were collected at different depths and prepared for the indoor EM test. The purpose of this test is to obtain the strength of shale rock core as a function of confining pressure, and
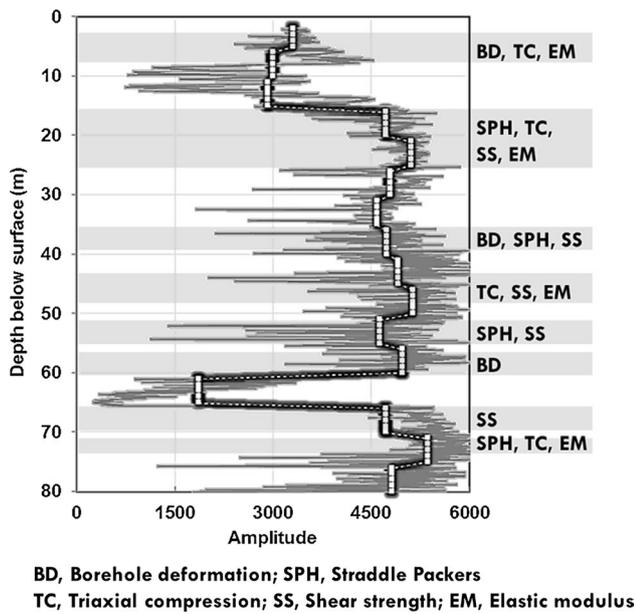
BD, Borehole deformation; SPH, Straddle Packers
TC, Triaxial compression; SS, Shear strength; EM, Elastic modulus

Fig. 5 Experimental depths of six in-situ and laboratory testing methods applied in this study

to estimate Young's static elasticity modulus (*E*) and static Poisson's ratio (*v*). Two types of tests were included, i.e,. the static EM test and the dynamic EM test. The first type determined the uniaxial compressive strength of rock core by using a tailored compression machine, and the strength-strain curve produced under the uniaxial state of stress was following the ASTM procedure 7012 and the suggested method of ISRM (Fairhurst and Hudson 1999). The second type estimated the dynamic moduli of rock based on the elastic wave velocity and density following the suggested method of ISRM (Bieniawski 1978).

The sampling and experimental depths of six in-situ and laboratory testing techniques applied in this study are illustrated in Fig. 5. With due consideration of spatial variability and uncertainty, the examined frequency in principle is set for every 10 m (33 feet) the borehole is advanced, and avoiding the section that is reported to be low in RQD. The estimated results of all above-mentioned hydraulic, deformation, elasticity, shear-strength parameters are summarized in Tables 1, 2, 3, 4, 5, 6 and 7. Clearly, except ATV logging data, all the in-situ and laboratory testing methods only give discrete data or regionally-averaged results.

Results and discussions

Exploratory data analysis

Firstly, attempting to provide an overview of the physical behavior of local geological formation, a contrastive analysis

Table 1 Summary of acoustic amplitude values and RQD index at different depth

Depth [m]	Acoustic amplitude	RQD
0–5	3308 ± 228	87–100
5–10	2992 ± 1058	40–85
10–15	2922 ± 1019	40–95
15–20	4722 ± 398	60–99
20–25	5109 ± 125	99–100
25–30	4793 ± 438	45–90
30–35	4588 ± 410	85–95
35–40	4733 ± 476	90–95
40–45	4907 ± 593	80–100
45–50	5136 ± 312	90–100
50–55	4625 ± 778	65–90
55–60	4976 ± 339	90–100
60–65	1862 ± 1176	20–95
65–70	4718 ± 1532	90–100
70–75	5536 ± 370	55–95
75–80	4758 ± 696	85–90

Acoustic amplitude range takes the arithmetic mean $\pm$ one standard deviation

RQD Rock Quality Designation index [%]

of all obtained parameters against depth in NH-01 is illustrated in Fig. 6, including the modulus deformation (BD-D) and modulus elasticity (BD-E) from borehole deformation test, hydraulic conductivity (SPH-*K*) and specific storage (SPH-*S*$_s$) from straddle-packer hydraulic test, peak cohesion (TC-*C*$_p$), peak and residual angle of friction (TC-ϕ_p, TC-ϕ_r) from triaxial compression test, apparent angle of friction (SS-ϕ_A) from shear strength test, static elastic modulus (SEM-*E*), dynamic elastic modulus (*E*$_d$) and shear modulus (*G*$_d$) from elastic modulus test.

As the results presented, the majority of estimates in NH-01 is consistent with the previous cognition and earlier findings. The estimated in-situ deformation and elastic modulus by borehole deformation test were found to increase with depth (associated with higher effective overburden stress). The increased magnitude of elastic modulus was much higher than deformation modulus, which corresponds to a rapid decrease of axial strain during the unloading process. The estimated hydraulic conductivity by straddle-packer hydraulic test was found to decrease with depth, ranging from 7.8E−5 m/s in the shallow depth (16.0–17.5 m), to 6.2E−10 m/s in the deeper horizon (73.0–74.5 m); the estimated magnitudes of specific storage were behaved differently, as shown, the estimated volume of water released from storage at the middle horizon was relatively low (≤ 0.000001).

The estimated peak cohesion and friction angle of the rock samples at different depths by triaxial compression test were observed between 4.5–7.9 Mpa and 49°–51°,
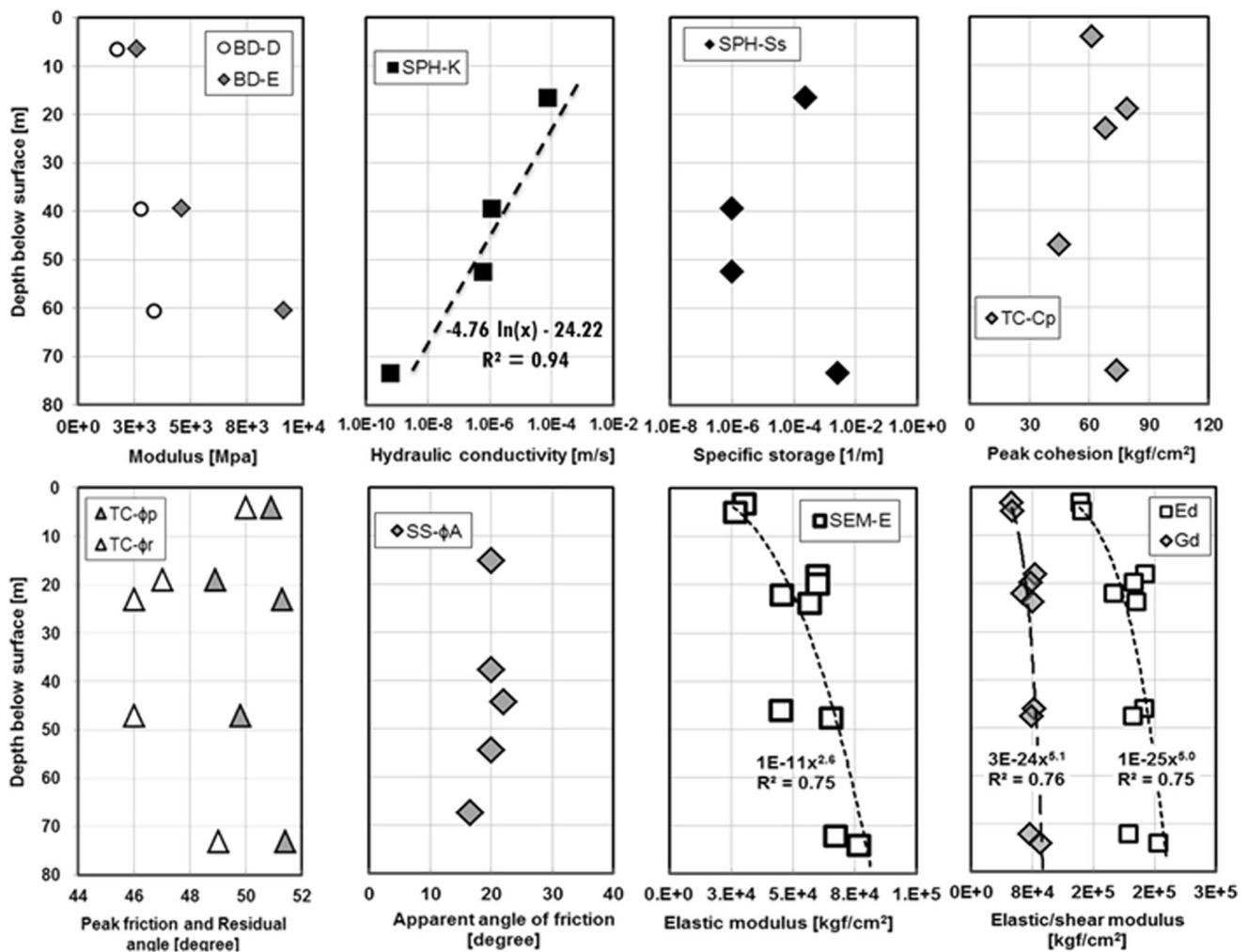
Fig. 6 The estimated parameters in NH-01 by different testing methodologies as a function of depth

Table 2 Summary of testing results from in-situ borehole deformation test

Depth [m]	Acoustic amplitude	Modulus deformation [MPa]	Modulus elasticity [MPa]
6–7	3701 ± 239	1706	2619
39–40	4678 ± 685	2782	4592
60–61	3527 ± 821	3338	9126

Table 3 Summary of testing results from in-situ straddle-packer hydraulic test

Depth [m]	Acoustic amplitude	Hydraulic conductivity, K [m/s]	Specific storage, S_s [1/m]
16–17.5	4668 ± 351	7.8E−5	2.3E−4
39–40.5	4769 ± 611	1.2E−6	1.0E−6
52–53.5	4335 ± 656	6.0E−7	1.0E−6
73–74.5	5418 ± 533	6.2E−10	2.7E−3

respectively. The residual cohesion of all samples was taken to be zero, and the residual angle was found to be less than the peak one. For the shear strength test, the magnitude of apparent cohesion was also taken to be zero and the apparent angle were found to be consistently around 20°–22°, except the sample obtained at the depth of 67.2–67.4 m showed a lower value of 16.5°, which could be associated with the effect of excess pore water pressure. By reviewing Fig. 5, a significant change regarding the amplitude of acoustic reflectivity is shown at the same depth, which accounts for a remarkable variation in formation hardness.

Estimation of the static modulus of elasticity shows that the magnitudes of elastic modulus were found to increase with depth, ranging from 26,400–29,000 kgf/cm^2 in the shallow depth (3.0–5.0 m), to 67,000–76,200 kgf/cm^2 in the deeper horizon (72.0–74.0 m), and the distribution of static Poisson's ratio ranged between 0.17 and 0.45. The same increasing tendency was also observed in the estimated magnitudes of dynamic elastic modulus (E_d) and shear modulus

Table 4 Summary of testing results from indoor triaxial compression test

Depth [m]	Acoustic amplitude	Peak cohesion C_p [kgf/cm^2]	Peak friction angle ϕ_p [°]	Residual angle of friction, ϕ_r [°]
3–5	3262 ± 257	61.0	50.9	49.9
18–20	4877 ± 148	79.1	48.9	47.1
22–24	5123 ± 129	68.0	51.3	46.2
46–48	5171 ± 280	44.7	49.8	45.7
72–74	5621 ± 309	73.8	51.4	49.1

Residual cohesion $C_r = 0$ [kgf/cm^2]

Table 5 Summary of testing results from indoor shear strength test

Depth [m]	Acoustic amplitude	Apparent angle of friction, ϕ_A [°]
15–15.2	3244 ± 267	20.0
37.6–37.8	4501 ± 346	20.0
44.3–44.5	3932 ± 544	22.0
54.3–54.5	4046 ± 1138	20.0
67.2–67.4	5218 ± 43	16.5

Apparent cohesion $C_A = 0$ [kgf/cm^2]

(G_d) from dynamic elastic modulus test, while the distribution of dynamic Poisson's ratio was found to be relatively homogeneous. On the whole, as consistent with the previous studies (Chen et al. 2009; Li and Popov 2018), in NH-01 a power law variation of the elastic and shear modulus with depth was obtained ($R^2 = 0.75$).

Cross comparison against acoustic amplitude

As the ATV logging enables continuous amplitude data collection along the depth, it is worthwhile to further investigate the specific correlation between all above-mentioned parameters and the corresponding acoustic amplitudes. By testing over the same datasets as reported from Tables 2, 3, 4, 5, 6 and 7, similar to the prior analysis for different parameters

in relation to depth, clear correlations are also found only between hydraulic conductivity estimated by straddle-packer hydraulic testing (SPH-K), static elastic modulus (SEM-E), dynamic elastic modulus (E_d), shear modulus (G_d) estimated from elastic modulus test, and the average acoustic amplitude values taken from the individual packed off section.

As shown in Fig. 7, the negative exponential correlation between SPH-K and the corresponding acoustic amplitude is presented. It implies that in NH-01 the section associated with a greater acoustic amplitude response is observed to be relatively low in hydraulic conductivity. Conversely, positive exponential correlations between elastic and shear modulus against the reflected acoustic amplitude are obtained. As can be seen in Fig. 8, both elastic and shear modulus are found to increase due to decrease in formation void ratio, which is a direct indication of the accumulation in loading pressure and as such the higher amplitude acoustic responses. The present correlation accounts for the fact that, in the current case, the variation of geometrical structure and acoustic amplitude is highly interlinked, which, to a certain degree, also implies the hydraulic conductivity nature of aquifer. This finding also affords us to expand the scope of thinking to explore the possible functional application of acoustic amplitude data.

Lastly, when re-visiting and evaluating the continuous ATV amplitude logging data as a function of depth, a holistic view in terms of the acoustic response (average of every

Table 6 Summary of testing results from indoor static elastic modulus test

Depth [m]	Acoustic amplitude	Elastic modulus E [kgf/cm^2]	Stress [kgf/cm^2]	Poisson's ratio v [dimensionless]
3.20–3.31	2948 ± 255	29,900	290.8	0.24
4.88–4.99	2938 ± 180	26,400	306.7	0.33
18.02–18.13	5041 ± 164	59,800	354.3	0.24
19.88–19.99	4890 ± 201	59,600	187.6	0.17
22.01–22.12	4909 ± 128	45,100	336.6	0.27
23.77–23.88	5139 ± 38	56,600	443.2	0.26
46.00–46.11	5359 ± 84	44,900	158.1	0.25
47.60–47.71	5134 ± 202	65,000	337.8	0.45
72.05–72.16	5625 ± 124	67,000	362.7	0.28
73.88–73.99	5245 ± 224	76,200	445.7	0.30

Table 7 Summary of testing results from indoor dynamic elastic modulus test

| Depth [m]	Acoustic amplitude	V_p [m/s]	V_s [m/s]	Dynamic elastic modulus, E_d [kgf/cm^2]	Dynamic shear modulus, G_d [kgf/cm^2]	Poisson's ratio v [dimensionless]
3.20–3.31	2948 ± 255	2946	1434	142,500	53,000	0.34
4.88–4.99	2938 ± 180	2955	1443	145,500	54,100	0.34
18.02–18.13	5041 ± 164	3837	1789	227,300	83,500	0.36
19.88–19.99	4890 ± 201	3901	1730	213,200	77,400	0.38
22.01–22.12	4909 ± 128	3688	1607	187,000	67,600	0.38
23.77–23.88	5139 ± 38	3673	1744	216,400	79,800	0.35
46.00–46.11	5359 ± 84	3934	1745	226,200	82,100	0.38
47.60–47.71	5134 ± 202	3387	1713	211,100	79,400	0.33
72.05–72.16	5625 ± 124	3455	1689	205,400	76,500	0.34
73.88–73.99	5245 ± 224	3876	1846	244,800	90,400	0.35
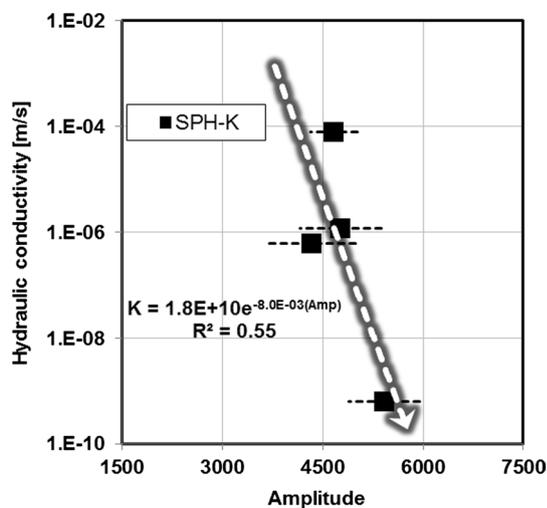
Fig. 7 The estimated hydraulic conductivity against acoustic amplitude in NH-01

5 m) of the formation configuration is shown. As shown in Fig. 9, to the current study case, the acoustic response pattern is found to be irregular, oscillated, and had no increasing tendency. Within the investigating range of 80 m, five flexure joints can be identified, including the depth of 10–15 m, 30–35 m, 50–55 m, 60–65 m, and 75–80 m at the bottom, at which the acoustic response are relatively lower. Geometrically speaking, they are the weakest points in the whole subsurface geological structure. Oppositely, with regards to the hydrogeological configuration, based on the correlation shown in Fig. 7, the presence of flexure joints should possibly be associated with a higher hydraulic conductivity to transmit water in comparison with the formation in the vicinity.

A holistic evaluation based on acoustic amplitude

With a view to analyzing the amplitude response curve as presented above, a better and proper interpretation would be needed in such case. A holistic evaluation on the geometry of rock formation by using ATV logging data, *i.e.,* the full reflection amplitude values, is proposed and classified by twofold considerations. In the first place, an internal evaluation can be made on the determination of baseline amplitude (as a benchmark reference), which should provide an a-priori indication on the intrinsic property of the targeted geometrical structure; the higher baseline amplitude, the higher intact rock strength properties. The magnitude of baseline amplitude can be calculated simply by taking the arithmetic mean of data across total length of drilling. In our case, for instance, the baseline amplitude is 4400, showed a trend for a moderately strong intact rock strength.

In the second place, an extrinsic evaluation is aimed at evaluating the degree of inflection for the full range of drill hole. The idea is derived from the empirical estimate of rock quality designation index (RQD), which gives a measure for the percentage of intact core pieces in the total length of core, and used to characterize the degree of jointing in a rock mass. In our study, a similar evaluation approach is dedicated to sum the lengths of all horizons that have higher amplitude response over the baseline, and divide it over the total length of drilling, the Aco-RQD (Acoustic-based Rock Quality Designation index) is therefore defined as the quotient showing below:

$$\text{AcoRQD} = \left(\frac{l_{\sum \text{ of horizons that has higher amplitude than the baseline value}}}{l_{\text{total length of drilling}}} \right) \times 100\%$$
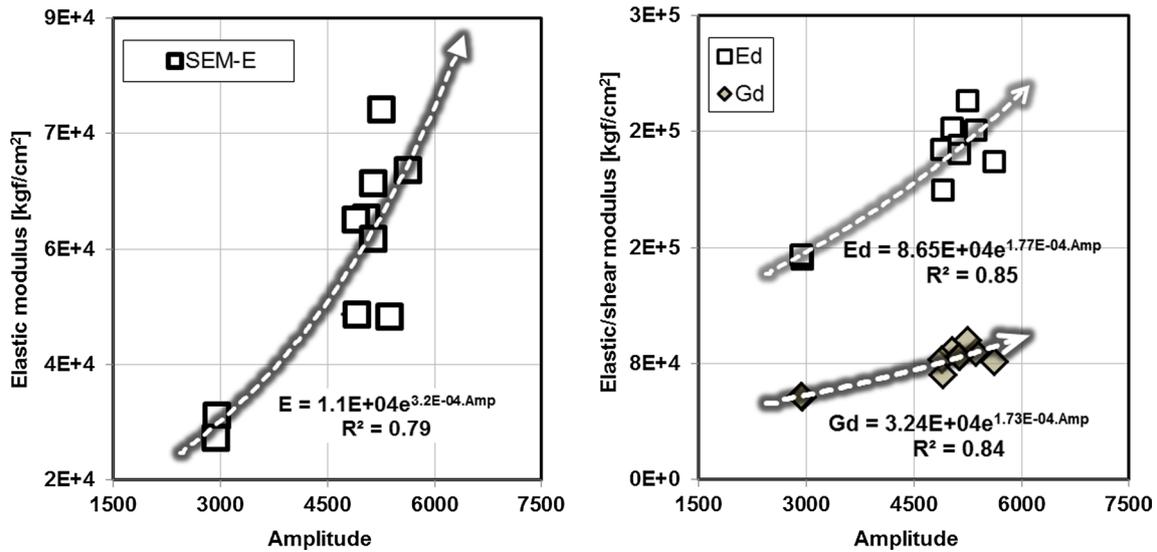
Fig. 8 The estimated elastic and shear modulus against acoustic amplitude in NH-01

A higher Aco-RQD value corresponds to a relatively stable and more rigid geological structure, which would likely contain less inflection sections, and a low Aco-RQD value stands for a porous and fractured one. As a necessary complement to the application of the RQD index, the calculation of Aco-RQD can be classified as follows: < 25% belongs to very poor rock mass structure, 25–50% poor rock mass structure, 51–75% moderate rock mass structure, 76–90% good intact rock mass structure, and 91–100% excellent intact rock mass structure. It can be easily applied in the field of engineering geology and all aspects of geotechnical

engineering. As long as the continuous ATV logging data are available to be received, a quick calculation is able to be undertaken on site without difficulty. By interpreting the result shown in Fig. 9, the Aco-RQD value in our case is 75%, which represents a moderately intact rock mass structure.

Such calculation provides an improved and holistic understanding for the geometry of rock mass under field condition. The regular and accustomed core-based RQD index is easily affected by core breakage upon handling and drilling performance, as such it only provides discrete subsurface information. With a full consideration to the geomechanical properties of rock, the proposed Aco-RQD approach provides a more objective evaluation for geotechnical investigation and design.
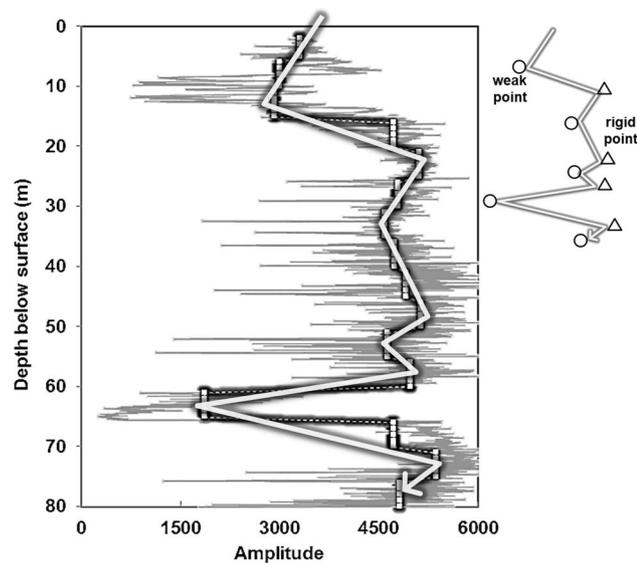
Conclusions and recommendations

Down-hole acoustic televiewer imaging (ATV logging) is a frequently used tool in geotechnical engineering and investigation. Except providing visual guides to the subsurface condition, which is already known from the published literature, this technique can also provide valuable information regarding rock-foundation performance in engineering practice.

The current study demonstrates the effectiveness of ATV logging that apparently has not yet been considered explicitly. A case study in southern Taiwan is taken up to demonstrate the presented approach, at which six commonly applied in-situ and laboratory testing techniques have been carried out. The significance of correlation between acoustic amplitude and different geotechnical parameters has been

Fig. 9 The variation of acoustic amplitude and response curve in NH-01

evaluated. The expected consistency as compared to the prior knowledge has been proven, which facilitates the introduction of a simple and reliable measure, Aco-RQD, for rating the degree of intact of a given rock mass geometry. It also accounts for the fact that ATV logging is able to provide timely and in-depth subsurface information in a field scale; if limited budget has been the main constraint, a holistic understanding of the rock mass structure is still possible when ATV logging data are accessible and meaningful. The result obtained in this study should be further validated in different geomechanical settings to expand its application.

Compliance with ethical standards

Conflict of interest The authors declare that they no conflict of interest.

References

ASTM D7012 (2014) Standard test methods for compressive strength and elastic moduli of intact rock core specimens under varying states of stress and temperatures. ASTM International, West Conshohocken

ASTM D4719 (2020) Standard test methods for preboredpressuremeter testing in soils. ASTM International, West Conshohocken

Bieniawski ZT (1978) Determining rock mass deformability: experience from case histories. Int J Rock Mech Min Sci GeomechAbstr. 15(5):237–247

Carnevale M, Domaracki M (2018) Limitations of acoustic televiewer measurements in large void spaces. In: Proceedings of the 31st symposium on the application of geophysics to engineering and environmental problems (SAGEEP), Nashville, Tennessee, USA

Cha SS, Lee JY, Lee DH, Amantini E, Lee KK (2006) Engineering characterization of hydraulic properties in a pilot rock cavern for underground LNG storage. EngGeol 84(3–4):229–243

Chang SP, Chuang SM (2001) Eutrophication study of twenty reservoirs in Taiwan. Water Sci Technol 44(6):19–26

Chen S, Yan C, Soh A (2009) Adhesive behavior of two-dimensional power-law graded materials. Int J Solids Struct 46(18–19):3398–3404

Chen HL, Tsou KW, Qiu CY (2014) Agricultural land use change in high risk area—a case study of Tsengwen and Nanhua Reservoir Watershed, Taiwan. Int Proc Chem Biol Environ Eng 79:47–51

Chou PY, Lo HC, Hsu SM, Lin YT, Huang CC (2012) Prediction of hydraulically transmissive fractures using geological and geophysical attributes: a case history from the mid Jhuoshuei River basin, Taiwan. Hydrogeol J 20(6):1101–1116

Comte JC, Cassidy R, Nitsche J, Ofterdinger U, Pilatova K, Flynn R (2012) The typology of Irish hard-rock aquifers based on an integrated hydrogeological and geophysical approach. Hydrogeol J 20(8):1569–1588

Deltombe JL, Schepers R (2004) New developments in real-time processing of full waveform acoustic televiewer data. J ApplGeophys 55(1–2):161–172

Dohnalik M, Kowalska S, Mikołajewski Z, Domonik A, Tabor Z (2015) Mineral composition of shales and the results of triaxial compression tests—a case study from the Ordovician and Silurian rocks of Poland. Nafta-Gaz 71(6):355–360

Fairhurst CE, Hudson JA (1999) Draft ISRM suggested method for the complete stress-strain curve for intact rock in uniaxial compression. Int J Rock Mech Min Sci 36(3):279–289

Gwynn XP, Brown MC, Mohr PJ (2013) Combined use of traditional core logging and televiewer imaging for practical geotechnical data collection. In: Proceedings of the international symposium on slope stability in open pit mining and civil engineering, Australian Centre for Geomechanics, pp 261–272

Hamm SY, Kim M, Cheong JY, Kim JY, Son M, Kim TW (2007) Relationship between hydraulic conductivity and fracture properties estimated from packer tests and borehole data in a fractured granite. EngGeol 92(1–2):73–87

ISRM (1978) Suggested methods for determining the strength of rock materials in triaxial compression. Int J Rock Mech Min Sci GeomechAbstr. 15(2):47–51

ISRM (1981) Suggested methods for determining shear strength. In: Brown ET (ed) Rock characterization testing and monitoring. Pergamon Press, Oxford, pp 129–140

Jo Y, Chang C, Ji SH, Park KW (2019) In situ stress states at KURT, an underground research laboratory in South Korea for the study of high-level radioactive waste disposal. EngGeol 259:105198

Kennel JR, Parker BL (2019) Acoustic televiewer amplitude data for porosity estimation with application to porewater conversion. GeolSocLond Spec Publ 479(1):177–185

Ku CY, Hsu SM, Chiou LB, Lin GF (2009) An empirical model for estimating hydraulic conductivity of highly disturbed clastic sedimentary rocks in Taiwan. EngGeol 109(3–4):213–223

Li Q, Popov VL (2018) Boundary element method for normal non-adhesive and adhesive contacts of power-law graded elastic materials. Comput Mech 61(3):319–329

Lin HH, Lin ML, Lu JH, Chi CC, Fei LY (2020) Deep-seated gravitational slope deformation in Lushan, Taiwan: transformation from cleavage-controlled to weakened rockmass-controlled deformation. EngGeol 264:105387

Liu WF, Leung YF (2018) Spatial variability of saprolitic soil properties and relationship with joint set orientation of parent rock: insights from cases in Hong Kong. EngGeol 246:36–44

Lo HC, Chou PY, Hsu SM, Chao CH, Wang CT (2012) Using borehole prospecting technologies to determine the correlation between fracture properties and hydraulic conductivity: a case study in taiwan. J Environ Eng Geoph 17(1):27–37

Nastev M, Morin R, Godin R, Rouleau A (2008) Developing conceptual hydrogeological model for Potsdam sandstones in southwestern Quebec, Canada. Hydrogeol J 16(2):373–388

Okko O, Front K, Anttila P (2003) Low-angle fracture zones in rapakivi granite at Hästholmen, southern Finland. EngGeol 69(1–2):171–191

Paillet FL, Crowder RE (1996) A generalized approach for the interpretation of geophysical well logs in ground-water studies—theory and application. Groundwater 34(5):883–898

Peyras L, Rivard P, Breul P, Millet A, Ballivy G (2015) Characterization of rock discontinuity openings using acoustic wave amplitude—application to a metamorphic rock mass. EngGeol 193:402–411

Shieh CL, Wang CM, Chen YS, Tsai YJ, Tseng WH (2010) An overview of disasters resulted from Typhoon Morakot in Taiwan. J Disaster Res 5(3):236–244

Tezuka K, Niitsuma H (2000) Stress estimated using microseismic clusters and its relationship to the fracture system of the Hijiori hot dry rock reservoir. EngGeol 56(1–2):47–62

Thill RE, D'Andrea DV (1976) Acoustic core logging in blast-damaged rock. EngGeol 10(1):13–36

Wang M, Kulatilake PHSW, Panda BB, Rucker ML (2001) Groundwater resources evaluation case study via discrete fracture flow modeling. EngGeol 62(4):267–291

Wang HW, Kondolf M, Tullos D, Kuo WC (2018) Sediment management in Taiwan's reservoirs and barriers to implementation. Water 10(8):1034

Williams JH, Johnson CD (2004) Acoustic and optical borehole-wall imaging for fractured-rock aquifer studies. J ApplGeophys 55(1–2):151–159

Zhang C, Chen Q, Qin X, Hong B, Meng W, Zhang Q (2017) In-situ stress and fracture characterization of a candidate repository for spent nuclear fuel in Gansu, northwestern China. EngGeol 231:218–229

Zhao XG, Wang J, Cai M, Ma LK, Zong ZH, Wang XY, Su R, Chen WM, Zhao HG, Chen QC, An QM, Qin XH, Ou MY, Zhao JS (2013) In-situ stress measurements and regional stress field assessment of the Beishan area, China. EngGeol 163:26–40

RESEARCH ARTICLE - HYDROLOGY

Effects of anisotropic changes on the conductivity of sedimentary aquifers, southeastern Niger Delta, Nigeria

N. J George[1] · **N. E Bassey[2]** · **A. M Ekanem[1]** · **J. E Thomas[1]**

Abstract

Sundry soils/rocks are characterized by electrical properties with clearly or obscurely expressed anisotropies. These anisotropic effects may be low, moderate or high depending on the coefficient of anisotropy (λ). The vertical electrical sounding technique employing Schlumberger electrode configuration and lithological information from boreholes were deployed to characterize the lithological diversity in homogeneous and anisotropic geologic units that serve as aquifer systems and their overlaying layers in the coastal region of Akwa Ibom State. Based on the λ, the assessed volume of sedimentary formation is classified into low anisotropy $\lambda < 1.2$, moderate anisotropy ($1.2 < \lambda \leq 1.3$) and high anisotropy ($\lambda > 1.3$) with alluvium (64.3%), inter-bedded shale and sandstone (14.3%) and shale and slate (21.3%). The estimated percentage of respective compositional coverage indicates that alluvium is dominant, while the blended inter-bedded shale and sandstone as well as the shale and slate are minor geologic units in the Benin Formation. Inferred index of spread of alluvium indicates that the homogeneous and anisotropic units assessed are intrinsic/microscopic in nature as identified by the impressed current that passed through geologic system. The results showcased that the plot between the strike-dependent resistivity $\left(\rho_\theta\right)$ at arbitrary chosen strike and geometric mean resistivity $\left(\rho_m\right)$ can be used as a yardstick for inferring the degree of consistency of geologic compositions in homogeneous and anisotropic media. Both the quantitative (graphic) and qualitative (contour) results portend the thin possibility of having anisotropy free geologic units. The finding proposes that ground resistivity measurements and interpretations of geologic structures should be constrained by borehole information in order to firm up the intended plans for obtaining clearer, defendable and well-resolved subsurface structures.

Keywords Anisotropy · Hydrogeological units · Diversity · Strike-dependent resistivity · Sensitivity

Introduction

Most ground inhomogeneity and anisotropy induce strong directional effects on resistivity distributions and sounding curves realized from measurements (Caglar and Avsar 2007). Lithological diversity, a dynamic process, which causes electrical anisotropy of soils/rocks, is an effect of alternating beddings of arenites and argillites within the subsurface. The arenite–argillite sequence can be regarded as a medium of anisotropic diversity characterized by vertical and horizontal resistivities. Such a medium has a conductivity sensitivity, which determines the shapes of geo-electric curves. Conductivity sensitivity is the ease of soil layer/bed to conduct current that passes through them. Specifically, geo-electric curves give insights into the local geology and structure of subsurface units penetrated by the injected current (Ekanem et al. 2019; Thomas et al. 2020). As the current finds penetrates the earth, information on the primary

✉ N. J George
nyaknnojimmyg@gmail.com; nyaknogeorge@aksu.edu.ng

N. E Bassey
nsikakbassey@aksu.edu.ng

A. M Ekanem
anny4mart@yahoo.com

J. E Thomas
jewelemem@gmail.com

[1] Department of Physics (Geophysics Research Group), Akwa Ibom State University, PMB 1162, Mkpat Enin, Uyo, Nigeria

[2] Department of Geology, Akwa Ibom State University, PMB 1162, Mkpat Enin, Uyo, Nigeria

geo-electrical indices, such as layer thickness (h), layer resistivity (ρ) and their corresponding depths (D) can be determined. Layer thickness is generally estimated from the depths of the layer parameters, which are initially gauged from the available VES curves. These primary geo-electric indices are estimates of importance and are used in evaluating transverse resistance ($\rho \cdot h$) and longitudinal conductance $\left(\frac{h}{\rho}\right)$ according to Hasan et al. (2019). These secondary geo-electric indices are called Dar Zarrouk parameters (DZP). Natural soils of various types have different electrical properties due to the composition, structure, water content, and temperature (George et al. 2010; Obiora et al. 2015). When the water content is below its optimum, the electrical conductivity of soils increases nonlinearly and the variation rate increases dramatically. However, when the water content is optimum, the degree of saturation, or dry density increases to a certain value and the electrical conductivity tends to be a constant. In addition, soil electrical conductivity increases with the increase in temperature, and it is observed that the electrical conductivity decreases with the increase in the number of wetting–drying cycles (Wei et al. 2013). Anisotropy is simply a condition where the earth resistivity/conductivity and hence the measured apparent resistivity/voltage is a function of the path of measurement. This anisotropy is common in clay, slate and shale, which are marked by divergent lineation or platy fabric (Greenhalgh et al. 2010). Anisotropy can exist in a macroscopic scale (Bala and Cichy 2015). Here, bands or a series of layers of divergent isotropic materials behave as they are single and equivalent anisotropic units. When anisotropy exists in a particular geologic units or minerals, it is referred to as micro-anisotropy or intrinsic anisotropy, which peculiarly depends on the crystal symmetry or texture of the material (Bala 2011). Besides the micro- and macro-anisotropies, rock cleavage, joints and fracturing, can also produce structural anisotropy peculiar to layers with alternating conductivities (rock and joint fill) Hobbs et al. (2009) and Bala and Cichy (2015). Such pseudo-anisotropy arises when the thickness of the individual isotropic bands or units is small relative to the electrode separation used for the measurement (Greenhalgh et al. 2010). The earth electrical specific resistance/resistivity deviates with the course or direction. These deviations are amply necessitated by the diversity of lithological units seen in the degree of electrical anisotropy. Permeability may be anisotropic when it changes in soils/rocks in different directions (Aissaoui et al. 2019). The effect of anisotropy exists in geologic units and if ignored, would lead to error in the interpretation of ground resistivities and the geologic structures. Some publications, which address the effects of anisotropy on surface resistivity measurements, include: Matias and Habberjam (1986), Matias (2002), Ekanem (2020). Borehole information remains the reliable means of constraining resistivity data interpretation (Asten 1974; Olasehinde and Bayewu 2011). This implies that the subsurface exhibits different electric properties in different directions (Karanth 1987; Karnkowski 1999; Yeboah-Forson and Whitman 2013). This study aims at exploring the veracity of electrical technique and geologic information in delineating the electrical conductivity sensitivity of sedimentary layers for effective understanding of the contribution of electrical resistivity in hydrogeophysical studies.

Theoretical insights

The diversity of specific resistance is based on 2 intrinsic anisotropic parameters, which are anisotropic coefficient (λ) and the average resistivity $\left(\rho_m\right)$ associated with the units of soils or rock considered (Aissaoui et al. 2019). Anisotropic diversity exists in many sedimentary formations due to permeability diversity (Bala and Cichy 2015). In homogeneously isotropic media of vertical resistivity $\left(\rho_v\right)$ and horizontal resistivity $\left(\rho_h\right)$ coefficient of anisotropy is equal to 1 as $\rho_m = \rho_h = \rho_v$ and lies between 1 and 2 for homogeneously anisotropic media characterized by anisotropy (Caglar and Avsar 2007). The coefficient of anisotropy in homogeneous and anisotropic geological unit is greater unity, because resistivity is always greatest in the transverse direction. A homogeneously isotropic unit filled with water, air or rockfield fractures is equally considered to demonstrate anisotropy. Practically, electrical current flowing perpendicularly to the soil bedding planes is largely opposed by high resistance (vertical resistivity) due a scenario that looks like series combination of resistors. Along the bedding planes, opposition to the current flow becomes comparatively reduced due to high resistance (horizontal resistivity) caused by a scenario that looks like a parallel combination of resistors. The former scenario has a higher resistivity than the later and hence explains the reason why vertical resistivity is proudly greater than the horizontal resistivity. In a formation that is homogeneous and anisotropic, with dipping angle, θ between the layer and the vertical to the direction of measurement, the electric field generates a potential difference (U) at an arbitrary point of the medium, which decreases inversely proportionally to the distance (r) from the source of the direct current, I as stated in Eq. 1

$$U = \frac{I \cdot \rho_m}{4\pi \cdot r \cdot \left(1 + \left(\lambda^2 - 1\right) \cdot \cos^2 \theta\right)^{\frac{1}{2}}} \tag{1}$$

where ρ_m is the geometric mean resistivity, the factor λ, is anisotropic coefficient defined by ρ_h and ρ_v, respectively, representing horizontal/longitudinal (minimum) resistivity and vertical/transverse (maximum) resistivity. Anisotropic coefficient (λ) is related with ρ_h and ρ_v in Eq. 2:

$$\lambda = \left(\frac{\rho_v}{\rho_h}\right)^{\frac{1}{2}} \tag{2}$$

The geometric mean resistivity (ρ_m) is gauged from the horizontal and vertical resistivities as given in Eq. 3. Practically, ρ_v is the electric resistivity measured vertically to the bedding, while ρ_h is the electric resistivity measured along the bedding.

$$\rho_m = \left(\rho_v \cdot \rho_h\right)^{\frac{1}{2}} \tag{3}$$

Based on the formation values of ρ_m and λ the electric potential difference depends on the strike direction, θ along which the potential difference changes with r. In collinear array, Kunz and Moran (1958); Dachnov (1967); Greenhalgh et al. (2009), theorized that a uniform and anisotropic medium is characterized by the measured resistivity $\rho(\theta)$ which depends on the anisotropic parameters given in Eq. 4:

$$\rho(\theta) = \frac{\lambda \cdot \rho_h}{\left(1 + \left(\lambda^2 - 1\right) \cdot \cos^2\theta\right)^{\frac{1}{2}}} = \frac{\rho_m}{\left(1 + \left(\lambda^2 - 1\right) \cdot \cos^2\theta\right)^{\frac{1}{2}}} \tag{4}$$

The expression above shows that $\rho_m = \lambda \cdot \rho_h$. When $\theta = 0°$ and $\theta = 90°$, respectively, Eqs. 5 and 6 result from Eq. 4

$$\rho_\theta^v = \frac{\lambda \cdot \rho_h}{\lambda} = \rho_h \tag{5}$$

$$\rho_\theta^h = \lambda \cdot \rho_h \tag{6}$$

Interestingly, Eqs. 5 and 6 indicate that $\rho_\theta^v < \rho_\theta^h$ by a factor of λ. This means that the apparent resistivity (ρ_θ^v) measured normal or vertical to the strike orientation is less than apparent resistivity (ρ_θ^h) measured along the strike direction. The observed inference is in opposition to the fact that the true resistivity (ρ_v) of anisotropic formation, normal to its stratification, is greater than the true resistivity (ρ_h), parallel to the plane of stratification. This condition is referred to as "the paradox of anisotropy" (Aissaoui et al. 2019).

Description of location and geology of the study area

The site of the survey has an area coverage of about 435 km^2 within the coastal fringe of Akwa Ibom State, southeastern Nigeria. The locations of soundings fall within latitudes 4°45′–4°35′ N and longitudes 7°30′–8°10′ E in the hinterland off the Atlantic coast (Fig. 1). The area has a semi-temperate climate with distinct seasons: wet and dry seasons. The wet season commences in April and ends in September, while the dry season begins from October and ends in March. The temperature during the dry season is about 2 °C–32 °C. Annual rainfall for the area ranges between 200 and 250 cm, and can ascend to ~ 320 cm at the peak period. The area,

Fig. 1 Schematic map showing **a** the location of Akwa Ibom State, which indicate the study area and **b** the study area showing the local geology, VES points, borehole cored sample points, borehole locations and the local government boundaries
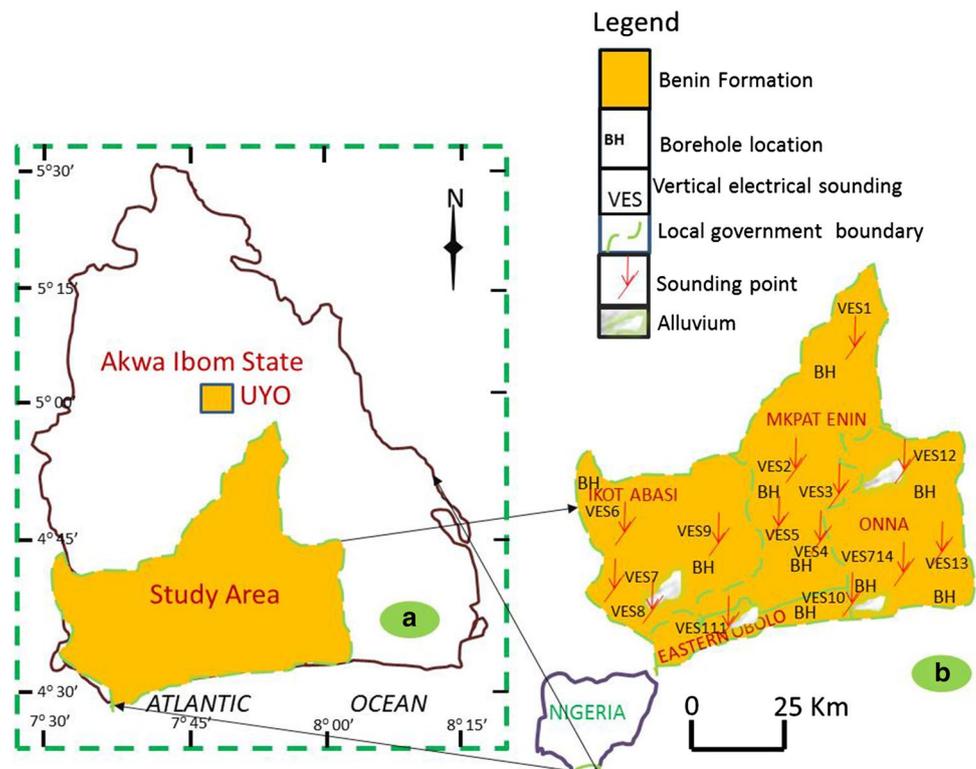
characteristically noted for its mangrove swamp is economically rich in crude oil and agricultural productions.

The area is geologically described by 3 lithostratigraphic formations. The oldest is Akata Formation (Eocene to recent) according to Peters et al. (1989), Akpan et al. (2013) and Ibanga and George (2016). This Formation exists as prodelta facies and serves as source rock for crude oil (Short and Stauble 1967). It is opined that the shales of this formation were formed during the initial development stages of Niger Delta Basin progradation and they are compacted and over-pressured with diapiric structures such as shale swells and ridges, which intrude into the overlying younger Agbada Formation. The Agbada Formation occurs throughout Niger Delta clastic wedge as the main reservoir and seal for crude oil accumulation in the basin. It is primarily paralic deltaic front facies with maximum thickness of about 13,000 feet or 4 km (Sort and Stauble 1967). The lithologies consist of alternating system of sands, silts and shales, arranged within about 10–90 feet successions and defined by progressive upward changes in grain size and bed thickness. The depositional environment of the Agbada Formation is interpreted to be fluvial–deltaic (Short and Stauble 1967; George et al. 2015). The base of the formation extends beyond 4.5 km or 4600 feet in some areas and is defined as earliest marine shale, while the surficial parts of the formation consist entirely of non-marine sand deposited in alluvial or upper coastal plain environments during progradation of the delta (Doust and Omatsola 1989). The youngest Benin Formation in which geologic units are considered in this work is the top part of the basin clastic wedge, from the Benin–Onitsha area in the north to the coastline in the south (Short and Stauble 1967). The subsurface structure in the Niger Delta Basin spans from simple rollover faults, multiple growth faults, antithetic faults and collapsed crest faults (Onuoha and Dim 2017). These sediments formed during the Late Eocene to Early Oligocene with the reservoirs mainly controlled by pre-and syn-sedimentary tectonic elements that responded to variable rates of subsidence and sediment supply (Doust and Omatsola 1989). The hydrogeological units here are mainly brownish and believed to be developed from moderately coarse textured alluvium. Habitually, the geologic units have grayish brown, and slightly finer texture occasionally intercalated with grits.

Materials and methods

The employment of vertical electrical sounding (VES) method whose interpretation is constrained by logged borehole information as geophysical prospecting tool is a standard geophysical prospecting method. Fourteen VES data (Fig. 1) were acquired near borehole in the study area to reduce the non-uniqueness problems associated with its

interpretation and at the same time to comfortably assess the deeper subsurface sedimentary electrical resistivity information. The Schlumberger electrode configuration was used to assess vertical and longitudinal electrical conductivity distributions at shallower and deeper depths through the use of IGIS signal enhancement resistivity meter SSP-MP-ATS resistivity meter. Even though the resistivity meter has the capability to average up to 32 cycles of values, measurement cycles were truncated after 4 stacks, provided the reading on the liquid crystal display correlates well with standard deviation < 10% when direct currents were artificially injected into the earth (Yadav and Singh 2007; George 2020). Half of the current electrode separation (AB/2) ranged between 1 and 150 m and that of the potential electrode (MN/2) varied from 0.25 to 30 m. The variation in the potential electrode separation was necessary to enhance the input signal strength. When the VES field survey was completed, the VES data were modeled using the RESIST code developed by Vander Velpen and Sporry (1993) and Vander Velpen (1988). Lithological data from nearby wells were used to constrain the 1D inverse models. After a couple of iterations, a satisfactory variation between the observed field data and theoretical data were represented by means of absolute root-mean square error, which was generally found to be less than 10% (Fig. 2).

Results and discussion

Primary and secondary geo-electrical indices

The results of the VES investigation, gave a 1-D electrical resistivity, thickness and depth of geo-electrical layers given in Table 1. The current at its maximum current electrode separations penetrated 4 sedimentary layers characterized with resistivity ranges of 95.2–3455.5 Ωm and mean of 834.8 Ωm in the first layer. The second layer ranged from 8.8 to 3606.6 Ωm with mean of 1327.8 Ωm. The third and fourth layers have ranges of 72.5–1464.5 Ωm with mean 657.2 Ωm and 117.1–4893.0 Ωm with mean 1551.2 Ωm, respectively. The 4th layer was not adequately delineated as current could not penetrate this layer at the maximum current electrode separation. With the use of lithological logs from nearby boreholes, VES data were manually and electrically modeled. The statistics of each of them showed characteristic four geo-electric layers with A, HA, KQ KH and HK cluster of curves. Noticeably, due to the information from wells, resistivity values and sizable thickness of geologic units for accumulation of water in the geologic aquifers, layers 2 and 3 were identified as potential aquifers. The respective inferred thicknesses and depths vary across the sounding points with ranges 0.5–19.6 m, 1.6–56.7 m, 16.1–89.2 m and 0.5–19.6 m, 2.1–76.3 m, 20.5 m for layers 1, 3 and 3. Table 1 indicates thickness soars with
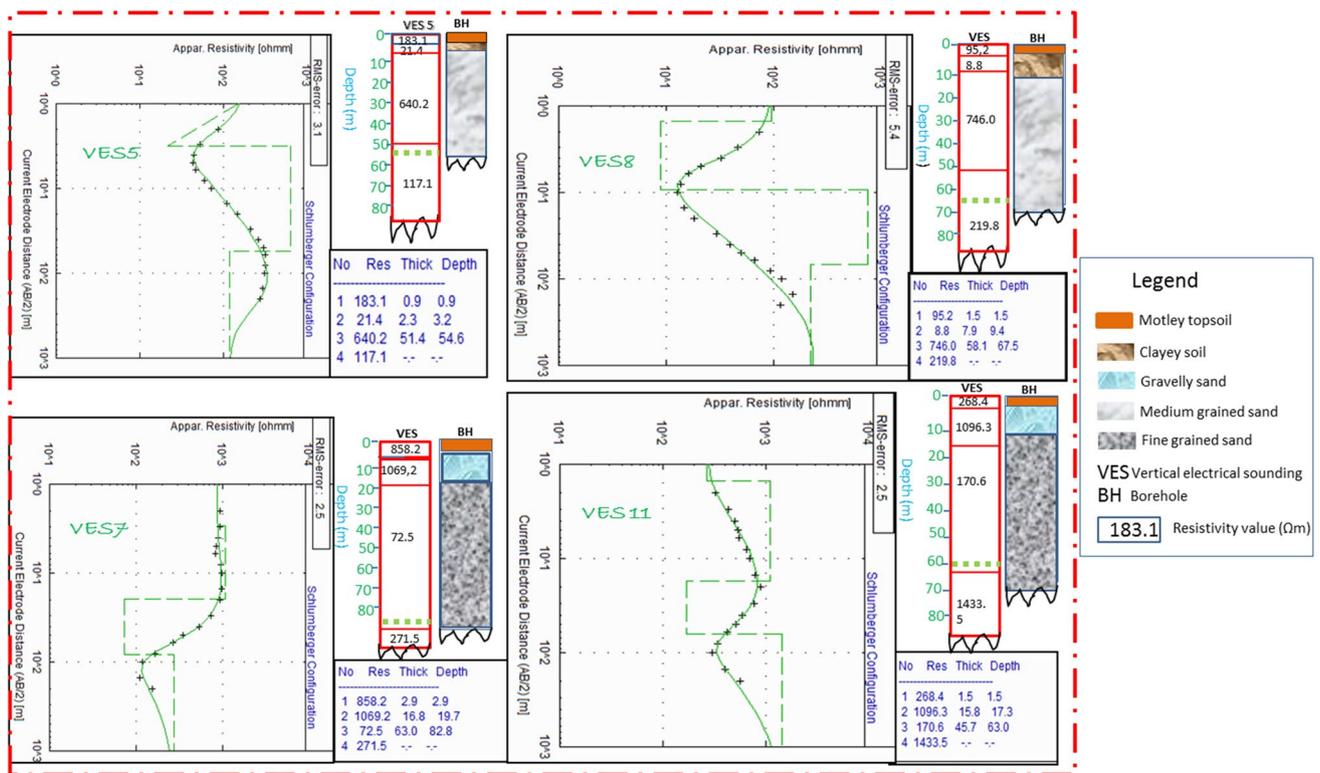
Fig. 2 Correlation of VES curves and the nearby lithology from borehole data

depth as the mean values attest. The geologic units penetrated by currents have noticeable electrical sensitivity of varying magnitudes due to intra and inter-bedded lithological diversity (Ibuot et al. 2013; George and Ekanem 2019). The hypothetical model of the subsurface in Fig. 3 identifies 2 distinct resistivities within the conducting earth units. These are the vertical (ρ_v) resistivity and the horizontal resistivity (ρ_v) estimated using Eqs. 7 and 8:

$$\rho_v = \frac{\sum_i^n h_i \rho_i}{\sum_i^n h_i} \tag{7}$$

$$\rho_h = \frac{\sum_i^n h_i}{\sum_i^n \frac{h_i}{\rho_i}} \tag{8}$$

where h_i, and ρ_i are the layer thickness and resistivity, respectively, for a given nth layer. These 2 directional parameters, defined by resistivity and thickness are directly proportional to each other. For homogeneous and anisotropic media, they are linearly related to each other by the square of average coefficient of anisotropy (Fig. 4). For the volume of homogenous and anisotropic medium that current passed through, the adjusted gradient ($\tan\theta = 0.9509$; $\theta \approx 44°$) for the line in

Fig. 4 is characterized by a highly correlated ($R = 0.8844$) expression given below:

$$\rho_v = 0.9508\rho_h + 340.78 \tag{9}$$

Coefficient of anisotropy

Although the medium has adjusted average approaching unity, the geological formation is not adjudged isotropic, because the estimated values of λ ranged from values 1.00 to 2.12 with 1.22 as the mean value. Besides, the scattering in Fig. 4 shows that the formation in the layers involved is inhomogeneous and anisotropic. The disparity as noted in Fig. 5, is a valid suggested reason for departure of λ from unity. Due to homogeneous and anisotropic disparities in the assessed geological units, classification of anisotropy in the area comparatively spans from low anisotropy ($\lambda < 1.2$), moderate anisotropy ($1.2 < \lambda \leq 1.3$) and high anisotropy ($\lambda > 1.3$) with percentage of 64.3, 14.3 and 21.3%, respectively. The lower disparity between the values of ρ_v and ρ_h is tended toward lower values of λ (Fig. 5). Hence, it can be inferred that on the average, anisotropy in the area is intrinsically or microscopically induced by virtue of the consequent sequences of

Table 1 Summary of geophysics survey in the study

VES	Location	Coordinate degree		No of Layer	Layer resistivity (Ohm m)				Layer thickness (m)			Layer depth (m)			Curve type
		Lat	Long		ρ_1	ρ_2	ρ_3	ρ_4	h_1	h_2	h_3	D_1	D_2	D_3	
1	MkpatEnin	4.7752	7.7854	4	666.5	3606.6	1054.6	375.7	5.3	24.9	32.4	5.3	30.2	62.6	KQ
2	MkpatEnin	4.7345	7.7733	4	401.6	1108.4	481.9	1195.9	1.1	27.5	50.5	1.1	28.6	79.1	KH
3	MkpatEnin	4.7034	7.8758	4	395.7	1567.1	510.3	1808.6	0.6	7.9	49.3	0.6	8.5	57.8	KH
4	MkpatEnin	4.6067	7.8167	4	660.4	1488.6	549.2	1248.7	1.3	3.6	47.8	1.3	4.9	52.7	KH
5	MkpatEnin	4.6535	7.7332	4	183.1	21.4	640.2	117.1	0.9	2.3	51.4	0.9	3.2	54.6	HK
6	IkotAbasi	4.6984	7.5511	4	949.9	2715.5	1464.5	3745.2	0.5	3.9	16.1	0.5	4.4	20.5	KH
7	IkotAbasi	4.6117	7.6317	4	858.2	1069.2	72.5	271.5	2.9	16.8	63.0	2.9	19.7	82.8	KH
8	IkotAbasi	4.5767	7.5684	4	95.2	8.8	746.0	219.8	1.5	7.9	58.1	1.5	9.4	67.5	HK
9	IkotAbasi	4.6184	7.7086	4	718.4	1844.9	1362.2	1440.8	0.5	1.6	80.8	0.5	2.1	82.9	KH
10	Eastern Obolo	4.5453	7.6359	4	268.4	1096.3	170.6	1433.5	1.5	15.8	45.7	1.5	17.3	63.0	KH
11	Onna	4.7172	8.0167	4	1721.4	118.2	611.2	4893.0	19.6	56.7	36.8	91.6	76.3	113.1	HA
12	Onna	4.5689	8.0278	4	3455.5	1825.6	393.9	2030.4	0.6	2.0	89.2	0.6	2.6	91.8	HA
13	Onna	4.5997	7.8832	4	1157.0	1318.3	735.6	1585.9	1.1	37.0	47.8	1.1	38.1	85.9	KH
14	Onna	4.6629	8.0098	4	156.4	800.8	407.8	1350.1	1.3	9.1	47.9	1.3	10.4	58.3	KH
Range					95.2–3455.5	8.8–3606.6	72.5–1464.5	117.1–4893.0	0.5–19.6	1.6–56.7	16.1–89.2	0.5–91.6	2.1–76.3	20.5–113.1	
Mean					834.8	1327.8	657.2	1551.2	2.76	15.5	51.2	7.9	18.3	69.5	

Fig. 3 Hypothetical geophysical model of a homogeneous and anisotropic earth due vertically dipping beds or fractures. Transverse/vertical resistivity (ρ_v) is perpendicular to the bedding planes or fractures, while longitudinal/horizontal resistivity (ρ_h) is parallel to the bedding planes or fractures; θ, strike direction
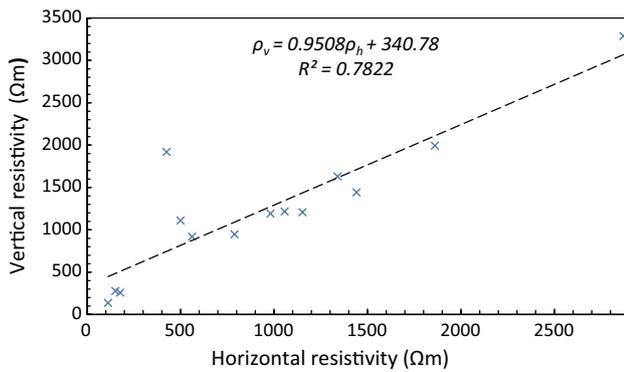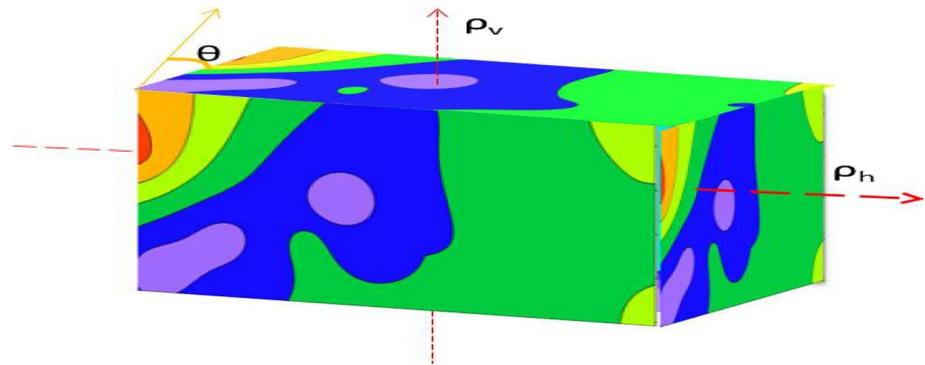

Fig.4 A graph showing the transverse/vertical resistivity against longitudinal/horizontal resistivity

argillaceous intercalations with arenaceous sedimentary units (Greenhalgh et al. 2010). Due to sensitivity of λ to geological diversity, very narrow and typical values of λ

are associated with divergent geologic units. Based on the low coefficient of anisotropy ($\lambda < 1.2$), which covers 64.3%, the geologic units that current passed through are mainly alluvia (intercalation of clay, silt or gravel of sedimentary origin) according to the range provided by (Hill 1972; Asten 1974; Greenhalgh et al. 2010). Comparatively, moderate range of coefficient of anisotropy ($1.2 < \lambda \leq 1.3$) and elevated range ($\lambda > 1.3$), with respective 14.3 and 21.3% coverage also define the study area. These geologic units are, respectively, described as interbedded shale and sandstone as well as shale and slate by Greenhalgh et al. (2010) and are all minor formations as their percentages comparatively reflect. The absence of extremely high values of coefficient of anisotropy ($\lambda > 2.1$) indicates that there is no intrusion of metasediments into the young sedimentary age from the deeply seated basement (Greenhalgh et al. 2010). The degree of anisotropy within the unit area is well expressed in the shape of longitudinal conductance versus transverse resistance curves for layers 1–3, respectively, as revealed in Fig. 6a–c. As

Fig. 5 A diagram showing the disparity between vertical/transverse and horizontal/longitudinal resistivities within the range of coefficient of anisotropy
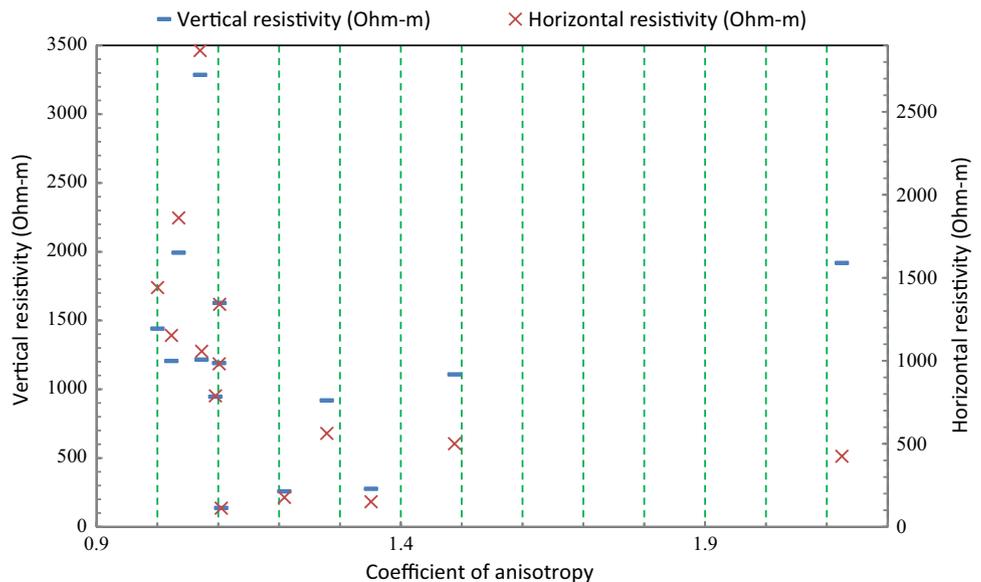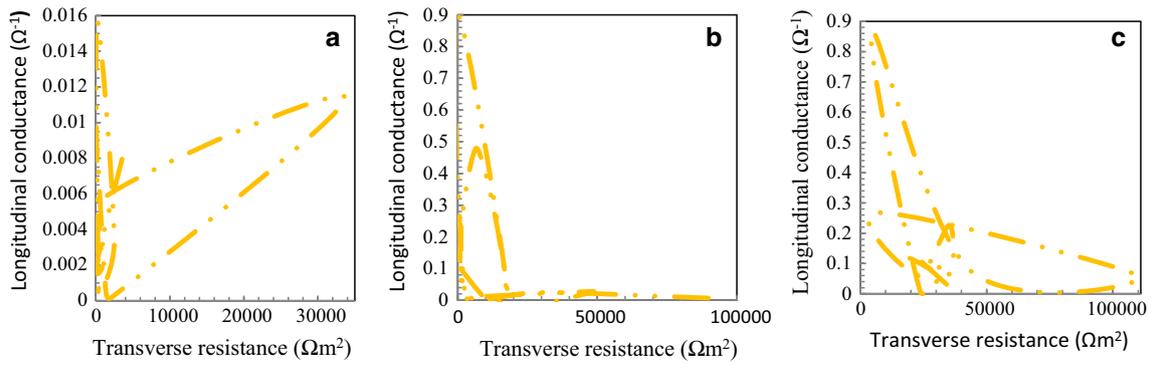
Fig. 6 Diagram showing variation of longitudinal conductivity versus transverse resistance for (**a**) layer 1, (**b**) layer 2 and (**c**) layer 3

observed in Fig. 6a–c, the values of DZP soar with depth and this showcases that thickness increases with depth. In 6a, the longitudinal conductance versus transverse resistance distribution is completely discordant with Fig. 6b and c, which are fairly consistent. The fair consistency accounts for cohesivity which is a function of depth of burial of geologic formation (George 2020).

Strike-dependent resistivity

The family of fairly collinear lines in Fig. 7 shows the relationship between resistivity, ρ_θ estimated as function of strike angle (Eq. 4) against the estimated geometric mean resistivity, ρ_m of homogeneous and anisotropic formation. The line can be summarily represented as Eq. 10:

Fig. 7 Graphs showing family of lines for strike dependence resistivity variations against mean resistivity
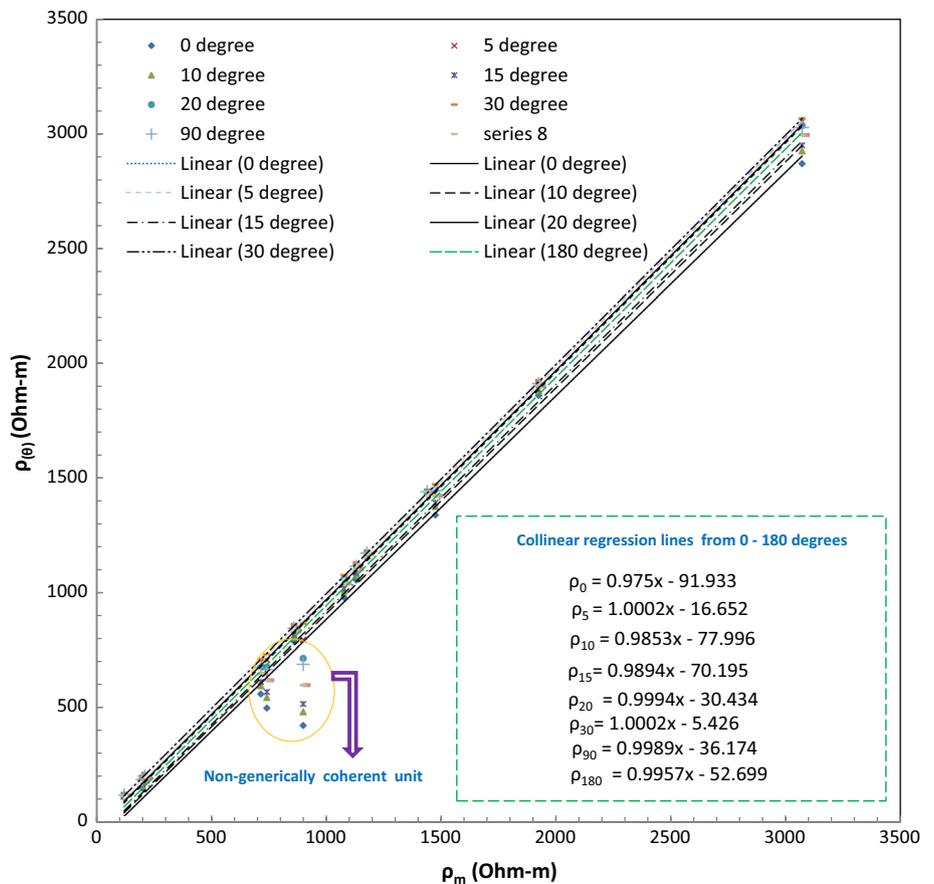
$$\rho_\theta = a\rho_m - b \qquad (10)$$

where a and b are constants representing slope and intercept. Due to consistency of intrinsic or microscopic anisotropy, the constant a appears to converge to unity as shown in the lines in Fig. 7 for strike directions θ arbitrarily chosen as 0°, 5°, 10°, 15°, 20°, 25°, 30°, 60°, 90° and 180°. The constant b becomes smaller as a converges to

unity and bigger as a marginally departs from unity. Therefore, the convergence and divergence in physical properties is function of anisotropic coefficient, which determines the strike direction. The segment of the lines with non-genetically consistent points indicate unit of non-geologically equivalent units. The contour map in Fig. 8a and b for vertical and horizontal resistivity shows consistency in the spread. This qualitative revelation is consistent with the quantitative analogy given in Fig. 4. The coefficient

Fig. 8 Diagram showing contour distributions of **a** ρ_v, **b** ρ_h and **c** λ in the stud area
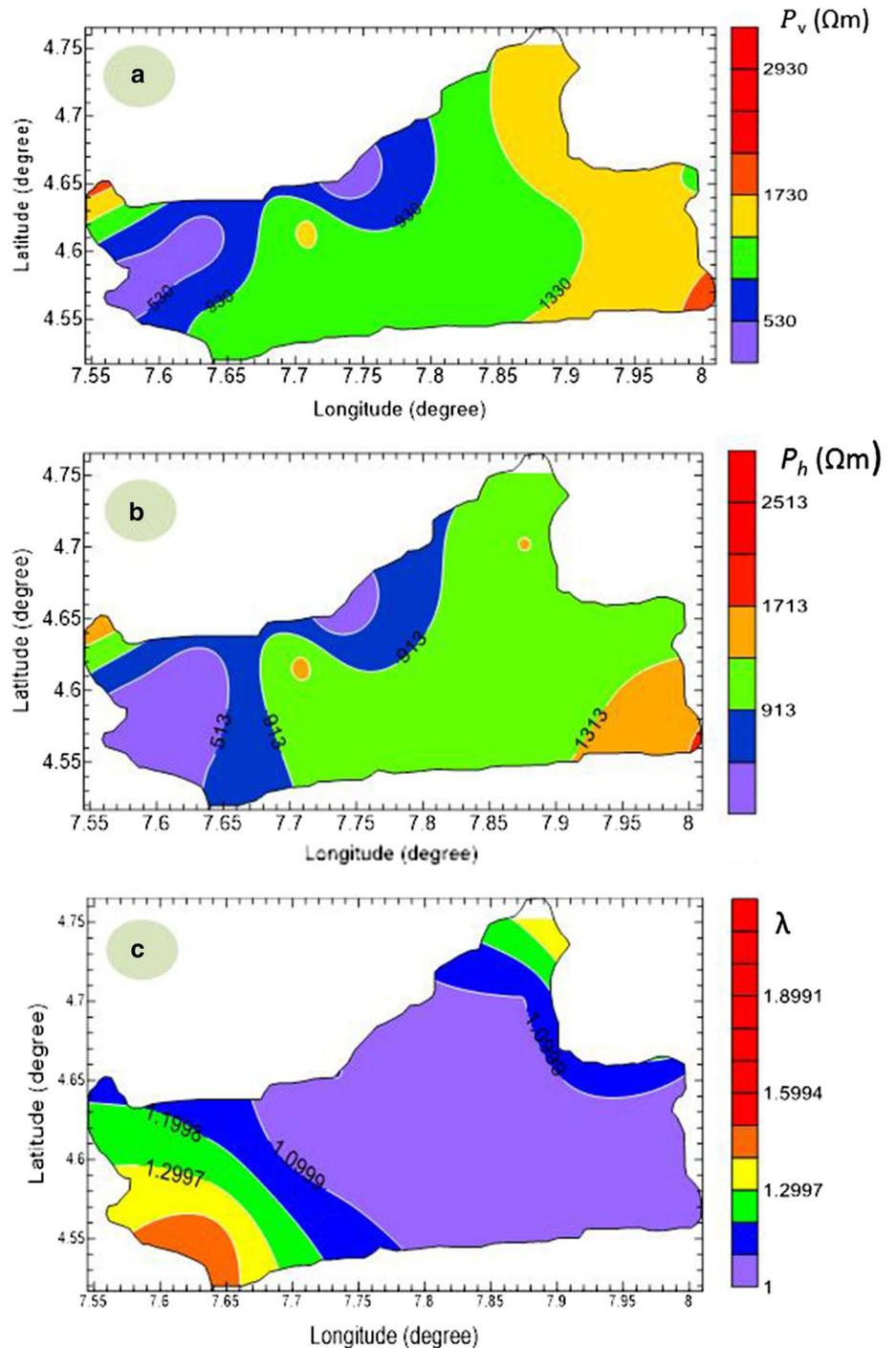
of anisotropy is Fig. 8c, shows some divergent attributes when compared to Fig. 8a and b. The displays in Fig. 8 also demonstrate that majority of the observed attributes are fairly correlated, while marginal component diverged as indicated in the color code. These graphic displays in contour map give the image of the homogeneous and anisotropic diversity, which gives room to false interpretation when in the absence of borehole information.

Conclusion

The VES technique and lithological information from boreholes were engaged to characterize the lithological diversity in homogeneous and anisotropic systems constituting the aquifer units and their overlaying layers in the hydrographic region of Akwa Ibom State. Based on the coefficient of anisotropy, the assessed volume of sedimentary formation is classified into low anisotropy ($\lambda < 1.2$), moderate anisotropy ($1.2 < \lambda \leq 1.3$) and high anisotropy ($\lambda > 1.3$) with alluvium (64.3%), inter-bedded shale and sandstone (14.3%) and shale and slate (21.3%). The respective percentage of coverage indicates that alluvium (intercalation of clay, silt or gravel of organic origin) is the dominant formation, while inter-bedded shale and sandstone as well as the shale and slate are minor geologic units. The dominant trend of alluvium indicates that the homogeneous and anisotropic units assessed are intrinsic or microscopic in nature due to the ensuing sequences of argillites intercalating with arenites in the existing hydrogeological units that the impressed current passed through. The sensitivity of current injected into the geologic units is noticeable as the delineation of fairly geologically consistent units is resolved. The absence of extremely high values ($\lambda > 2.1$) indicates that there is no intrusion of metasediments into the young sedimentary age (Greenhalgh et al. 2010). The results show that the plot between the strike-dependent resistivity versus geometric mean resistivity can be used as the basis for determining the degree of consistency of homogeneous and anisotropic geologic compositions. From this plot, it was possible to conclude the slopes of families of curves converge to unity with reduced intercepts, while slopes that marginally depart from unity are marked with high intercepts. Both the quantitative (graphic) and qualitative (contour) results portend the thin possibility of having anisotropy free geologic units. Since the effect of anisotropy really exists in geologic units either prominently or obscurely and can falsify the measured ground resistivities and geologic structure during interpretation ignored isotropic homogeneous and isotropic medium only exist in principle but not in practice. The revelations of prominent and obscured anisotropies, which coexist in hydrogeological units as revealed in Figs. 6, 7 and 8, suggest that every ground resistivity measurement and interpretation of geologic structures be accompanied or constrained by borehole information for reliability of results to be guaranteed.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the there is no conflict of interest.

References

Aissaoui R, Bounif A, Zeyen H, Messaoudi S (2019) Evaluation of resistivity anisotropy parameters in the Eastern Mitidja basin, Algeria, uses azimuthal electrical resistivity tomography, Near Surface. Geophysics 17:359–378. https://doi.org/10.1002/nsg.12048

Akpan AE, Ugbaja AN, George NJ (2013) Integrated geophysical, geochemical and hydrogeological investigation of shallow groundwater resources in parts of the Ikom-Mamfe Embayment and the adjoining areas in Cross River State. Niger Environ Earth Sci 70(3):1435–1456. https://doi.org/10.1007/s12665-0132232-3

Asten MW (1974) The influence of electrical anisotropy on mise a la masse surveys. Geophys Prospect 22:238–245

Bala M (2011) Evaluation of electric parameters of anisotropic sandyshaly miocene formations on the basis of resistivity logs. Acta Geophy 59(5):954–966

Bala M, Cichy A (2015) Evaluating electrical anisotropy parameters in miocene formations in the cierpisz deposit. Acta Geophys 63(5):1296–1315

Caglar I, and Avsar U (2007) Characterization of electrical anisotropy for cuttability of rocks using geophysical resistivity measurements, near surface in geophysics. In: 13th European meeting of environmental and engineering geophysics, Istanbul, p 3–5

Dachnov WN (1967) Electric and magnetic methods of logging. Fundament of theory, Nedra

Doust H, Omatsola E (1989) Niger Delta. In: Edwards JD, Santogrossi PA (eds) Divergent/passive margin basins, vol 48. Tulsa, American association of petroleum geologists memoir, pp 201–238

Ekanem AM (2020) Georesistivity modelling and appraisal of soil water retention capacity in Akwa Ibom State university main campus and its environs, Southern Nigeria. Model Earth Syst Environ. https://doi.org/10.1007/s40808-020-00850-6

Ekanem AM, George NJ, Thomas JE, Nathaniel EU (2019) Empirical relations between aquifer geohydraulic-geoelectric properties derived from surficial resistivity measurements in parts of Akwa Ibom State, Southern Nigeria. Nat Resour Res. https://doi.org/10.1007/s11053-019-09606-1

George NJ (2020) Appraisal of hydraulic flow units and factors of the dynamics and contamination of hydrogeological units in the littoral zones: a case study of Akwa Ibom State University and its environs, Mkpat Enin L.G.A, Nigeria. Nat Resour Res. https://doi.org/10.1007/s11053-020-09673-9

George NJ, Ekanem AM (2019) Indices of energy and appraisal for electrical current signal at polarising frequency using electrical drilling: a novel approach. In: Sundararajan N, Eshagh M, Saibi H, Meghraoui M, Al-Garni M, Giroux B (eds) On significant applications of geophysical methods: advances in science, technology and innovation (IEREK Interdisciplinary Series for Sustainable Development). Springer, Cham. https://doi.org/10.1007/978-3-030-01656-2_21

George JG, Ekanem AM, Ibanga JI, Udosen NI (2017) Hydrodynamic implications of aquifer quality index (AQI) and flow zone indicator (FZI) in groundwater abstraction: a case study of coastal hydro-lithofacies in South-eastern Nigeria. J Coast Conserv 21(4):759–776. https://doi.org/10.1007/s11852-017-0535-3

George NJ, Ibanga JI, Ubom AI (2015) Geoelectrohydrogeological indices of evidence of ingress of saline water into freshwater in parts of coastal aquifers of IkotAbasi. South Niger J Afri Earth Sci 109:37–46

George NJ, Obianwu VI, Akpan AE, Obot IB (2010) Assessment of shallow aquiferous units and their coefficients of anisotropy in the coastal plain sands of Southern Ukanafun local government area, Akwa Ibom State. South Niger Arch Phys Res 2:118–128

Greenhalgh SA, Marescot L, Zhou B, Greenhalgh M, Wiese T (2009) Electric potential and Frechet derivatives for a uniform anisotropic medium with a tilted axis of symmetry. Pure Appl Geophys 166:673–699

Greenhalgh S, Wiese T, Marescot L (2010) Comparison of DC sensitivity patterns for anisotropic and isotropic media. J Appl Geophys 70(2010):103–112. https://doi.org/10.1016/j.jappgeo.2009.10.003

Hasan M, Yanjun S, Gulraiz A, Weijun J (2019) Application of VES and ERT for fresh-saline interface in alluvial aquifers of Lower Bari Doab, Pakistan. J Appl Geophys 164:200–213

Hill DG (1972) A laboratory investigation of electrical anisotropy in Precambrian rocks. Geophysics 37:1022–1038

Hobbs B, Werthmuller D, Engelmark F (2009) The effect of resistivity anisotropy on earth impulse responses. ASEG Ext Abstr 2009:1–5

Ibanga JI, George NJ (2016) Estimating geohydraulic parameters, protective strength, and corrosivity of hydrogeological units: a case study of ALSCON, Ikot Abasi, southern Nigeria. Arab J Geosci 9:363. https://doi.org/10.1007/s12517-016-2390-1

Ibuot JC, Akpabio GT, George NJ (2013) A survey of the repository of groundwater potential and distribution using geo-electrical resistivity method in Itu Local Government Area (L.G.A), Akwa Ibom State, Southern Nigeria. Cent Eur J Geosci 5(4):538–547. https://doi.org/10.2478/s13533-012-0152-5

Karanth KR (1987) Groundwater assessment, development and management. Tata-McGraw Hill, New Delhi

Karnkowski P (1999) Oil and gas deposits in Poland. Geosynoptics Society (GEOS), Krakow

Kunz KS, Moran JH (1958) Some effects of formation anisotropy on resistivity measurements in boreholes. Geophysics 23(4):770–794. https://doi.org/10.1190/1.1438527

Matias MJS (2002) Square array anisotropy measurements and resistivity sounding interpretation. J Appl Geophys 49:185–194

Matias MJS, Habberjam GM (1986) The effect of structure and anisotropy on resistivity measurements. Geophysics 51:964–971

Obiora DN, Ibuot JC, George NJ (2015) Evaluation of aquifer potential, geoelectric and hydraulic parameters in Ezza North, southeastern Nigeria, using geoelectric sounding. J Sci Technol Int. https://doi.org/10.1007/s13762-015-0886-y

Olasehinde PI, Bayewu OO (2011) Evaluation of electrical resistivity anisotropy in geological mapping: a case study of Odoara, West Central Nigeria. Afr J Environ Sci Technol 5(7):553–566

Onuoha KM, Dim CIP (2017) Developing unconventional petroleum resources in Nigeria: an assessment of shale gas and shale oil prospects and challenges in the inland Anambra Basin. In: Onuoha KM (ed) Advances in petroleum geoscience research in Nigeria, Chapter 16, 1st edn. University of Nigeria, Nsukka, pp 313–328

Petters SW (1989) Akwa Ibom State: physical background, soil and landuse and ecological problems. In: Technical report for government of Akwa Ibom State, pp 603

Short KC, Stauble AJ (1967) Outline of the geology of Niger Delta. Assoc Pet Geol Bull 54:761–779

Thomas JE, George NJ, Ekanem AM, Nsikak EE (2020) Electrostratigraphy and hydrogeochemistry of hyporheic zone and water-bearing caches in the littoral shorefront of Akwa Ibom State University, Southern Nigeria. Environ Monit Assess 192:505. https://doi.org/10.1007/s10661-020-08436-6

Vander-Velpen BPA (1988) A computer processing package for D.C. resistivity interpretation for an IBM compatibles. ITC, The Netherlands, p 4

Vander Velpen BPA, Sporry RJ (1993) RESIST: a computer program to process resistivity sounding data on PC compatibles. Comput Geosci 19(5):691–703

Wei B, Lingwei K, Aiguo G (2013) Effects of physical properties on electrical conductivity of compacted lateritic soil. J Rock Mech Geotech Eng 5(5):406–411

Yadav GS, Singh SK (2007) Integrated resistivity surveys for delineation of fractures for ground water exploration in hard rock areas. J Appl Geophys 62(3):301–312

Yeboah-Forson A, Whitman D (2013) Electrical resistivity characterization of anisotropy in the Biscayne Aquifer. Groundwater 52(5):728–736

Impact of convection and stability parameters on lightning activity over Andhra Pradesh, India

N. Umakanth[1] · G. Ch. Satyanarayana[1] · B. Simon[2] · Prashant Kumar[2] · M. C. Rao[3]

Abstract

This paper brings out the interconnection of flash rate density (FRD) with convection and stability parameters over Andhra Pradesh (AP), India. The convection parameters include rainfall, relative humidity, specific humidity, surface air temperature (SAT) and air temperature (at 850 mb). The stability parameters include convective available potential energy (CAPE), lifted index, K-index, total totals index (TTI), humidity index and total precipitable water. Both convective and stability parameters indicate good correlation with lightning activity. SAT and AT 850 mb have shown good correlations with lightning, which is a clear indication of interaction between warm air and dry air. CAPE and TTI have shown strong positive correlation with lightning activity. The correlation coefficient between FRD and CAPE is 0.81. We have also studied the influence of convective and stability parameters during lightning and no lightning activity. Later, we also attempted the estimation of lightning activity by using artificial neural network model. By using convection and stability parameters, six learning algorithms were used for training the artificial neural network. Scaled conjugate gradient backpropagation training algorithm has given the better estimation, whereas resilient backpropagation training algorithm has shown the poor estimation of FRD.

Introduction

Lightning is a natural and common phenomenon seen in the tropical countries which results in many deaths every year throughout the world during pre-monsoon months (March, April and May) (Weng et al. 2010). Globally, 6000 to 24,000 deaths occurred due to lightning strikes every year all over the world (Gomes and Kadir 2011). Human fatalities have been increasing in the recent years due to lightning activity. According to NDMA annual report 2018, out of all natural disasters 39% of deaths are due to lightning during the time period of 1967–2012 (Source: https://ndma.gov.in/images/pdf/Draft-Guidelines-thunderstorm.pdf). There were more

lightning deaths in 2004 (Singh and Singh 2015). The extent of lightning disaster is large as it costs more than 2000 deaths every year in India (Vaddadi et al. 2015; Das 2015). Out of all seasons, pre-monsoon records high frequency for thunderstorms over Andhra Pradesh (AP) and Telangana (TA) states (Tyagi, 2007). Lightning usually occurs either in one thunder cloud or between two clouds or between cloud and the ground (Weng et al. 2010). Therefore, lightning has a devastating effect on our society in numerous ways. In general, the thunderstorm comprises lightning and thunder which take place simultaneously at the region of occurrence. The electric charge which is observed in thunderstorm is referred as lightning, whereas the observed sound in thunderstorm is known as thunder (Johns and Doswell 1992). In order to monitor that lightning activity, ground-based lightning detectors have been very useful in studying severe weather conditions like thunderstorms, tornadoes, etc. Globally, there are many lightning detectors which are active. Earth network and worldwide lightning network were playing crucial role in real-time lightning analysis.

Various studies were available on lightning activity over India which was done based on LIS-TRMM satellite. Studies

✉ N. Umakanth
   nandivadaumakanth@gmail.com

1 Department of Atmospheric Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522502, India

2 Atmospheric and Oceanic Sciences, Space Applications Centre (SAC), Ahmedabad 380023, India

3 Department of Physics, Andhra Loyola College, Vijayawada 520008, India

by Murugavel et al. (2014) and Saha et al. (2017) indicated the importance of satellites and also the influence of convection on lightning activity. A recent study by Liou and Kar (2010) explained that CAPE and lightning are interconnected over Taiwan. Some researchers tried to analyze the lightning activity and its relationship with CAPE, rainfall, air temperature, relative humidity and specific humidity (Manohar et al. 1999; Kandalgaonkar et al. 2005a, b; Manohar and Kesarkar 2005; Tinmaker et al. 2010; Tinmaker et al. 2015). Factors such as surface air temperature, rainfall and CAPE showed good correlation which indicates their influence on lightning activity over Maharashtra, India (Tinmaker et al. 2015). The lightning flashes and thunder days have positive correlation in summer season. If lightning flashes are increasing, the thunderstorm-occurring days are also increasing (lavigne et al. 2019). From this motivation, we made an attempt to study the monthly variations in lightning activity from LIS-TRMM satellite.

After studying the lightning activity, we have used the concept of artificial neural network (ANN). This mathematical model working is similar to the functioning of neurons present in human brain. We know that neurons play a crucial role in performing any task by the humans. One neuron interacts with other for solving any task. In the same manner, ANN also uses input values to compute numerically in hidden neuron layer and give the outputs. The ANN is used as a tool for problem solving which are inherently difficult for both humans and computers. Chaudhuri and Chattopadhyay (2005) explained the usefulness of multilayer feed-forward artificial neural network for predicting surface temperature and relative humidity. Chattopadhyay (2007) explained the usefulness of feed-forward ANN with one hidden layer for predicting the southwest monsoon rainfall over India which was better than the linear regression forecast. Litta et al. (2013) showed the forecasting of thunderstorms using ANN based on meteorological parameters. Levenberg–Marquardt has mastered the other algorithms which are used to predict thunderstorm statistically. This work gave us the idea to estimate the lightning using various parameters (Litta et al. 2013).

In this paper, we tried to study the link between lightning activity and the convection parameters such as rainfall (RF), relative humidity (RHUM), specific humidity (SHUM), surface air temperature (SAT) and air temperature (at 850 mb). We also attempted to study the bond between lightning activity and stability parameters such as convective available potential energy (CAPE), lifted index (LI), K-index (KI), total totals index (TTI), humidity index (HI) and total precipitable water (TPW). We also studied the daily influence of convective and stability parameters during lightning and no lightning activities. For monthly variations, we took 20 years, i.e., 1995–2014. For daily variations, we took only one year, i.e., 2004. Later, the experiments were conducted

with ANN model to estimate flash rate density (FRD) using convection and stability parameters over Andhra Pradesh as shown in Fig. 1a, b. Cape is included in both the trainings of ANN as it is associated with both atmospheric stability and atmospheric convection. The performance of six training algorithms, namely Levenberg–Marquardt, resilient backpropagation, one-step secant method, scaled conjugate gradient backpropagation, gradient descent backpropagation and random order incremental, was evaluated using convection and stability parameters to estimate FRD. The accuracy of the estimations was evaluated by the correlation coefficient (CC), the root-mean-square error (RMSE), mean absolute error (MAE), index of agreement (IOA) and BIAS.

Data and methodology

Data

Lightning imaging sensor (LIS) instrument was embedded on tropical rainfall measuring mission (TRMM) satellite which was developed by NASA, USA. It is an earth observing system. The main purpose of using this instrument was to identify the lightning flashes. This instrument was launched to overcome the accuracy and sensitivity issues of optical transient detector (OTD). Another important reason for selecting the LIS data for our study was its high efficiency (90%) and less bias. This efficiency was possible because of its inclination, i.e., 35°. This instrument on TRMM satellite monitors lightning activity globally at lesser altitudes providing valuable information required for analyzing lightning activity over tropical countries (Boccippio et al. 2002).

This work is implemented on Andhra Pradesh region as shown in Fig. 2. The monthly flash rate data with 0.5° resolution were collected for Andhra Pradesh region from LIS-TRMM satellite for a time period of 20 years (1995–2014). The Web site used for the collection was (http://thunder.msfc.nasa.gov/data/data_lis.html) (Cecil et al. 2006). The daily data of temperature, relative humidity, rainfall, CAPE, specific humidity and TPW parameters are collected at 0.5° resolution for 20-year period (1995–2014) from ERA-Interim ECMWF re-analysis data (http://apps.ecmwf.int/datasets/data/interim_full_daily/). Temperature and relative humidity data were collected at different pressure levels. First, the collected data were extracted from netcdf (.nc) format to .data format using MATLAB program. The temperature and relative humidity values which were extracted were used to compute dew point temperature parameter at all pressure levels. By using temperature and dew point temperature values, we calculated LI, KI, TTI and HI parameters for every day using the formulae given below. The daily values of all the parameters were averaged

Fig. 1 a Algorithm used for the estimation of lightning flash rate density (FRD) using artificial neural networks with the help of atmospheric convective parameters as inputs, **b** algorithm used for the estimation of lightning flash rate density (FRD) using artificial neural networks with the help of atmospheric stability parameters as inputs
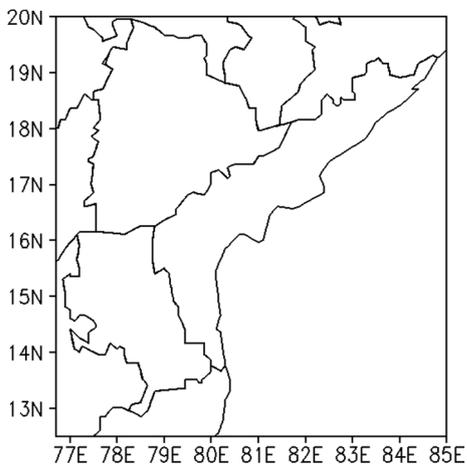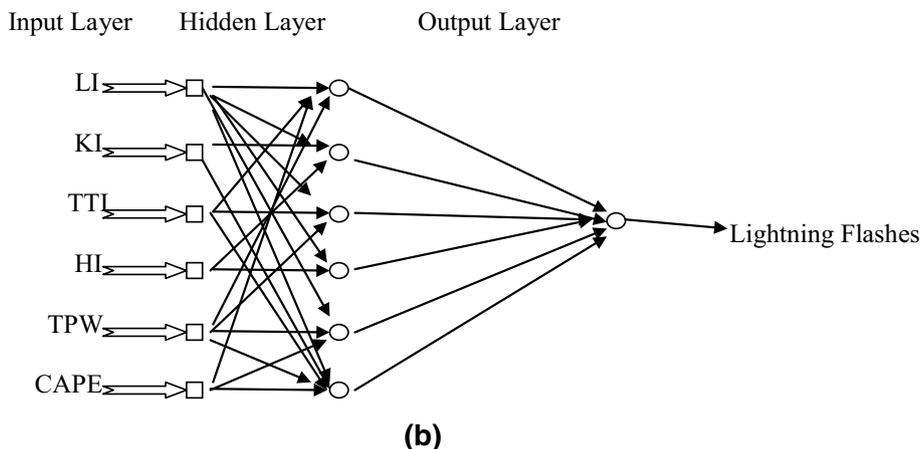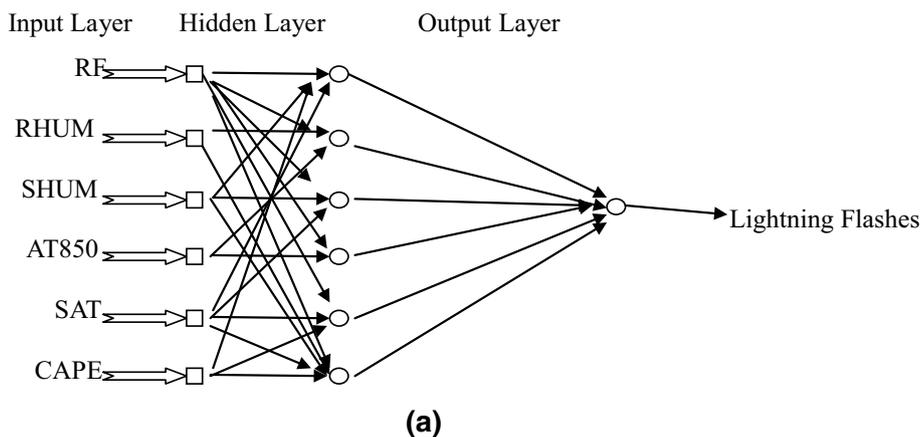

(a)

(b)

Fig. 2 Study region of Andhra Pradesh

for each month separately over a period of 20 years. The monthly variations in every parameter were discussed in this study. There are four seasonal divisions in India, namely (a) winter (December–February), (b) summer or pre-monsoon (March–May), (c) rainy or southwest monsoon (June–September) and (d) post-monsoon (October–November).

Methodology

The formulae used to calculate the different indices that are analyzed in this study are given below.

(i) K-index (KI)

The K-index is the subtraction of temperature and dew point temperature at different pressure levels of the atmosphere as shown below (George 1960):

$$KI = (Temp850 - Temp500) + Tempd850 \\ - (Temp700 - Tempd700) \tag{1}$$

where Temp represents temperature and Tempd represents dew point temperature. The possibility of thunderstorm occurrence is based on the threshold values given below.

KI	Chances for thunderstorm formation
Less than 288 K	Almost 0% chance for occurrence
288 to 293 K	20% chance for occurrence
294 to 298 K	20–40% chance for isolated thunderstorm occurrence
299 to 303 K	40–60% chance for scattered thunderstorm occurrence
304 to 308 K	60–80% chance for widespread thunderstorm occurrence
309 to 313 K	80–90% chance for numerous thunderstorms
Greater than 313 K	More than 90% chance for thunderstorm occurrence

(ii) Total Totals Index (TTI)

The total totals index is obtained by differencing the temperature and dew point temperatures at 850 and 500 hpa pressure levels. The threshold values of TTI values are shown below (Miller, 1967):

Cross totals, $CT = Tempd850 - Temp500$ Vertical totals,

$VT = Temp850 - Temp500$ Total totals,

$$TT = CT + VT = Temp850 + Tempd850 - 2Temp500 \tag{2}$$

where Temp denotes temperature and Tempd denotes dew point temperature.

The risk of occurrence of severe thunderstorm is given below

TTI (K)	Chances for thunderstorm formation
44 to 45 K	Occurrence of isolated moderate thunderstorm
46 to 47 K	Occurrence of scattered moderate/few heavy thunderstorms
48 to 49 K	Occurrence of scattered moderate/isolated severe thunderstorms
50 to 51 K	Occurrence of scattered severe thunderstorms
52 to 55 K	Occurrence of scattered to numerous heavy thunderstorms
Greater than 55 K	Occurrence of very severe thunderstorms.

(iii) Lifted Index (LI)

This index is calculated to analyze the atmospheric conditions in the lower layers of troposphere. The LI threshold values are discussed below (Galway 1956):

$$\text{Lifted index (LI)} = \text{Temp}_{500} - \text{Temp}_{parcel} \tag{3}$$

where Temp is the air temperature and Temp_{parcel} is the parcel temperature that is lifted from surface to 500 hpa pressure level.

LI (K)	Chances for thunderstorm formation
Greater than 2	No thunderstorm occurrence
0 to 2	Possibility of a small thunderstorm
−2 to 0	Possibility of moderate thunderstorm
−4 to −2	Possibility of moderate to severe thunderstorms
Less than −4	Possibility of very severe thunderstorm

(iv) Humidity Index (HI)

Humidity index is a measure of water vapor availability at 850, 700 and 500 hPa pressure levels (Jacovides and Yonetani 1990). It is one of few indices which show the significance of relative humidity in the severe thunderstorm occurrences:

$$\begin{aligned} HI = &(Temp - Tempd)850 + (Temp - Tempd)700 \\ &+ (Temp - Tempd)500 \end{aligned} \tag{4}$$

where Temp is temperature and Tempd is dew point temperature.

(v) Convective available potential energy (CAPE)

CAPE is determined by the below formula defined by (Kunz 2007)

$$CAPE = \int_{Z_f}^{Z_n} g \left[\frac{T_{v,parcel} - T_{v,env}}{T_{v,env}} \right] dz \tag{5}$$

where $T_{v,parcel}$ and $T_{v,env}$ represent the virtual temperature of the parcel and environment, respectively, and Z_f and Z_n denote the level of free convection and neutral buoyancy.

The threshold values of CAPE which interpret the thunderstorm occurrence are given below.

Less than 300 J/kg	Little or no convective potential
300 to 1000 J/kg	Weak convective potential
1000 to 2500 J/kg	Moderate convective potential
Greater than 2500 J/kg	Strong convective potential

The ANN network which we considered was feed-forward, multilayer perceptron network. It has two layers: one is hidden layer and the other one is output layer. The ANN network used in the study is trained and tested using MATLAB. The input layer of ANN consists of six atmospheric

inputs. Each layer is activated by log-sigmoidal activation function. The design of our ANN model is shown in Fig. 1. The ANN model needs a complete data of inputs. In this model, after the data are loaded to the network, associated weights are adjusted to estimate output based on the inputs. We gave convective and stability parameters as inputs to ANN to estimate FRD as output as shown in Fig. 1.

We used six training algorithms for ANN model as discussed below.

(i) Levenberg–Marquardt training algorithm (LM): The LM algorithm is one of the most commonly used trusted algorithms for linear and nonlinear parameters. It uses finite differencing schemes which are approximated by Jacobian. This is all related to the first-order partial derivative function. It is used for the optimization of weights of the inputs. The main benefit of using LM is its performance does not deviate from the default gradient search for better neural computing. This algorithm is also used as a good technique for solving nonlinear least squares (Chen et al. 2017).

(ii) Resilient backpropagation training algorithm (RP): RP is used to train the neural network to implement complex algorithm which is better than backpropagation (Cömert and Kocamaz 2017):

$$\Delta x_k = -\text{sign}\left(\frac{\Delta E_k}{\Delta x_k}\right)\Delta k \tag{6}$$

Whereas Δx_k denotes the current changes in weight vectors, ΔE_k denotes the error function E at k and Δk denotes the increment in bias.

(iii) One-step secant method training algorithm (OSS): OSS requires small storage for computation per epoch than quasi-Newton backpropagation algorithm. This fills up the gap between conjugate gradient algorithms and quasi-Newton algorithms (Hagan et al. 1994).

(iv) Scaled conjugate gradient backpropagation training algorithm (SCG): SCG is a second-order algorithm. It is applied to feed-forward neural network such as LM. It saves the time for running each iteration (Moller 1993)

(v) Gradient descent backpropagation training algorithm (GD): GD provides the first-order gradient term in the neural performance. This method helps us to find the minimum value of the given data. It is used mainly for those parameters which cannot be calculated analytically such as algebra. This algorithm adapts weights to predict outputs. The main goal of using this algorithm is to estimate the target func-

tion that coincides with input and output data (S.L. Pinjare and Arun Kumar 2012).

(vi) Random-order incremental training algorithm (R): Incremental training is usually applied for static and dynamic networks. This training updates the weights and biases of the inputs to obtain the projected targets. R is usually applied to the data defined at multiple time series. The whole network was updated at a time for each step in the training (Huang et al. 2006).

Statistical metrics are comprised of the following terms: (a) correlation coefficient (CC), (b) root-mean-square error (RMSE), (c) mean absolute error (MAE), (d) index of agreement (IOA) and (e) BIAS. The above terms were calculated to evaluate the ANN model performance in the estimation of lightning flashes during the period of study. Formulae (Wilks 2006) are as follows:

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right) \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - o_i\right| \tag{8}$$

$$CC = \frac{\sum_{i=1}^{n}\left(f_i - f'\right)\left(o_i - o'\right)}{\sqrt{\sum_{i=1}^{n}\left(f_i - f'\right)^2}\sqrt{\sum_{i=1}^{n}\left(o_i - o'\right)^2}} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(f_i - o_i\right)^2}{n}} \tag{10}$$

$$IOA = 1.0 - \frac{\sum_{i=1}^{n}\left(f_i - o_i\right)^2}{\sum_{i=1}^{n}\left(\left|f_i - o'\right| + \left|o_i - o'\right|\right)^2} \tag{11}$$

where o_i and f_i correspond to observations and forecast values.

Results and discussion

For this study, the LIS-TRMM satellite flash rate data were analyzed during 1995–2014. For our convenience, we have divided all the parameters into two categories which influence lightning activity. One is convection parameters, and the second is stability parameters. Both affect the lightning activity. ERA-Interim ECMWF re-analysis daily data were utilized for the analysis of stability and thermodynamic parameters. Convection parameters include rainfall, relative humidity, specific humidity, surface air temperature and air temperature at 850 mb. Stability parameters include

convective available potential energy, lifted index, K-index, total totals index, humidity index and total precipitable water. We tried to analyze the relationships of lightning activity with each parameter of convection and stability parameters over Andhra Pradesh, India. Later, we also attempted estimation of lightning activity by training artificial neural network.

The monthly variations in mean sea surface air temperature (SAT) and FRD are shown in Fig. 3a. We can clearly understand that the FRD values uniformly increased from January to May and started to decline from June to December. The minimum FRD value of 0.99 fl km -2 month -1 is seen in January, whereas the maximum FRD value of 18.23 fl km -2 month -1 is seen for May month. We can also observe that surface air temperature also followed the same consistency of increased values from January to May and started to decrease from June to December as similar to FRD. The minimum SAT value of 24.76 °C is seen in December, whereas the maximum SAT value of 32.94 °C is seen for May month. We can see the same kind of variations between FRD and air temperature (AT 850 mb) as shown in Fig. 3b. As high air temperatures are seen in May month,

there is good heating of land surface. This makes atmosphere unstable leading to intense convection. The latent heat released during this convection process triggers the development of cumulonimbus clouds. During pre-monsoon, SAT and at 850 mb are higher, which helps the air to pick up the moisture in the lower levels of atmosphere. When there is cold air in the upper levels, the moist air in lower levels interacts leading to severe or moderate thunderstorms. The correlations between FRD and SAT, FRD and at 850 mb are 0.87 and 0.85, respectively. So the variations in air temperatures are good indicator for lightning activity. The main reason for using air temperature at 850 mb is to understand the advection between moist air and warm air which makes atmosphere unstable. This moist air is being lifted up in the atmosphere for the occurrence of cloud formation which helps for the thunderstorm development leading to lightning.

The plot of monthly mean variations in flash rate density and specific humidity as shown in Fig. 3c exhibits the highest values in monsoon months, especially August. The correlation coefficient between flash rate density and specific humidity is found to be 0.68. The monthly mean variations plot of flash rate density and relative humidity as shown in
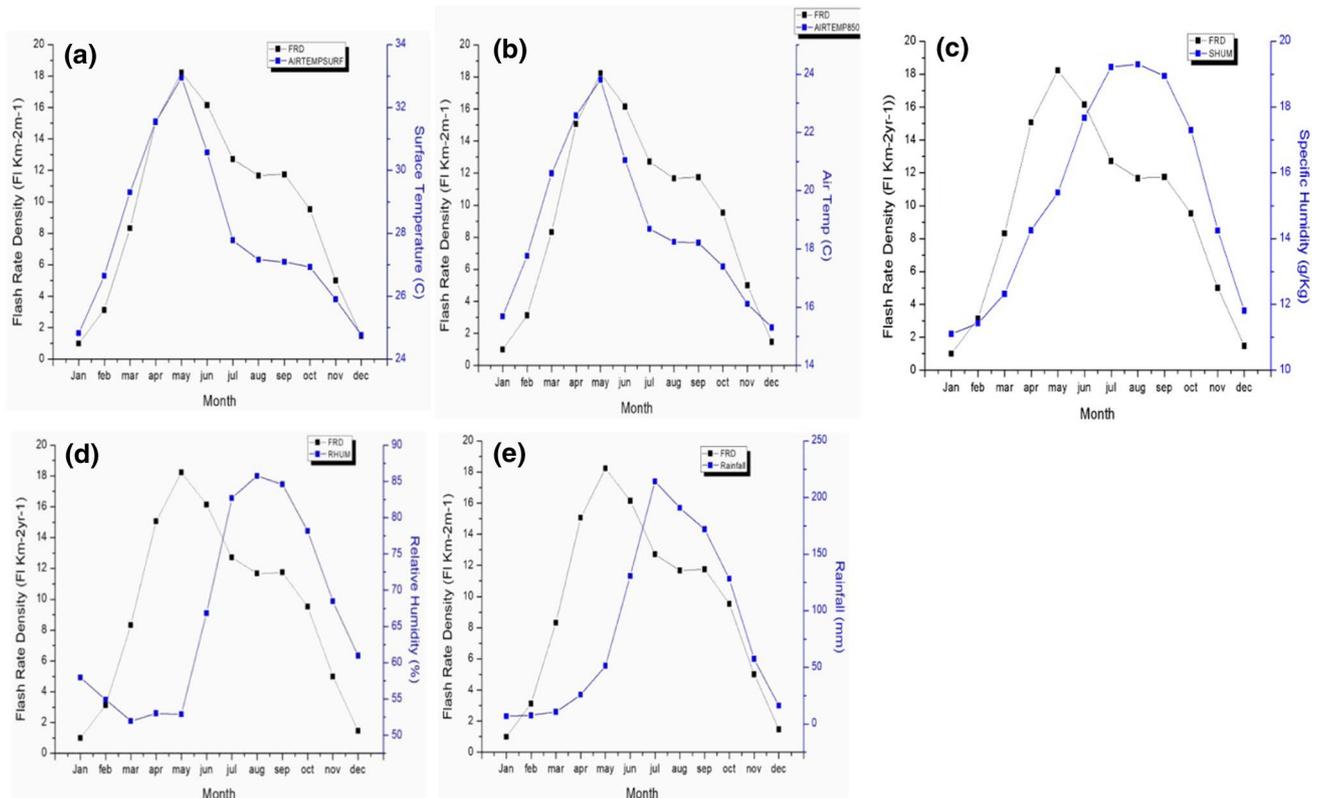

Fig. 3 a Monthly variations in FRD and mean sea surface air temperature. b Monthly variations in FRD and air temperature (AT) (850 mb). c Monthly variations in FRD and mean specific humidity. d Monthly variations in FRD and mean relative humidity. e Monthly variations in FRD and mean rainfall (RF)

Fig. 3d also indicated the highest values during monsoon months as specific humidity. The correlation coefficient between flash rate density and relative humidity is found to be 0.67. The reason behind to analyze both humidities is to study the influence of temperature on moist air and the flash rate density. Specific humidity is influencing the flash rates more than the relative humidity during pre-monsoon season. Figure 3e shows monthly mean variations in FRD and RF over the Andhra Pradesh region during the study period. The highest amounts of rainfall were seen during monsoon months, especially July. The correlation coefficient between flash rate density and rainfall is found to be 0.50. The May month has shown highest rainfall in pre-monsoon months which is an indication of rains leading as flash floods with lightning.

We have also tried to study the correlation between stability indices and the flash rate density. These indices play a vital role in analyzing the thunderstorm severity. Figure 4a shows monthly mean variations in FRD and CAPE over the Andhra Pradesh region during the study period. The highest values of CAPE were seen during pre-monsoon months such as April and May. The highest CAPE values for May month indicate its influence on lightning.

The correlation coefficient between flash rate density and CAPE is found to be 0.81. Figure 4b shows monthly mean variations in FRD and LI over the Andhra Pradesh region during the study period. FRD and LI parameters have strong negative correlation with each other. The correlation coefficient between flash rate density and LI is found to be -0.72. The LI values are lowest, the higher the chances for severe lightning. The monthly variations in mean K-index and FRD are shown in Fig. 4c. Highest values were seen over pre-monsoon months which are a clear indication of high lightning strikes over Andhra Pradesh. Highest KI mean value of 324 K was seen for May month. The correlation coefficient between flash rate density and K-index is found to be 0.63. The monthly variations in mean TTI and FRD are shown in Fig. 4d. Highest values were seen over pre-monsoon months which are a clear indication of high lightning strikes over Andhra Pradesh. Highest TTI mean value of 47 K was seen for May month. The correlation coefficient between flash rate density and TTI is found to be 0.83. The monthly variations in mean HI and FRD are shown in Fig. 4e. Lowest values were seen over monsoon and post-monsoon months which is a clear indication of high availability of moisture leading
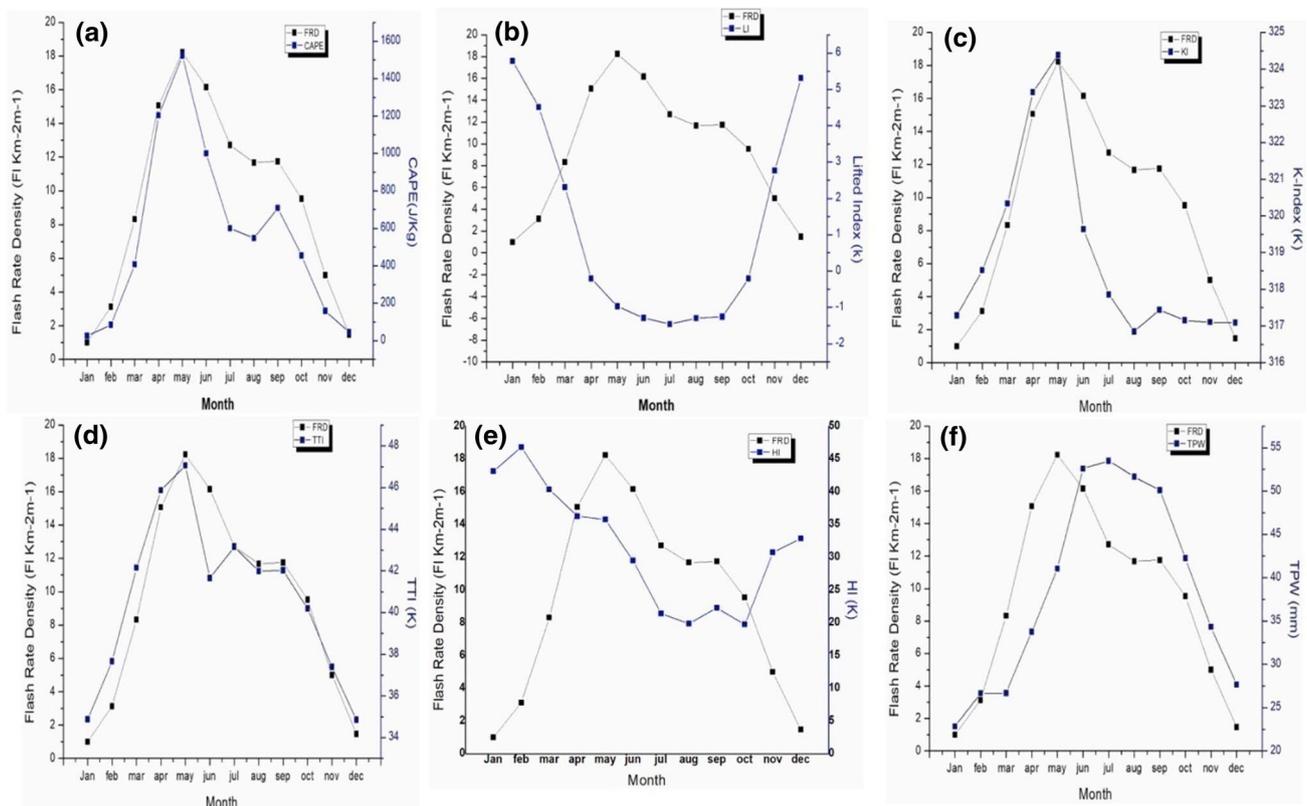

Fig. 4 a Monthly variations in FRD and mean CAPE. **b** Monthly variations in FRD and mean LI. **c** Monthly variations in FRD and mean KI. **d** Monthly variations in FRD and mean TTI. **e** Monthly variations in FRD and mean HI. **f** Monthly variations in FRD and mean TPW

Fig. 5 ANN-based six training models used for the estimation of FRD using atmospheric stability parameters
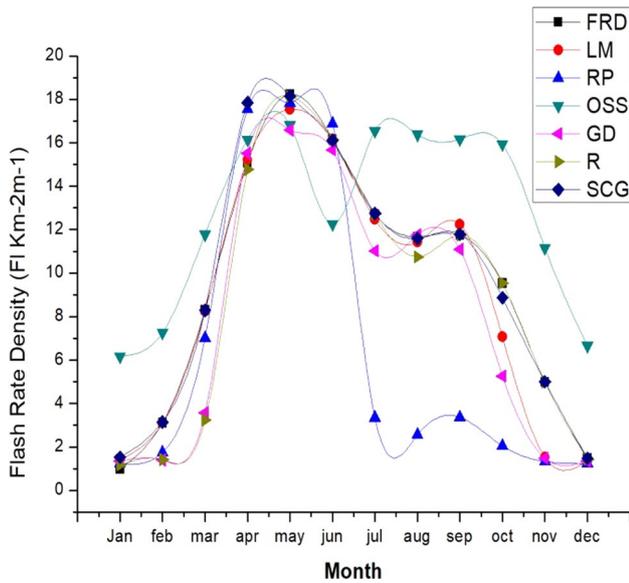

Fig. 6 ANN-based six training models used for the estimation of FRD using atmospheric convective parameters

toward rains over Andhra Pradesh. The mean HI values for pre-monsoon months are 6.3, 6.1 and 5.6 K. These HI threshold values also contribute for high lightning activity over A.P. The correlation coefficient between flash rate density and HI is found to be 0.43.

The monthly variations in mean TPW and FRD are shown in Fig. 4f. Highest values were seen over monsoon and post-monsoon months which are a clear indication of high lightning strikes over Andhra Pradesh. Highest TPW mean value of 53.48 K was seen for July month. May and April months are also showing good thresholds for convection. The correlation coefficient between flash rate density and TPW is found to be 0.71. We trained convective parameters as inputs by applying six different training algorithms such as LM, RP, OSS, GD, R and SCG to obtain the estimated flash rate density (FRD) as shown in Fig. 5. LM, GD, R and SCG have shown good estimations for lightning. SCG has shown the highest correlation for the estimation of lightning activity, whereas RP showed the lowest correlation. We also trained stability parameters as inputs by applying six different training algorithms such as LM, RP, OSS, GD, R and SCG to obtain the predicted flash rate density (FRD) as shown in Fig. 6. LM, GD, R and SCG have shown good estimations for lightning. LM and R have shown the highest correlation for the estimation of lightning activity, whereas RP showed the lowest correlation. By the analysis of Table 1 and Table 2, we understood that LM, SCG and R have better correlation and index of argument. LM has very less BIAS which is useful for lightning estimation.
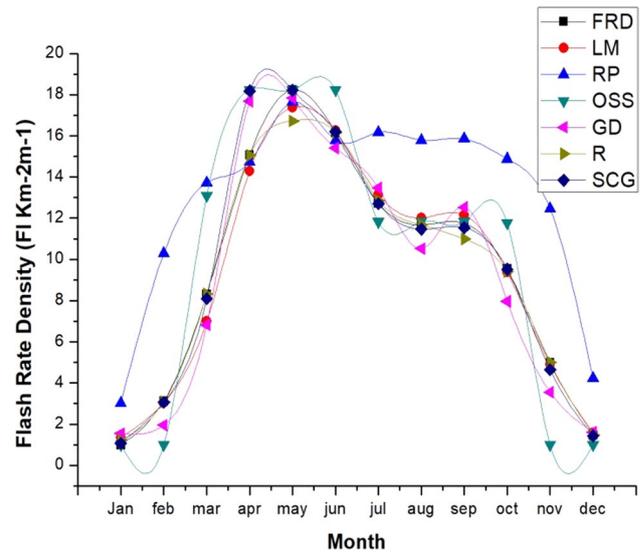
Table 1 Statistical metrics of the evaluation of the ANN models-predicted lightning flash density using stability parameters over Andhra Pradesh during 1995–2014

	BIAS	MAE	RMSE	CC	IOA
LM	0.132	− 0.132	0.556	0.959	0.997
RP	− 3.380	3.380	4.310	0.876	0.875
OSS	− 0.415	0.415	2.302	0.951	0.971
GD	0.260	− 0.260	1.242	0.929	0.990
R	0.172	− 0.17	0.496	0.917	0.998
SCG	− 0.175	0.17	0.912	0.909	0.994

Table 2 Statistical metrics of the evaluation of the ANN models predicted lightning flash density using convective parameters over Andhra Pradesh during 1995–2014

	BIAS	MAE	RMSE	CC	IOA
LM	0.522	− 0.522	1.262	0.910	0.990
RP	3.154	− 3.154	5.171	0.784	0.860
OSS	− 3.272	3.27	4.439	0.852	0.850
GD	1.496	− 1.496	2.282	0.927	0.968
R	0.663	− 0.663	1.572	0.930	0.984
SCG	− 0.211	0.211	0.841	0.936	0.995

The basic objective behind this estimation was to check the accuracy of the algorithms. Correlation analysis between FRD values (observed) and estimated FRD values was carried out for both the categories, i.e., convective (Fig. 7a) and stability (Fig. 7b) parameters. Figure 7 gives the scatter plot between estimated and observed data for pre-monsoon season during 2004. We have computed correlation coefficient

(CC), root-mean-square error (RMSE) and BIAS for convective and stability parameters as shown in Tables 1 and 2. The CC is 0.86; i.e., 86% of FRD estimated values from convective parameters using all six algorithms were matching with observed data. The CC is 0.93; i.e., 93% of FRD estimated values from stability parameters using all six algorithms were matching with observed data. From this, we found that both convective and stability parameters were so helpful in FRD estimation. FRD estimated from stability parameters are having better correlation with LIS TRMM FRD data.

The lightning activity and CAPE parameter are almost 90% correlated. As we do not have the daily lightning data availability, we tried to compare stability and convection parameters with CAPE parameter. Here, we have taken CAPE thresholds as an indication of lightning activity. In 2004, nearly 1,40,000 lightning flashes were observed and 350 lightning deaths were recorded (Singh and Singh 2015). Based on O. Singh and J. Singh, 2015 research work, we have selected year 2004 for studying daily variations. Figures 8, 9, 10 and 11 show the daily mean values for whole Andhra Pradesh. We have taken Andhra Pradesh region as

Fig. 7 **a** Scatter plot between convective parameters estimated data and observed data for March–May months during the year 2004. **b** Scatter plot between stability parameters estimated data and observed data for March–May months during the year 2004

(a)

(b)
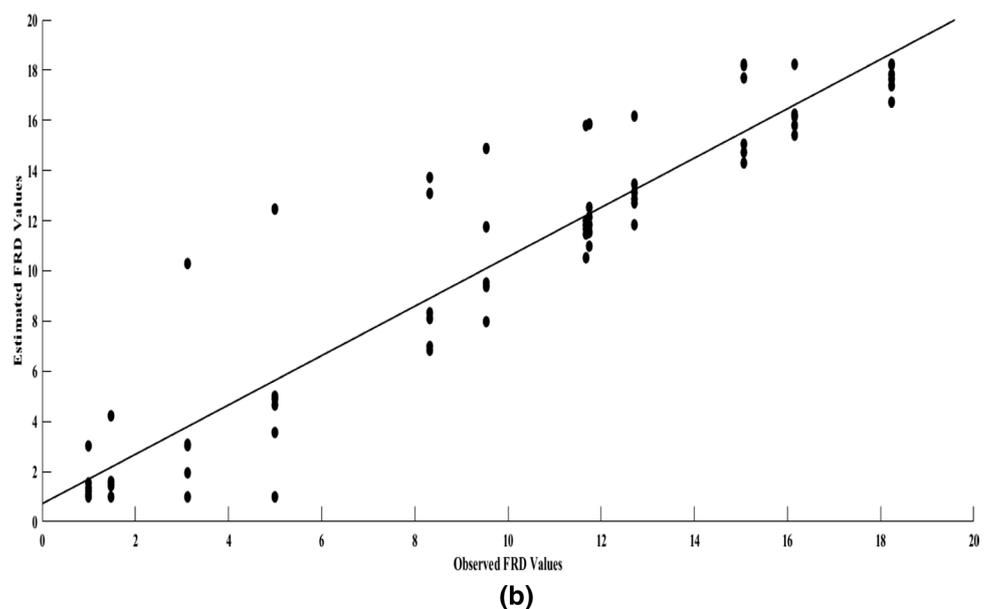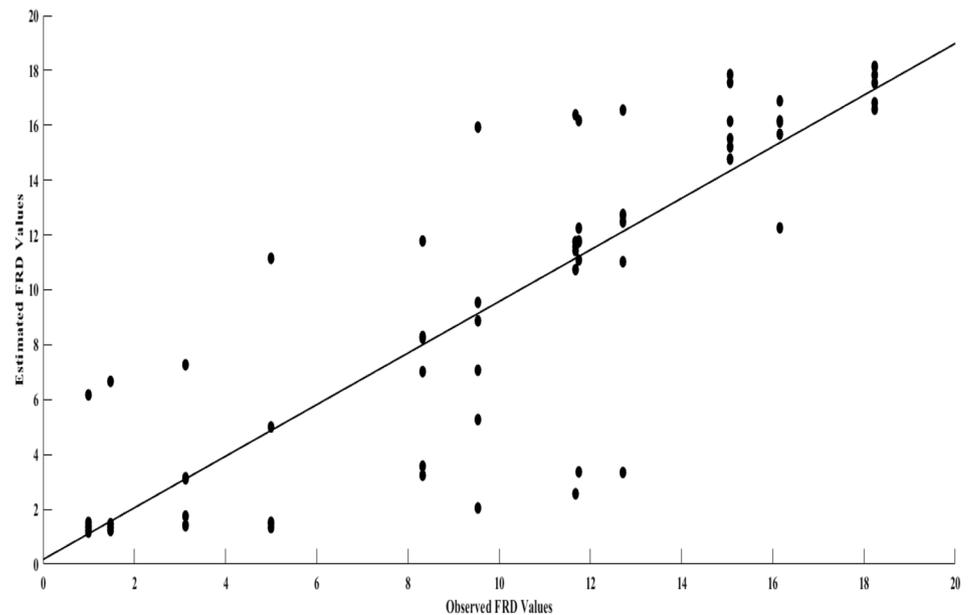
one box, and the daily mean of all those grids in that box is shown. Figure 8 shows the winter seasonal daily variations for 2004 year. In Fig. 8a, we can see daily mean variations in CAPE and LI over the Andhra Pradesh region during the year 2004. Both CAPE and LI have shown no signs of lightning activity. The LI values are positive. The highest values of CAPE were 260 J/kg which took place during the last week of November. The TPW values were also low, and only in the last week of November, there was a sudden increase as shown in Fig. 8b. The specific humidity values were also indicating no signs of lightning activity shown in Fig. 8c. Apart from the last week of November, the KI values were also indicating no signs of lightning activity as shown in Fig. 8d. The TTI values were also indicating no signs of lightning activity apart from November 20 to 27 as shown in Fig. 8e. The HI values were low indicating high availability of moisture as shown in Fig. 8f. Figure 8g, h indicates that CAPE fluctuations are linked with SAT and at 850 mb. Figure 8i indicates low RHUM values when there is no lightning activity. Figure 8j indicates that daily variation in CAPE and rainfall relationship. From Fig. 8, we conclude that lightning activity is less in winter. Though

the threshold values were not indicating severe lighting activity, the parameters such as HI, RHUM, TPW, TTI and KI have shown a sudden increase in November (20–27th) favoring rainfall. The threshold values of parameters were HI < 35 K; RH: 60–65%; TPW: 27–30 mm; TTI: 42–45 K; and KI: 295–300 K.

Figure 9 shows the pre-monsoon seasonal daily variations over the whole Andhra Pradesh region for 2004 year. In Fig. 9a, we can see daily mean variations in CAPE and LI over Andhra Pradesh region during the year 2004. Both cape and LI have shown good signs of lightning activity. The LI values are negative, especially in April and May months. CAPE values also show high thresholds for lightning in April and May. The highest values of CAPE were 1900 J/kg which took place during the May month. The TPW values were also indicating good availability of moisture in May month as shown in Fig. 9b. The specific humidity values were also indicating good signs of lightning activity shown in Fig. 9c. Apart from March month, the KI values were also indicating good signs of high lightning activity as shown in Fig. 9d. The TTI values were also indicating good signs of lightning activity as shown in Fig. 9e. The HI values were
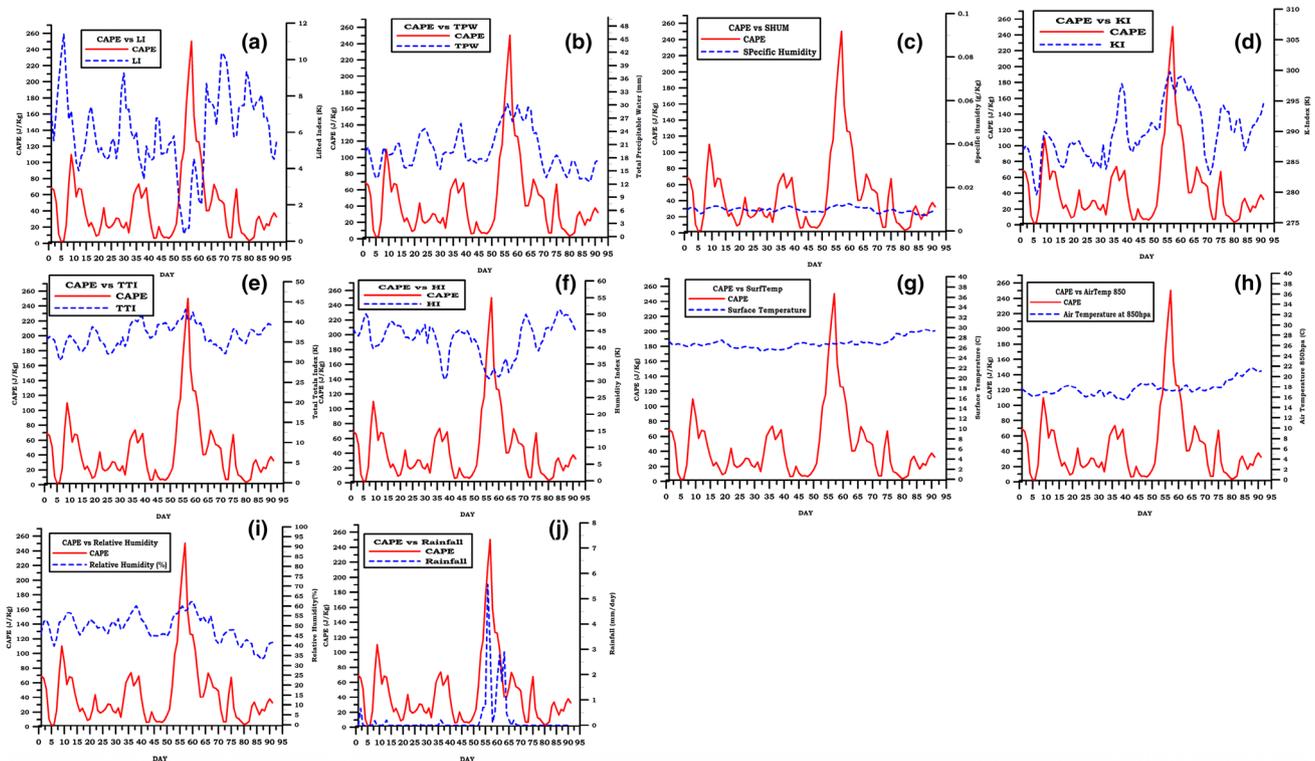

Fig. 8 Winter season daily variations in **a** CAPE and LI; **b** CAPE and TPW, **c** CAPE and specific humidity, **d** CAPE and KI; **e** CAPE and TTI; **f** CAPE and HI; **g** CAPE and surface temperature; **h** CAPE and air temperature (850 hpa); **i** CAPE and relative humidity; and **j** CAPE and rainfall for the year 2004
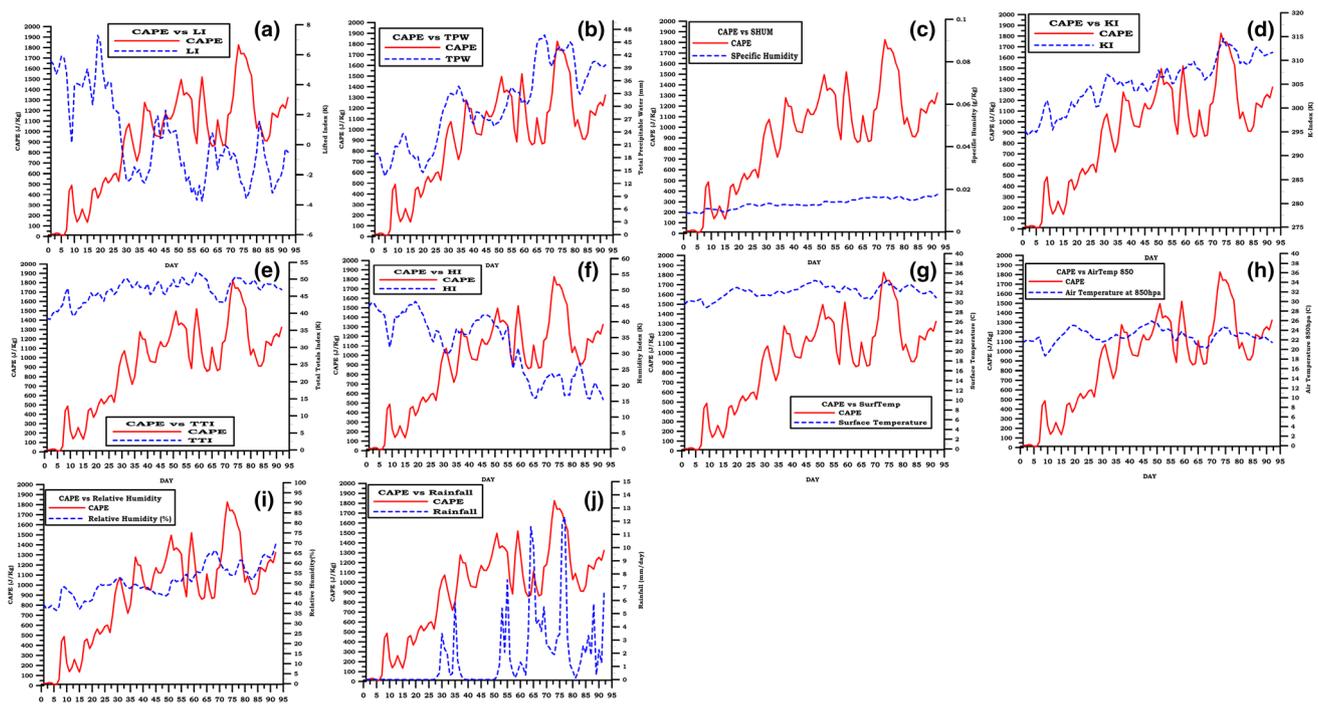
Fig. 9 Pre-monsoon season daily variations in **a** CAPE and LI, **b** CAPE and TPW, **c** CAPE and specific humidity, **d** CAPE and KI, **e** CAPE and TTI, **f** CAPE and HI, **g** CAPE and surface temperature, **h** CAPE and air temperature (850 hpa), **i** CAPE and relative humidity, **j** CAPE and rainfall for the year 2004

low indicating high availability of moisture as shown in Fig. 9f. Apart from lighting day, the HI values are again higher. Figure 9g, h indicates that CAPE fluctuations are linked with SAT and at 850 mb. Because of these variations, the atmosphere gets unstable. Figure 9i indicates low RH values when there is no lightning activity. But when there is a lightning activity, the RHUM values were suddenly increased. Figure 9j indicates that daily variation in CAPE and rainfall relationship. When there was high CAPE which indicates good lightning activity, there was sudden rainfall, especially in the months April and May. From Fig. 9, we conclude that lightning activity is high in pre-monsoon. When there is a lightning activity, the threshold values of parameters such as HI, RH, TPW, TTI and KI were indicating severe lighting activity, and when there is no lightning activity, the parameters showed no signs of lightning. The threshold values of parameters which indicated lightning activity were LI: -2 to -4 K; HI < 25 K; RH: 55–65%; TPW: 39–48 mm; TTI: 45–50 K; and KI: 310–320 K. During Pre-monsoon season, cumulonimbus clouds develop during the afternoon and evening hours which lead to heavy and intense thunderstorm occurrences. In monsoon season, the

influence of large-scale synoptic systems reduces lightning activity (Tinmaker et al. 2015).

Figure 10 shows the southwest monsoon seasonal daily variations for 2004 year. In Fig. 10a, we can see daily mean variations in CAPE and LI over the Andhra Pradesh region during the year 2004. In the starting week, i.e., first 10 days of June, the CAPE values have indicating moderate to severe thunderstorm possibility. After one week of June, you can see the chances for moderate lightning was also very less. Both cape and LI have shown very low lightning activity. The LI values ranging from -2 to -4 indicate weak to moderate chances for lightning. The highest values of CAPE were 1690 J/kg which took place during the first week of June. The TPW values were high values which is an indication of high moisture availability due to winds. This shows the good chances for rainfall activity than lightning as shown in Fig. 10b. The specific humidity values were also indicating less signs of lightning activity as shown in Fig. 10c. Apart from the first week of June, the KI values were constant indicating less signs of severe lightning activity as shown in Fig. 10d. The TTI values were also indicating small signs of moderate lightning activity apart from
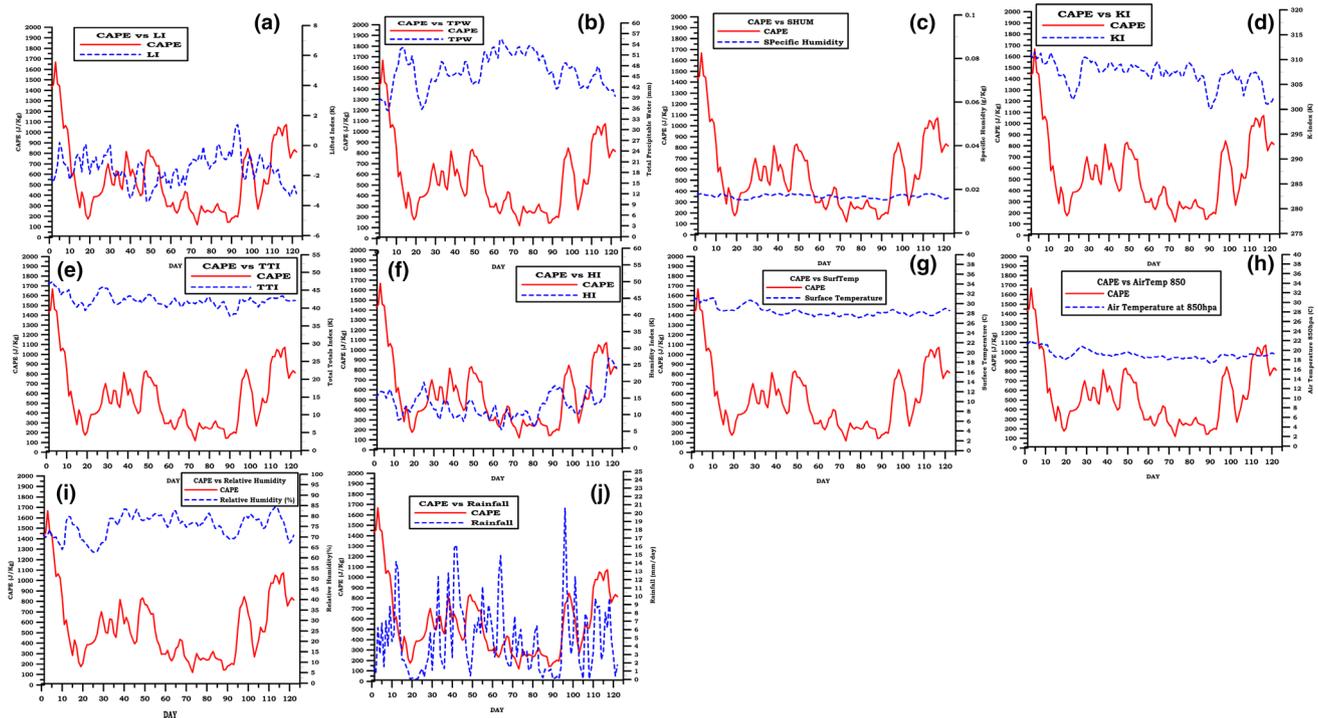
Fig. 10 Monsoon season daily variations in **a** CAPE and LI, **b** CAPE and TPW, **c** CAPE and specific humidity, **d** CAPE and KI, **e** CAPE and TTI, **f** CAPE and HI, **g** CAPE and surface temperature, **h** CAPE and air temperature (850 hpa), **i** CAPE and relative humidity, **j** CAPE and rainfall for the year 2004

01 to 05 June as shown in Fig. 10e. The HI values were low indicating high availability of moisture as shown in Fig. 10f. Figure 10g, 10 h indicates that CAPE fluctuations are linked with SAT and at 850 mb. Figure 10i indicates high RHUM values which favor high rainfall activity as shown in Fig. 10j. From Fig. 10, we conclude that lightning activity is less in monsoon season. Though the threshold values of CAPE were not indicating severe lighting activity, the parameters such as HI, RH, TPW, TTI and KI have shown good signs for monsoonal rainfall and lightning activity. The threshold values of parameters were HI < 20 K; RH: 70–85%; TPW: 48–55 mm; TTI: 37–44 K; and KI: 300–310 K.

Figure 11 shows the post-monsoon seasonal daily variations for 2004 year. In Fig. 11a, we can see daily mean variations in CAPE and LI over the Andhra Pradesh region during the year 2004. Both cape and LI have shown no signs of lightning activity. The LI values are positive. The highest values of CAPE were 1050 J/kg which took place during the mid week of October. The TPW values were also low, and only in October, there was a good moisture availability as shown in Fig. 11b. The specific humidity values were also indicating no signs of lightning activity shown in Fig. 11c. Apart from the first week of October, the KI

values were also indicating low signs of weak to moderate lightning activity as shown in Fig. 11d. The TTI values were also indicating low signs of moderate lightning activity as shown in Fig. 11e. The HI values were low in first 10 days of October indicating high availability of moisture which leads to rainfall as shown in Fig. 11f. Figure 11g, 11 h indicates that CAPE fluctuations are linked with SAT and at 850 mb. Figure 11i indicates low RH values when there is no lightning activity. Figure 11j indicates that daily variation in CAPE and rainfall relationship. From Fig. 11, we conclude that lightning activity is less in post-monsoon. Though the threshold values were not indicating severe lighting activity, the parameters such as HI, RHUM, TPW, TTI and KI have shown a sudden increase in October (01–15th) favoring rainfall. The threshold values of parameters which indicated sudden increase with no lightning activity were HI < 20 K; RH: 70–85%; TPW: 39–48 mm; TTI: 40–42 K; and KI: 302–310 K. The rainfall activity is highest in monsoon season. Though pre-monsoon season has received good showers, the highest rainfall activity was seen in monsoon season, especially in July month. When we observed CAPE values, highest values were observed in pre-monsoon season, especially in May month, and low CAPE values were seen in
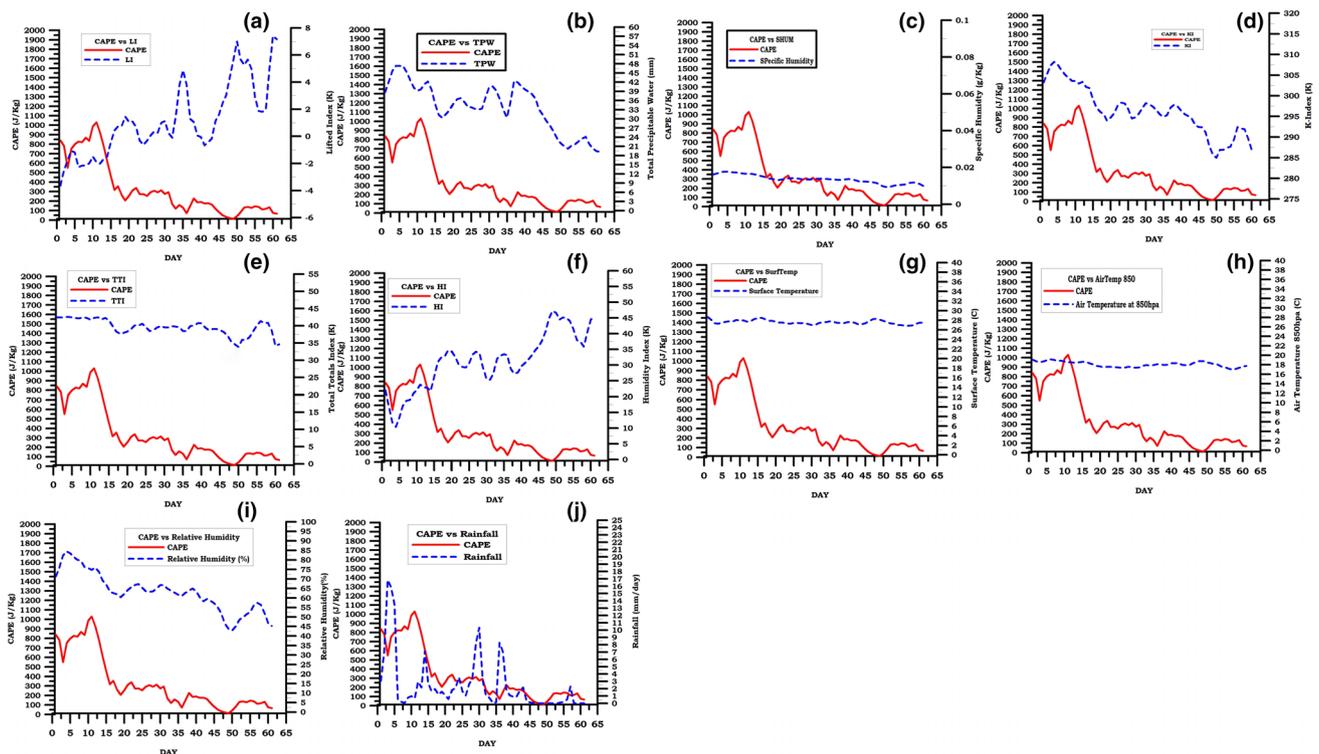
Fig. 11 Post-monsoon season daily variations in **a** CAPE and LI, **b** CAPE and TPW, **c** CAPE and specific humidity, **d** CAPE and KI, **e** CAPE and TTI, **f** CAPE and HI, **g** CAPE and surface temperature, **h** CAPE and air temperature (850 hpa), **i** CAPE and relative humidity, **j** CAPE and rainfall for the year 2004

monsoon season. Even though there is high rainfall activity in monsoon season, low CAPE values were seen.

During winter season, the mean KI values were ranging between 317 and 320 K over all regions of AP and Telangana. The Rayalaseema region and Telangana regions record 319–321 K mean KI values during 1995–2014. These threshold values are good enough for thunderstorms. During pre-monsoon season, the mean KI values were ranging between 323 and 326 K over all regions of AP and Telangana. The Telangana region records 325 K mean KI values during 1995–2014. These threshold values over Telangana and AP indicate that intense thunderstorms may occur. During monsoon season, the mean KI values are ranging from 318 to 319 K which is reasonably a good indication for thunderstorm occurrence. The KI parameter influence is almost same throughout the AP and TA regions indicating moderate chances for thunderstorm occurrence. During post-monsoon season, the mean KI values are ranging from 316 to 319 K which is reasonably a good indication for thunderstorm possibility. Out of all seasons, pre-monsoon season KI values are higher and it is a clear indication for the occurrence of

intense thunderstorms. The other seasons also indicated good thresholds for thunderstorm formation as shown in Fig. 12 during 1995–2014.

During winter season, the mean TTI values were ranging between 33 and 39 K over most regions of AP and TA. These threshold values are not good enough for the thunderstorm formation. During pre-monsoon season, the mean TTI values were ranging between 44 and 48 K over all regions of AP and TA. These threshold values indicate the possibility for intense thunderstorm occurrences. During monsoon season, the mean TTI values are ranging from 42 to 43 K which indicates very few chances for the thunderstorm occurrence. During post-monsoon season, the mean TTI values are ranging from 38 to 40 K which shows that very few chances for small isolated thunderstorm. Out of all seasons, pre-monsoon season TTI values are higher and it is a clear indication for the occurrence of severe intense thunderstorms as shown in Fig. 13 during 1995–2014.

During winter season, the mean HI values were ranging between 30 and 42 K over most regions of AP and TA. These threshold values indicate very few chances for thunderstorm
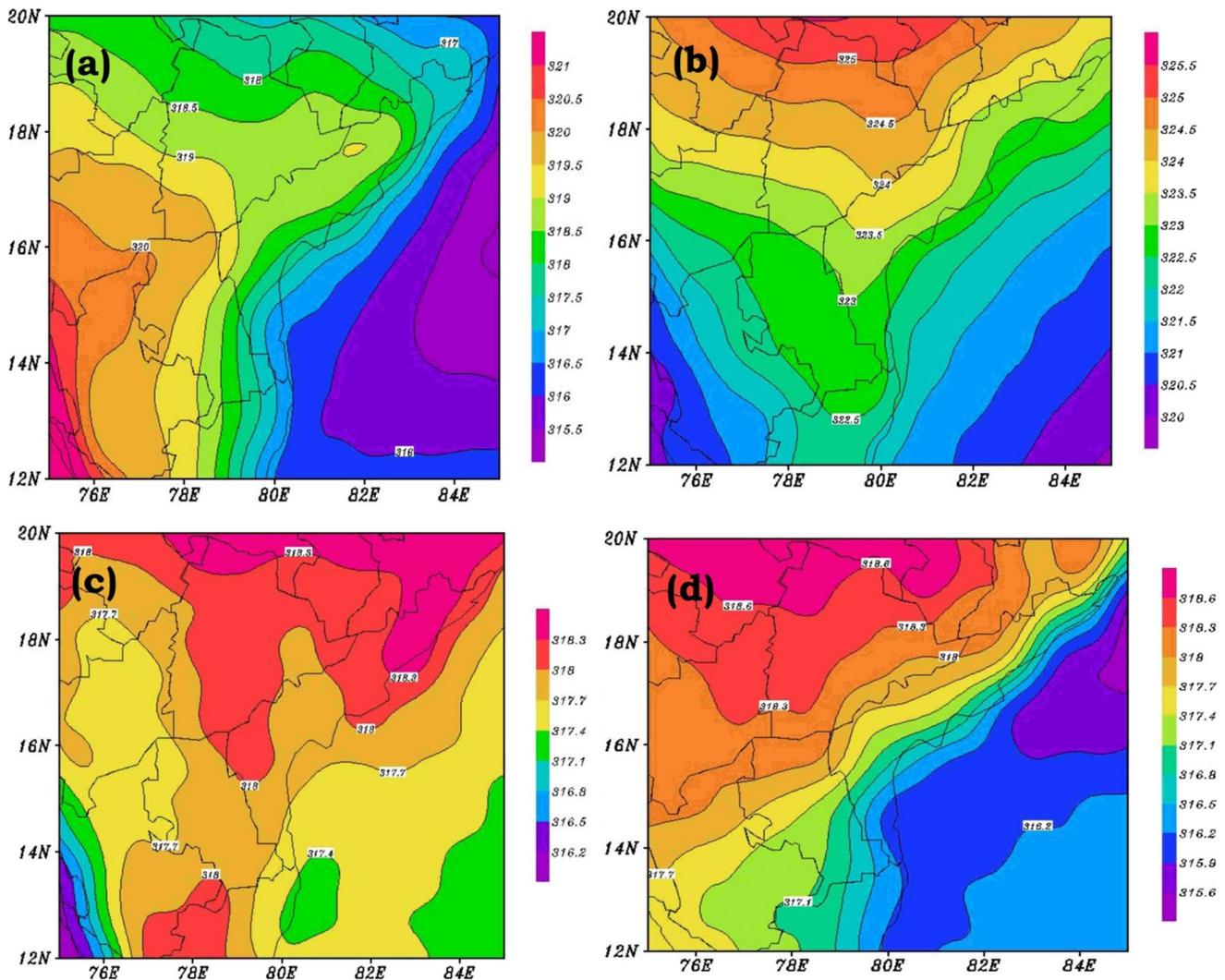
Fig. 12 Spatial distribution of mean KI parameter for **a** winter season, **b** pre-monsoon season, **c** monsoon season and post-monsoon season during the time period of 1995–2014

occurrence. During pre-monsoon season, the mean HI values were ranging between 33 and 37 K over all regions of AP and TA during 1995–2014. These threshold values also indicate very few chances for thunderstorm occurrence. During monsoon season, the mean HI values are ranging from 14 to 18 K which indicates high chances for the thunderstorm occurrence. In both pre-monsoon and monsoon seasons, north AP has low HI values which indicates the high chances for thunderstorm occurrences. During post-monsoon season, the mean HI values are ranging from 24 to 34 K which is reasonably a good indication for moderate thunderstorm possibility. Rayalaseema region has less HI values which shows high moisture availability during this season. Out of all seasons, monsoon season

HI values were below 30 K which is a clear indication for the occurrence of severe intense thunderstorms as shown in Fig. 14 during 1995–2014.

During winter season, the mean LI values were ranging between 3 to 7 K over most regions of AP and TA which is a good indication of no thunderstorm activity during 1995–2014. During pre-monsoon season, the mean LI values were ranging between 1 and − 2 K. These threshold values indicate very few chances for the thunderstorm formation. During monsoon season, the mean LI values are ranging from − 0.3 to − 2.1 K which indicates a good chances for the thunderstorm occurrence, especially in northeastern regions of AP. During post-monsoon season, the mean LI values
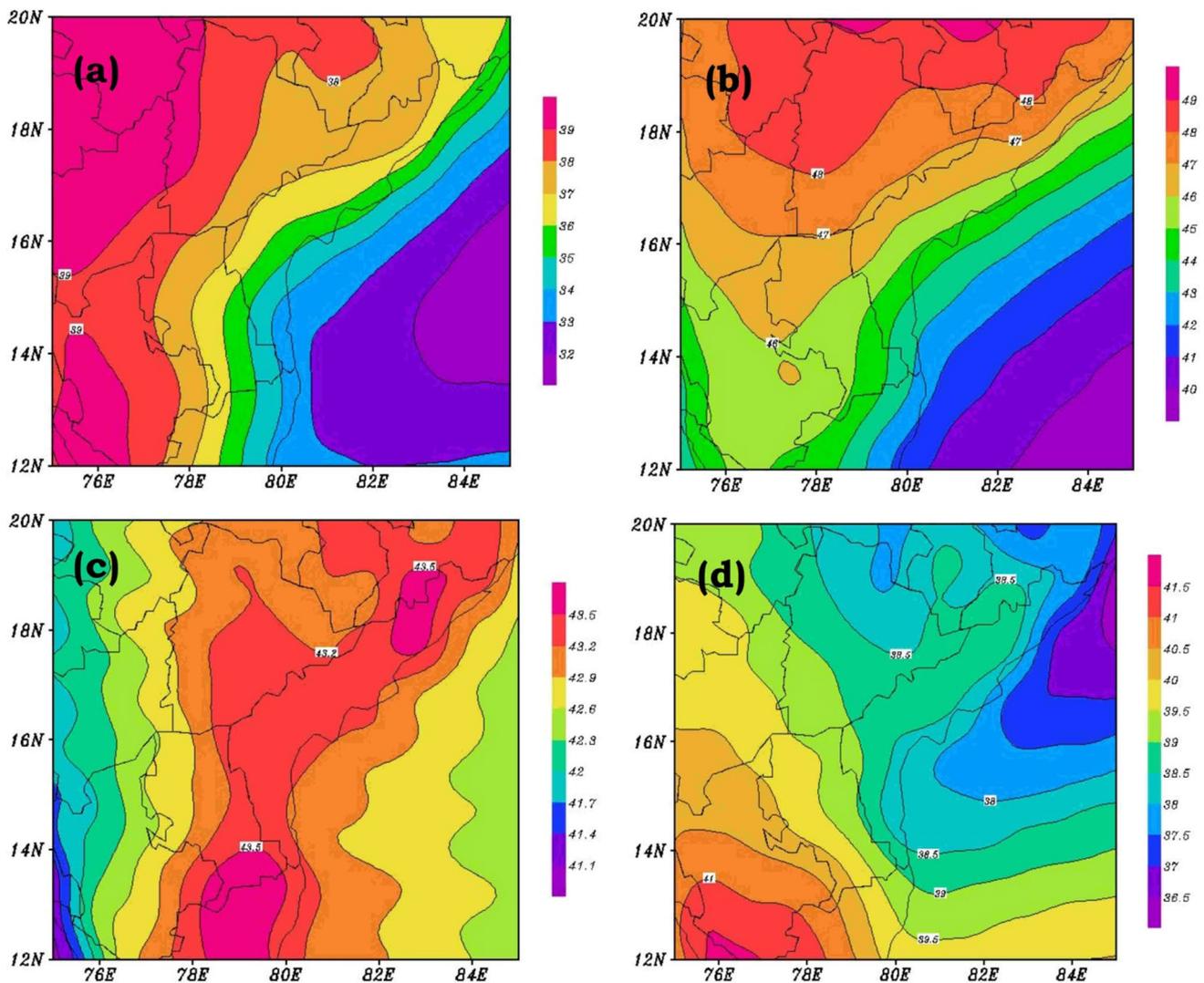
Fig. 13 Spatial distribution of mean TTI parameter for **a** winter season, **b** pre-monsoon season, **c** monsoon season and post-monsoon season during the time period of 1995–2014

are ranging from 4 to −0.5 K over AP and TA. On TA, the positive LI values indicate very less thunderstorm chances, whereas negative values over southeastern region of AP indicate good chances for moderate thunderstorm formation as shown in Fig. 15 during 1995–2014.

As shown in Fig. 16, the mean CAPE values were ranging between 0 and 100 J/kg over most regions of AP and TA during winter season. These threshold values are not good enough for the thunderstorm formation. During pre-monsoon season, the mean CAPE values were ranging between 300 and 1200 J/kg over all regions of TA and Rayalaseema regions. The mean CAPE values were higher over coastal AP, i.e., 1200–2700 J/kg. The CAPE values showed nearly

3000 J/kg over north AP. These threshold values indicate the possibility of intense thunderstorm occurrences. During monsoon season, the mean CAPE values were ranging from 300 to 1000 J/kg which indicates very few chances for the thunderstorm occurrence. During post-monsoon season, the mean CAPE values were ranging from 150 to 600 J/kg which shows that very few chances for small isolated thunderstorm. Out of all seasons, in the pre-monsoon season CAPE values are higher and it is a clear indication of the occurrence of severe intense thunderstorms during 1995–2014.

During winter season, the mean TPW values were ranging between 21 and 28 mm over most regions of AP and TA. These threshold values indicate very few chances for
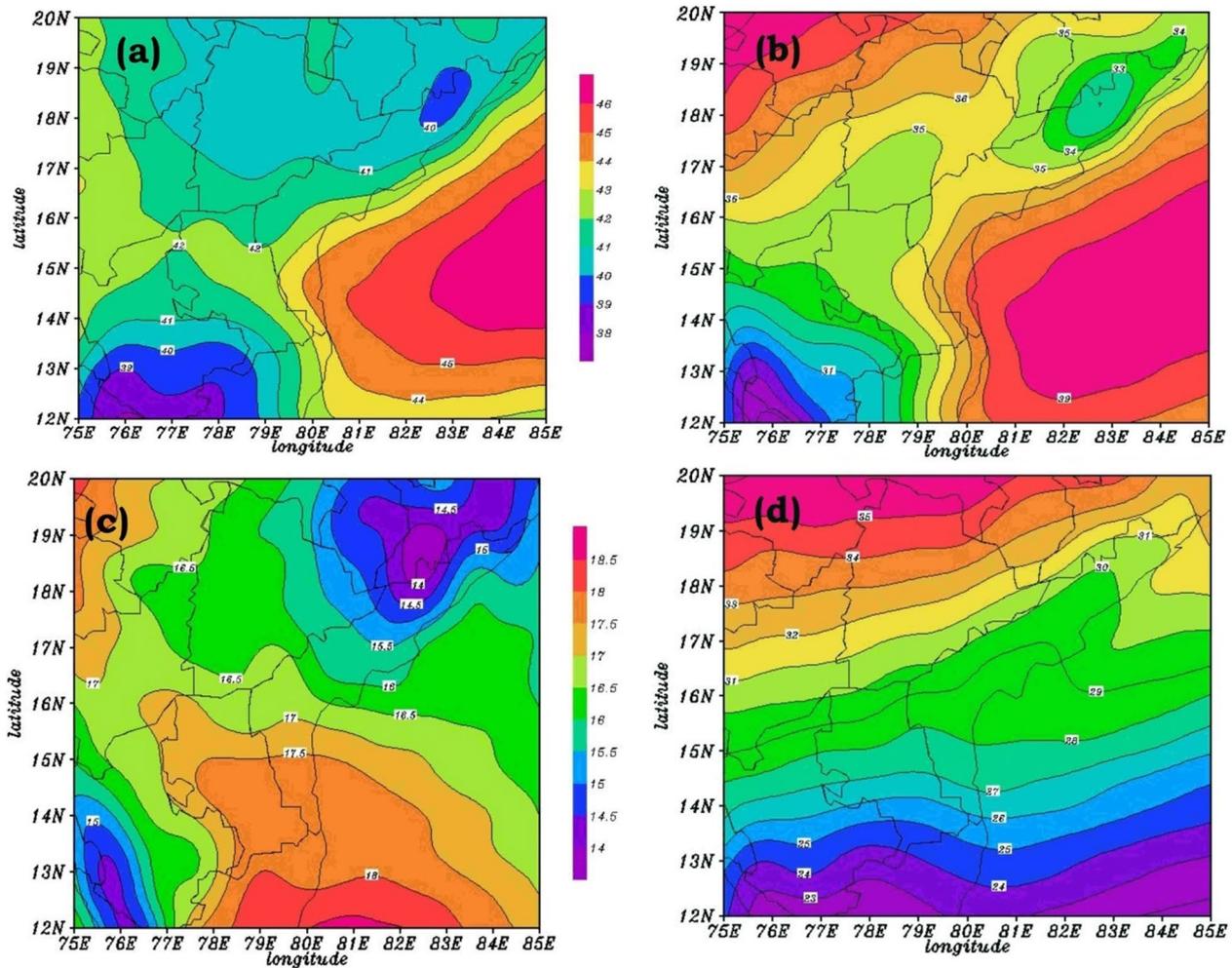
Fig. 14 Spatial distribution of mean HI parameter for **a** winter season, **b** pre-monsoon season, **c** monsoon season and post-monsoon season during the time period of 1995–2014

thunderstorm occurrence. During pre-monsoon season, the mean TPW values were ranging between 30 and 42 mm over all regions of AP and TA during 1995–2014. These threshold values also indicate good chances for isolated small thunderstorm occurrence. During monsoon season, the mean TPW values are ranging from 46 to 60 mm which indicates high chances for the thunderstorm occurrence. Both in pre-monsoon and monsoon seasons, north AP has high TPW values which indicates the high chances for thunderstorm occurrences. During post-monsoon season, the mean TPW values were ranging from 30 to 46 mm which is reasonably a good indication of moderate thunderstorm possibility. Rayalaseema region also exhibits high TPW values which show high moisture availability during monsoon season. Out of all seasons, in monsoon season TPW values were above 42 mm which is a clear indication of the occurrence of severe intense thunderstorms as seen in Fig. 17 during 1995–2014.

During winter season, the mean FRD values were ranging between 0 to 10 over most regions of AP and TA except north AP which is a good indication of no thunderstorm activity during 1995–2014. During pre-monsoon season, the mean FRD values were ranging between 20 and 100. These threshold values indicate very high chances for the thunderstorm formation. During monsoon season, the mean FRD values were ranging from 20 to 80 which indicate good chances for the thunderstorm occurrence, especially in northeastern and southeastern regions of AP. During post-monsoon season, the mean FRD values were ranging from 10 to 20 over AP and TA. On TA and the FRD values indicate very less thunderstorm chances, whereas FRD values
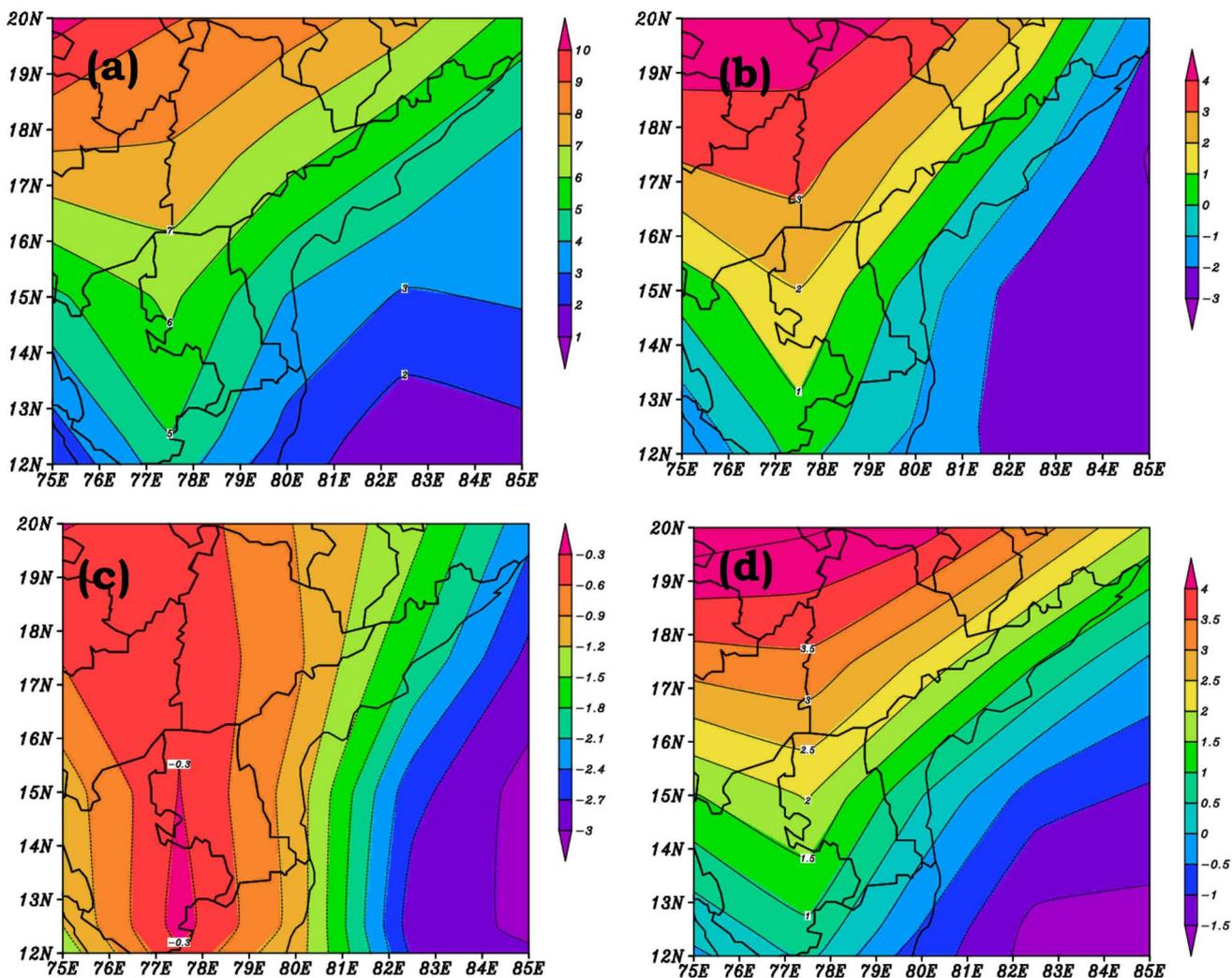
Fig. 15 Spatial distribution of mean LI parameter for **a** winter season, **b** pre-monsoon season. **c** Monsoon season and post-monsoon season during the time period of 1995–2014

over southeastern region of AP indicate good chances for moderate thunderstorm formation as seen in Fig. 18 during 1995–2014.

As shown in Fig. 19, the mean RF values were ranging between 10 and 40 mm over most regions of AP and TA. These RF values indicate very few chances for the convection. During pre-monsoon season, the mean RF values were ranging between 10 and 40 mm over most regions of AP and TA during 1995–2014. The north-east AP receives RF ranging between 40 and 80 mm due to intense thunderstorm activity in summer season. During monsoon season, the mean RF values are ranging from 100 to 400 mm which indicates high rainfall activity. Both in pre-monsoon and monsoon seasons, north AP has high RF values which indicates the high convection. During post-monsoon season, the mean RF values were ranging from 40 to 100 mm over

TA which is reasonably a good indication for convection, whereas AP receives 100–250 mm rainfall, especially over south-east Rayalaseema region. This is due to medium-scale synoptic systems during this season. Out of all seasons, monsoon season records high rainfall over all parts of TA and AP regions during 1995–2014.

Conclusions

The present paper is an analysis of convective and stability parameters and their influence on FRD over Andhra Pradesh during 20-year period (1995–2014). The analysis has shown the following principal results. The monthly
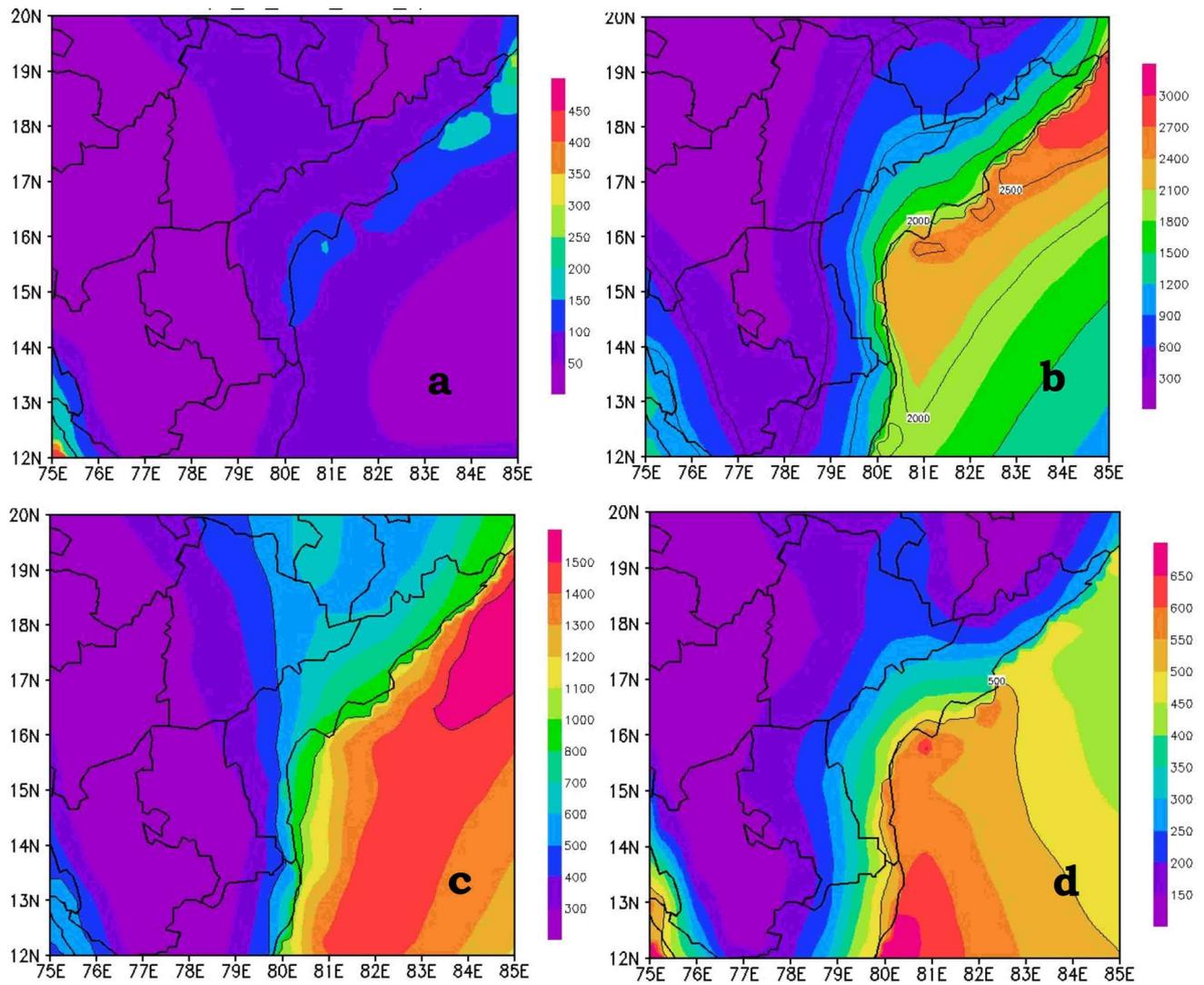
Fig. 16 Spatial distribution of mean CAPE parameter for **a** winter season, **b** pre-monsoon season, **c** monsoon season and post-monsoon season during the time period of 1995–2014

variation in FRD and SAT increased uniformly from January to May and started to decline from June to December. The maxima FRD is seen in May with a value 18.23 fl km − 2 month − 1, and minima FRD is seen over January with a value 0.99 fl km − 2 month − 1. The minimum SAT value of 24.76 °C is seen in December, whereas the maximum SAT value of 32.94 °C is seen for May month. The correlations between FRD and SAT, and FRD and AT (at 850 mb) are 0.87 and 0.85, respectively. The daily variations in SAT and AT also showed a uniform increase with respect to CAPE. The monthly variation in FRD and RF exhibits

highest amounts of rainfall during monsoon months, especially July. The May month has shown highest rainfall in pre-monsoon months. The daily variation in CAPE and RF also suggests that when CAPE values were ranging between 800 and 2000 J/kg, there was good rainfall in pre-monsoon, i.e., RF values were ranging from 2 to 13 mm which is mainly due to lightning activity. The monthly variation in FRD and CAPE reveals highest values of CAPE were seen during pre-monsoon months such as April and May. The highest CAPE values for May month indicate its influence on lightning. The monthly variation in FRD
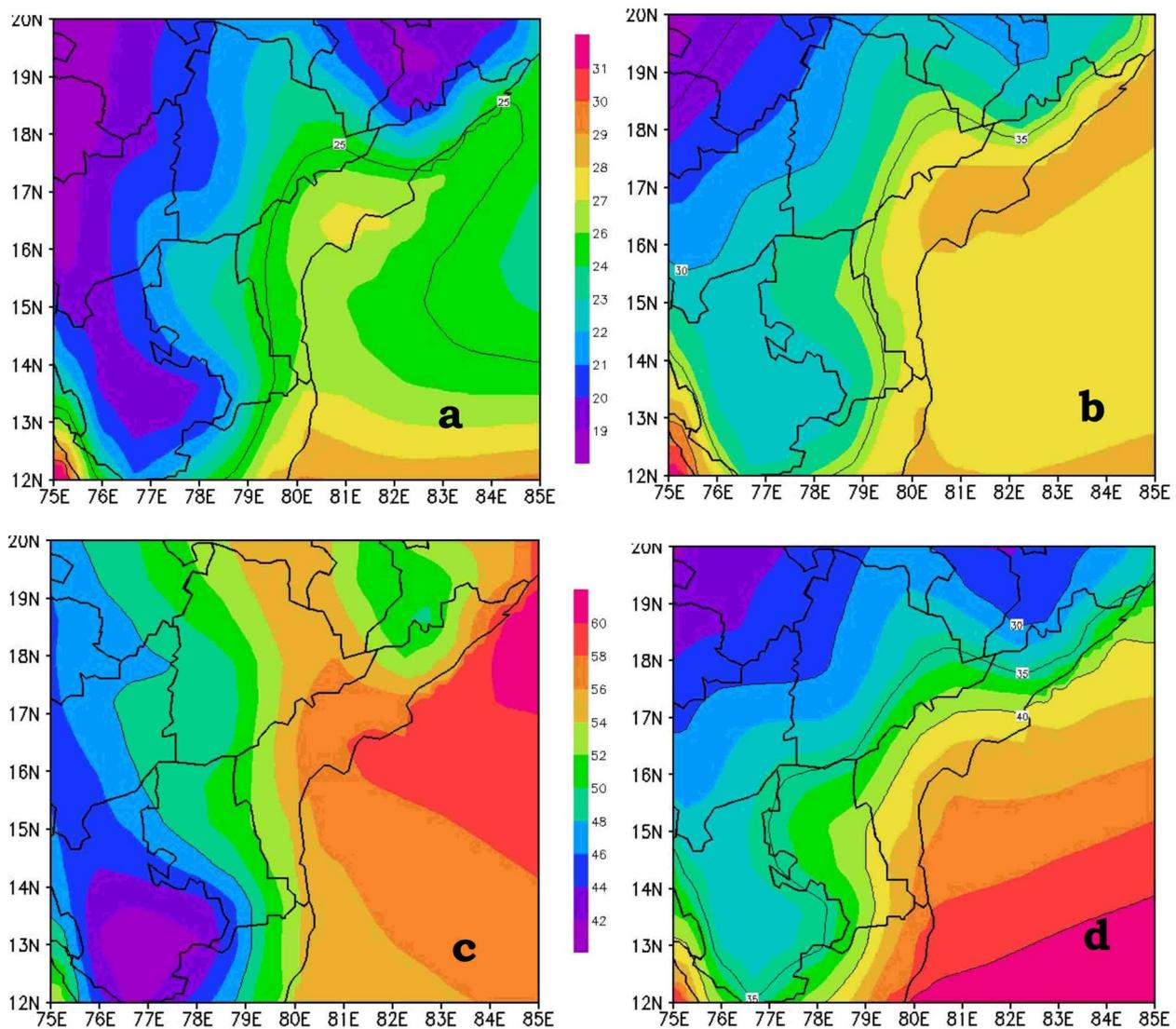
Fig. 17 Spatial distribution of mean TPW parameter for **a** winter season, **b** pre-monsoon season, **c** monsoon season and post-monsoon season during the time period of 1995–2014

and LI shows strong negative correlation. The monthly variations in FRD with KI and TTI show highest values over pre-monsoon months, which is a clear indication of high lightning strikes over Andhra Pradesh. Highest KI mean value of 324 K was seen for May month, whereas highest mean TTI value is 47 K. The monthly variation in FRD and HI displays the lowest values over monsoon and post-monsoon months which is a clear indication of high moisture availability. The mean HI values for pre-monsoon months show good threshold values which contribute to high lightning activity over Andhra Pradesh. The monthly variation in FRD and TPW exhibits highest values over

monsoon and post-monsoon months. Highest TPW mean value of 53.48 K was seen for July month. May and April months are also showing good thresholds for convection. In this paper, a multilayer perceptron ANN model network was utilized for the estimation of FRD. We used six training learning algorithms of artificial neural network to estimate FRD. The ANN model setup was almost same for the usage of six algorithms. When we gave convection parameters as inputs to ANN model, SCG has shown the highest correlation for the estimation of lightning activity, whereas ANN model with RP showed the lowest correlation. We also attempted the training of ANN with stability
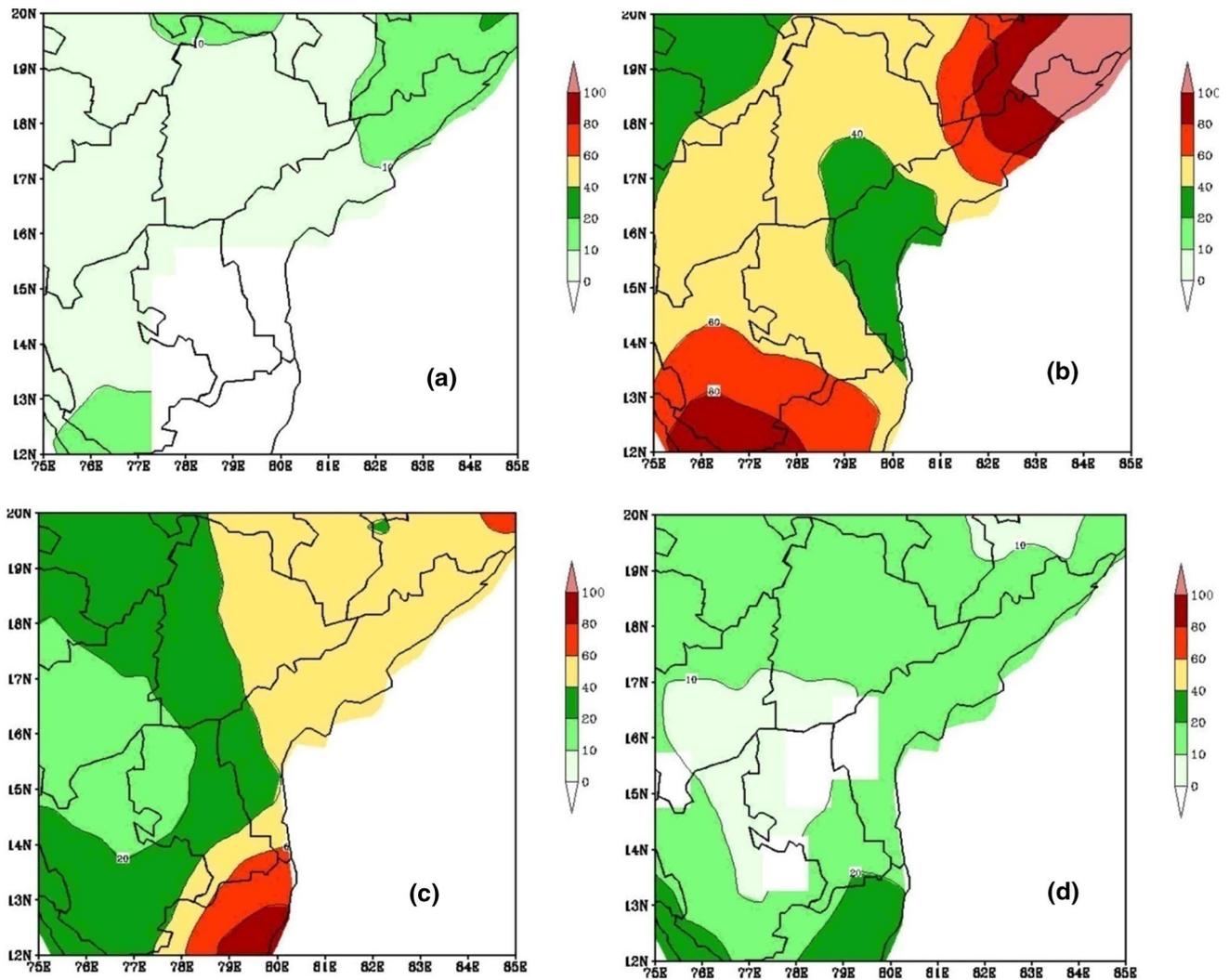
Fig. 18 Spatial distribution of mean FRD parameter for **a** winter season, **b** pre-monsoon season, **c** monsoon season and post-monsoon season during the time period of 1995–2014

parameters; LM and R have shown the highest correlation for the estimation of lightning activity. RP showed the lowest correlation. The estimation results reveal that the ANN model, which is trained by SCG, can accurately estimate the FRD, and it is having better estimation when compared to other learning algorithms. Daily variations in parameters such as HI, RH, TPW, TTI and KI have shown good signs for lightning activity when they have touched the thresholds for severe thunderstorm. When there is no lightning activity, the values were normal. In future prospective, we like to use more algorithms to study and estimate the lightning activity.

Based on FRD, the values of all parameters which account for severe convection are expressed in terms of threshold values. SAT > 28 C; at 850 mb > 20 C; SHUM > 15 g/kg; RHUM > 65%; RF > 0.5 mm; CAPE > 1500 J/kg; LI < −0.5 K; KI > 320 K; TTI > 42 K; HI < 35 K and TPW > 40 mm. All parameters showed good thresholds for thunderstorm occurrence. On the other hand, the values of SAT < 24 C; at 850 mb < 16 C; SHUM < 10 g/kg; RHUM < 50%; RF < 0.5 mm; CAPE < 600 J/kg; LI > 0.5 K; KI < 316 K; TTI < 36 K; HI > 45 K and TPW < 30 mm showed no convective activity.
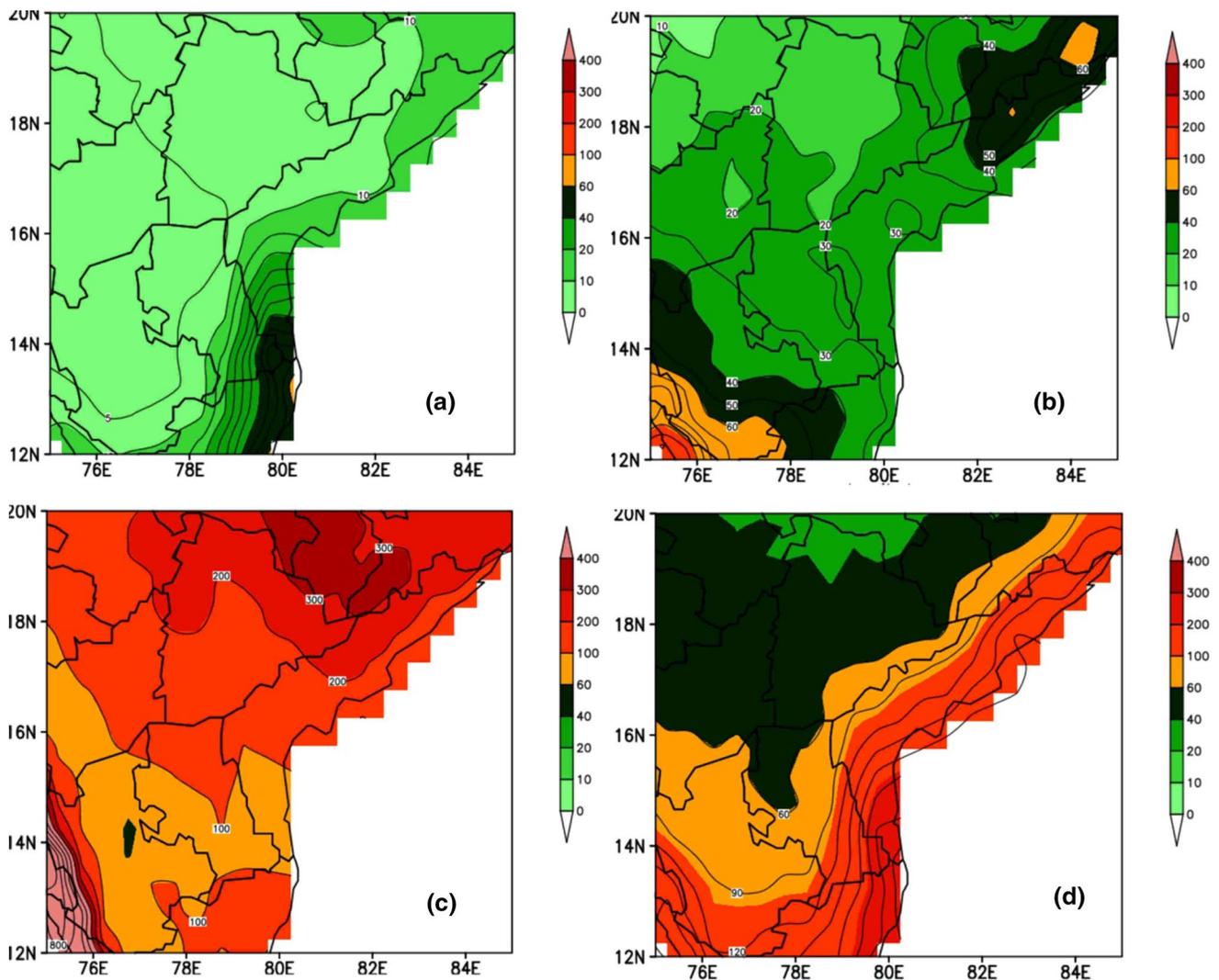
Fig. 19 Spatial distribution of mean RF parameter for **a** winter season, **b** pre-monsoon season, **c** monsoon season and post-monsoon season during the time period of 1995–2014

Compliance with ethical standards

Conflict of Interest The authors declare that there is no conflict of interests regarding the publication of this paper.

References

Boccippio DJ, Koshak WJ, Blakeslee RJ (2002) Performance assessment of the optical transient detector and lightning imaging sensor. Part I: predicted diurnal variability. J Atmos Ocean Technol 19:1318–1332

Cecil, Daniel J (2006) *LIS/OTD 0.5 degree high resolution monthly climatology (HRMC)* [indicate subset used]. Dataset available online from the NASA Global Hydrology Resource Center DAAC, Huntsville, Alabama, U.S.A. DOI: http://dx.doi.org/10.5067/LIS/LIS-OTD/DATA303

Chattopadhyay S (2007) Feed forward artificial neural network model to predict the average summer-monsoon rainfall in India. Acta Geophys 55(3):369–382

Chaudhuri S, Chattopadhyay S (2005) Neuro-computing based short range prediction of some meteorological parameters during the pre-monsoon season. Soft Comput 9(5):349–354

Christian HJ, Blakeslee RJ, Boccippio DJ, Boeck WL, Buechler DE, Driscoll KT, Goodman SJ, Hall JM, Koshak WJ, Mach DM, Stewart MF (1999) Global frequency and distribution of lightning as observed from space by the Optical Transient Detector. J. Geophysical Research. 108(D1):4005

Chen LV, Xing Y, Zhang J, Na X, Li Y, Liu T, Cao D, Wang F-Y (2017) Levenberg-Marquardt back propagation training of multilayer neural networks for state estimation of a safety critical cyber-physical system. IEEE Trans Ind Inf 99:1–1

Cömert Z, Kocamaz AF (2017) A study of artificial neural network training algorithms for classification of cardiotocography signals. Bitlis Eren Univ J Sci Technol 7(2):93–103

Das Y (2015) Some aspects of thunderstorm over India during pre-monsoon season: a preliminary report-I. J Geosci Geomat 3(3):68–78

Galway GJ (1956) The Lifted index as a predictor of latent instability. Bull Am Meteor Soc 37:528–529

George GJ (1960) Weather forecasting for aeronautics. Academic Press, New York, p 673

Gomes C, Kadir MZAA (2011) A theoretical approach to estimate the annual lightning hazards on human beings. Atmos Res 101:719–725

Huang G-B (2006) Universal approximation using incremental constructive feed forward networks with random hidden nodes. IEEE Trans Neural Netw 17(4):1–4

Jacovides CP, Yonetani T (1990) An evaluation of stability indices for thunderstorm prediction in greater cyprus. Wether Forcast 5(4):559–569

Johns RH, Doswell CA III (1992) Severe local storms forecasting. Wea Forecast 7:588–612

July Hagan MT, Menhaj MB (1994) Training feedforward networks with the Marquardt algorithm. IEEE Trans Neural Netw IEEE 5(6):989–993

Kandalgaonkar SS, Tinmaker MIR, Kulkarni JR, Nath A, Kulkarni MK (2005a) Spatio-temporal variability of lightning activity over the Indian region. J Geophys Res 110(D11):110–118

Kandalgaonkar SS, Tinmaker MIR, Nath AS, Kulkarni MK, Trimbake HK (2005b) Study of thunderstorm and rainfall activity over the Indian region. Atmosfera 18:91–101

Kunz M (2007) The skill of convective parameters and indices to predict isolated and severe thunderstorms. Nat Hazards Earth Syst Sci 7:327–342

Lavigne T, Liu C, Liu N (2019) How does the trend in thunder days relate to the variation of lightning flash density? J Geophys Res Atmos 124(9):4955–4974

Liou YA, Kar SK (2010) Study of cloud-to-ground lightning and precipitation and their seasonal and geographical characteristics over Taiwan. Atmos Res 95:115–122

Litta A, Mary Idicula S, Mohanty UC (2013) Artificial neural network model in prediction of meteorological parameters during premonsoon thunderstorms. Int J Atmos Sci 10:525–535

Manohar GK, Kesarkar AP (2005) Climatology of thunderstorm activity over the Indian region: III. Latitudinal and seasonal variation. Mausam 56(3):581–592

Manohar GK, Kandalgaonkar SS, Tinmaker MIR (1999) Thunderstorm activity over India and the Indian southwest monsoon. J Geophys Res 104(D4):4169–4188

Miller RC (1967) Notes on analysis and severe storm forecasting procedures of the Military Weather Warning Center. Tech. Report 200, AWS, USAF.[Headquarters, AWS, Scott AFB, 1L 62225]

Moller MF (1993) A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw 6:525–533

Murugavel P, Pawar SD, Gopalakrishan V (2014) Climatology of lightning over Indian region and its relationship with convective available potential energy. Int J Climatol 34:3179–3187

Pinjare LS, Arun Kumar M (2012) Implementation of neural network back propagation training algorithm on FPGA. Int J Comput Appl 52(6):1

Saha UD, Singh SK, Midya RP, Singha AK, Kumar S (2017) Spatio-temporal variability of lightning and convective activity over South/South-East Asia with an emphasis during El Niño and La Niña. Atmos Res 197:150–156

Singh O, Singh J (2015) Lightning fatalities over India: 1979–2011. Meteorol Appl 22(4):770–778

Tinmaker MIR, Ali K, Beig G (2010) Relationship between lightning activity over Peninsular India and sea surface temperature. J Appl Meteorol Climatol 49:828–835

Tinmaker MIR, Aslam MY, Chate DM (2015) Lightning activity and its association with rainfall and convective available potential energy over Maharashtra, India. Nat Hazards 77:293

Tyagi A (2007) Thunderstorm climatology over Indian Region. Mausam 58(2):189–212

Vaddadi S, Remella VV, Vaddadi RS (2015) Dermatological manifestation of lightning injury: a case report. Am J Med Sci 2(1):30–33

Weng LY, Omar J, Siah YK, Ahmed SK, Abidin IZ (2010) Lightning forecasting using ANN-BP & Radiosonde. In: Proceeding of International conference onintellegent computing and cognitive informatics, Kulala Lumpur, Malaysia, pp 153–155

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Academic Press, London

Comparison of emissivity retrieval methods from ASTER data using Fourier-Transform Infrared Spectroscopy

Silvia Beatriz Alves Rolim[1] · Bijeesh Kozhikkodan Veettil[2,3] · Pâmela Suélen Käfer[1] · Atilio Efrain Bica Grondona[1] · María Luján Iglesias[1] · Lucas Ribeiro Diaz[1] · Cristiano Lima Hackmann[1]

Abstract

Land surface emissivity retrieval is important for the remote identification of natural materials and can be used to identify the presence of silicate minerals. However, its estimation from passive sensors involves an undetermined function related to radiance data, which is influenced by the atmosphere. We tested three methods for temperature emissivity retrieval in a dune field composed of 99.53% quartz (SiO_2) using Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) imagery. The tested methods were the reference channel method (RCM), emissivity normalization method (ENM), and temperature emissivity separation (TES) method. An average quartz reference spectrum for the dune samples was calculated from an emissivity database based on temperature and used to evaluate the emissivity products of four ASTER images. In general, the three tested methods had a good approximation when analysed the emissivity reference curve, especially for longer wavelengths that ranged between 2 and 4% of emissivity. The RCM and ENM produced very similar results with the coefficients of determination (R^2) as 0.9960 (RMSE 0.0184) and 0.9959 (RMSE 0.0185), respectively. RCM method presented superior results (R^2: 0.9960, RMSE: 0.0184), compared to the TES method (R^2: 0.9947, RMSE: 0.0197). The TES method showed good results only for shorter wavelengths and, hence, to identify specific targets using ASTER data, such as silicate minerals, it is better to use the RCM method. The emissivity value selected at the saturation point of the spectral library based on temperature is fundamental in acquiring more reliable data.

Introduction

Land surface emissivity (LSE) is an important physical property in geological mapping, particularly in the exploration of ore deposits associated with silicates in mafic and ultramafic rocks (Christensen et al. 2000). These minerals do not always present expressive diagnostic features in their spectral characteristics in the visible and shortwave infrared regions (Christensen et al. 2000). Land surface temperature and emissivity are independent variables, and the thermal infrared spectral radiance measured in remote sensing is dependent on both (Dash et al. 2002; Gillespie et al. 2011). Therefore, the inverse Planck equation is underdetermined, with two unknowns and a single measurement. Practical inversion algorithms designed to calculate temperature (T) and emissivity (ε) from the measurements cannot do the separation (between T and ε) accurately because the recovered temperature and emissivity may co-vary. The estimation of temperature and emissivity from satellite data is influenced by the atmosphere. Furthermore, it is necessary to consider the possible instrument noise and calibration errors in addition to the effects of a spectral mixture on a sub-pixel basis (Collins et al. 2001).

Several methods have been developed for retrieving LSE from satellite imagery as well as the atmospheric correction for the thermal infrared (TIR) data (Constantino and

✉ Bijeesh Kozhikkodan Veettil
bijeesh.veettil@tdtu.edu.vn

1. Geologic Remote Sensing Laboratory (LabSRGeo), Centro Estadual de Pesquisas em Sensoriamento Remoto e Meteorologia, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande Do Sul, Brazil

2. Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam

3. Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Angelini 2016). According to Li et al. (2013a, b), emissivity retrieval methods are broadly classified into three types: (1) Semi-empirical (based on the relationship among the information in the visible, shortwave infrared and thermal infrared spectral regions); (2) Multi-channel emissivity separation (directly from emitted radiance); and (3) Physics-based (restrictions based on the additional unknowns due to the spectral absorption and emission in the intervening atmosphere). In types (1) and (3), it is assumed that the atmospheric effects on the radiances measured at the top of atmosphere (TOA) have been accurately corrected or that the radiances are measured at ground level (Li et al. 2013a, b). However, these three types of emissivity retrieval methods have been developed for specific applications and the results may depend on sensor type and the target analysed (Rolim et al. 2016). Most of the geological studies use multi-channel emissivity retrieval methods and the common ones are the reference channel method (RCM), emissivity normalization method (ENM), and temperature emissivity separation (TES) (Gillespie et al. 1986; Watson 1992; Kealy and Gabell 1990). TES method is mainly used to derive temperature (T) and emissivity (ε) for soil and mineral substrates.

The objective of this study was to evaluate the performance of the RCM, ENM and TES algorithms in the retrieval of temperature and emissivity using ASTER thermal infrared data. The analysis was carried out in a sub-tropical coastal dune field, which is composed of 99.53% of quartz (SiO2), using a temperature-based spectral library of quartz (SiO2) mineral as a reference curve (Käfer et al. 2019).

Study area

The study area is one of the remaining transgressive dunes at Cidreira Beach, extending approximately 30 km^2, and located on the North Coast of Rio Grande do Sul state in southern Brazil (Fig. 1). The dune field has inter-dune regions that generally present undergrowth and temporary water bodies, mainly in the winter periods, due to the nearby presence of the groundwater table (Tomazelli et al. 2008). The dunes are characterized by the presence of fine sand (125–250 μm) with different grain shapes, such as sub-rounded (68%), rounded (18%) and sub-angular (14%), and are composed of quartz (99.53%) and heavy minerals (0.47%).

Data and methodology

ASTER satellite data

ASTER was launched in December 1999 on-board NASA's Terra platform (JPL 2019). Despite the discontinuity of its shortwave infrared (SWIR) sensor, its TIR and VNIR (visible and near infrared) channels are still widely used (NASA 2019). We used four cloud-free ASTER Level 1T scenes acquired between 2011 and 2018 as well as the corresponding Level 2 AST_05 data (Table 1). The ASTER Level 1 Precision Terrain Corrected Registered At-Sensor Radiance (AST_L1T) data contains calibrated at-sensor radiance. This multi-file product includes data in VNIR, SWIR and

Fig. 1 Map of the study area in the northern coast of the Rio Grande do Sul state with dune field highlighted with the red polygon (RGB colour composite of ASTER image acquired on 09/29/2015)
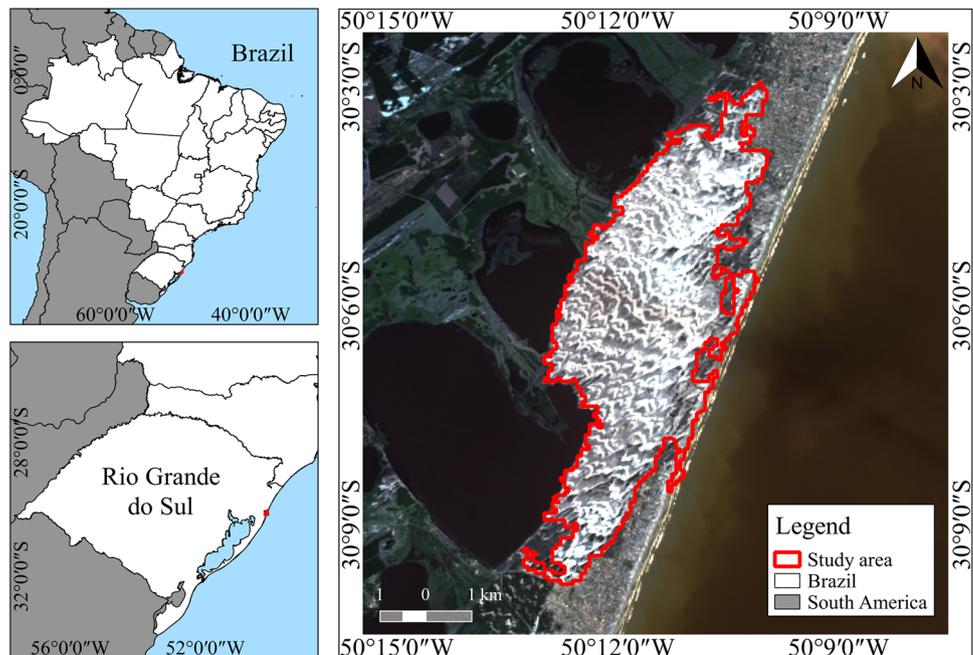
Table 1 Details of the ASTER scenes used for this study

Date of acquisition	Path	Row	Sun elevation	Acquisition time	Relative humity (%)
10 December 2018	220	081	66.9	13:36	44
10 October 2016	220	081	57.9	13:30	71
29 September 2015	220	081	54.6	13:36	72
22 May 2011	220	081	33.3	13:29	74

TIR channels with spatial resolutions of 15, 30 and 90 m, respectively (NASA LP DAAC 2015). The Level 2 ASTER Surface Emissivity (AST_05) data is an on-demand product produced using the Temperature/Emissivity Separation (TES) algorithm for the five Thermal Infrared (TIR) bands (NASA LP DAAC 2001).

Laboratory data

In most of the studies, when retrieving earth surface temperature (EST) and LST, the emissivity is obtained from the ASTER spectral library (http://speclib.jpl.nasa.gov), which provides more than 2300 spectra in wavelengths covering from the visible to TIR region. However, errors may be introduced when the values are not representatives from the area evaluated (Chen et al. 2016). In addition, the available spectral libraries do not have temperature information and different temperatures might have different emissivities for the same land surface types (Li et al. 2013a, b).

We performed emissivity measurements in a controlled environment based on temperatures by using the Fourier-transform infrared spectroscopy (FT-IR) Model 102F. This was essential for the instrument to acquire stability and perform well the calibration process eluding errors in the calculation of the quartz reference curve.

The atmospheric downwelling radiance ($L\downarrow \lambda$) is obtained by measuring a golden reference panel with an emissivity of 0.04. The sample spectral emissivity ε is calculated as follows:

$$\varepsilon_\lambda = \frac{L\lambda - B\lambda(Ts)}{B\lambda(Ts) - L^\downarrow\lambda} \tag{1}$$

where $L\lambda$ is the spectral radiance and $B\lambda(Ts)$ refers to Planck's equation, given as:

$$B\lambda(Ts) = \frac{C1\lambda^{-5}}{\exp(C2/\lambda T) - 1} \tag{2}$$

where C1 and C2 are constants (C1 = 1.191×10^8 W μm^4 sr^{-1} m^{-2}, C2 = 1.439×10^4 μm K). Assuming that $\varepsilon = 1$ between 7.5 and 8 μm, the sample temperature can be obtained. This wavelength range was chosen because the emissivity and temperature retrieval is the most accurate at the maximum

emissivity value (Korb et al. 1999). The downwelling radiance measurements were carried out before and after the sample measurements. If the amount of downwelling radiation changes between measurements, the properties determined for the material will have an associated error. Therefore, the accuracy in the downwelling radiance measurements determines the accuracy in obtaining the emissivity and temperature spectra.

Emissivity retrieval methods for remote geological mapping

The use of infrared vibrational spectroscopy allows the study of mineral compositions with non-destructive techniques. In geologic remote sensing, identification of minerals can be done by analysing emissivity absorption features in lower energy levels, in which silicate and non-silicate minerals show well defined spectral signatures (Hook and Kahle 1996).

The correlation between stretching force constants and bond characteristics of the vibrational mode are evident and most common in quartz (SiO_2). Quartz has a minimum emissivity associated with the fundamental molecular vibration, which is related to changes in the bond length or bond angle between Si–O-Si, and forms asymmetric stretching between 8 and 12 μm (Reststrahlen feature) and a maximum emissivity at approximately 7.5 μm and 12.3 μm (Christiansen feature) (Salisbury et al. 1987). The minimum emissivity occurs at approximately 8.5 μm for some framework silicates (quartz and feldspar) and at progressively longer wavelengths for minerals having sheet, chain, and isolate $(SiO_4)^{4-}$ tetrahedron (Christensen et al. 2000).

Reference channel method (RCM)

This method, developed by Kahle et al. (1980), assumes that the emissivity in a specific channel is constant for all pixels (Li et al. 2013a, b). An approximate surface temperature is then derived from atmospherically corrected radiances and is used to retrieve the emissivity values for the remaining channels (Kahle et al. 1980; Sobrino and Jiménez-Muñoz 2002).

The algorithm calculates the image temperature ($T_{i,r}$) based on the reference emissivity value for each pixel i of the specified channel r as follows:

$$T_{i,r} = \frac{C_2}{\lambda_r}\left(\ln\left(\frac{(c_1\,\varepsilon_{\max})}{\pi\,L_r\,\lambda_r^5} + 1\right)\right)^{-1} \quad (3)$$

where λ_r is the wavelength of the reference channel (μm), L_r is the radiance measured for the reference channel, $\varepsilon_{\max}$ is the target emissivity

T_i is used to calculate the emissivity (ε_n) in the remaining bands (Eq. 4), generating the emissivity spectrum of the target (Kahle and Alley 1992; Andrade 2011).

$$\varepsilon_{i,n} = \frac{L_{i,n}}{B_n(T_i)} \quad (4)$$

where, N being the total number of channels and $n \neq r$. The channel selection and the emissivity value are conditioned by previous knowledge of the target (Rolim et al. 2016). However, the RCM has a 1% uncertainty in the emissivity of the reference channel, which can result in an error of 0.5 K in the land surface temperature (LST) retrieved, as well as errors of 1–2% in the emissivity of the other channels (Li et al. 1999, 2013a, b).

Emissivity normalization method (ENM)

This method was first described by Gillespie (1985), which is similar to the previous method and is, in fact, an improvement of the RCM. The ENM requires a priori knowledge of the target, but not of a specific channel (reference channel). The emissivity value is applied in all channels and then the temperatures are calculated for each channel. The highest temperature value for a specific pixel i, among the N calculated temperatures, is taken and applied in the remaining channels of the pixel (Gillespie et al. 1998) as follows:

$$T_{i,n} = \frac{C_2}{\lambda_n}\left(\ln\left(\frac{(c_1\,\varepsilon)}{\pi\,L_n\,\lambda_n^5} + 1\right)\right)^{-1},$$
$$T_{i,H} = \max(T_{i,n}) \text{ with } n = 1, \dots, N \quad (5)$$

where $T_{i,n}$ is the calculated temperature for each channel in the pixel i. Finally, the emissivity is calculated for all channels using the temperature values $T_{i,H}$ (Eq. 6), generating the emissivity spectrum of the target for all channels.

$$\varepsilon_{i,n} = \frac{L_{i,n}}{B_n(T_{i,H})} \quad (6)$$

As stated, the ENM does not require the specification of a band for the chosen emissivity and the maximum temperatures can be determined in different channels, resulting in a superior performance compared to the RCM. However, the ENM requires an appropriate atmospheric correction, which implies the precision of ε and T retrieval.

Temperature and emissivity separation (TES) method

The TES Method is a hybrid method that uses the land-leaving radiance (L') and downwelling sky irradiance ($S\downarrow$) as inputs, providing an initial estimate of surface temperatures and emissivities using the ENM module. The results are improved by applying two other modules—the RATIO module (RM) and the Minimum–Maximum Difference (MMD) module. The output data consists of the $N-1$ emissivity band and a single temperature band (Gillespie et al. 1998; Sobrino et al. 2008). The three modules of the TES algorithm are described below.

(a) ENM Module: retrieves the temperature as described above (Eq. 5). An initial value of 0.99 is assumed as $\varepsilon_{\max}$ for low contrast surfaces (e.g. water, snow, and vegetation). For high contrast surfaces (e.g. rock or soil), a value of 0.96 is used.

(b) RATIO Module: the relative emissivities (β_b) are found by calculating the ratio between the ENM emissivity values and their average:

$$\beta_b = \varepsilon_b 5\left[\sum \varepsilon_b\right]^{-1}, \quad b = 10, 14 \quad (7)$$

(c) MMD Module: The β spectrum must be then scaled to actual emissivity values, and the surface temperatures must be recalculated from these new emissivity values and from the atmospherically corrected radiances (L' and $S\downarrow$). An empirical relationship predicting a minimum emissivity ($\varepsilon_{\min}$) is used to convert β_b to ε_b. The first step in the MMD module is to calculate the spectral contrast:

$$\text{MMD} = \max(\beta b) - \min(\beta b) \quad (8)$$

where the minimum emissivity is predicted and used to find the TES emissivity (Gillespie et al. 1999).

$$\varepsilon_{\min} = 0.994 - 0.687 * \text{MMD}^{0.737}; \quad \varepsilon_{b=}\beta_b\left(\frac{\varepsilon_{\min}}{\min(\beta_b)}\right) \quad (9)$$

Lastly, in a quality control module, the accuracy and precision of T and ε were validated (Gillespie et al. 1999).

Data analysis

Automatic image processing algorithms was implemented in a MATLAB environment. The RCM and ENM were applied to four ASTER TIR (AST L1T) scenes of the studied coastal dune field (Fig. 1). Furthermore, four TES scenes were used as AST_05 on-demand emissivity products. These twelve emissivity products from RCM, ENM and AST_05 were compared to an average spectral signature of quartz samples obtained from the study area.

The five ASTER Level 1T bands were atmospherically corrected for transmittance and upwelling and downwelling radiance, with the use of the Radiative Transfer Equation—RTE (Eq. 10). The parameters were calculated using the PcModWin5© (ONTAR 2009) radiative transfer model user interface environment of the MODerate spectral resolution atmospheric TRANsmittance algorithm and computer (MODTRAN5®) (Berk et al. 2011). The atmosphere was considered plane-parallel, in thermodynamic equilibrium, and cloud-free.

$$L_{sur,i} = \frac{L_{sen} - L_i^{\uparrow}}{\tau_i} + (\varepsilon_i - 1)\frac{F_i^{\downarrow}}{\pi} \tag{10}$$

where $L_{sur,i}$ is the surface radiance (W/m^2 sr μm) for channel i, L_{sen} is the at-sensor radiance (W/m^2 sr μm), τ is the atmospheric transmittance, ε_i is the emissivity, $Li\uparrow$ is the upwelling radiance (W/m^2 sr μm), and $Fi\downarrow$ is the hemispheric downwelling sky irradiance (W/m^2 sr μm).

Launching a radiosonde is one of the best ways to obtain the amount of water vapour in the atmosphere. However, this method is expensive, time-consuming and not possible when the information needs to be known in an operational way for many regions around the globe. In addition, radiosonde data are not always available for specific sites in the real world (Sobrino et al. 2004; Coll et al. 2010; Wang et al. 2015).

Costa et al. (2018) tested the Weather Research and Forecasting (WRF) (Skamarock et al. 2008) model to reproduce primary meteorological data in the same ASTER scene corresponding to the study area (dune field) and they compared the results with the atmospheric profile data obtained from the Porto Alegre international airport (30°S, 51°18′W), located approximately 80 km west of the study area. A strong positive correlation (0.94 for relative humidity, 0.99 for air temperature, and 0.99 for dew point temperature) was observed between the profiles generated by the WRF model and those from the radiosonde. Therefore, we used the radiosonde data from the Porto Alegre international airport in this study. The radiosonde registered the physical variables from an initial geopotential height of 0.003 km to an altitude of approximately 24 km. Profiles of temperature, pressure and relative humidity are used in the model. The default profiles of Ozone and CO_2 during mid-latitude

summer were applied. The parameters for transmittance (τ), upwelling radiance ($Li\uparrow$) and downwelling radiance ($Li\downarrow$) were resampled for the five TIR bands using the sensor spectral response function (SRF). Emissivity values from each channel were selected from the temperature-based spectral library in the Geologic Remote Sensing Laboratory (LabSRGeo) at the Federal University of Rio Grande do Sul, Brazil.

The average spectral signature used as a reference in this study was calculated from quartz mineral samples (size: 125–250 μm), which were collected from the study area (Fig. 1). We selected samples from the whole dune area in order to homogenize the sampling. In the laboratory, the samples were heated and measured in a controlled environment with 58% relative humidity using a Fourier-transform infrared spectrometer (μFT-IR) Model 102F (Korb et al. 1996). The instrument was placed facing the quartz samples at angles close to the nadir. The standard optical input is one-inch in diameter with a 4.8° expanding field of view.

The first step in obtaining measurements is using the μFT-IR to calibrate the instrument against any blackbodies. The calibration process consists of choosing two temperatures: a temperature for a cold blackbody (below room temperature) and a temperature for a warm blackbody (above the sample temperature). This leads to a more accurate calibration because the instrument response is linear over the range of radiances generated by these black body temperatures (Korb et al. 1996). Therefore, the calibration range is assumed as between 10 and 40 °C (283–313 K).

The quartz samples were heated up to 60 °C (333.15 K) and the measurements were taken. Measurements ceased when the sample temperatures reached 26 °C (1–2 measurements per minute). We selected all of the emissivity measurements up to 40 °C (313.15 K), which totalled 37 measurements as there was not much variation in emissivity values with variations in temperature (Käfer et al. 2019). The instrument calibration was carried out twice throughout the whole process. Afterwards, we weighted all the curves by the sensor spectral response function (SRF) of ASTER and averaged the results, which is considered the reference curve (Fig. 2).

Results

The results generated by the methods and the spectral signature are shown in Fig. 3, which shows the emissivity spectra retrieved from the RCM, ENM and TES, and the quartz reference spectrum obtained from the LabSRGeo Spectral Library (LAB). We calculated the RMSE and R^2 in order to quantify the comparison among the algorithms and the reference curve.

In general, all of the tested methods performed well in longer wavelengths (channels 13 and 14) with lower
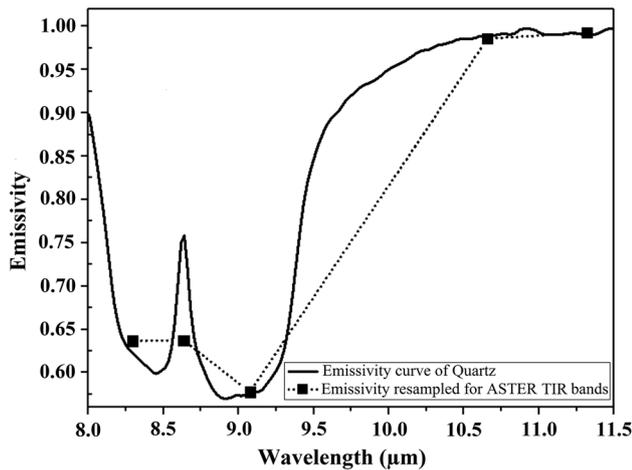
Fig. 2 Emissivity curve of quartz (solid line) from the LabSRGeo spectral library resampled for the TIR-ASTER spectral response function (dashed line with squares corresponding to 8.291 μm; 8.634 μm; 9.075 μm; 10.657 μm; 11.318 μm)

standard deviations (σ) ranging between 2 and 4%. On the other hand, for shorter wavelengths, all these methods poorly retrieved the emissivity, resulting in higher σ values. The average reference curve (LAB) had a higher σ for the longer wavelengths, especially in channel 14, where there are significant absorption bands due to the presence of CO_2 in the atmosphere. In addition, the μFT-IR spectrometer may show instability in situations when the emissivity value is close to 1.

The RCM and ENM methods demonstrated similar performances for the TIR-ASTER images. However, these two methods overestimated the emissivity values for shorter wavelengths (channels 10, 11 and 12), except for the date in 2016, where they achieved the best results, especially for channels 11 and 12. For longer wavelengths, the two methods consistently underestimated the emissivity values. When analysing the longer wavelengths, the RCM produced results closer to the reference curve (LAB) compared to the ENM for some images.

The TES algorithm overestimated the emissivity values for shorter wavelengths (channels 10, 11 and 12) for all of the images. In some cases, such as the scenes acquired in 2011 and 2018, it was noted that TES had a better performance for the shorter wavelengths when compared to RCM and ENM. In channel 10, TES had results very close to the reference curve (LAB). Nonetheless, these results may not be entirely accurate due to an edge band, which can be influenced due to atmospheric effects. According to Gillespie et al. (2011), the greater atmospheric absorption at wavelengths above 11 μm and below 8.5 μm introduces higher errors in emissivity retrieval. All estimated averages and standard deviations can be checked in Table 2.

Table 3 displays the quantitative results for the methods compared with the reference curve (LAB). As mentioned above, the RCM and ENM produced very similar results, with the coefficient of determination (R^2) 0.9959 and 0.9960, respectively. The root mean square errors (RMSE) for RCM and ENM were 0.0184 and 0.0185, respectively, which is also very similar. The RCM showed a slightly higher R^2 and a slightly lower RMSE. As illustrated in the previous section, TES had superior results for two of the four images. However, it can be observed that the TES method presented the worst results in terms of R^2 (0.9947) and RMSE (0.0197).

Discussion

Land surface emissivity is an important physical property in geological mapping, particularly in the exploration of ore deposits related to silicates in mafic and ultramafic rocks. These minerals do not always exhibit expressive diagnostic features between the visible and shortwave infrared regions. Therefore, we used TIR vibrational spectroscopy to study geologic targets in order to support field campaigns. Most geological studies use RCM, ENM, and TES methods (Gillespie et al. 1986; Watson 1992; Kealy and Gabell 1990) for this purpose. The RCM and ENM were tested in ENVI software package and TES is an on-demand product (AST_05) acquired directly from NASA LP DAAC.

Kealy and Hook (1993) reported that ENM has a superior performance when compared to RCM. This is because, in the first method, an emissivity channel is not specified and the temperature can be different between pixels. However, in this study, RCM showed slightly closer results in relation to the quartz reference spectrum. We used 0.96 as an initial emissivity for the RCM (standard value), and verified that the RCM may be used to identify specific targets with silicate minerals with superior performance relative to the ENM.

Li et al. (2013a, b) mentioned that it is difficult to find a unique emissivity value that is appropriate for all surface materials in one reference channel. However, we worked with a pseudo-invariant target in a very homogeneous region (99.53% SiO_2 and it was verified that the RCM is superior in identifying specific targets such as silicate minerals.

Gillespie et al. (2011) demonstrated the presence of errors in TES method in its final products due to the iterative processes of estimating downwelling radiation, which is seen in the results (see Table 3) as TES performance is inferior to the other two methods. Gillespie et al. (2011) also pointed out that these errors lead to increased emissivity in all bands, which can be observed in our experiment for shorter wavelengths (channels 10, 11 and 12). However, in two images (acquired in 2011 and 2018), the results of TES method were even superior to RCM and ENM.
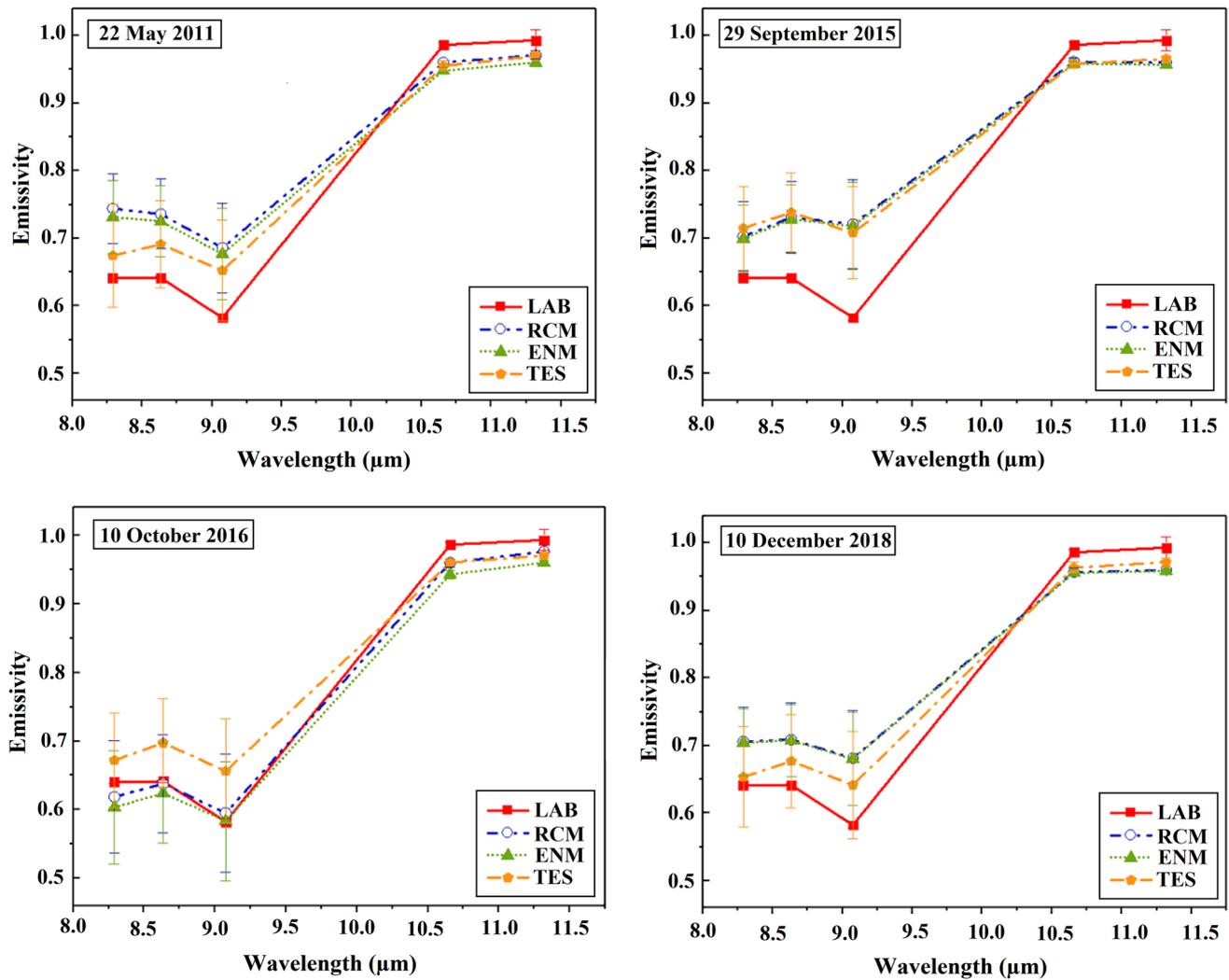
Fig. 3 Comparison between the average spectral profiles of emissivity generated from the RCM, ENM and TES methods, and the reference curve (LAB) for all dates considered in this study. The symbols represent the ASTER channels

Because of its simplicity and the lack of any need for prior knowledge of the surface, the TES method has already got significant attention from the scientific community (Dash et al. 2005; Li et al. 2013a, b). It is important to emphasize that, in addition to the land surface type, the user must always consider applicability and accuracy. An important problem associated with the TES algorithm is that it classifies radiometric noise as real spectral contrast, which has a significant impact on the extraction of the emissivity information since TES relies on an empirical relationship between the emissivity minimum and the MMD (Coll et al. 2007).

Considering the mean values (Table 3), the quantitative analysis demonstrates that it is better to use RCM or ENM instead of the TES standard product AST_05. Despite the minimal differences between the RCM and ENM results, the RCM produced results closer to the reference curve (LAB)

for longer wavelengths, thus being the best option for emissivity retrieval.

Conclusions

We compared ε and T retrieval methods for geologic targets, considering the lack of controlled field experiments (on-site radiosonde and radiance measurements) and commercial availability of algorithms. The RCM, ENM, and TES methods were evaluated from the ASTER radiance data for an experiment conducted in a coastal dune field. The pre-processing steps were performed in the µFT-IR, which improved the instrument stability and resulted in reliable measurements for 37 quartz samples from the study site. The calibration against the instrument blackbodies was carried out twice throughout the whole process, and it was decisive

Table 2 Emissivity (ε) and standard deviation (σ) of the RCM, ENM and TES methods (average values for the dune field mask) and laboratory spectral curve

Date	Channel	λ (µm)	ε_{RCM}	ε_{ENM}	ε_{TES}	ε_{LAB}	σ_{RCM}	σ_{ENM}	σ_{TES}	σ_{LAB}
	10	8.284	0.744	0.731	0.674	0.641	0.052	0.054	0.076	0.003
	11	8.664	0.726	0.715	0.691	0.641	0.052	0.053	0.065	0.002
22 May 2011	12	9.079	0.686	0.677	0.652	0.582	0.066	0.067	0.074	0.004
	13	10.665	0.960	0.948	0.955	0.986	0.000	0.005	0.007	0.005
	14	11.281	0.971	0.960	0.969	0.993	0.005	0.000	0.006	0.016
	10	8.284	0.703	0.700	0.715	0.641	0.052	0.050	0.062	0.003
	11	8.664	0.731	0.728	0.738	0.641	0.053	0.051	0.059	0.002
29 September 2015	12	9.079	0.721	0.718	0.708	0.582	0.066	0.064	0.068	0.004
	13	10.665	0.961	0.958	0.958	0.986	0.005	0.003	0.006	0.005
	14	11.281	0.960	0.957	0.966	0.993	0.000	0.004	0.005	0.016
	10	8.284	0.619	0.604	0.67	0.641	0.082	0.082	0.07	0.003
	11	8.664	0.638	0.624	0.70	0.641	0.071	0.072	0.06	0.002
10 October 2016	12	9.079	0.595	0.584	0.66	0.582	0.086	0.087	0.08	0.004
	13	10.665	0.960	0.943	0.96	0.986	0.006	0.006	0.01	0.005
	14	11.281	0.976	0.960	0.97	0.993	0.000	0.001	0.01	0.016
	10	8.284	0.706	0.704	0.654	0.641	0.051	0.050	0.074	0.003
	11	8.664	0.709	0.708	0.677	0.641	0.055	0.053	0.069	0.002
10 December 2018	12	9.079	0.681	0.680	0.642	0.582	0.070	0.069	0.080	0.004
	13	10.665	0.957	0.956	0.963	0.986	0.006	0.005	0.008	0.005
	14	11.281	0.960	0.959	0.971	0.993	0.000	0.003	0.007	0.016

These values were calculated for the response filter function of ASTER bands

Table 3 Comparison among the emissivities (ε) generated by different methods

Method	Date	R^2	RMSE
	10 December 2018	0.9990	0.0104
RCM	10 October 2016	0.9981	0.0144
	29 September 2015	0.9886	0.0353
	22 May 2011	0.9984	0.0133
	Mean	0.9960	0.0184
	10 December 2018	0.9989	0.0106
ENM	10 October 2016	0.9977	0.0159
	29 September 2015	0.9883	0.0356
	22 May 2011	0.9987	0.0119
	Mean	0.9959	0.0185
	10 December 2018	0.9969	0.0183
TES	10 October 2016	0.9974	0.0169
	29 September 2015	0.9962	0.0204
	22 May 2011	0.9886	0.0231
	Mean	0.9947	0.0197

in the acquisition of an average emissivity spectrum for pure quartz, which was used as a reference to compare the three retrieval methods.

The study area, which is a coastal dune, approximately 30 km² in size and composed of 99.53% SiO_2, is an ideal study area where one can simulate several scenarios in the field and laboratory in order to improve the identification of mafic and ultramafic silicate minerals. The average spectrum of almost pure quartz, used as reference from a temperature-based spectral library, is a good alternative to test algorithms when there is no experimental field data available.

RCM and ENM (implemented in ENVI software package) and TES, which is an on-demand product showed similar performances and can be applied for emissivity retrieval with good approximations. Although RCM and ENM showed similar results, the RCM presented the best fit in relation to the quartz reference signature. In general, the RCM values were the closest to the quartz reference curve. Moreover, the quantitative results of R^2 and RMSE affirmed this performance. The TES method may work better for shorter wavelengths; however, the R^2 and RMSE values for TES were the worst results. As pointed out by Gillespie et al. (2011), the TES method tends to overestimate emissivity due to an inaccurate atmospheric correction, which might have occurred in the shorter wavelengths. In addition, TES final products have an error associated with the iterative processes of estimating downwelling radiation.

In general terms, TES is commonly used in retrieval methods, specifically for soils and rocks (Gillespie et al. 1998, 2011). In this work, TES performed better in several image channels. However, if we consider the details, the RCM had a better performance in specific regions characterized by lower atmospheric absorption (channel 13). The only exception was the 2018 scene, where the relative humidity of 44% was much lower than the other three images (71–74%).

In this case, the TES method showed the best fit (2.2%) compared to the RCM (2.9%).

The emissivity retrieval of quartz targets in a dune field composed of 99.53% SiO_2 is essential for remote geological mapping of mafic and ultramafic rocks, as the fundamental frequencies of most minerals correspond to longer wavelengths (greater than 3 μm). In this context, anion groups of SiO_2 produce specific diagnostic features in silicates at 8.5 μm (quartz and feldspars), and at progressively longer wavelengths for those with sheet, chain and isolated SiO_4 tetrahedral (Hunt and Vincent 1980; Christensen et al. 2000).

Other targets are being tested in a temperature-based spectral library in order to expand the study on emissivity retrieval. Additionally, as pointed out by Li et al. (1999), possible instrument noise and calibration errors should be considered if there are more targets in the area. In a future perspective, it is recommended to improve the classification and validation systems, and consider the low emissivity contrast between targets and the nonlinear effects of a spectral mixture on a sub-pixel basis.

References

Andrade L (2011) Estimativa de temperatura e emissividade com imagens do sensor HSS (Hyperspectral Scanner System) e suas relações com materiais urbanos. Master's Thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brazil

Berk A, Anderson GP, Acharya PK, Shettle EP (2011) MODTRAN®5.2.1 user's manual. Burlington, MA, USA: Spectral Sciences Inc

Chen F, Yang S, Su Z, Wang K (2016) Effect of emissivity uncertainty on surface temperature retrieval over urban areas: investigations based on spectral libraries. ISPRS J Photogram Remote Sens 114:53–66. https://doi.org/10.1016/j.isprsjprs.2016.01.007

Christensen PR, Bandfield JL, Hamilton VE, Howard DA, Lane MD, Piatek JL, Ruff SW, Stefanov WL (2000) A thermal emission spectral library of rock-forming minerals. J Geophys Res 105:9735–9739. https://doi.org/10.1029/1998JE000624

Coll C, Caselles V, Valor E, Niclòs R, Sánchez JM, Galve JM, Mira M (2007) Temperature and emissivity separation from ASTER data for low spectral contrast surfaces. Remote Sens Environ 110:162–175. https://doi.org/10.1016/j.rse.2007.02.008

Coll C, Galve JM, Sanchez JM, Caselles V (2010) Validation of Landsat-7/ETM+thermal-band calibration and atmospheric correction with ground-based measurements. IEEE Trans Geosci Remote Sens 48:547–555. https://doi.org/10.1109/TGRS.2009.2024934

Collins EF, Roberts Dar A, Borel CC (2001) Spectral mixture analysis of simulated thermal infrared spectrometry data: an initial temperature estimate bounded TESSMA search approach.

IEEE Trans Geosci Remote Sens 39:1435–1446. https://doi.org/10.1109/36.934075

Constantino D, Angelini MG (2016) Thermal monitoring using an ASTER image. J Appl Remote Sens 10(4):046031. https://doi.org/10.1117/1.JRS.10.046031

Costa BSC, Rocha NS, Silva SCS, Molmann Junior RA, Munchow GB, Silveira VC, Rolim SBA, Alves RCM, Kafer P, Diaz LR (2018) The use of the Weather Research and Forecasting Model to estimate the vertical profile of meteorological data. In: international geoscience and remote sensing symposium, IGARSS 2018, the 38th annual symposium of the IEEE Geoscience and Remote Sensing Society (GRSS), 2018, Valencia, Spain

Dash P, Göttsche FM, Olesen FS, Fischer H (2002) Land surface temperature and emissivity estimation from passive sensor data: theory and practice-current trends. Int J Remote Sens 23(13):2563–2594. https://doi.org/10.1080/01431160110115041

Dash P, Gottsche FM, Olesen FS, Fischer H (2005) Separating surface emissivity and temperature using two-channel spectral indices and emissivity composites and comparison with a vegetation fraction method. Remote Sens Environ 96:1–17. https://doi.org/10.1016/j.rse.2004.12.023

Exelis Visual Information Solutions I (2015) Exelis visual information solutions I. ENVI software version 5.3. Boulder, CO, USA

Gillespie AR (1985) Lithologic mapping of silicate rocks using TIMS. In: Abbott EA (ed) Proceedings of the TIMS data user's workshop. Jet Propulsion Laboratory, Pasadena, pp 29–44

Gillespie A, Kahle A, Walker R (1986) Color enhancement of highly correlated images. I. Decorrelation and HSI contrast stretches. Remote Sens Environ 20:209–235. https://doi.org/10.1016/0034-4257(86)90044-1

Gillespie A, Rokugawa S, Matsunaga T, Cothern J, Hook S, Kahle A (1998) A temperature and emissivity separation algorithm for Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) images. IEEE Trans Geosci Remote Sens 36:1113–1126. https://doi.org/10.1109/36.700995

Gillespie AR, Rokugawa S, Hook SJ, Matsunaga T, Kahle B (1999) Temperature/emissivity separation algorithm theoretical basis document, version 2.4, contract NAS5-31372, NASA, Washington, DC, USA

Gillespie A, Abbott E, Gilson L, Hulley G, Jimenez-Munoz J, Sobrino J (2011) Residual errors in ASTER temperature and emissivity products AST08 and AST05. Remote Sens Environ 115:3681–3694. https://doi.org/10.1016/j.rse.2011.09.007

Hook SJ, Kahle AB (1996) The Micro Fourier Transform Interferometer (μFTIR): a new field spectrometer for acquisition of infrared data of natural surfaces. Remote Sens Environ 56:172–181. https://doi.org/10.1016/0034-4257(95)00231-6

Hunt G, Vincent R (1980) The behavior of spectral features in the infrared emission from particulate surfaces of various grain sizes. J Geophys Res 73:6039–6046. https://doi.org/10.1029/jb073i018p06039

Käfer PS, Rolim SB, Iglesias ML, Rocha NS, Diaz LR (2019) Land surface temperature retrieval by LANDSAT 8 thermal band: applications of laboratory and field measurements. IEEE J Select Top Appl Earth Obser Remote Sens. https://doi.org/10.1109/JSTARS.2019.2913822

Kahle A, Alley R (1992) Separation of temperature and emittance in remotely sensed radiance measurements. Remote Sens Environ 42:107–111. https://doi.org/10.1016/0034-4257(92)90093-Y

Kahle A, Madura D, Soha J (1980) Middle infrared multispectral aircraft scanner data: analysis for geological applications. Appl Opt 19:2279–2290. https://doi.org/10.1364/AO.19.002279

Kealy P, Gabell A (1990) Estimation of emissivity and temperature using alpha coefficients. In: Proceedings of 2nd TIMS workshop, jet propulsion laboratory, Pasadena, California, pp 90–55

Kealy P, Hook S (1993) Separating temperature and emissivity in thermal infrared multispectral scanner data: implications for recovering land surface temperatures. IEEE Trans Geosci Remote Sens 31:1155–1164. https://doi.org/10.1109/36.317447

Korb A, Dybwad P, Wadsworth W, Salisbury J (1996) Portable FTIR spectrometer for field measurements of radiance and emissivity. Appl Opt 35:1679–1692. https://doi.org/10.1364/AO.35.001679

Korb A, Salisbury JW, D'Ária DM (1999) Thermal-infrared remote sensing and Kirchhoff's law 2. Field measurements. J Geophys Res-Solid Earth 104:339–350. https://doi.org/10.1029/97JB03537

Li Z, Becker F, Stoll M, Wan Z (1999) Evaluating six methods for extracting relative emissivity spectra from thermal infrared images. Remote Sens Environ 69:197–214. https://doi.org/10.1016/S0034-4257(99)00049-8

Li ZL, Tang BH, Wu H, Ren H, Yan G, Wan Z, Trigo IF, Sobrino JA (2013a) Satellite-derived land surface temperature: current status and perspectives. Remote Sens Environ 131:14–37. https://doi.org/10.1016/j.rse.2012.12.008

Li Z, Wu H, Wang N, Qiu S, Sobrino J, Wan Z, Tang B, Yan G (2013b) Land surface emissivity retrieval from satellite data. Int J Remote Sens 34:3084–3127. https://doi.org/10.1080/01431161.2012.716540

NASA (2019) https://asterweb.jpl.nasa.gov/swir-alert.asp. Accessed 23 May 2019

NASA LP DAAC (2015) ASTER level 1 precision terrain corrected registered at-sensor radiance V003 [Terra ASTER]. NASA EOSDIS Land Processes DAAC. doi: https://doi.org/10.5067/ASTER/AST_L1T.003

NASA/METI/AIST/Japan Spacesystems, and U.S./Japan ASTER Science Team. ASTER Level 2 Emissivity Product, distributed by NASA EOSDIS Land Processes DAAC (2001) https://doi.org/10.5067/aster/ast_05.003

ONTAR (2009) PcModWin© Manual. Version 5.0 v2ro. North Andover, MA, USA: Ontar Corporation

Rolim SBA, Grondona A, Hackmann CL, Rocha C (2016) A review of temperature and emissivity retrieval methods: applications and restrictions. Am J Environ Eng 6:119–128. https://doi.org/10.5923/s.ajee.201601.18

Salisbury JW, Hapke B, Eastes J (1987) Usefulness of weak bands in midinfrared remote sensing of particulate planetary surfaces. J Geophys Res-Solid Earth 92:702–710. https://doi.org/10.1029/JB092iB01p00702

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda MG, Huang X-Y, Wang W, Powers JG (2008) A description of the advanced research WRF Version 3. NCAR Tech. Note NCAR/TN-475 + STR

Sobrino JA, Jiménez-Muñoz JC (2002) Surface emissivity retrieval from digital airborne imaging spectrometer data. J Geophys Res 107(D23):4729. https://doi.org/10.1029/2002JD002197

Sobrino JA, Jiménez-Muñoz JC, Paolini L (2004) Land surface temperature retrieval from LANDSAT TM5. Remote Sens Environ 90:434–440. https://doi.org/10.1016/j.rse.2004.02.003

Sobrino J, Jiménez-Muñoz J, Sória G, Romaguera M, Guanter L, Moreno J, Plaza A, Martínez P (2008) Land surface emissivity retrieval from different VNIR and TIR sensors. IEEE Trans Geosci Remote Sens 48:316–327. https://doi.org/10.1109/TGRS.2007.904834

Tomazelli L, Dillenburg S, Guimarães E, Correa M (2008) Geomorfologia e potencial de preservação dos campos de dunas transgressivos de Cidreira e Itaipeva, Litoral Norte do Rio Grande do Sul. Rev Pesquisas em Geociências 32:47–55

Wang F, Qin Z, Song C, Tu L, Karnieli A, Zhao S (2015) An improved mono-window algorithm for land surface temperature retrieval from landsat 8 thermal infrared sensor data. Remote Sens 7:4268–4289. https://doi.org/10.3390/rs70404268

Watson K (1992) Spectral ratio method for measuring emissivity. Remote Sens Environ 42:113–116. https://doi.org/10.1016/0034-4257(92)90094-Z

RESEARCH ARTICLE - ATMOSPHERIC & SPACE SCIENCES

Space and time variability of meteorological drought in Syria

Safwan Mohammed[1] · Karam Alsafadi[2] · Talal Al-Awadhi[3] · Youssef Sherief[3,4] · Endre Harsanyie[1,5] ·
Ahmed M. El Kenawy[3,6]

Abstract
This study assesses the spatial and temporal characteristics (e.g., frequency, intensity, spatial extent) of meteorological drought in Syria. Specifically, drought was characterized using the observed rainfall data from 36 rain gauges spanning the period between 1990 and 2010 and covering the main climatic regions in Syria (i.e., Mediterranean, arid, semiarid and mountainous). Meteorological drought was assessed using the standardized precipitation index (SPI) at 12-month timescale, allowing for detecting the impacts of climate variability on agricultural droughts. The dominant modes of drought were defined using an S-mode of the principal component analysis. To assess the links between meteorological drought evolution and vegetation greening in Syria, the time series of SPI were correlated with the normalized difference vegetation index (NDVI). Time series of NDVI were retrieved from the remotely sensed National Oceanic and Atmospheric Administration Advanced Very High-Resolution Radiometer (NOAA/AVHRR) sensor at a spatial resolution of 25 km for the common period 1990–2010. Trend analysis suggests a statistically significant increase in the frequency and intensity of drought at 12-month timescale. The observed intensification of meteorological drought is mostly associated with the increase in mild and moderate droughts, relative to extremes and very extreme droughts. Results also suggest a statistically significant decrease ($p < 0.05$) in vegetation greening over Syria during the study period, especially in the eastern parts of the country. Our results demonstrate that the decrease in vegetation cover can directly be linked to the anomalous drought events, with Pearson' r coefficients generally above 0.6. This dependency was more highlighted during wintertime for the Mediterranean vegetation and in northeastern portions of the country. Overall, the increase in the frequency and intensity of meteorological drought, combined with a series of unrest and political instability, have drastic impacts on the agricultural sector in Syria, with serious implications for crop yield.

Keywords SPI · NDVI · Spatiotemporal patterns · Climate change · NOAA-AVHRR · Vegetation · Syria

✉ Ahmed M. El Kenawy
kenawy@mans.edu.eg

1 Institution of Land Utilization, Technology and Regional Planning, University of Debrecen, Debrecen 4032, Hungary

2 Department of Geography and GIS, Faculty of Arts, Alexandria University, Alexandria 25435, Egypt

3 Department of Geography, Sultan Qaboos University, Al Khoud, Muscat, Oman

4 Department of Geography, Faculty of Arts, Zagazig University, Zagazig, Egypt

5 National Agricultural Research and Innovation Center, Institute of Agricultural Engineering, Tessedik Sámuel 4, Gödöllő H-2100, Hungary

6 Department of Geography, Mansoura University, Mansoura 35516, Egypt

Introduction

In the era of climate change, the Mediterranean region has frequently witnessed severe drought episodes, making it one of the most "hot spot" regions worldwide (Giorgi 2006; Naumann et al. 2018). Historical droughts have been identified in the Mediterranean, as evidenced by documentary and proxies (Nicault et al. 2008; Domínguez-Castro et al. 2010; Cook et al. 2016; Ljungqvist et al. 2016). The decline of several ancient civilizations in the Mediterranean was partially being attributed to particular mega-droughts (Schneider and Adalı 2014; Manning et al. 2017). In their assessment of drought variability and trends in the Northern Hemisphere, Caloiero et al. (2018a) indicated that the Mediterranean basin is among the most vulnerable regions to severe drought. Numerous works confirmed that the

tropical sea surface temperatures (SSTs) anomalies, associated with El Niño, La Niña and the El Niño–Southern Oscillation (ENSO), have accelerated drying over many regions worldwide, including the Middle East (e.g., Cai et al. 2014; Barlow et al. 2016; El Kenawy et al. 2016a; Schubert et al. 2016).

In a changing climate, there is an increased demand for moisture in the atmosphere, which will likely alter the atmospheric circulation patterns, with a trend toward more dryness (Cai et al. 2011; Drumond et al. 2011; Ahlström et al. 2015; Azorin-Molina et al. 2015; Stojanovic et al. 2018). Under future scenarios, the recently observed warmer and drier conditions in the Mediterranean are projected to continue (Gao and Giorgi 2008; Hoerling et al. 2012; Ruffault et al. 2014). It is evident that the probability of exceptional warm periods, similar to the record summer of 2003, is likely to increase in the region in the future. The IPCC (2013) suggested a rise in summer temperature from 2 to 6 °C in the Mediterranean by the end of the century. This drying may accelerate under the projected decreases in regional precipitation, combined with an increase in evaporation. Numerous studies have suggested more drying in the Mediterranean under increased greenhouse gas forcing (e.g., Giorgi and Lionello 2008; IPCC 2013; Seager et al. 2014; Spinoni et al. 2018). Nonetheless, there is almost a degree of uncertainty in the magnitude to which these changes could be realized. These changes, combined with other important processes (e.g., land use changes, overgrazing, water resources management, overpopulation, urbanization), could have drastic impacts on local communities. These impacts can include crop failure, water scarcity, food supply, biodiversity loss, among others (Nelson et al. 2014; Mukherjee et al. 2019; Orians et al. 2019; Rashid and Beecham 2019). A representative example is the drought episode in the Mediterranean and the Middle East region in 2000, which caused losses of US$900 million in Morocco and US$160 million in Jordan (De Pauw 2005).

Recalling the wide web of environmental, social, and economic impacts of drought on both natural and human environments (IPCC 2013; Solh and van Ginkel 2014; Khayyati and Aazami 2016; Eklund and Thompson 2017), the vulnerability to drought can vary considerably from one region to another. As a Mediterranean country, Syria has suffered from frequent and intense drought events during the past decades (Selby et al. 2017; Salah et al. 2018). These events had drastic impacts on internal migration, water supply, and crop failure. Nonetheless, these impacts have been much pronounced in most recent decades, mainly due to the rapid population growth, water scarcity, land degradation and desertification, weak infrastructure, low adaptive capacity, besides latest political conflicts (Lelieveld et al. 2012; Terink et al. 2013; Salameh and Fallah 2018; Abdo 2018; Mohammed et al. 2020b, Abdo 2018). A representative example

is the drought episode that hit the eastern part of Syria between 2006 and 2008, which impacted wheat production and forced the national government to import wheat for the first time since the 1990s (Kelley et al. 2015). Gleick (2014) indicated that, due to climate change and poor planning and management of environmental resources, the agricultural sector witnessed rapid changes in the past decades. Specifically, almost 800,000 people lost their livestock and roughly one million suffered from food insecurity. Also, Cook et al. (2016) showed that the Levant 1998–2012 drought, mostly represented in Cyprus, Israel, Jordan, Lebanon, Palestine, Syria, and Turkey, was the driest on record over the past 900 years in the eastern Mediterranean. The impacts of this recent drought episode in Syria were documented by Selby et al. (2017) who assessed its impacts on agricultural failure and demographic changes (e.g., forced migration from eastern parts of the country to mainland. Taken these impacts into consideration, a range of studies have indicated that climate change in general and drought in particular can be seen as a key driver of the recent Syrian conflict (e.g., Butler 2018; Gilmore et al. 2018; Ide 2018).

The increase in drought severity and frequency has been verified for the whole Mediterranean basin (e.g., Hoerling et al. 2012; Vicente-Serrano et al. 2014; Caloiero et al. 2018a), and for specific sub-regions in the basin (e.g., Vicente-Serrano and López-Moreno 2006; Pascoa et al. 2017a, b; Kostopoulou et al. 2017; Vidal et al. 2010; Dabanli et al. 2017; Vicente-Serrano and Cuadrat-Prats 2007; Buttafuoco et al. 2015; Vasiliades et al. 2011). Unfortunately, a detailed spatial assessment of drought in the eastern Mediterranean in general and Syria in particular is still lacking. This can simply be explained by data limitations. Specifically, the region suffers from a paucity and scarcity of observational climate network, with few high-quality records extending back several decades. The majority of the relevant literature for drought characteristics are available either within a Mediterranean context or discussed from a sociopolitical perspective (e.g., Kelley et al. 2015; Gleick 2014). Even for some hydroclimatic studies (e.g., Mahfouz et al. (2016) for Lebanon, Haensel and Zurba (2015) for Palestine, Mustafa and Rahman (2018) and Aladaileh et al. (2019) for Jordan, and Hameed et al. (2018) for Iraq), a very limited number of meteorological observatories are employed for drought characterization. In Syria, drought has almost been discussed within a socioeconomic context, with poor reference to climatic conditions responsible for drought variability. For example. Kelley et al. (2015) and Gleick (2014) linked drought in Syria to the poor governance and the mismanagement of natural resources. Similarly, Selby et al. (2017) and De Châtel (2014) stressed that the Syrian conflict could respond to a variety of economic, political and social factors, with a minimized role of climate change. Recently, Mathbout et al. (2018) assessed the

spatial and temporal characteristics of drought in Syria for the period 1961–2012. They indicated a significant negative trend in the values of the Standardized Precipitation Index (SPI) and the Standardized Precipitation Evapotranspiration Index (SPEI) values across the whole country. Taken into consideration these previous studies, our study is unique in that it provides a more detailed spatial assessment of drought behavior in Syria, with a main focus on the links between drought and vegetation conditions and crop failure. With the current advances in remote sensing, information provided from earth observation satellites has been freely available, allowing for quantifying the impacts of drought on different disciplines, including vegetation (e.g., Breshears et al. 2005; Vicente-Serrano et al. 2019) and agriculture (e.g., Winkler et al. 2017). Unfortunately, such an analysis is still lacking over Syria.

Given this background, the overriding aims of this study are to (1) assess recent changes in meteorological drought characteristics (e.g., frequency and intensity) over Syria during the period 1990–2010 using a common drought indicator (SPI) and employing the most dense network (N = 36) of meteorological data available for the Syrian territory, (2) link changes in drought characteristics with vegetation greening, as represented by the Normalized Difference Vegetation Index (NDVI), and (3) evaluate the impact of meteorological drought on crop failure in Syria from 1990 to 2010.

Methods

Study area

Syria is located in southwestern Asia between the latitudes of 32.24° and 37.23° N and the longitudes of 35.78° and 42.74° E (Fig. 1). Syria is characterized by complex topography, with a gradient from southeast to northwest. Due to the complex topography, latitude and land–sea interactions, climate is highly variable over space and time. With a total area of 185, 180 km², the climate in Syria is generally continental, with hot summers (> 40 °C) and cold winters (< 0 °C). Indeed, this continentality is less pronounced close to the Mediterranean coast and in mountainous regions, compared to the mainland and eastern portions of the country. Rainfall in Syria is mainly concentrated from October to April, with annual totals ranging between 250 and 1000 mm (Fig. 2). Rainfall is highly variable over space, as a function of distance to maritime influences (i.e., the Mediterranean Sea) and topography. In particular, while the eastern regions of the country are almost dry (< 25 mm/year), rainfall is more intense in coastal and mountainous regions (750–1000 mm/year) and to a less extent in mainland regions (250–500 mm/year). Overall, following Köppen climate classification, Syria is divided into three main climate groups: hot arid

desert (BWh), semiarid hot desert (BSh), and Mediterranean hot summer (Csa). In Syria, the agricultural lands occupy more than 75% of the total land, being the main income source for more than 48% of the population (Alkhalil 2009; Al-Youssef et al. 2016; Mohammed et al. 2019a, b). Importantly, it consumes 87% of the national water supply (Moradi et al. 2011). Furthermore, agriculture system is a mixed crop/livestock agro-ecosystems. With almost 27% of the Syrian territory being classified as arable land (6.5 million hectares), agriculture contributes to more than 17.6% of the Gross Domestic Product (Mohammed et al. 2020c). The total area of rain-fed agriculture is almost three times that of irrigated agriculture, making this vital sector more sensitive to climate change. Spatially, the irrigated agriculture is mainly distributed over the eastern parts of the country. The most common crops are wheat, barley, cotton and olive.

Dataset description

Monthly rainfall data were collected from 36 ground meteorological stations for the period 1990–2010. Data were provided by the Syrian Ministry of Agriculture (MoA). All stations were subjected to a rigorous procedure to check data completeness, quality, and homogeneity. Following the guidelines of the World Meteorological Organization (WMO 1986), data were checked first to remove any erroneous data that may result from digitizing and achieving errors (e.g., negative values of precipitation). Then, the external consistency of the data was verified by comparing data at each candidate station with best-correlated neighboring stations. The main characteristics of these rain gauges are listed in Table 1. As illustrated in Fig. 1, the stations are evenly distributed over space, covering the main climatic regions in Syria, albeit with less density over central and southeastern regions of the country.

Drought characterization

Drought can be grouped into four main categories: meteorological, hydrological, agricultural, and socioeconomic (Wilhite and Glantz 2009). Meteorological drought is characterized by precipitation deficits for a period of time, while hydrological drought is more linked to shortage in surface and groundwater flows (Mishra and Singh 2010). Rather, agricultural drought occurs when the available amount of moisture for soil and plant is much lower than atmospheric demand for evapotranspiration (Woli et al. 2012). The socioeconomic drought is associated with the different environmental and socioeconomic impacts caused by water scarcity (Mishra and Singh 2010). In this study, meteorological drought was characterized using SPI, which is a well-established drought index (McKee et al. 1993). This index is recommended by the World Meteorological Organization
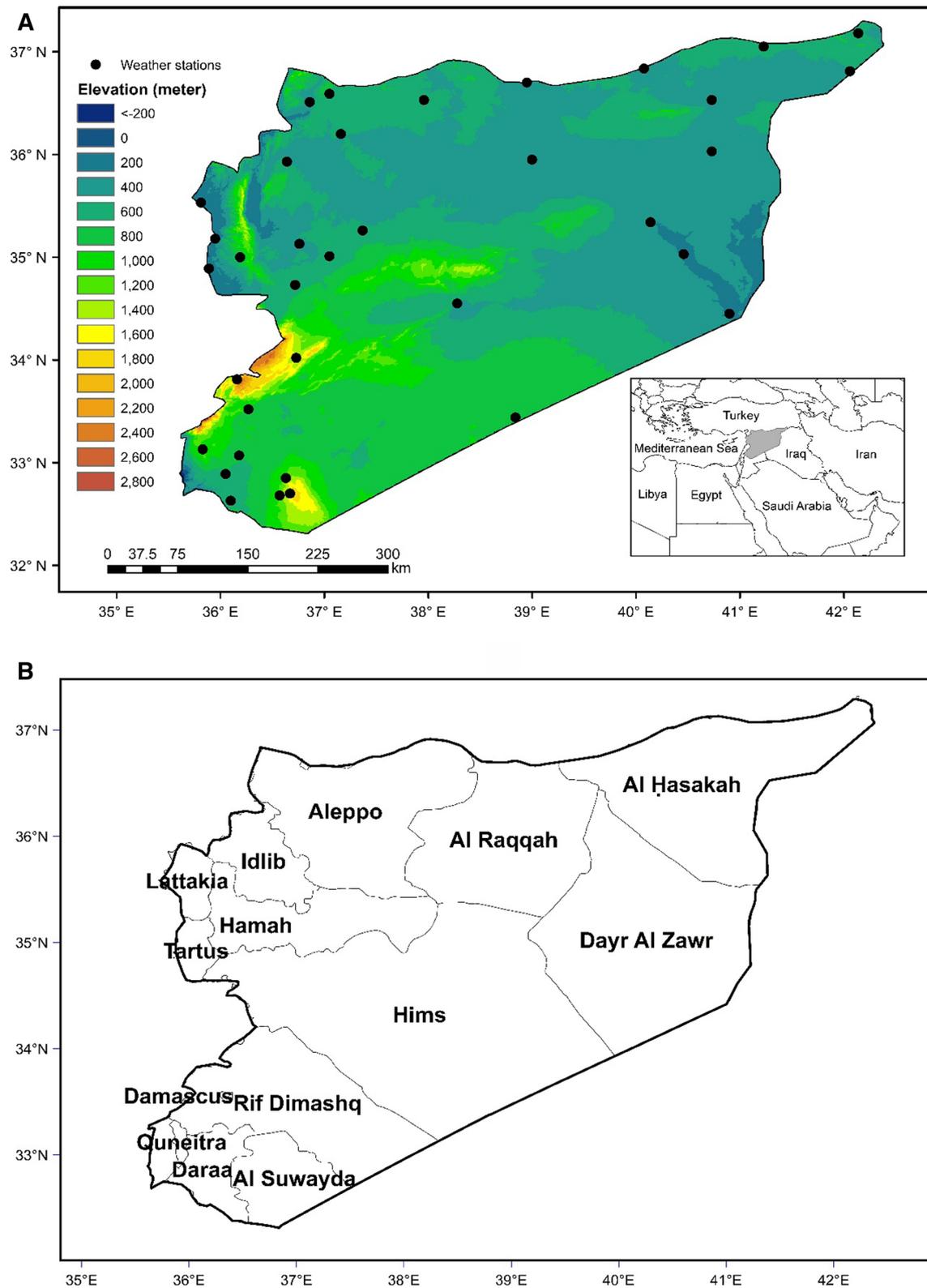
Fig. 1 **a** Topography of Syria and the spatial distribution of the rain gauges (black dots) used in this study, and **b** the main administrative divisions of the country

Fig. 2 Seasonal cycle of rainfall over Syria. Monthly averages were calculated for the base period 1990–2010. The central dotted lines indicate the mean, while the central black lines denote the median. The lower and upper dots show the 5th and 95th percentiles of the monthly temperature distribution. Black circles denote the highest and lowest values
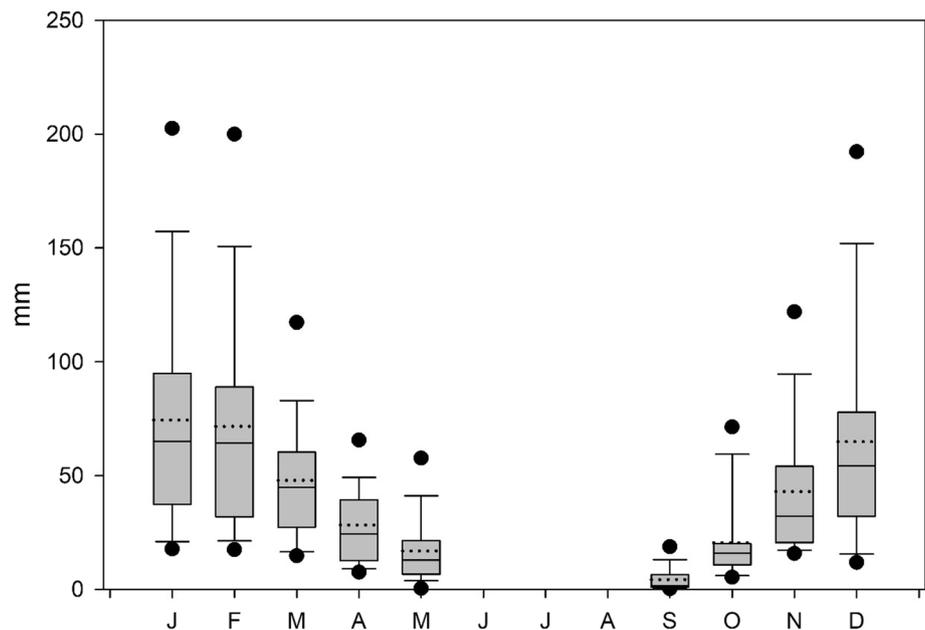
(WMO) for assessing drought characteristics (Sobral et al. 2019). In literature, there exists a broad range of widely-recognized drought indices, including Palmer Drought Severity Index (PDSI) (Palmer 1965), SPEI (Vicente-Serrano et al. 2010), Surface Water Supply Index (SWSI) (Shafer and Dezman 1982), Soil Moisture Drought Index (SMDI) (Hollinger et al. 1993), Regional Deficiency Index (RDI) (Stahl 2001), Streamflow Drought Index (SDI) (Nalbantis and Tsakiris 2008), Rainfall Anomaly Index (RAI) (van Rooy 1965), and the Standardized Runoff Index (SRI) (Shukla and Wood 2008). For better characterization of hydrological and agricultural droughts, indices that consider important hydroclimatic variables (e.g., soil moisture, runoff and stream discharge) or those impacting heat and energy fluxes (e.g., air temperature, wind speed, and evapotranspiration) are favorable than the indices that consider precipitation only (e.g., SPI). However, recalling that large number of meteorological variables necessary to quantify these indices and their data access limitations in our study domain, our preference was given to the use of the SPI (which requires rainfall data only). This decision should also be seen in the context that, as opposed to humid regions, drought in arid and semiarid regions is largely driven by rainfall deficiency. Overall, SPI is recommended for drought monitoring, particularly meteorological droughts (Hayes et al. 2011; WMO 2012). Similar to other multi-variables indices (e.g., SPEI, PDSI), SPI can also be computed at different timescales, allowing for delineating different drought categories (e.g., meteorological, agricultural, and hydrological). Furthermore, it allows a direct comparison among regions with different precipitation magnitude and climatic conditions. Moreover, multi-scalar character of this index allows to

define the response to drought, which can vary considerably within the same region in response to drought timescale (Sheffield et al. 2009; González-Hidalgo et al. 2018). Mishra and Singh (2010) provided a comprehensive review of the main features of the different drought indices. Overall, due to its simplicity, ease of computation, multi-scalarity, and low data requirements, SPI has been used in this work.

Statistically, SPI converts precipitation data into probabilities (e.g., gamma, Pearson III, log-Logistic) computed at different timescales. Herein, the Pearson III distribution was applied to convert precipitation data into standardized series, with an average of zero and a variance of one. This normalization allows for comparison among regions of different geographic and climatic characteristics. Herein, the SPI values were computed based on the Pearson III distribution and L-moment statistics following the procedure described by Vicente-Serrano (2006).

As SPI is a multi-scalar index, we restricted our analysis to the 12-month timescale. While shorter timescales (e.g., 1-, 3-, and 6-month scales) can give indications on meteorological droughts, longer timescales (e.g., 12 months) can inform on the impacts of climate change on annual changes in precipitation that can impact important sectors in the region (e.g., agriculture and crop yield). Numerous studies have adopted the 12-month timescale to characterize meteorological drought in different regions worldwide (e.g., Moradi et al. 2011; Sahoo et al. 2015; El Kenawy et al. 2016a, 2020). In this context, it is well-recognized that hydrological and agricultural droughts cannot be measured immediately due to the delay between its manifestation and the eventual determination of the magnitude (via yield measurements made at the end of the growing season) (Tan

Table 1 The main characteristics (i.e., latitude, longitude, elevation above mean sea level) of the rain gauges used in this study

Station	Longitude (°E)	Latitude (°N)	Elevation (m)	Annual total precipitation (mm)
Afrien	36.51	36.86	262	423.1
Ain Arb	32.70	36.67	1557	534.0
Aleppo	36.20	37.16	397	380.7
Azzaz	36.59	37.05	555	406.2
Banyas	35.18	35.95	31	782.9
Boukamal	34.45	40.90	195	113.0
Damascus	33.52	36.27	759	124.1
Dier-zour	35.34	40.14	213	147.3
Draa	32.63	36.10	519	241.4
Hamma	35.13	36.76	288	336.9
Haskah	36.53	40.73	314	230.3
Homs	34.73	36.72	510	429.9
Idleb	35.93	36.64	431	502.0
Lattakia	35.52	35.79	30	756.9
Manbej	36.53	37.96	460	255.6
Miadeen	35.03	40.46	187	177.0
Mlkiah	37.18	42.14	486	484.7
Nabk	34.02	36.73	1288	113.6
Nawa	32.89	36.05	554	404.9
Palmera	34.55	38.28	408	97.0
Qamishlih	37.05	41.23	456	365.2
Qnitara	33.13	35.83	943	668.8
Raqqa	35.95	39.00	246	170.7
Rass-Ain	36.85	40.07	349	266.4
Salmiah	35.01	37.05	459	275.6
San	35.26	37.37	449	215.1
Sanamine	33.07	36.18	637	253.7
Shadadah	36.03	40.73	274	230.3
Shahba	32.85	36.63	1078	309.8
Srghaia	33.81	36.16	1400	124.1
Swieda	32.68	36.57	1024	332.2
Tall-Abiad	36.70	38.95	351	233.2
Tanf	33.44	38.84	716	97.0
Tartous	34.89	35.89	20	880.2
Wadi-Eouin	35.00	36.19	649	1277.2
Yaroubiah	36.81	42.06	393	333.7

The annual total precipitation was calculated for the base period 1990–2010

et al. 2015). As such, below-normal rainfall may not cause an immediate decline in crop health, given that plants can adapt to meteorological drought using short-term adaptation mechanisms to deal with water stress (Rockström and Oweis 2009). Herein, SPI was calculated at 12-month timescale to capture this delayed impact of rainfall deficit on crop yield. Agricultural droughts are commonly assessed using drought indices at these long timescale (e.g., 18 and 24 months), as these longer timescales allow to capture the accumulated effect of meteorological drought that can affect soil moisture and stream discharge. Several studies employed SPI at a long

timescale to characterize meteorological drought in many regions worldwide (e.g., El Kenawy et al. 2016b; Gumus and Algin 2017; Kubiak-Wójcicka and Bąk 2018; Cammalleri et al. 2019). SPI calculation was made using SPI package within R platform following the guidelines of the World Meteorological Organization (WMO). Following the scheme of Agnew (2000), the values of the SPI were categorized into different classes, as listed in Table 2.

With the aim of defining the dominant spatial patterns of meteorological drought variability in Syria, we employed the principal component analysis (PCA) in an

Table 2 Drought categories according to Agnew's scheme (2000)

SPI values	Drought category
Less than 0	No drought
Less than − 0.50	Mild drought
Less than − 0.84	Moderate drought
Less than − 1.28	Severe drought
Less than − 1.65	Extreme drought
More than − 1.65	Very extreme drought

S-mode (Richman 1986). This multivariate statistic can produce meaningful spatial patterns of drought, which summarizes the high variations of climate in Syria on one hand and facilitates the development of appropriate adaptation strategies to cope with issues arising from climate change, especially at local and regional scales, on the other hand. The PCA has commonly been used to obtain homogenous spatial and temporal patterns of climate (e.g., Wang 2005; El Kenawy et al. 2013), including drought (e.g., Vicente-Serrano 2006; El Kenawy et al. 2016b; Mathbout et al. 2018). We employed a correlation matrix to conduct the PCA and obtained the components in the original correlation coefficient values using a weighted coefficient of each meteorological station in each component. Following the results of the empirical orthogonal functions (EOFs), we maintained only the most significant components: i.e., those with eigenvalue greater than one (Kaiser 1960). To reduce data dimensions and capture the maximum variability of SPI data, the retained components were rotated using the Varimax method. Then, each meteorological station was assigned to a unique component based on the highest correlation coefficient found between each station and all components.

In order to extend the temporal analysis in each region, we identified the most representative observatory of each component (based on factorial loadings). The aim was to analyze the interannual variability of drought characteristics corresponding to each retained component. Changes in SPI series were assessed using the ordinary least squares method, while their statistical significance was tested using the modified Mann- Kendall statistic (Hamed and Ramachandra 1998). Basically, this is a non-parametric statistic, which does not assume a priori knowledge of data distribution. Importantly, this statistic accounts for the presence of serial autocorrelation in the series, limiting its possible influence on trend significance. Herein, it is noteworthy to indicate that some uncertainty in the trend calculation can be originated from the relatively short span of this study (1990–2010), as this period might not be sufficient to characterize a stationary climate, i.e., a climate that presents variability, but that is not changing significantly over time. From a statistical perspective, as opposed to parametric tests, this test is less sensitive

to the sample size. The term "trend" hereinafter will be used to refer to the available 21 years of data.

Links of drought variability to NDVI

Remote sensing products provide spectral radiance through a wide range of the electromagnetic spectrum, allowing for defining indices that characterize vegetation conditions (Tucker 1979).

In this study, we assessed the links between meteorological drought variability and changes in vegetation cover using NDVI. The NDVI is one of the widely recognized vegetation indices that measure the fractional absorbed photosynthetically active radiation (Myneni et al. 1995a, b). The NDVI has been widely employed in several studies as an indicator of green biomass (e.g., Lu 2006), vegetation cover (e.g., Deng et al. 2018), and the Net Primary Production (NPP) (e.g., Xu et al. 2012). We retrieved NDVI from the National Oceanic and Atmospheric Administration Advanced Very High-Resolution Radiometer (NOAA/AVHRR) sensor (https://search.earthdata.nasa.gov) (Roy, et al. 2002). NDVI series were obtained for the whole Syrian territory at 1 km spatial resolution for the period 1990–2010. Although some global space-based products, such as Satellite Pour l´Observation de la Terre (SPOT) and Moderate-Resolution Imaging Spectroradiometer (MODIS), can provide NDVI at an analog horizontal resolution (1 km), our preference was given to use NOAA/AVHRR data set. This is simply because the temporal coverage of these fine-resolution products is generally short (mostly from the early 2000s), while NOAA/AVHRR records extend back to the early 1980s. Also, with the high-resolution spectral information from passive sensors, a wide range of remote sensing vegetation indices have been employed to characterize vegetation dynamics. Xue and Su (2017) provided a comprehensive review of the advantages and limitations of remotely sensed vegetation indices. However, NDVI is one of the well-established vegetation indices that measures the fractional absorbed photosynthetically active radiation (e.g., Myneni et al. 1995). Herein, Pearson's r correlation was computed to evaluate the direction and strength of association between drought variability and NDVI evolution over the common period 1990–2010. This association was computed for the whole Syrian domain as well as for the leading components defined following PCA results (Sect. 2.3). This procedure gives indications on the impacts of meteorological drought variability on vegetation greening at different spatial scales in Syria. The statistical significance of the correlation was assessed at the 95% level ($p < 0.05$). Prior to calculating correlation, the time series of SPI and NDVI were detrended. The rationale behind this procedure is to limit the possible influence of the trends presented in the series and accordingly separate the observed patterns of variability from any global warming

signal presented in the data. A similar procedure was applied in previous works (e.g., El Kenawy et al. 2016a).

Links of drought variability to crop yield

Also, we assessed the impacts of meteorological drought characteristics using some available sectorial data for Syria. Specifically, we linked drought with agricultural data obtained from the Ministry of Agriculture (MoA) for the period 2000–2010. The data included the annual production (in tons) of the main cereal crops and grazing, including barley, wheat, grazing clover, grazing barley, grazing flowering.

Results

Changes in drought intensity and vulnerability

Figure 3 illustrates the regional series of SPI at 12-month timescale. The regional series was calculated for the whole Syria using data of all station-based SPI series. As depicted, there is a statistically significant decrease in the SPI index, on the order of -1.1.9 Z unit/decade, suggesting an increase in the intensity of meteorological drought events, compared to wet conditions. Notably, meteorological drought was more intensified in the early (1999–2001) and late 2000s (2007–2010), while wet events predominated mainly from 1992 to 1998, being much enhanced from 2002 to 2005. The most prolonged dry episode extended for 37 consecutive months from November 1998 to November 2001, followed by a 34-month dry episode from January 2007 to October 2009. These two episodes were generally mild and moderate, with an average SPI value of -1.01 and -0.92, respectively. On the other hand, two main wet episodes were identified: from March 1996 to October 1998 and from December 2001 to January 2005.

The vulnerability of the Syrian territory to the different drought classes is illustrated in Fig. 4. As depicted, the percentage of stations impacted by all drought classes exhibited an increase during the study period. Nonetheless, this increase was found statistically significant ($p < 0.05$) only for mild and severe droughts. In contrast, the percentage of stations influenced by wet conditions (i.e., SPI > 0) showed a strong decline from 1990 to 2010, on the order of -8.9% per decade ($p < 0.05$). A comparison between the different drought classes suggests that the increases in the percentage of stations impacted by mild, moderate, and severe droughts were much stronger than those of extreme and very extreme droughts. For example, the percentage of stations influenced by mild drought increased by 3.8% per decade, compared to 2.6% and 1.3% for severe and moderate droughts, respectively. Notably, the stations affected by very extreme drought (SPI < −1.65) were almost half of all stations during the extraordinary drought episode from May 1999 to January 2000, whereas more than 20% of the stations were triggered by very extreme drought from May to November 2008.

Spatial and temporal variability of drought

An S-mode of the PCA was employed to SPI series to define the dominant spatial patterns of drought variability in the study domain. Following the results of the scree plot (Fig. 5), we defined seven components that together account for 83.9% of the total explained variance of drought. The first (PC1) and second (PC2) components account together for almost 48.3% of the total variance, while the remaining components account only for less than 10% of the total variance for each, with PC7 explaining only 4.1% of the variance.

Figure 6 illustrates the spatial distribution of the loadings corresponding to each retained component. A quick inspection of the spatial structure of the different components
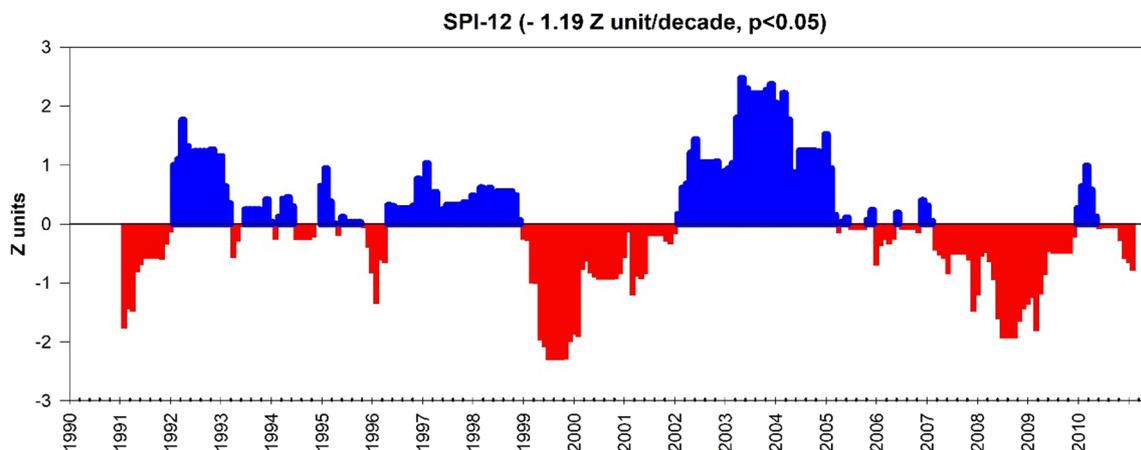

Fig. 3 Temporal evolution of the SPI values at 12-month timescale. The amount of change is expressed in Z units per decade
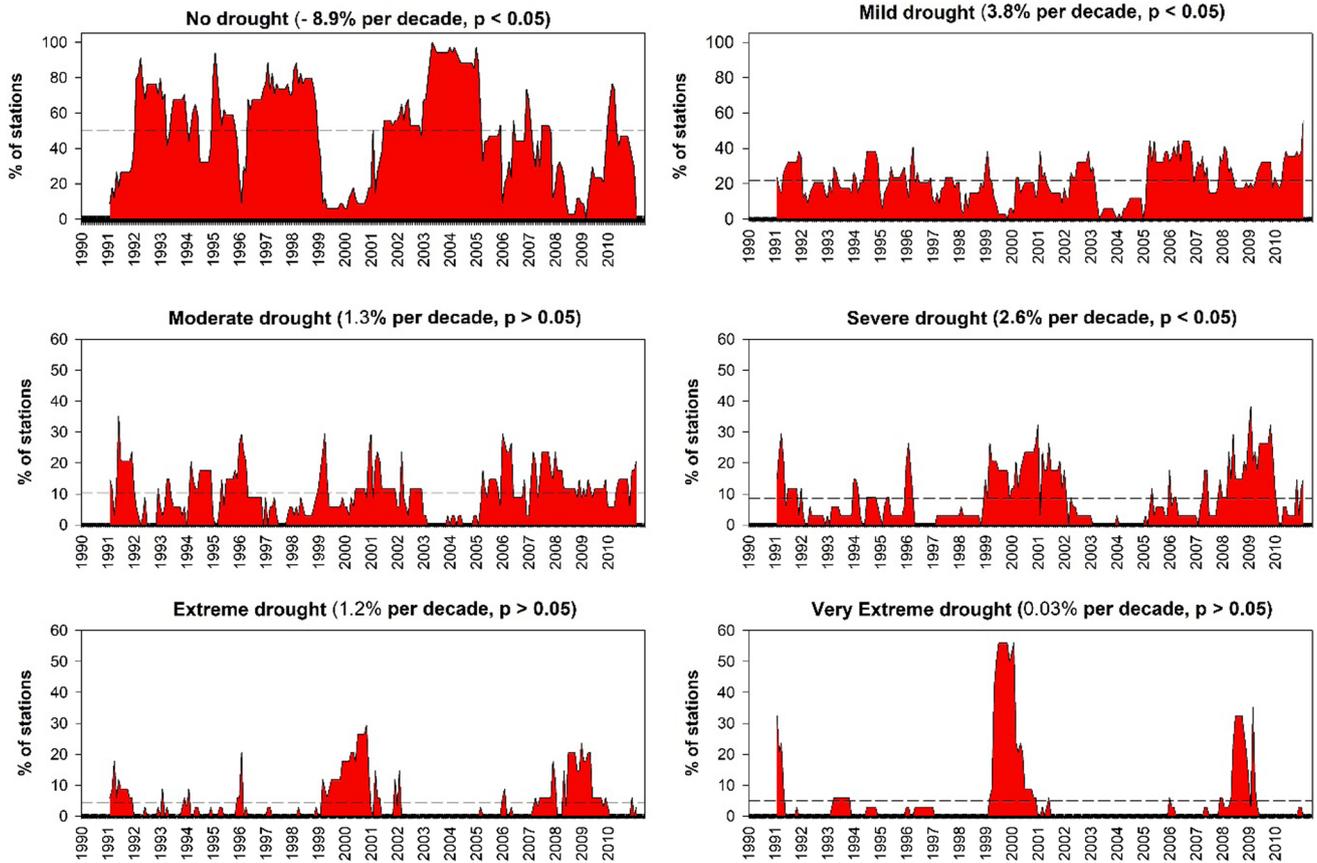
Fig. 4 Percentage of stations impacted by each drought category for the period 1990–2010. Magnitude of change given in bold is statistically significant at the 95% level ($p < 0.05$). The dashed line indicates the mean of this percentage for the whole study period

Fig. 5 A scree plot showing the cumulative variance explained by the different components following the PCA results. The seven components were retained based on eigenvalues greater than one (as indicated by the horizontal red line)
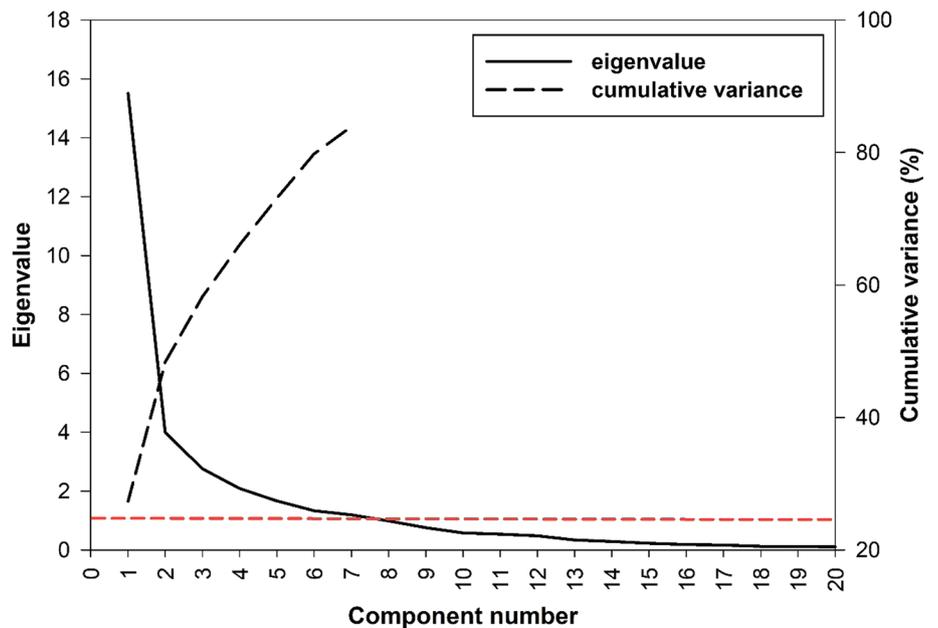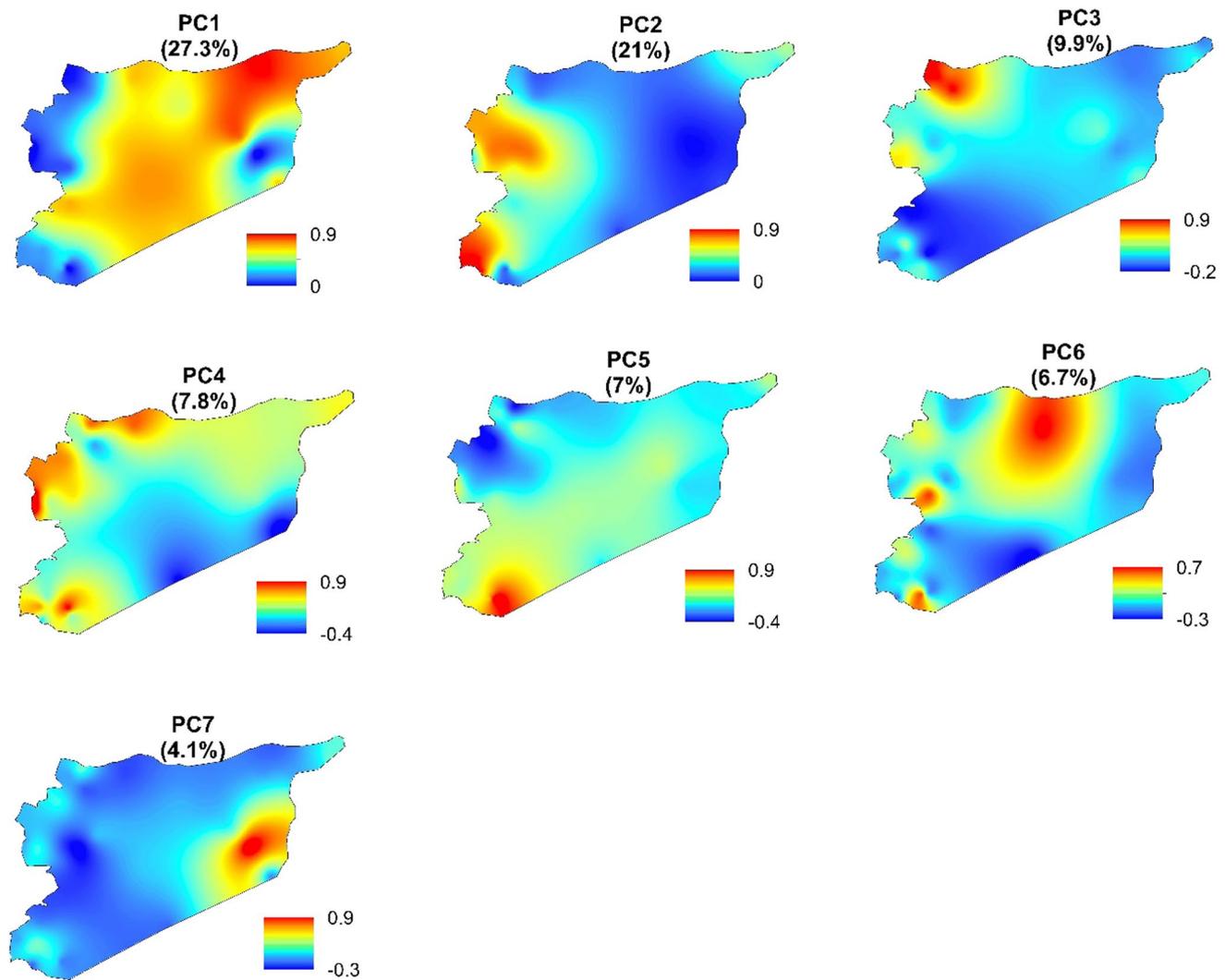
Fig. 6 Spatial distribution of PC loadings (correlations between each observation and each retained component, with values ranging theoretically between −1 and 1). The number between parentheses indicates the amount of variance explained by each component

reveals that they are spatially distinct. For example, PC1 corresponds to regions with low annual rainfall (< 200 mm), while PC2 is mainly assigned to the Mediterranean region of Syria with annual rainfall totals exceeding 800 mm. In the same context, PC5 summarizes drought variability in regions with moderate rainfall (200–400 mm). Overall, the first component (PC1), which explains 27.3% of the total variance, showed higher loadings over central and northeastern portions (especially in Al Hasakah). On the other hand, the highest loadings of PC2 were particularly distributed along the Mediterranean coast, mainly in Lattakia, Tartus and Qunietra. The remaining components, with less than 10% of variance for each, were assigned to specific small regions in the country, indicating that these components capture the local variability of meteorological drought in Syria. A representative example is PC3 whose highest loadings

were recorded in the northwestern portions of Aleppo region. Similarly, PC5 exhibited its maximum loadings in Al Suwayda, while the loadings corresponding to PC6 were greater in Al Raqqah.

Figure 7 illustrates the temporal evolution of meteorological drought corresponding to each PC. Results reveal important differences among the components. Although some components showed an increase in SPI values (e.g., PC3 and PC5), others exhibited a decrease (e.g., PC1, PC2, PC4, PC6, and PC7). Apart from PC5 and PC7, changes in SPI values of all components were statistically significant at the 95% level ($p < 0.05$). The largest decline in SPI values (i.e., increase in drought intensity) was found for PC1 (−0.74 Z unit/decade), PC4 (−0.55 Z unit/decade) and PC2 (−0.45 Z unit/decade). On the other hand, the tendency toward an increase in SPI values (i.e., less intense drought)
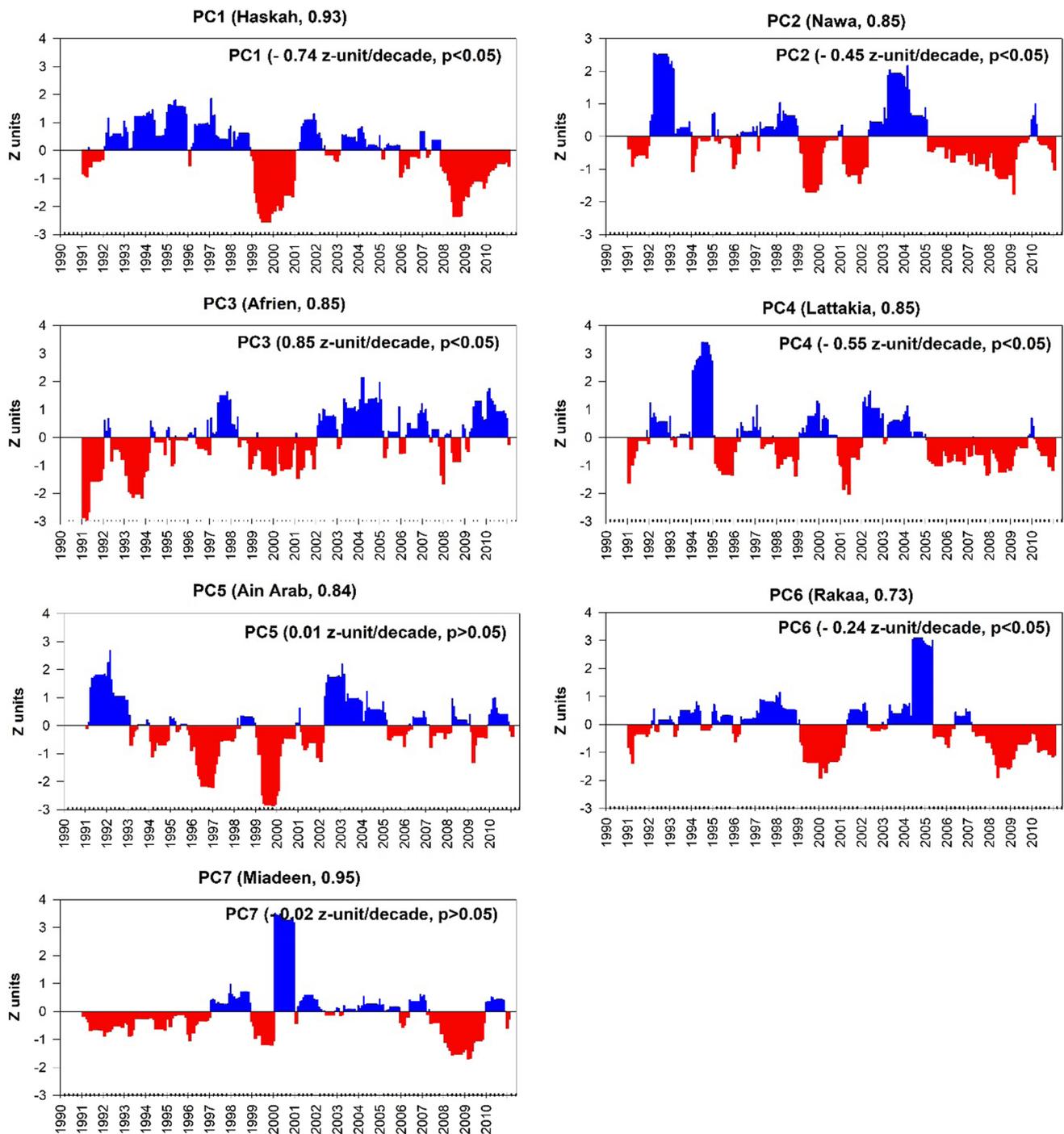
Fig. 7 Temporal evolution of the scores corresponding to the best-correlated observatory with each PC (the name of this observatory and its factorial loading is included). The amount of change is expressed in z unit per decade. Statistically significant ($p < 0.05$) and nonsignificant ($p > 0.05$) trends are shown

was markedly noted for PC3, with magnitudes of change of 0.85 Z unit/decade. The increase in wet conditions for PC5 was statistically nonsignificant ($p > 0.05$). These findings suggest a high spatial variability of drought in the study domain. Apparently, the tendency toward drier conditions

was more pronounced along the Mediterranean coast and in Al Hasakah district to the northeast. In contrast, northern Aleppo and Al Suwayda were the regions that exhibited an increase in wet conditions over the study period. PC7, whose higher loadings were assigned mainly to Dayr Alzawr,

showed a decrease in SPI values (i.e., drier conditions), though being statistically nonsignificant ($p > 0.05$). Nonetheless, the regions corresponding to this component witnessed exceptional wet episodes, mainly during 2000/2001. A similar pattern was observed for PC6 (Al Raqqah), which witnessed an extraordinary wet episode from April 2004 to April 2005, with SPI values exceeding 3.

Links between drought variability and vegetation greening

Figure 8 summarizes Pearson's correlation coefficients calculated between the detrended time series of SPI at 12-month timescale and NDVI for the common period 1990–2010. This association was assessed for the whole Syria and the seven significant PCs, described in Sect. 3.2. As expected, the association between SPI and NDVI was generally positive for all months either for the whole country or its subregions, with few exceptions (e.g., PC5 in February, March, October, November, and December). The weakest correlation between NDVI and SPI was markedly found for PC4 and PC5, with r values generally below 0.2. In contrast, the maximum correlation was noted for PC1, PC2, and PC3, with the highest r value found for PC1 in April (r = 0.7). Notably, the association between drought and vegetation was much stronger during the rainy season (November- May), while it was weaker during summertime (June–August). A comparison between the different PCs indicates some spatial differences. Specifically, vegetation cover in PC1 (mainland and northeastern region) responded strongly to drought in

the late winter and early spring (February to April), while its response was weaker in the early winter (December–January). A different pattern is observed for PC3 (Aleppo), with stronger responses during wintertime (December–February). In comparison to other PCs, PC2 (close to the Mediterranean coast) shows high intra-variability in its response to drought variability, with stronger feedback at the end of spring (May) and early summer (June). Conversely, NDVI of the Mediterranean region of Syria shows low, and even negative, correlation with meteorological drought variations during wintertime, mainly from November to February. NDVI corresponding to PC7 (the most continental region in Syria; Dayr Alzawr) shows low intra-annual variability in its response to drought, with r values ranging between 0.1 and 0.3 for majority of months. As illustrated in Fig. 8, PCs that explain much of the meteorological drought variability (i.e., PCs 1, 2 and 3) and extend over vast areas of Syria (refer to Fig. 6) were the only PCs that showed statistically significant correlations between SPI and NDVI ($p < 0.05$). This correlation was generally above 0.5, especially at the end of springtime. In contrast, PCs that explained lower amount of drought variability ($< 10\%$) such as PCs 5, 6 and 7 exhibited nonsignificant impact of meteorological drought variations on NDVI. On the national scale, the best correlation between NDVI and SPI was found during wintertime (December-January), May and June. Contrarily, the lowest association was noted in October and November (r ≈ 0) and during summer months (r ≈ 0.35).

The statistically significant positive correlation found between NDVI and SPI, especially for the leading PCs (PCs
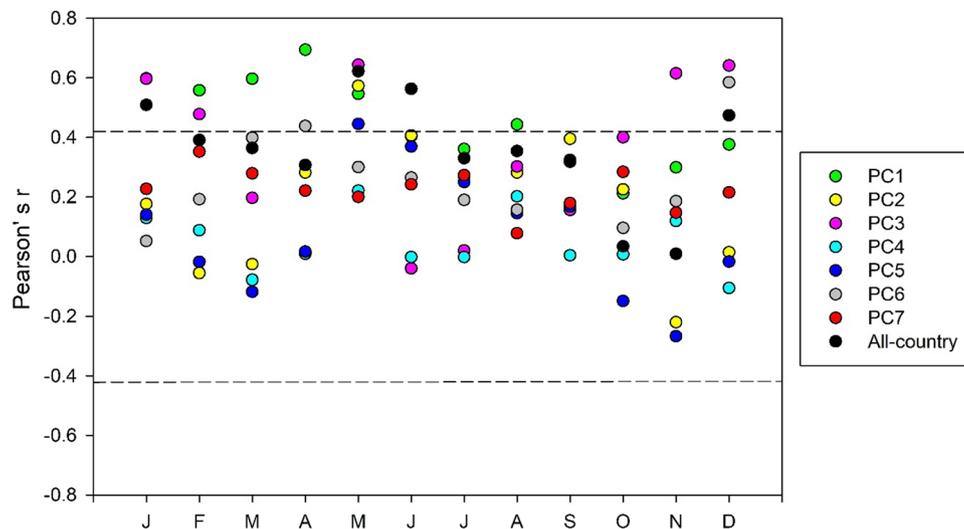

Fig. 8 Pearson's correlation coefficient calculated between SPI 12-month series for the best-correlated observatory with each PC and NDVI series extracted from the nearest grid cell to the best-correlated observatory. The correlations were calculated on the monthly scale for each component and for the whole domain from 1990 to

2010. The regional series of SPI 12 months and NDVI for the whole domain were computed using an arithmetic average for all stations (SPI 12 months) and grids (NDVI) spanning Syria. The dashed line indicates the statistical significance threshold of Pearson's r

1, 2 and 3), was confirmed at a detailed spatial scale (1 km assessment) using Moderate-Resolution Imaging Spectroradiometer (MODIS). Figure 9 illustrates the differences in NDVI values, retrieved from MOD13A3 V.6 (https://search.earthdata.nasa.gov), between April 2003 and April 2008. April was selected as it represents the average conditions of spring season: the season in which NDVI showed a stronger response to meteorological drought variability. In the same context, the years 2003 and 2008 were selected as representatives of the anomalous wet (2003) and dry (2008) conditions in Syria over the study period (1990–2010) (Fig. 3). Interestingly, the results indicate a significant decrease in the vegetation cover between 2003 and 2008, on the order of 19.2% of the total vegetation area in the country (Table 3). Spatially, the northeastern parts of Syria, which correspond to PC1 and explain 27.3% of drought variability, exhibited the largest decline in vegetation cover. The correlation between NDVI, though being retrieved from NOAA/AVHRR, and SPI recorded its highest value for PC1 in April ($r = 0.7$). Interestingly, the northern central areas of Syria (Al Raqqah; PC6) also witnessed a significant decrease in vegetation cover (Fig. 9). The correlation between NOAA/AVHRR-NDVI and PC6 was quite strong and statistically

significant ($r = 0.44$) for April. Interestingly, PC6 ranked second, after PC1, among all sub-regions in terms of the association between NOAA/AVHRR-NDVI and SPI during April, which was also evident using NDVI data retrieved from MODIS product. Notably, the Mediterranean region of Syria (PC2) exhibited less changes in NDVI between 2003 and 2008.

Links between drought variability and crop failure

We simply analyzed the impacts of meteorological drought variability on crop failure in Syria by assessing changes in the relative anomalies of the national yield of the main cereal
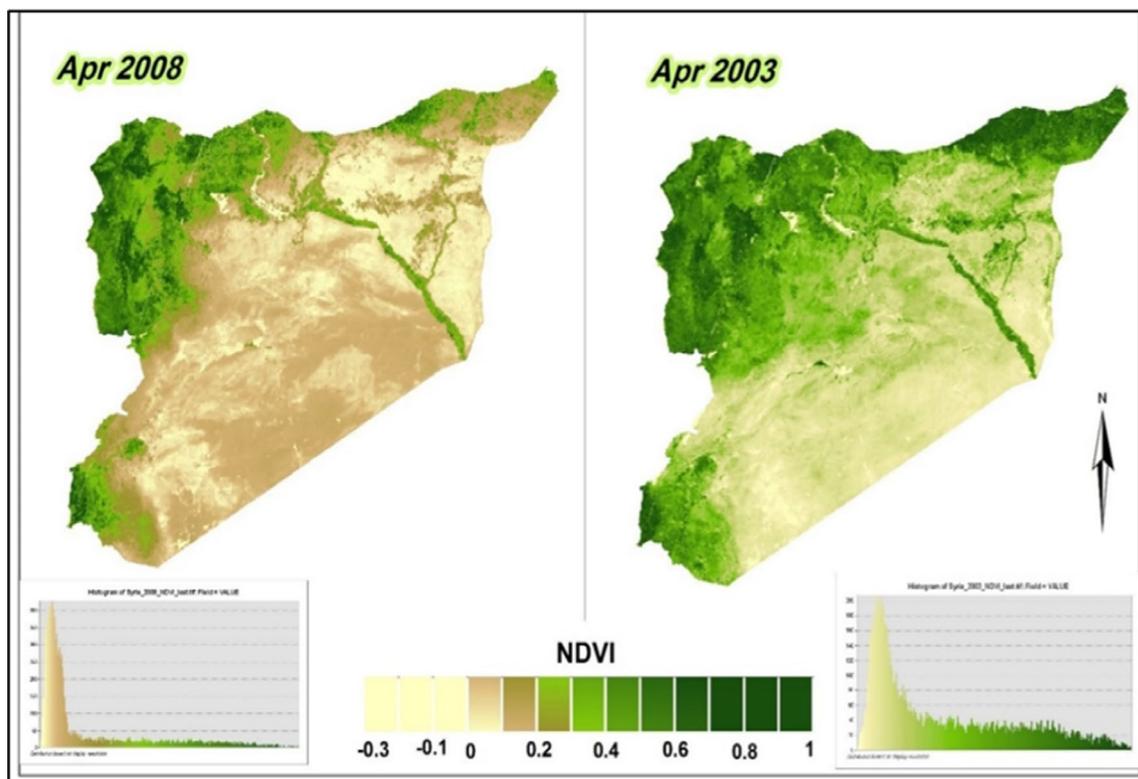
Table 3 Changes in vegetation coverage over Syria using differences in the NDVI values between April 2003 and April 2008

NDVI threshold	April 2003		April 2008		% Changes
	Area (km2)	%	Area (km2)	%	
≤ 0.3	122470.32	60.93	161,069.10	80.14	19.20
> 0.3	78,516.08	39.07	39,917.33	19.86	–19.20

Fig. 9 Spatial and temporal differences in NDVI values over Syria between April 2003 and April 2008. Herein, NDVI values were retrieved from the Moderate-Resolution Imaging Spectroradiometer (MODIS) at a spatial resolution of 1 km, allowing for a detailed spatial comparison of NDVI differences between April 2003 and April 2008. NDVI was retrieved using a 16-day composite from MOD13A3 V. 6 satellite product (https://search.earthdata.nasa.gov; Roy et al. 2002)

crops (e.g., wheat, barley) and grazing products (e.g., clover, flowering) for the period 2000–2010 (Fig. 10). For cereal crops, results showed a noticeable decline in the national yield during the two mega-drought episodes in 2000s: i.e., 2000–2001 and 2007–2010. However, the impacts of meteorological drought on cereal crop yield were notably more enhanced for the period 2007–2010. In 2008, Syria witnessed a remarkable crop failure, where the national yield of wheat, barley, lentils and chickpeas declined by − 47.3, − 69.3, − 72.4, and − 54.6%, respectively, compared to the 2000–2010 yield average. In 2000, the decline in crop yield was noted primarily for barley (− 75.1%), lentils (− 41%) and wheat (− 23.5%). On the other hand, results did not reveal a dominant pattern of changes in grazing production, albeit a significant decline in grazing clover (− 32.8%) and grazing flower (− 43.4%) is noted in 2008.

Discussion

Drought is a key natural hazard, with serious environmental, economic and social implications (IPCC 2013; Solh and van Ginkel 2014; Khayyati 2016). Over the past few decades, characterization of drought has received an ongoing interest in hydroclimatic research, particularly with the accelerated impact of recent climate change (e.g., Solh and van

Ginkel 2014; Vicente-Serrano et al. 2014; Khayyati 2016; El Kenawy et al. 2016b; Dabanli et al. 2017; Alsafadi et al. 2020).

In this work, we assessed the spatial and temporal variability of meteorological drought in Syria from 1990 to 2010 using observational rainfall data from 36 meteorological stations. We employed the SPI to diagnosis drought characteristics at 12-month timescale. Due to its ease of computation and low data requirements, SPI has increasingly been used in a wide range of studies to assess spatiotemporal variability and trends of drought at local (e.g., Vicente-Serrano and López-Moreno 2006; Shah et al. 2015; Mohammed et al. 2020a), regional (e.g., Li et al. 2008; Liu et al. 2015; El Kenawy et al. 2016b; Mohammed and Harsányi 2019), continental (e.g., Winkler et al. 2017; Caloiero et al. 2018a) and global scales (e.g., Dutra et al. 2014; Mo and Lyon 2015; Spinoni et al. 2019), including the Middle East region (e.g., Al-Qinna et al. 2011; Golian et al. 2015; Mahfouz et al. 2016; Abu Hajar et al. 2019; Mohammed et al. 2019a, b). Using a multi-variable drought index (e.g., SPEI) is definitely desired for a proper assessment of drought characteristics, given that it accounts for important climatic variables that can have significant influence on drought evolution (e.g., air temperature, wind speed) (Vicente-Serrano et al. 2010). According to Van Lanen and Van Loon (2012), wet-to-dry-season droughts are predominant in the Mediterranean area.

Fig. 10 Variations in the yield of winter main crops (upper) and winter grazing (lower) over Syria from 2000 to 2010. The anomalies were computed relative to the base period 2000–2010 and expressed in percentages (%)
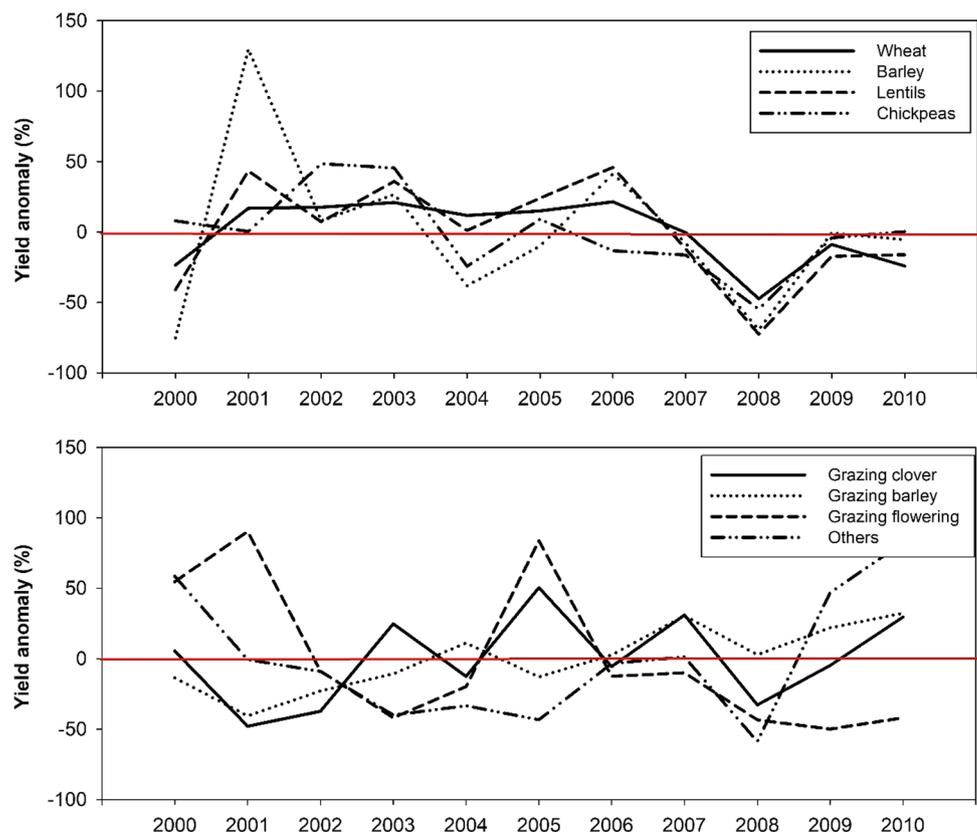
While the development of this type of drought is mostly controlled by precipitation, the recovery is

controlled by both precipitation and temperature (Van Loon 2015). However, such indices need a large amount of data, which might not be available or accessible in developing countries, with poor instrumental infrastructure, like Syria. Nonetheless, several studies indicate that the findings of SPI and other multi-variable indices (e.g., SPEI) are generally comparable over space and time (e.g., Mathbout et al. 2018).

This study demonstrates a statistically significant decline in SPI values over Syria (on the order of − 1.1.9 Z unit/decade), suggesting an acceleration of meteorological drought frequency and intensity from 1990 to 2010. Temporarily, results suggest two main drought episodes, with anomalous SPI values: 1999–2001 and 2007–2010. Although Syria has been impacted by frequent drought episodes over the past decades (Solh and van Ginkel 2014), the most recent drought episode of 2007–2010 was recorded as one of the driest episodes in recent decades (Kelley et al. 2015; Ide 2018; Caloiero et al. 2018a). Cook et al. (2016) indicated that the 1998–2012 mega-drought that hit the Levant region, including Syria, was the driest on record over the past nine centuries. This anomalous event was interpreted within the context of recent warming trend and anthropogenic influences (Kelley et al. 2015; Caloiero et al. 2018a). According to Tanarhte et al. (2012) and Mathbout et al. (2018), since the mid-1980s, the Fertile Crescent region exhibited a statistically significant increase in air temperature, which accompanied by a significant decline in precipitation. Similarly, in their assessment of changes in aridity conditions, El Kenawy et al. (2016a) found that the Fertile Crescent witnessed the most rapid increase in aridity over the Middle East and North Africa (MENA) region. This large-scale trend was also confirmed by several regional assessments over Jordan (e.g., Rahman et al. 2015), Syria (e.g., Salameh and Fallah 2018, Mohammed and Fallah 2019), and Iraq (e.g., Robaa and Al-Barazanji 2015). For instance, Rahman et al. (2015) found a statistically significant decline in annual rainfall over Jordan, at an average rate of 12 mm decade^{-1}. This finding was confirmed by Mariotti and Dell'Aquila (2012) for the whole Mediterranean during the second half of the 20th century. Herein, it should be emphasized that a longer temporal span of data (> 21 years) is highly desired to provide a more robust trend assessment of meteorological drought changes in Syria, which allows for detecting whether the observed changes reveal "real" trends or merely a part of interannual to multidecadal dry/wet cycles that can be induced by circulation variability (e.g., El Nino Southern Oscillation [ENSO], Northern Atlantic Oscillation [NAO], Atlantic Oscillation [AO] phases). However, the recent drought episodes in Syria must be seen within a broader extent that covers the Fertile Crescent in specific and the Mediterranean

in general. For example, a trend toward more drying has been evident in numerous studies spanning the Fertile Crescent, including Lebanon (e.g., Mahfouz et al. 2015), Jordan (e.g., Aladaileh et al. 2019), Iraq (e.g., Hameed et al. 2018), Palestine (e.g., Haensel and Zurba 2015), and Israel (e.g., Turco et al. 2017). For example, Mahfouz et al. (2015) found an increase in drought frequency over Lebanon, with more drought acceleration during winter and spring. Using the SPI and SPEI, Turco et al. (2017) provided strong evidences on an enhancement of drought conditions over Israel from 1982 to 2011. Recently, Mathbout et al. (2018) found a statistically significant increase in drought frequency in Syria from 1961 to 2012 using SPI and SPEI. At a broader spatial scale, this drying trend was also observed across the Western Mediterranean and the Middle East, albeit with reliance on relatively longer climatic datasets, including Oman (e.g., El Kenawy et al. 2020), Turkey (e.g., Saglamoglu and Irvem 2020), Algeria (e.g., Merabti et al. 2018), Italy (e.g., Caloiero et al. 2018b), and Spain (e.g., Vicente-Serrano and López-Moreno 2006; Vicente-Serrano and Cuadrat-Prats 2007). Our results indicate that the recent anomalous drought episodes over Syria (especially in the 2000s) are mostly associated with the increase in the frequency of mild, moderate and severe droughts, recalling that changes in extreme and very extreme droughts were generally weak and statistically nonsignificant ($p > 0.5$).

Spatially, the PCA suggests much heterogeneity of drought variability in Syria, where seven components (PCs) explained 83.9% of the total variance of drought over the 21-year study period. Three components (PCs 1 to 3) explained together 58.2% of the total variance, compared to less than 10% of the explained variance for each of the remaining components. This finding suggests a strong spatial variability of meteorological drought in Syria, stressing that any drought management and adaptation strategies should adopt regional and even local standards rather than national standards. Herein, it should be emphasized that it is highly desired to retain a few and distinguishable number of components, which could explain the majority of the variance in meteorological drought. This is simply because this low number of components allows for better explanation of the physical configurations responsible for drought variability corresponding to each independent component. However, from both statistical and practical perspectives, this is not the case for many hydroclimatic applications, as the number of components reveals the complexity of climatic and geographic conditions in any study domain. As such, for particular regions or variables, a larger number of components could be introduced in response to a high spatial and/or temporal variability of climate. This is typically the case for Syria, given the complex interactions among the region's topography, synoptic-scale processes, and anthropogenic forcing (Mathbout et al. 2018). In Syria, climate is impacted

by systems of major importance from a climatological perspective (i.e., the Mediterranean, Sahara, Indian Monsoon, and the central Asia configurations). Also, climate is influenced by strong topographical and maritime-continental gradients. All these conditions make the climate highly variable and the complex behavior of drought can thus be expected, with varying regional and even local responses. Our results demonstrate that the most leading component (PC1), which explained alone 27.3% of the total variance, had higher loadings over the mainland Syria and in northeastern portions. This region contributes to more than 40% of the total arable land in Syria, from which almost one-third is rain-fed. On the other hand, the second component (PC2) explained 21% of the total variance, mainly assigned to the Mediterranean region of Syria.

With the current advances in space-based sciences, satellites can provide invaluable hydroclimatic information at fine spatial scale. In this work, we assessed the impacts of meteorological drought variability on vegetation cover in Syria using the relatively long-term NDVI data retrieved from NOAA/AVHARR, as well as the spatially detailed NDVI data from MODIS. Results reveal clear temporal and spatial differences in the response of vegetation cover to drought. Higher positive correlations (r > 0.5) between SPI and NDVI series were found from 1990 to 2010 for components that explained much of the meteorological drought variability (i.e., PCs 1, 2, 3), suggesting a decline in vegetation cover in response to the anomalous drought episode. This was clearly evident when comparing NDVI differences between April 2003 and April 2008 using the MODIS detailed spatial data (1 km). Notably, the Mediterranean vegetation was drastically impacted by the anomalous dry episode starting in 2007, registering a decline of 19.2% of its total area in Syria. On the other hand, the association was generally weak and statistically nonsignificant for components explaining low amounts of variability (i.e., PCs 4, 5, 6 and 7). The association between meteorological drought and vegetation was stronger during the rainy season, mainly in winter (December-January) and late spring (April–May). Similarly, the correlation was weak (< 0.15) during summertime, due to absence of precipitation on one hand and lack of vegetation on the other hand. In the literature, several works have employed space-based data to assess and monitor droughts in the Mediterranean using vegetation indices, such as the NDVI, temperature vegetation drought index (TVDI), and vegetation condition index (VCI) (Gouveia et al. 2017; Vicente-Serrano et al. 2019). This assessment was made for the whole Mediterranean region (e.g., Gouveia et al. 2017), as well as for specific sub-regions such as Spain (e.g., Vicente-Serrano et al. 2019). These studies agree that the impacts of drought variability on vegetation greening are likely to be stronger at long drought timescale (e.g., 12, 18, and 24 months) than at shorter timescales (e.g., 3 and 6

months). This can be seen in the context that the vegetation responds better to the long-term accumulated rainfall deficit, especially in regions with high intra-annual and inter-annual variability of climate like the Mediterranean. The strong association between rainfall and NDVI was confirmed in the Fertile Crescent, especially in coastal regions (e.g., Al-Bakri and Suleiman 2004; Abdo 2018). However, this dependency is controlled largely by several environmental variables (e.g., soil type, land use), especially in inner areas with lower rainfall.

This work gives a strong evidence on crop failure in Syria in response to recent severe drought episodes. In 2008, the yield of main cereal crops in Syria witnessed a drastic decline, reaching − 47.3%, − 69.3% for wheat and barley, respectively. These losses can directly be linked to drought episodes from 2007 to 2010, characterized with anomalous temperature rise and rainfall deficit, impacting crop phenology, soil moisture, water supply, and even changes in the distribution of pests and plant diseases. Herein, we should stress that although crop failure can not only be attributed to rainfall deficit (as represented by SPI), mainly because crop failure can be impacted by other important hydroclimatic and anthropogenic variables. This may include heat and energy fluxes and human activities. In general, the arid and semiarid regions are typically characterized by higher temperatures and strong evaporative demand, which could inevitably aggravate soil moisture shortage. According to Vicente-Serrano et al. (2020), the role of AED in drought severity is more in dry regions characterized by low precipitation and soil moisture deficit. Albeit with this possible impact, the temporal variability of drought is mostly driven by precipitation, which is undoubtedly the main controller of drought variability and accordingly crop failure. In this context, our study illustrates the degree to which agriculture sector is sensitive to changes in the frequency, intensity or duration of meteorological drought. These impacts are expected to be more enhanced in water-scarce regions, like the mainland and northeastern portions of the country. Qatna (2010) summarized the direct impacts of 2007–2010 drought episode on local communities in Syria, indicating that more than 500,000 Syrian families working in agriculture and livestock sectors were impacted, accompanied by losses of 25–80% of the total animals in the country. El-Shaer et al. (1997) indicated that wheat crop losses, due to climate warming, are estimated at − 20% to − 50%, more pronounced in the southern and eastern Mediterranean. In Syria, the intensification of drought can indirectly influence the country's economy, as it can aggravate or create conflicts among different water-demanded sectors (e.g., grazing, industry, and tourism), posing further challenges to water management strategies and policies in the country. In the near future, Syria is likely to suffer from food insecurity situation, particularly with high rates of population

growth, agricultural land degradation due to climate change besides other anthropogenic effects (e.g., land use changes), trans-boundary water disputes (especially with Turkey), in addition to rapid global alterations in crop prices. All these challenges should also be seen in the context of the low adaptive capacity to the impacts of climate change in Syria, especially with poor infrastructure and current civil wars and political and socioeconomic instability. This situation can be complicated further in the future under enhanced future greenhouse gases, which are likely to occur in the Middle East (Evans 2009; Hoerling et al. 2012; Lelieveld et al. 2012, 2016). Hoerling et al. (2012) pointed out that a warming of 0.5 °C in SST over the Indian Ocean is projected to enhance drying over the eastern Mediterranean. Similarly, Lelieveld et al. (2012) indicated that future climate change over the eastern Mediterranean and the Middle East could have drastic impacts on agriculture in the region, especially in water-scarce regions like Egypt and Syria. Future enhanced warming and the projected rainfall deficit would increase irrigation demands and may cause a migration of traditional climatic zones.

Several attempts have been made to provide physical explanations of drought variability in the Mediterranean region in general and in the Fertile Crescent in particular (e.g., Teuling et al. 2013; Vicente-Serrano et al. 2014; Páscoa et al. 2017a, b). Drought intensification in the Mediterranean region is mostly linked to lack of precipitation, air temperature rise, and enhanced atmospheric evaporative demand. A representative example is Vicente-Serrano et al. (2014) who attributed drought severity over the Southern Mediterranean during the past decades to an increase of atmospheric evaporative demand, which was driven largely by temperature rise, especially during summertime. Similarly, Teuling et al. (2013) found a significant association between evapotranspiration and drought episodes in the Western Europe. In the Fertile Crescent, Salah et al. (2018) linked drought episodes with changes in storm track activities over the eastern Mediterranean. This finding concurs with Philandras et al. (2011) who found that the strong dipole of the NAO phase controls storm tracks over the Mediterranean, favoring for drier conditions in the Middle East. Similarly, Páscoa et al. (2017a, b) associated drier conditions over the Mediterranean with SST warming, in addition to other external forces like aerosols. While understanding the physical mechanisms of drought variability in Syria is beyond the scope of this study. It can be emphasized that climatic variability and change has a major impact on drought evolution in Syria. However, it is also noteworthy to indicate that anthropogenic forces, e.g., land cover/use changes and water regulation, could—to some extent—accelerate drought severity and intensify its negative impacts on local communities. In Syria, the Mediterranean forests have largely been impacted by rapid and strong land cover changes during the past decades, in response to rapid urbanization rates (Ozturk et al. 2008).

Recalling the current drought impacts, shrubs and forests witnessed a noticeable decline, especially close to the Mediterranean coast, where many large cities are located (e.g., Tartus, Lattakia, Idleb, and Hamah). Other factors may also include water management practices, particularly with the dominance of the traditional agriculture in Syria. Overall, while the statistical results (e.g., PCA, Pearson's correlation) confirm the intensification of meteorological droughts in Syria in the past few decades with drastic impacts on vegetation cover and crop yield, further efforts are needed to confirm to explain and confirm these impacts from a physical perspective. This can be an avenue for a future research.

Herein, it is noteworthy indicating that the results of this work should be seen in the context of the characteristics of the climatic dataset employed in this work, which still shows specific limitations in terms of spatial density and temporal coverage. Unfortunately, this uneven distribution of climate records is a typical characteristic of many climatic datasets in developing countries, including Syria. Nonetheless, in the meantime, the agreement between our findings and those of other long-term regional drought assessments in the Fertile Crescent and the Middle East region (e.g., Iraq, Lebanon, Jordan, Oman, Turkey and Iran) suggests that recent drought episodes over Syria is likely more linked to climate signals rather than natural climate variability induced by changes in large-scale atmospheric and oceanic circulation patterns (e.g., ENSO and NAO). However, this study also highlights the need for a dense network of meteorological records at high temporal resolution, which is desired for a multidecadal assessment of drought variability in the Fertile Crescent region in general and Syria in particular.

Concluding remarks

This work analyzed spatial and temporal characteristics of meteorological drought over Syria from 1990 to 2010 using a relatively dense network ($N = 36$) of observational rainfall data. SPI was employed to diagnosis drought at 12-month timescale. Drought characteristics were linked to changes in vegetation cover, as retrieved from the NOAA/AVHRR data. Some available sectorial data, mainly for cereal crops and grazing flowers, were also linked to drought variability. The key findings of this study can be summarized, as follows:

- Syria witnessed a statistically significant increase in the frequency and intensity of meteorological drought over the study period, with two major drought episodes (1999–2001 and 2007–2010).
- The increase in meteorological drought severity is mostly linked to the increase in mild and moderate droughts,

with less impact from extreme and very extreme droughts.

- Spatially, northeastern parts of the country, the mainland and the Mediterranean region were more prone to drought impacts, with less influence in northern and southern portions of the country.
- Meteorological drought variability showed strong influence on vegetation cover, especially the Mediterranean forests and shrubs. This influence was more pronounced during the rainy season, especially in December, January, April and May.
- Drought had drastic impacts on the agricultural sector in Syria during the 2007–2010 drought episode, with a significant crop failure (e.g., wheat, barley).
- Although there is a widespread global debate regarding the political impacts of recent climate change, the Syrian conflict may be partly linked to climate change. The mega-drought episode that hit Syria from 2007 to 2010 may be seen as a contributing factor to this political crisis, especially with its drastic impacts on crop failure and water resources availability.

Overall, this study stresses that drought evolution over Syria is highly variable over space, suggesting that future drought management and adaptation plans should be made at local and sub-regional scales rather than at a national scale (especially for water resources management and food security).

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abdo HG (2018) Impacts of war in Syria on vegetation dynamics and erosion risks in Safita area, Tartous, Syria. Region Environ Change 18:1707–1719

Abu Hajar HA, Murad YZ, Shatanawi KM, Al-Smadi BM, Abu Hajar YA (2019) Drought assessment and monitoring in Jordan using the standardized precipitation index. Arab J Geosci 12:417

Agnew CT (2000) Using the SPI to identify drought. Drought Network News 12:6–12

Ahlström A, Raupach MR, Schurgers G, Smith B, Arneth A, Jung M et al (2015) The dominant role of semi-arid ecosystems in the trend and variability of the land CO2 sink. Science 348:895–899

Aladaileh H, Al Qinna M, Karoly B, Al-Karablieh E, Rakonczai J (2019) An investigation into the spatial and temporal variability of the meteorological drought in Jordan. Climate 7:82

Al-Bakri JT, Suleiman AS (2004) NDVI response to rainfall in different ecological zones in Jordan. Int J Remote Sens 25:3897–3912

Al-Qinna MI, Hammouri NA, Obeidat MM, Ahmad FY (2011) Drought analysis in Jordan under current and future climates. Clim Change 106:421–440

Alsafadi K, Mohammed SA, Ayugi B, Sharaf M, Harsányi E (2020) Spatial-temporal evolution of drought characteristics over Hungary between 1961 and 2010. Pure appl Geophys 1:1–18

Al-Youssef A, Shaaban A, Suliman A, Mazid A, Naal Y, Khoja S (2016) Economics of wheat production in the work area of the project of enhancing food security in Arab Countries/Syria. Syrian J Agric Res 3(2):75–87 **(in Arabic)**

Asadi Zarch MA, Sivakumar B, Sharma A (2015) Droughts in a warming climate: a global assessment of Standardized precipitation index (SPI) and Reconnaissance drought index (RDI). J Hydrol 526:183–195

Azorin-Molina C, Vicente-Serrano SM, Sanchez-Lorenzo A, McVicar TR, Morán-Tejeda E, Revuelto J et al (2015) Atmospheric evaporative demand observations, estimates and driving factors in Spain (1961–2011). J Hydrol 523:262–277

Barlow M, Zaitchik B, Paz S, Black E, Evans J, Hoell A (2016) A review of drought in the middle east and southwest Asia. J Clim 29:8547–8574

Breshears DD, Cobb NS, Rich PM, Price KP, Allen CD, Balice RG et al (2005) Regional vegetation die-off in response to global-change-type drought. Proc Natl Acad Sci USA 102:15144–15148

Butler CD (2018) Climate change, health and existential risks to civilization: a comprehensive review (1989–2013). Int J Environ Res Public Health 15:2266

Buttafuoco G, Caloiero T, Coscarelli R (2015) Analyses of drought events in Calabria (Southern Italy) using standardized precipitation index. Water Resour Manage 29:557–573

Cai W, van Rensch P, Borlace S, Cowan T (2011) Does the Southern Annular Mode contribute to the persistence of the multidecade-long drought over southwest Western Australia? Geophys Res Lett 1:38

Cai W, Purich A, Cowan T, Pv Rensch, Weller E (2014) Did climate change-induced rainfall trends contribute to the Australian millennium drought? J Clim 27:3145–3168

Caloiero T (2018) SPI trend analysis of New Zealand applying the ITA technique. Geosciences 8:101

Caloiero T, Veltri S, Caloiero P, Frustaci F (2018a) Drought analysis in europe and in the mediterranean basin using the standardized precipitation index. Water 10:1043

Caloiero T, Sirangelo B, Coscarelli R, Ferrari E (2018b) Occurrence probabilities of wet and dry periods in southern Italy through the SPI evaluated on synthetic monthly precipitation series. Water 10:336

Cammalleri C, Barbosa P, Vogt JV (2019) Analysing the relationship between multiple-timescale SPI and GRACE terrestrial water storage in the framework of drought monitoring. Water 11(8):1672

Cook BI, Anchukaitis KJ, Touchan R, Meko DM, Cook ER (2016) Spatiotemporal drought variability in the Mediterranean over the last 900 years. J Geophys Res Atmos 121:2060–2074

Dabanli İ, Mishra AK, Şen Z (2017) Long-term spatio-temporal drought variability in Turkey. J Hydrol 552:779–792

De Châtel F (2014) The role of drought and climate change in the syrian uprising: untangling the triggers of the revolution. Middle Eastern Studies 50:521–535

De Pauw E (2005) Monitoring agricultural drought in the near east. In: Boken VK, Cracknell AP, Heathcote RL (eds) Monitoring and predicting agricultural drought. Oxford University Press, New York

Deng Y, Wang S, Bai X, Tian Y, Wu L, Xiao J et al (2018) Relationship among land surface temperature and LUCC, NDVI in typical karst area. Sci Rep 8:641

Didan K, Munoz AB, Solano R, Huete A (2015) MODIS vegetation index user's guide (MOD13 Series). University of Arizona, Vegetation Index and Phenology Lab

Domínguez-Castro F, García-Herrera R, Ribera P, Barriendos M (2010) A shift in the spatial pattern of Iberian droughts during the 17th century. Clim Past 6:553–563

Drumond A, Nieto R, Hernandez E, Gimeno L (2011) A Lagrangian analysis of the variation in moisture sources related to drier and wetter conditions in regions around the Mediterranean Basin. Nat Hazards Earth Syst Sci 11:2307–2320

Dutra E, Pozzi W, Wetterhall F, Di Giuseppe F, Magnusson L, Naumann G et al (2014) Global meteorological drought—part 2: seasonal forecasts. Hydrol Earth Syst Sci 18:2669–2678

Eklund L, Thompson D (2017) Differences in resource management affects drought vulnerability across the borders between Iraq, Syria, and Turkey. Ecology and Society 22(4):1

El Kenawy A, López-Moreno JI, Vicente-Serrano SM (2013) Summer temperature extremes in northeastern Spain: spatial regionalization and links to atmospheric circulation (1960–2006). Theoret Appl Climatol 113:387–405

El Kenawy AM, McCabe MF, Vicente-Serrano SM, Robaa SM, López-Moreno JI (2016a) Recent changes in continentality and aridity conditions over the Middle East and North Africa region, and their association with circulation patterns. Climate Res 69:25–43

El Kenawy AM, McCabe MF, Vicente-Serrano SM, López-Moreno JI, Robaa SM (2016b) Changes in the frequency and severity of meteorological droughts over Ethiopia from 1960 to 2013. Cuadernos De Investigación Geográfica 42(2):1

El Kenawy AM, Al Buloshi A, Al-Awadhi T, Al Nasiri N, Navarro-Serrano F et al (2020) Evidence for intensification of meteorological droughts in Oman over the past four decades. Atmos Res 246:105126

Esau I et al (2016) Trends in normalized difference vegetation index (NDVI) associated with urban development in northern West Siberia. Atmos Chem Phys 16:9563–9577

Evans JP (2009) 21st century climate change in the Middle East. Clim Change 92(3–4):417–432. https://doi.org/10.1007/s10584-008-9438-5

Fern RR et al (2018) Suitability of NDVI and OSAVI as estimators of green biomass and coverage in a semi-arid rangeland. Ecol Ind 94:16–21

Gao X, Giorgi F (2008) Increased aridity in the Mediterranean region under greenhouse gas forcing estimated from high resolution simulations with a regional climate model. Global Planet Change 62:195–209

Gilmore EA, Herzer Risi L, Tennant E, Buhaug H (2018) Bridging research and policy on climate change and conflict. Current Climate Change Reports 4:313–319

Giorgi F (2006) Climate change hot-spots. Geophys Res Lett 33:1

Giorgi F, Lionello P (2008) Climate change projections for the Mediterranean region. Global Planet Change 63:90–104

Gleick PH (2014) Water, drought, climate change, and conflict in Syria. Weather, Climate, and Society 6:331–340

Golian S, Mazdiyasni O, AghaKouchak A (2015) Trends in meteorological and agricultural droughts in Iran. Theoret Appl Climatol 119:679–688

González-Hidalgo JC, Vicente-Serrano SM, Peña-Angulo D, Salinas C, Tomas-Burguera M, Beguería S (2018) High-resolution spatio-temporal analyses of drought episodes in the western Mediterranean basin (Spanish mainland, Iberian Peninsula). Acta Geophys 66:381–392

Gouveia CM, Trigo RM, Beguería S, Vicente-Serrano SM (2017) Drought impacts on vegetation activity in the Mediterranean region: an assessment using remote sensing data and multi-scale drought indicators. Global Planet Change 151:15–27

Gumus V, Algin HM (2017) Meteorological and meteorological drought analysis of the Seyhan—Ceyhan River Basins. Turkey. Met. Apps 24:62–73

Haensel S, Zurba K (2015) Precipitation characteristics and trends in the Palestinian territories during the period 1951–2010. FOG-FREIBERG Online Geosci. 39:103–130

Hamed KH, Ramachandra Rao A (1998) A modified Mann-Kendall trend test for autocorrelated data. J Hydrol 204:182–196

Hameed M, Ahmadalipour A, Moradkhani H (2018) Apprehensive drought characteristics over Iraq: results of a multidecadal spatiotemporal assessment. Geosciences 8:58

Hayes M, Svoboda M, Wall N, Widhalm M (2011) The lincoln declaration on drought indices: universal meteorological drought index recommended. Bull Am Meteor Soc 92(4):485–488

Hoerling M, Eischeid J, Perlwitz J, Quan X, Zhang T, Pegion P (2012) On the increased frequency of mediterranean drought. J Clim 25:2146–2161

Hollinger SE, Isard SA, Welford MR (1993) A new soil moisture drought index for predicting crop yields. In: Preprints, Eighth Conference on Applied Climatology, Anaheim, CA, Am Meteor Soc, pp 187–190

Ide T (2018) Climate war in the middle east? Drought, the Syrian civil war and the state of climate-conflict research. Current Climate Change Reports 4:347–354

IPCC (2013) Climate change 2013: the physical science basis. In: Stocker TF et al. (ed) Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, p 1535

Jain VK, Pandey RP, Jain MK, Byun H-R (2015) Comparison of drought indices for appraisal of drought characteristics in the Ken River Basin. Weather Climate Extremes 8:1–11

Jr RRH (2002) A review of twentieth-century drought indices used in the United States. Bull Am Meteor Soc 83:1149–1166

Kaiser HF (1991) Coefficient alpha for a principal component and the Kaiser-Guttman Rule. Psychol Rep 68:855–858

Kelley CP, Mohtadi S, Cane MA, Seager R, Kushnir Y (2015) Climate change in the Fertile Crescent and implications of the recent Syrian drought. Proc Natl Acad Sci 112:3241–3246

Khayyati M (2016) Drought impact assessment on rural livelihood systems in Iran. Ecol Ind 69:850–858

Kostopoulou E, Giannakopoulos C, Krapsiti D, Karali A (2017) Temporal and spatial trends of the standardized precipitation index (SPI) in Greece using observations and output from regional climate models. Springer, Cham, pp 475–481

Kubiak-Wójcicka K, Bąk B (2018) Monitoring of meteorological and meteorological droughts in the Vistula basin (Poland). Environ Monit Assess 190(11):691

Lelieveld J, Hadjinicolaou P, Kostopoulou E, Chenoweth J, El Maayar M, Giannakopoulos C et al (2012) Climate change and impacts in the Eastern Mediterranean and the Middle East. Clim Change 114:667–687

Li W, Fu R, Juárez RIN, Fernandes K (2008) Observed change of the standardized precipitation index, its potential cause and implications to future climate change in the Amazon region. Philos Trans R Soc Lond B Biol Sci 363:1767–1772

Liu X, Wang S, Zhou Y, Wang F, Li W, Liu W (2015) Regionalization and Spatiotemporal Variation of Drought in China Based on Standardized Precipitation Evapotranspiration Index (1961–2013). Adv Meteorol 2015:18

Ljungqvist FC, Krusic PJ, Sundqvist HS, Zorita E, Brattström G, Frank D (2016) Northern Hemisphere hydroclimate variability over the past twelve centuries. Nature 532:94

Lu D (2006) The potential and challenge of remote sensing-based biomass estimation. Int J Remote Sens 27(7):1297–1328. https://doi.org/10.1080/01431160500486732

Mahfouz P, Mitri G, Jazi M, Karam F (2016) Investigating the temporal variability of the standardized precipitation index in Lebanon. Climate 4:27

Manning JG, Ludlow F, Stine AR, Boos WR, Sigl M, Marlon JR (2017) Volcanic suppression of Nile summer flooding triggers revolt and constrains interstate conflict in ancient Egypt. Nat Commun 8:900

Mariotti A, Dell'Aquila A (2012) Decadal climate variability in the Mediterranean region: roles of large-scale forcings and regional processes. Clim Dyn 38:1129–1145

Mathbout S, Lopez-Bustins JA, Martin-Vide J, Bech J, Rodrigo FS (2018) Spatial and temporal analysis of drought variability at several time scales in Syria during 1961–2012. Atmos Res 200:153–168

McKee TB, Doesken NJ, Kleist J (1993) The relationship of drought frequency and duration to time scales. In: Proceedings of the 8th conference on applied climatology, American Meteorological Society, pp 179–184

Merabti A, Martins DS, Meddi M et al (2018) Spatial and time variability of drought based on SPI and RDI with various time scales. Water Resour Manage 32:1087–1100. https://doi.org/10.1007/s11269-017-1856-6

Mishra AK, Singh VP (2010) A review of drought concepts. J Hydrol 391:202–216

Mo KC, Lyon B (2015) Global meteorological drought prediction using the North American Multi-Model ensemble. J Hydrometeorol 16:1409–1424

Mohammed SA, Fallah RQ (2019) Climate change indicators in Alsheikh-Badr Basin (Syria). Geogr Environ Sustainabil 12(2):87–96

Mohammed SA, Harsányi E (2019) Drought cycle tracking in Hungary using Standardized Precipitation Index (SPI). Acta Agraria Debreceniensis 2:97–101

Mohammed SA, Alkerdi A, Harsányi E, János N (2019a) Syrian crisis repercussions on the agricultural sector: case study of wheat, cotton and olives. Region Sci Policy Practice 1:1. https://doi.org/10.1111/rsp3.12222

Mohammed S, Alsafadi K, Mohammad S, Mousavi N (2019). Drought trends in Syria from 1900 to 2015. In: Proceedings of the 4th international congress of developing agriculture, natural resources, environment and tourism of Iran, Tabriz Islamic Art University In cooperation with Shiraz University and Yasouj University, Tabriz, Iran, vol 14

Mohammed S, Alsafadi K, Daher H, Gombos B, Mahmood S, Harsányi E (2020a) Precipitation pattern changes and response of vegetation to drought variability in the eastern Hungary. Bull Natl Res Centre 44:1–10

Mohammed S, Alsafadi K, Talukdar S, Kiwan S, Hennawi S, Alshiehabi O et al (2020b) Estimation of soil erosion risk in southern part of Syria by using RUSLE integrating geo informatics approach. Rem Sens Appl Soc Environ 1:100375

Mohammed S, Khallouf A, Kiwan S, Alhenawi S, Ali H, Harsányi E et al (2020c) Characterization of major soil orders in Syria. Euras Soil Sci 53:420–429

Moradi HR, Rajabi M, Faragzadeh M (2011) Investigation of meteorological drought characteristics in Fars province, Iran. CATENA 84:35–46

Mukherjee A, Wang S-YS, Promchote P (2019) Examination of the climate factors that reduced wheat yield in Northwest India during the 2000s. Water 11:343

Mustafa A, Rahman G (2018) Assessing the spatio-temporal variability of meteorological drought in Jordan. Earth Syst Environ 2:247–264

Myneni RB, Hall F, Sellers P, Marshak A (1995) The interpretation of spectral vegetation indexes. IEEE Trans Geosci Remote Sens 33:481–486

Nalbantis I, Tsakiris G (2009) Assessment of meteorological drought revisited. Water Resour Manage 23:881–897

Naumann G, Alfieri L, Wyser K, Mentaschi L, Betts RA, Carrao H et al (2018) Global changes in drought conditions under different levels of warming. Geophys Res Lett 45:3285–3296

Nelson GC, Valin H, Sands RD, Havlík P, Ahammad H, Deryng D et al (2014) Climate change effects on agriculture: economic responses to biophysical shocks. Proc Natl Acad Sci 111:3274–3279

Nicault A, Alleaume S, Brewer S, Carrer M, Nola P, Guiot J (2008) Mediterranean drought fluctuation during the last 500 years based on tree-ring data. Clim Dyn 31:227–245

Orians CM, Schweiger R, Dukes JS, Scott ER, Müller C (2019) Combined impacts of prolonged drought and warming on plant size and foliar chemistry. Ann Bot 124:41–52

Ozturk M, Gucel S, Sakcali S, Gork C, Yarci C, Gork G (2008) An overview of plant diversity and land degradation interactions in the eastern mediterranean. Nat Environ Cult Mediterr Region 1:215–239

Palmer WC (1965) Meteorological drought. Research Paper No. 45, 1965, p 58, U.S. Dep. Commer. Weather Bur. Washington, DC. Research

Parece TE, Campbell JB (2017) Assessing urban community gardens' impact on net primary production using NDVI, Urban Agriculture & Regional Food Systems

Pascoa P, Gouveia CM, Russo A, Trigo RM (2017) Drought trends in the Iberian Peninsula over the last 112 years. Adv Meteorol 2017:13

Philandras CM, Nastos PT, Kapsomenakis J, Douvis KC, Tselioudis G, Zerefos CS (2011) Long term precipitation trends and variability within the Mediterranean region. Nat Hazards Earth Syst Sci 11(12):3235–3250. https://doi.org/10.5194/nhess-11-3235-2011

Qatna H (2010) The drought affected the north-eastern region of Syria between 2008–2009. Organization of Syrian economics science, Damascus, Syria (**In Arabic**)

Rashid MM, Beecham S (2019) Development of a non-stationary Standardized Precipitation Index and its application to a South Australian climate. Sci Tot Environ 657:882–892

Richman MB (1986) Rotation of principal components. J Climatol 6:293–335

Robaa S, Al-Barazanji Z (2015) Mann-Kendall trend analysis of surface air temperatures and rainfall in Iraq. Idojaras 119:493–514

Roy DP et al (2002) The MODIS Land product quality assessment approach. Remote Sens Environ 83:62–76

Ruffault J, Martin-StPaul NK, Duffet C, Goge F, Mouillot F (2014) Projecting future drought in Mediterranean forests: bias correction of climate models matters! Theoret Appl Climatol 117:113–122

Saglamoglu A, Irvem A (2020) Regional drought analysis for the eastern Mediterranean region of Turkey using run and SPI method. Fresenius Environ Bull 26:1256–1262

Sahoo AK, Sheffield J, Pan M, Wood EF (2015) Evaluation of the tropical rainfall measuring mission multi-satellite precipitation analysis (TMPA) for assessment of large-scale meteorological drought. Remote Sens Environ 159:181–193

Salah Z, Nieto R, Drumond A, Gimeno L, Vicente-Serrano SM (2018) A Lagrangian analysis of the moisture budget over the Fertile Crescent during two intense drought episodes. J Hydrol 560:382–395

Salameh A, Fallah R (2018) Changes in air temperature and precipitation over the Syrian Coastal Region. Cuadernos Geográficos 57(3):140–151

Schneider AW, Adalı SF (2014) "No harvest was reaped": demographic and climatic factors in the decline of the Neo-Assyrian Empire. Clim Change 127:435–446

Schubert SD, Stewart RE, Wang H, Barlow M, Berbery EH, Cai W et al (2016) Global meteorological drought: a synthesis of current understanding with a focus on SST drivers of precipitation deficits. J Clim 29:3989–4019

Seager R, Liu H, Henderson N, Simpson I, Kelley C, Shaw T et al (2014) Causes of increasing aridification of the mediterranean region in response to rising greenhouse gases. J Clim 27:4655–4676

Selby J, Dahi OS, Fröhlich C, Hulme M (2017) Climate change and the Syrian civil war revisited. Polit Geogr 60:232–244

Shafer BA, Dezman LE (1982) Development of a surface water supply index (SWSI) to assess the severity of drought conditions in snowpack runoff areas. In: Proceedings of the West. Snow Conference, pp 164–175

Shah R, Bharadiya N, Manekar V (2015) Drought index computation using standardized precipitation index (SPI) method for Surat District, Gujarat. Aquat Procedia 4:1243–1249

Sheffield J, Andreadis KM, Wood EF, Lettenmaier DP (2009) Global and continental drought in the second half of the Twentieth century: severity–area–duration analysis and temporal variability of large-scale events. J Clim 22:1962–1981

Shukla S, Wood AW (2008) Use of a standardized runoff index for characterizing hydrologic drought. Geophys Res Lett 35:1

Sobral BS, Oliveira-Júnior JFd, de Gois G, Pereira-Júnior ER, Terassi PMdB, Muniz-Júnior JGR et al (2019) Drought characterization for the state of Rio de Janeiro based on the annual SPI index: trends, statistical tests and its relation with ENSO. Atmos Res 220:141–154

Solh M, van Ginkel M (2014) Drought preparedness and drought mitigation in the developing world's drylands. Weather Clima Extremes 3:62–66

Spinoni J, Vogt JV, Naumann G, Barbosa P, Dosio A (2018) Will drought events become more frequent and severe in Europe? Int J Climatol 38:1718–1736

Spinoni J, Barbosa P, De Jager A, McCormick N, Naumann G, Vogt JV et al (2019) A new global database of meteorological drought events from 1951 to 2016. J Hydrol Region Stud 22:100593

Stahl K (2001) Meteorological drought—a study across Europe. Albert-Ludwigs Universität Freiburg, Freiburg

Stojanovic M, Drumond A, Nieto R, Gimeno L (2018) Variations in moisture supply from the mediterranean sea during meteorological drought episodes over Central Europe. Atmosphere 9:278

Tan C, Yang J, Li M (2015) Temporal-spatial variation of drought indicated by SPI and SPEI in Ningxia Hui Autonomous region. China. Atmosphere 6(10):1399–1421

Tanarhte M, Hadjinicolaou P, Lelieveld J (2012) Intercomparison of temperature and precipitation data sets based on observations in the Mediterranean and the Middle East. J Geophys Res Atmos 117(D12):1. https://doi.org/10.1029/2011JD017293

Terink W, Immerzeel WW, Droogers P (2013) Climate change projections of precipitation and reference evapotranspiration for the Middle East and Northern Africa until 2050. Int J Climatol 33:3055–3072

Teuling AJ, Van Loon AF, Seneviratne SI, Lehner I, Aubinet M, Heinesch B, Bernhofer C, Grünwald T, Prasse H, Spank U (2013) Evapotranspiration amplifies European summer drought. Geophys Res Lett 40:2071–2075. https://doi.org/10.1002/grl.50495

Tucker CJ (1979) Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ 8:127–150

Turco M, von Hardenberg J, AghaKouchak A, Llasat MC, Provenzale A, Trigo RM (2017) On the key role of droughts in the dynamics of summer fires in Mediterranean Europe. Sci Rep 7:81

Van Lanen HAJ, Van Loon AF (2012) A process-based typology of hydrological drought. Hydrol Earth Syst Sci 16:1915–1946. https://doi.org/10.5194/hess-16-1915-2012

Van Loon AF (2015) Hydrological drought explained. WIREs. Water 2:359–392. https://doi.org/10.1002/wat2.1085

van Rooy MP (1965) A rainfall anomaly index independent of time and space. Notos 14:43–48

Vasiliades L, Loukas A, Liberis N (2011) A water balance derived drought index for pinios river Basin, Greece. Water Resourc Manag 25:1087–1101

Vicente-Serrano SM (2006) Differences in spatial patterns of drought on different time scales: an analysis of the Iberian Peninsula. Water Resour Manage 20:37–60

Vicente-Serrano SM, Cuadrat-Prats JM (2007) Trends in drought intensity and variability in the middle Ebro valley (NE of the Iberian Peninsula) during the second half of the twentieth century. Theoret Appl Climatol 88:247–258

Vicente-Serrano SM, López-Moreno JI (2006) The influence of atmospheric circulation at different spatial scales on winter drought variability through a semi-arid climatic gradient in Northeast Spain. Int J Climatol 26:1427–1453

Vicente-Serrano SM, Beguería S, López-Moreno JI (2010) A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. J Clim 23:1696–1718

Vicente-Serrano SM, López-Moreno JI, Lorenzo-Lacruz J, Kenawy AE, Azorin-Molina C, Morán-Tejeda E et al (2011) The NAO impact on droughts in the mediterranean region. In: Vicente-Serrano SM, Trigo RM (eds) Hydrological, socioeconomic and ecological impacts of the North Atlantic oscillation in the mediterranean region. Springer, Dordrecht, pp 23–40

Vicente-Serrano SM, Lopez-Moreno JI, Beguería S, Lorenzo-Lacruz J, Sanchez-Lorenzo A, García-Ruiz JM, Azorin-Molina C, Morán-Tejeda E, Revuelto J, Trigo R, Coelho F, Espejo F (2014) Evidence of increasing drought severity caused by temperature rise in southern Europe. Environ Res Lett 9:1. https://doi.org/10.1088/1748-9326/9/4/044001

Vicente-Serrano SM, Azorin-Molina C, Peña-Gallardo M, Tomas-Burguera M, Domínguez-Castro F, Martín-Hernández N et al (2019) A high-resolution spatial assessment of the impacts of drought variability on vegetation activity in Spain from 1981 to 2015. Nat Hazards Earth Syst Sci 19:1189–1213

Vicente-Serrano SM, McVicar TR, Miralles DG, Yang Y, Tomas-Burguera M (2020) Unraveling the influence of atmospheric evaporative demand on drought and its response to climate change. Rev Clim Chang 11:1. https://doi.org/10.1002/wcc.632

Vidal JP, Martin E, Franchistéguy L, Habets F, Soubeyroux JM, Blanchard M et al (2010) Multilevel and multiscale drought reanalysis over France with the Safran-Isba-Modcou hydrometeorological suite. Hydrol Earth Syst Sci 14:459–478

Wang B, An S-I (2005) A method for detecting season-dependent modes of climate variability: S-EOF analysis. Geophys Res Lett 32

Wang Y et al (2019) Spatio-temporal analysis of phenology in Yangtze River Delta based on MODIS NDVI time series from 2001 to 2015. Front Earth Sci 13:92–110

Wilhite DA, Glantz MH (2009) Understanding: the drought phenomenon: the role of definitions. Water International 10:111–120

Winkler K, Gessner U, Hochschild V (2017) Identifying droughts affecting agriculture in Africa based on remote sensing time series between 2000–2016: rainfall anomalies and vegetation condition in the context of ENSO. Remote Sens 9:831

Woli P, Jones JW, Ingram KT, Fraisse CW (2012) Agricultural reference index for drought (ARID). Agron J 104:287–300

World Meteorological Organization (2012) Guidelines on the selection of Reference Climatological Stations (RCSs) from the existing Climatological Station Network. 1986. WMO/TD-No. 130. World Climate Programme, WMO, Geneva, Switzerland.World Meteorological Organization. Standardized Precipitation Index User Guide. (M. Svoboda, M. Hayes and D. Wood). WMO-No. 1090. Geneva. ISBN 978-92-63-11091-6. 2012, p 16

Xu C, Li Y, Hu J, Yang X, Sheng S, Liu M (2012) Evaluating the difference between the normalized difference vegetation index and net primary productivity as the indicators of vegetation vigor assessment at landscape scale. Environ Monit Assess 184:1275–1286

Xue J, Su (2017) Significant remote sensing vegetation indices: a review of developments and applications. J Sens 1353691

Zargar A, Sadiq R, Naser B, Khan FI (2011) A review of drought indices. Environ Rev 19:333–349. https://doi.org/10.1139/a11-013

Zhang Y et al (2019) Mapping annual forest cover by fusing PALSAR/PALSAR-2 and MODIS NDVI during 2007–2016. Rem Sens Environ 224:74–91