Auto-adjoint time domain elastic full waveform inversion: FWI accelerated with CUDA using adjoint sources calculated with automatic differential

Tianze Zhang, Kris Innanen, Daneil Trad, Jian Sun

ABSTRACT

We propose the Auto-adjoint time domain elastic full waveform inversion in this report, which is a FWI framework accelerated with CUDA using adjoint sources calculated with automatic differential method. In this FWI framework, the forward modeling and the adjoint modeling are accelerated by CUDA, and the adjoint sources are calculated by the automatic differential method. These two features allows us to perform time domain FWI with GPU acceleration and explore how different kinds of objective functions can influence the inversion results effectively. We study the objective function behavior for the ℓ-2 norm. ℓ1-norm, Global-correlation based, Envelope based, objective function, and ℓ-1 norm between the real and imaginary part of the synthetic data and the observed data (ℓ1 RI objective function). According to the numerical test we did in this paper, the ℓ1 RI objective function has better ability to tolerate the noise when poor initial model is used for inversion, compared with all the other objective functions we considered.

INTRODUCTION

Full-waveform inversion (FWI) is a challenging data-fitting procedure based on full-wavefield modeling to extract quantitative information from seismograms. For the gradient-based FWI, the most important components are the efficient forward and backward modeling methods, and the optimization methods to update the parameter models. The forward and backward modeling method could be achieved using the same finite difference method if we use self-adjoint wave equation. The gradient of the parameters can be calculated by using the zero-lag correlation between the forward modeling wavefields generated with the wavelet source, and the back propagated wavefield generated with the adjoint-sources Tarantola (1986). The utilization of the different kinds of objective functions accounts for using the corresponding kinds of adjoint-sources to perform the backpropagation resulting in different kinds of gradients for updating the parameter models. Thus, the study of objective functions is an important aspect of FWI.

The least-squares objective function remains the most commonly used criterion in FWI, although it theoretically suffers from poor robustness in the presence of large isolated and non-Gaussian errors Brossier et al. (2010). The ℓ-1 norm is not based on the Gaussian statistics in data space, and it was first introduced in the time domain with Tarantola (2005). Djikpéssé and Tarantola (1999) used the ℓ1-norm to successfully invert the field data in the Gulf of Mexico. Huber (1973) combines the advantages of the ℓ1 and ℓ2 norm to release the nonlinearity of the objective function. The small residuals would be updated according to the ℓ-2 norm, and the large residuals would be updated with the ℓ1- norm. Therefore, the hybrid objective function has the ability to tolerate the outliers' noise in data. The correlation-based Liu et al. (2017), instantaneous phase-based Choi and Alkhalifah (2015), Envelope based objective functions (Wu et al., 2014) are all introduced by researchers to overcome the difficulties of FWI when the initial models are poor, and there is noise present in the observed data. In this study, we will first give a brief introduction to the commonly used objective functions and then perform the numerical tests to perform the inversion with poor initial models and the occurrence of noise in the observed data.

OBJECTIVE FUNCTIONS

$\ell 2$ norm objective function

The $\ell 2$ norm objective function calculates the $\ell 2$ norm between the observed data $\mathbf{d}_{\text{obs}}(t)$, and the synthetic data $\mathbf{d}_{\text{syn}}(\mathbf{x}_r, t, \mathbf{m})$, where the $\mathbf{x}_r$ represents the position of the receivers, and $\mathbf{m}$ represents the model parameters.We can also consider this as the waveform difference (WD) objective function. t represents the discrete time. The mathematical expression of the $\ell 2$ norm is:

$$\phi(\mathbf{m}) = \frac{1}{2}\sum_{\mathbf{x_r}}\int_0^T ||\mathbf{d}_{\text{syn}}(\mathbf{x_r}, t, \mathbf{m}) - \mathbf{d}_{\text{obs}}(t)||^2 \text{dt} \tag{1}$$

The $\ell 2$ norm objective function is one of the most commonly used objective functions. The $\ell 2$ norm objective function is easy to calculate, and we assume that the data noise is Gaussian distributed when the $\ell 2$ norm objective function is used to calculate the data misfit function. This assumption is very hard to avoid due to the Central Limited Theorem.

$\ell 1$ norm objective function

The $\ell 1$ norm objective function calculates the absolute error between the observed data and the synthetic data. When using the $\ell 1$ norm objective function, we assume that the data noises are Laplacian distributed. The mathematical expression of the $\ell 1$ norm objective function is formulated as:

$$\phi(\mathbf{m}) = \sum_{\mathbf{x_r}}\int_0^T ||\mathbf{d}_{\text{syn}}(\mathbf{x_r}, t, \mathbf{m}) - \mathbf{d}_{\text{obs}}(t)|| \text{dt} \tag{2}$$

The data residuals are normalized according to their amplitudes, which gives clear insight into why the $\ell 1$ norm is expected to be less sensitive to large data residuals.

Huber norm objective function

Huber (1973) introduce the Huber norm by introducing a threshold that controls the transition between the $\ell 1$ norm and $\ell 2$ norm. The Huber loss function is strongly convex in a uniform neighbourhood of its minimum; at the boundary of this uniform neighbourhood, the Huber loss function has a differentiable extension to an affine function. These properties allow it to combine much of the sensitivity of the mean-unbiased, minimum-variance estimator of the mean (using the quadratic loss function) and the robustness of the median-unbiased estimator (using the absolute value function).

$$\phi(\mathbf{m}) = \begin{cases} \frac{1}{2}\sum\limits_{\mathbf{x_r}}\int_0^T ||\mathbf{d}_{\text{syn}}(\mathbf{x_r}, t, \mathbf{m}) - \mathbf{d}_{\text{obs}}(t)||^2 \text{dt}, & \text{if } |\alpha| \le \epsilon \\ \sum\limits_{\mathbf{x_r}}\int_0^T ||\mathbf{d}_{\text{syn}}(\mathbf{x_r}, t, \mathbf{m}) - \mathbf{d}_{\text{obs}}(t)|| \text{dt}, & \text{if } |\alpha| > \epsilon \end{cases}, \tag{3}$$

where α is the absolute difference between the obsvered data and the synthetic data.

Global correlation objective function

The global correlation norm for waveform inversion was used Sen and Stoffa (1991), and Stoffa and Sen (1991), who applied it to separate-source inversion with simulated annealing and genetic algorithms. Choi and Alkhalifah (2012) used the global correlation between observed and modelled data as an alternative objective function for the inversion of the marine streamer data. Through the global correlation norm, we measure the similarity between the observed and modelled data and update the velocity model in the direction of maximizing the similarity. The Global correlation objective function based objective function is formulated as:

$$\phi(\mathbf{m}) = -\sum_{\mathbf{x_r}} \int_0^T \frac{\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r}, t, \mathbf{m}) * \mathbf{d}_{\mathrm{obs}}(\mathbf{x_r}, t, \mathbf{m})}{\sqrt{E_{\mathrm{obs}} E_{\mathrm{syn}}}} \mathrm{dt}, \tag{4}$$

where the E_{obs} and E_{syn} are the energy of the observation data trace and the synthetic data-trace: $E_{\mathrm{obs}} = \int_0^T \mathbf{d}^2_{\mathrm{obs}}(\mathbf{x_r}, t, \mathbf{m})\mathrm{dt}$. $E_{\mathrm{syn}} = \int_0^T \mathbf{d}^2_{\mathrm{syn}}(\mathbf{x_r}, t, \mathbf{m})\mathrm{dt}$. Choi and Alkhalifah (2012) suggested that the global correlation-based objective function is theoretically the same as the least-square norm of the normalized wavefields.

Zero mean global correlation objective function

(Dong et al., 2020) propose Zero mean global correlation objective function (ZMGC) for the FWI, to tackle the situation when the observed data lack low-frequency components or when the estimation for the wavelet is incorrect, which can be considered as a variation of GC-based objective function. The formula of the ZMGC objective function is:

$$\phi(\mathbf{m}) = -\sum_{\mathbf{x_r}} \int_0^T \frac{(\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r}, t, \mathbf{m}) - \bar{d_{obs}}) * (\mathbf{d}_{\mathrm{obs}}(\mathbf{x_r}, t, \mathbf{m}) - \bar{d_{syn}})}{\sqrt{E_{\mathrm{robs}} E_{\mathrm{rsyn}}}} \mathrm{dt}. \tag{5}$$

$E_{\mathrm{robs}} = \int_0^T (\mathbf{d}_{\mathrm{obs}}(\mathbf{x_r}, t, \mathbf{m}) - \bar{\mathbf{d}_{\mathrm{obs}}})^2 \mathrm{dt}$. $E_{\mathrm{rsyn}} = \int_0^T (\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r}, t, \mathbf{m}) - \bar{\mathbf{d}_{\mathrm{syn}}})^2 \mathrm{dt}$. $\bar{d_{syn}}$ and d_{obs} are the mean value of the traces.

Envelope based objective function

Wu et al. (2014) found that the envelope fluctuation and decay of seismic records carry ultra-low-frequency (ULF, i.e., the frequency below the lowest frequency in the source spectrum) signals that can be used to estimate the long-wavelength velocity structure. We then developed envelope inversion for the recovery of low-wavenumber components of media (smooth background) so that the initial model dependence of waveform inversion can be reduced.

$$\phi(\mathbf{m}) = \frac{1}{2} \sum_{\mathbf{x_r}} \int_0^T ||\mathcal{A}_{\mathrm{syn}}(\mathbf{x_r}, t, \mathbf{m}) - \mathcal{A}_{\mathrm{obs}}(\mathbf{x_r}, t, \mathbf{m})||^2 \mathrm{dt}, \tag{6}$$

where $\mathcal{A}_{\mathrm{syn}}$ and $\mathcal{A}_{\mathrm{obs}}$ are the envelope of $\mathbf{d}_{\mathrm{syn}}$ and $\mathbf{d}_{\mathrm{obs}}$ are the envelope of the signal. The envelope of the signal can be calculated by using the Hilbert transform:

$$\mathcal{A}(t) = \sqrt{\mathbf{d}^2(t) + \mathbf{d}_H^2(t)}^{\,p}, \tag{7}$$

where $\mathbf{d}(t)$ is any arbitrary data vector which varies with time t, and $\mathbf{d}_H(t)$ is the Hilbert transform of data $\mathbf{d}$. $\mathcal{A}(t)$ is the envelope of the data $\mathbf{d}(t)$. p is the power for the envelope data, and it can be any positive number. According to Wu et al. (2014), $p = 2$ is a relatively good choice for large-scale seismic waveform inversion problems.

Milti-scale Z transform objective function

The multi-scale Z transform objective function is formulated as:

$$\phi(\mathbf{m}) = \frac{1}{2} \sum_{\mathbf{x_r}} \int_0^{\omega_{max}} ||\Re(\mathcal{Z}_{\mathrm{syn}}) - \Re(\mathcal{Z}_{\mathrm{obs}})||^2 \mathrm{d}\omega \tag{8}$$

where $\mathcal{Z}$ is the discrete Z transform of the data, and $\Re$ means to take the real part of the transformed data. The discrete $\mathcal{Z}$ transform of the and time series data $\mathbf{d} = [d(0), \ldots d(N)]$, where the N is the length of the data vector $\mathbf{d}$, can be calculated with:

$$\mathcal{Z} = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & {z_d}^{-1} e^{-2\pi i/N} & {z_d}^{-2} e^{-4\pi i/N} & \ldots & {z_d}^{-N+1} e^{-2\pi i(N-1)/N} \\ 1 & {z_d}^{-1} e^{-4\pi i/N} & {z_d}^{-2} e^{-8\pi i/N} & \ldots & {z_d}^{-N+1} e^{-4\pi i(N-1)/N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & {z_d}^{-1} e^{-2\pi i(N-1)/N} & {z_d}^{-2} e^{-4\pi i(N-1)/N} & \ldots & {z_d}^{-N+1} e^{-2\pi i(N-1)^2/N} \end{bmatrix} \begin{bmatrix} d(0) \\ d(1) \\ d(2) \\ \vdots \\ d(N) \end{bmatrix}. \tag{9}$$

$z_d \geq 1$ is the damping factor. If $z_d = 1$, this operation calculates the cross-correlation of the seismic data with a series of different frequencies sine and cosine functions, which makes the $\mathcal{Z}$ the discrete Fourier transform. If $z_d > 1$, we are calculating the cross-correlation of the seismic data with a series of different frequencies sine and cosine functions with a damped amplitude which is the Laplacian transform. A larger value of z_d represents a larger damped property. By controlling the values of the damping factor z_d, we can achieve the multi-scale inversion strategy for full waveform inversion.

$\ell 1$ RI objective function

The motivation for introducing this objective function is to better take the phase information into the inversion to tackle the local minimum and the cycle skipping problem. The phase information is closely related to the kinematics property in the wavefield, so it has great potential to be used in FWI for overcoming the local minima problem. The phase of the observed data and the synthetic data can be calculated with the following equations:

$$\phi_{\mathrm{syn}} = \arctan \frac{\Re(\mathcal{F}_{\mathrm{syn}})}{\Im(\mathcal{F}_{syn})},\ \phi_{\mathrm{obs}} = \arctan \frac{\Re(\mathcal{F}_{\mathrm{obs}})}{\Im(\mathcal{F}_{obs})}, \tag{10}$$

where $\mathcal{F}_{obs}$ and $\mathcal{F}_{syn}$ are the Fast Fourier transform of the observed data and the synthetic data. $\Re$ means taking the real part of the data, and $\Im$ means taking the imaginary part of the data. However, the phase suffers from the wrapping problem, which is introduced by numerical calculation using equation 10. In order to reduce this problem; we try to avoid the arc-tangent calculation. Due to the fact that the arc-tangent function is a monotonically increasing function. Thus we propose the $\ell 1$ RI objective function:

$$\phi(\mathbf{m}) = \sum_{\mathbf{x_r}} \int_0^{\omega_{max}} ||\Re(\mathcal{F}_{\text{syn}}) - \Re(\mathcal{F}_{\text{obs}})|| + ||\Im(\mathcal{F}_{\text{syn}}) - \Im(\mathcal{F}_{\text{obs}})|| \mathrm{d}\omega. \tag{11}$$

We can see that the $\ell 1$ RI objective function calculates the $\ell 1$ norm distance between the real and the imaginary part of the Fourier transform between the synthetic data and the observed data. From the above equation, we can see that the calculation of the objective function performed in the frequency domain can calculate the $\ell 1$ norm distance, for the real and the imaginary part of the data, between the synthetic data and the observed data.

AUTO-ADJOINT EFWI FRAMEWORK

Figure 1 shows the inversion framework of the Auto-adjoint EFWI. F is the forward modelling operator. The forward modeling operator is achieved in the form of the recurrent neural network (RNN) accelerated by CUDA. $\mathrm{D_{syn}}$, and $\mathrm{D_{obs}}$ are the synthetic data and observed data respectively. OBJ is the objective function value calculated according to the objective function we choose. $\mathrm{ADengine}$ is the automatic differential engine which is used to calculate the adjoint sources for backpropagation. B stands for the backpropagation operator. which is accelerated by CUDA, which is also programmed in the form of the RNN. The only difference between the F and B are the source terms. OPT is the optimization method which gives the direction for the elastic models for model updating. And $\mathbf{m}$ is the updated model which can be used for the next iteration. The major difference between the Auto-adjoint EFWI and the conventional FWI AD engine, which allow us to efficiently formulate and test different kinds of objective functions performance for the FWI.

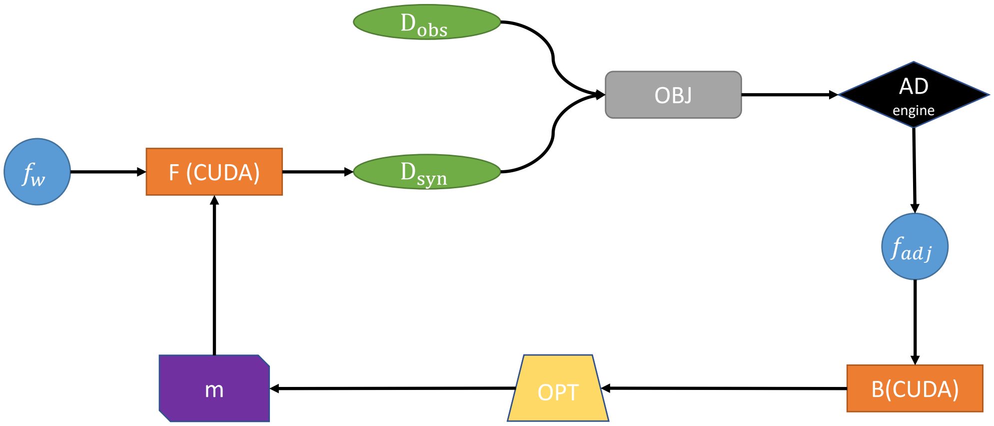

FIG. 1. The inversion framework of Auto-adjoint EFWI.

Objective function list

Name	Acronym	Expression $\mathcal{X}$
$\ell2$ norm	WD	$\frac{1}{2}\sum\limits_{\mathbf{x_r}}\int_0^T \|\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})-\mathbf{d}_{\mathrm{obs}}(t)\|^2\mathrm{d}t$
$\ell1$ norm	ABSWD	$\sum\limits_{\mathbf{x_r}}\int_0^T \|\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})-\mathbf{d}_{\mathrm{obs}}(\mathbf{x_r},t,\mathbf{m})\|\mathrm{d}t$
Huber norm	HWD	$\begin{cases}\frac{1}{2}\sum\limits_{\mathbf{x_r}}\int_0^T \|\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})-\mathbf{d}_{\mathrm{obs}}(t)\|^2\mathrm{d}t, & \text{if } \lvert\alpha\rvert\le\epsilon \\ \sum\limits_{\mathbf{x_r}}\int_0^T \|\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})-\mathbf{d}_{\mathrm{obs}}(\mathbf{x_r},t,\mathbf{m})\|\mathrm{d}t, & \text{if } \lvert\alpha\rvert>\epsilon\end{cases}$ α is the absolute difference between observed data and synthetic data. ϵ is a hyper-parameter threshold.
Global Correlation	GC	$-\sum\limits_{\mathbf{x_r}}\int_0^T \frac{\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})*\mathbf{d}_{\mathrm{obs}}(t)}{\sqrt{E_{\mathrm{obs}}E_{\mathrm{syn}}}}\mathrm{d}t$,where the E_{obs} and E_{syn} are the energy of the trace. $E_{\mathrm{obs}}=\int_0^T \mathbf{d}^2_{\mathrm{obs}}(t)\mathrm{d}t$. $E_{\mathrm{syn}}=\int_0^T \mathbf{d}^2_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})\mathrm{d}t$.
Zero mean Global Correlation	ZMGC	$-\sum\limits_{\mathbf{x_r}}\int_0^T \frac{(\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})-\bar{d}_{obs})*(\mathbf{d}_{\mathrm{obs}}(t)-\bar{d}_{syn})}{\sqrt{E_{\mathrm{robs}}E_{\mathrm{rsyn}}}}\mathrm{d}t$. $E_{\mathrm{robs}}=\int_0^T(\mathbf{d}_{\mathrm{obs}}(t)-\bar{d}_{\mathrm{obs}})^2\mathrm{d}t$. $E_{\mathrm{rsyn}}=\int_0^T(\mathbf{d}_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})-\bar{d}_{\mathrm{syn}})^2\mathrm{d}t$. $\bar{d}_{syn}$ and $\bar{d}_{obs}$ are the mean value of the traces.
$\ell1$ real and imaginary	$\ell1$ RI	$\sum\limits_{\mathbf{x_r}}\int_0^{\omega_{max}} \|\Re(\mathcal{F}_{\mathrm{syn}})-\Re(\mathcal{F}_{\mathrm{obs}})\|+\|\Im(\mathcal{F}_{\mathrm{syn}})-\Im(\mathcal{F}_{\mathrm{obs}})\|\mathrm{d}\omega$, where $\mathcal{F}_{\mathrm{syn}}$ and $\mathcal{F}_{\mathrm{obs}}$ are the FFT of $\mathbf{d}_{\mathrm{syn}}$ and $\mathbf{d}_{\mathrm{obs}}$. $\Re$ stands for real part. $\Im$ stands for imaginary part.
Envelope	EN	$\frac{1}{2}\sum\limits_{\mathbf{x_r}}\int_0^T \|\mathcal{A}_{\mathrm{syn}}(\mathbf{x_r},t,\mathbf{m})-\mathcal{A}_{\mathrm{obs}}(t)\|^2\mathrm{d}t$, where $\mathcal{A}_{\mathrm{syn}}$ and $\mathcal{A}_{\mathrm{obs}}$ are the envelope of $\mathbf{d}_{\mathrm{syn}}$ and $\mathbf{d}_{\mathrm{obs}}$.
Multi-scale Z transform	MZ	$\frac{1}{2}\sum\limits_{\mathbf{x_r}}\int_0^{\omega_{max}} \|\Re(\mathcal{Z}_{\mathrm{syn}})-\Re(\mathcal{Z}_{\mathrm{obs}})\|^2\mathrm{d}\omega$, where $\mathcal{Z}_{\mathrm{syn}}$ and $\mathcal{Z}_{\mathrm{obs}}$ are the discrete Z transform of $\mathbf{d}_{\mathrm{syn}}$ and $\mathbf{d}_{\mathrm{obs}}$ with damping coefficient $z_d>1$.

OBJECTIVE FUNCTION BEHAVIORS FOR FWI

In this section, we use the methodology introduced by Gholami et al. (2013) to display how the different kinds of objective functions vary with respect to parameters. We will display how the objective functions based on $\ell2$-norm, $\ell1$-norm, GC-based objective function, EN-based objective function, the $\ell1$ RI objective function, and the MZ objective function vary with the isotropic elastic media parameters V_p, V_s and ρ. The contour of the Huber norm objective function is very similar to the $\ell1$-norm, and the contour of ZMGC-based objective function is similar to the GC-based objective function, so the contour of these two objective functions is not displayed.

Figure 2 shows the contour plot of the objective function behaviours for the variation of V_p and ρ, when we regard V_s as the constant value. In this figure, ρ changes from $500kg/m^3$ to $2500kg/m^3$ and V_p changes from $2250km/s$ to $4500km/s$. Figure 2 (a)-(f) are the objective function for the WD, ABSWD, GC, $\ell 1$ RI, EN, MZ based objective function respectively. The center of the contour shows the global minimum of the inversion. We observe that the new objective function, the $\ell 1$ RI, correctly reflects the global minimum of the model. Also, we can see that all the objective functions show fewer local minimums. The objective function for the EN and MZ method is less steep compared with the others. The $\ell 1$ RI objective function has the steepest convergence rate, indicating that it could have a faster convergence rate for the V_p and ρ parameter when V_s is considered as constant. In Figure 2, the second row shows how the objective function changes when we fix V_p, V_s and changes the ρ. We observe that all the objective functions are very smooth in this example. The same phenomenon could be seen when we change the value of V_p and fix the value of ρ and V_s which is the third row of figure 2. Few local minimums can be observed on all of the contours, indicating that with the absence of V_s variation, the simultaneous inversion of the V_p and ρ shows less likely to be trapped into the local minima.

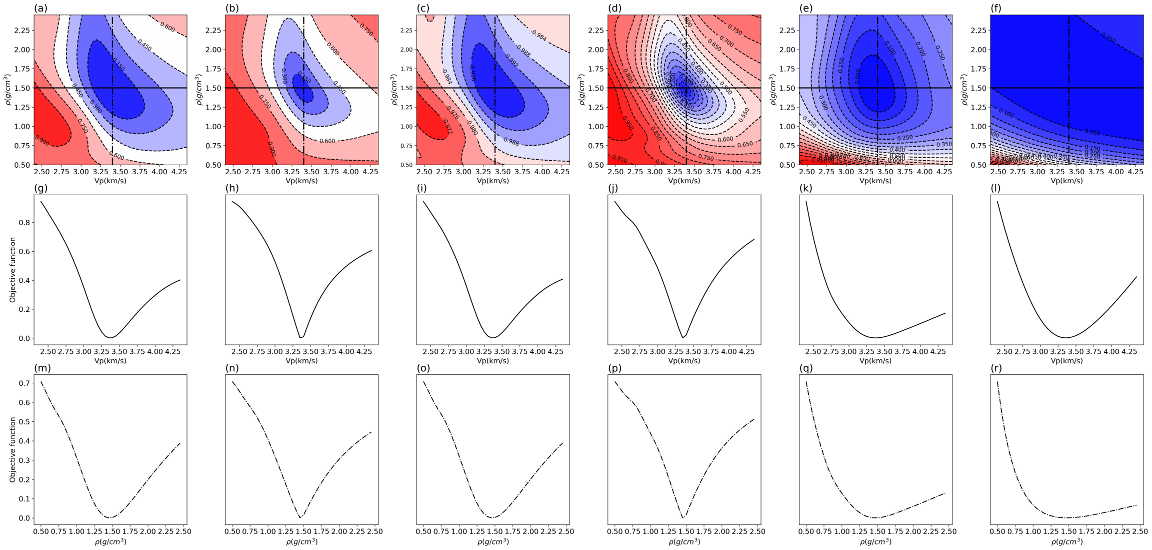

FIG. 2. The Contour of FWI different objective functions when the value of V_s is fixed ($1.35km/s$) and vary the value of V_p (from $2.25km/s$ to $4.5km/s$) and ρ (from $0.5g/cm^3$ to $2.5g/cm^3$). (a)-(f) are the 2D contour of the WD, ABSWD, GC, $\ell 1$ RI objective function, EN, and the MZ based objective function. (g)-(l) 1D profile of normalized different objective function when ρ and V_s are fixed and varies the value of V_p. (m)-(r) 1D profile of normalized different objective function when V_p and V_s are fixed and varies the value of ρ.

Figure 3 (a)-(f) illustrates the contour plot of how the objective functions of WD, ABSWD, GC, $\ell 1$ RI, EN, and MZ based objective function, when we fix the value of ρ and vary the value of V_p (from $2250m/s$ to $4500m/s$) and V_s (from $400m/s$ to $2500m/s$) in EFWI. The center of the contour shows the global minimum of the inversion. We can observe several local minimums around the global minimum in the contour map of WD, ABSWD, GC, and $\ell 1$ RI objective function, while no local minimum is observed in Figure 3 (e) and (f), indicating that the inversion using the EN and MZ objective function has the ability to release the local minimum problem in EFWI, as abundant low-frequency information is contained in the adjoint sources to generate gradient with large scale information.

When V_s is varying from $0.4km/s$ to $1.25km/s$, we observe few local minimums for the $\ell 1$ RI objective function compared with WD, ABSWD, GC based objective function, indicating that V_s could be better recovered by using the $\ell 1$ RI objective function. If we fix V_s and ρ and change the value of V_p, all the objective functions in the second row of Figure 3 are very smooth and have no local minimum. In Figure 3, the third row illustrates the 1D objective function variation when we fix V_p and ρ, and changes the value of V_s. We can see that the WD, ABSWD, GC, and $\ell 1$ RI-based objective function have several local minimums, while the value of EN and MZ objective function has few. It means that the V_s has much more influence on the presents of the local minimums in the objective functions compared with V_p.

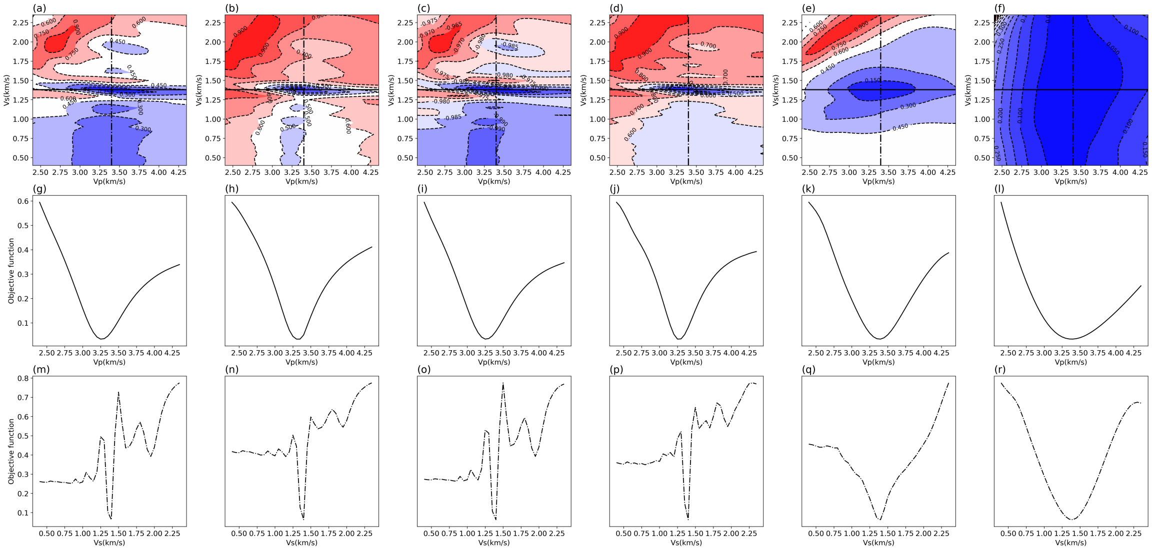

FIG. 3. The Contour of FWI different objective functions when the value of ρ is fixed $(1.5g/cm^3)$ and vary the value of V_s $(0.4km/s$ to $2.5km/s)$ and V_p (from $2.25km/s$ to $4.5km/s$). (a)-(f) are the 2D contour of the WD, ABSWD, GC, $\ell 1$ RI, EN, and the MZ objective function. (g)-(l) 1D profile of normalized different objective function when ρ and V_s are fixed and varies the value of V_p. (m)-(r) 1D profile of normalized different objective function when V_p and ρ are fixed and varies the value of ρ.

Figure 4 shows how the objective function responds to the variation of V_s and ρ if we regard V_p as the fixed value. We can see a lot of local minimums on the contour in the figures 4 (a), (b), and (c) for WD, ABSWD, GC, and $\ell 1$ RI as well. Also, the "true valley" in the center of the image in the contour is very narrow. This can make inversion wonder out of the true convergence region if we do not have a good starting point. However, compared the WD, ABSWD, and GC, the $\ell 1$ RI objective function has fewer local minimums when V_s varies from $0.4km/s$ to $1.25km/s$ indicating that $\ell 1$ RI objective function is less likely to be trapped into the local minimum during the update of the V_s. The contour given by the EN and MZ objective function shows a wider convergence region and showing no local minimum. The second row in Figure 4 shows how the objective functions respond to the variation of density when we fix V_s and V_p. It shows that there are no local minimums for all these objective functions. However, if we change the value of V_s and fix the V_p and density, only the MZ objective function contains no local minimums in the third columns of Figure 4. Thus, with all the discussions above, we can conclude that the variation of V_s

plays a much more critical role in the presence of the local minimums than other isotropic elastic parameters.

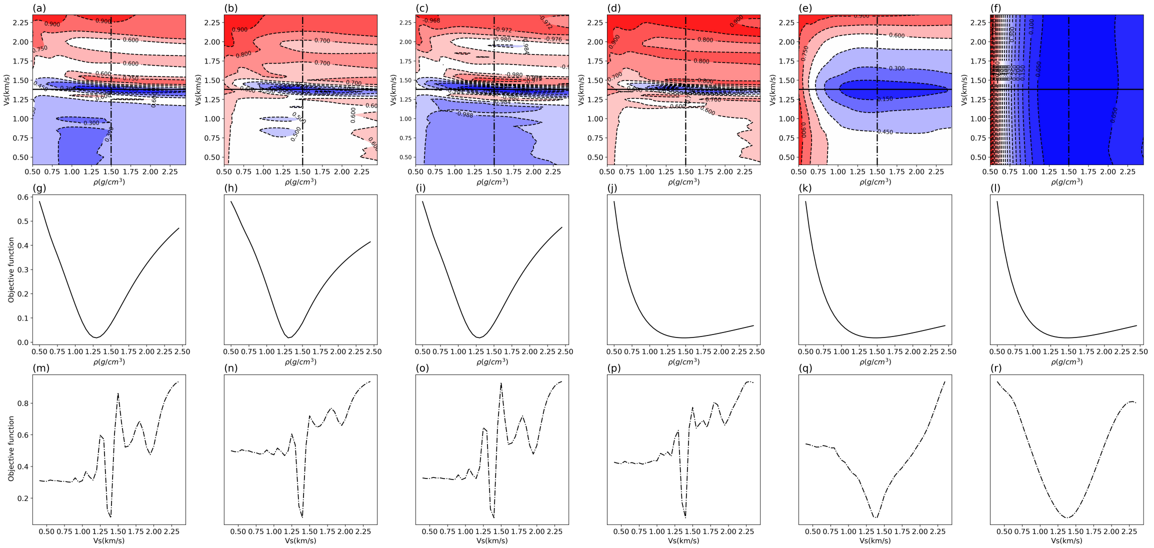

FIG. 4. The Contour of FWI different objective functions when the value of V_p is fixed ($3.3km/s$) and vary the value of V_s (from $0.4km/s$ to $2.5km/s$) and ρ from ($0.4g/cm^3$ to $2.5g/cm^3$). (a)-(f) are the 2D contour of the WD, ABSWD, $\ell 1$ RI, EN, and the MZ objective function. (g)-(l) 1D profile of normalized different objective function when V_p and V_s are fixed and varies the value of ρ. (m)-(f) 1D profile of normalized different objective function when V_p and ρ are fixed and varies the value of V_s.

OPTIMIZATION

gradient calculation

FWI can be considered as an iterative optimization method, which uses the first-order gradient information, or the second-order Hessian information to update the model parameter iteratively to obtain the inversion results. To obtain FWI results, we need to calculate the gradients first. According to Tarantola (1986); Köhn (2011), the sensitivity kernels, which is the partial derivative of the objective function with respect to the model parameters, for the density ρ, shear modulus μ, and bulk modulus κ, can be expressed as:

$$K_\rho(\mathbf{x}) = -\int_0^T \rho(\mathbf{x})\partial_t \mathbf{s}^\dagger(\mathbf{x}, T-t)\cdot\partial_t \mathbf{s}(\mathbf{x}, t)\mathrm{dt}, \tag{12}$$

$$K_\mu(\mathbf{x}) = -\int_0^T 2\mu(\mathbf{x})\partial_t \mathbf{D}^\dagger(\mathbf{x}, T-t)\cdot\partial_t \mathbf{D}(\mathbf{x}, t)\mathrm{dt}, \tag{13}$$

$$K_\kappa(\mathbf{x}) = -\int_0^T \kappa(\mathbf{x})[\nabla\cdot\mathbf{s}^\dagger(\mathbf{x}, T-t)][\nabla\cdot\mathbf{s}(\mathbf{x}, t)]\mathrm{dt}, \tag{14}$$

where the $\mathbf{s}(\mathbf{x}, t)$ is the forward modeling wavefield, and $\mathbf{s}^\dagger(\mathbf{x}, T-t)$ is the adjoint wavefield. $\mathbf{D} = \frac{1}{2}(\nabla\mathbf{s}+\nabla\mathbf{s})^\dagger - \frac{1}{3}(\nabla\cdot\mathbf{s})\mathbf{I}$ and $\mathbf{D}^\dagger$ are the traceless strain deviator and its adjoint, respectively. The adjoint wavefields are generated with the adjoint sources according to the objective function we use. In this study, we focus on the inversion for parameter V_P, V_S and density ρ. The sensitivity of the ρ is illustrated in equation 12, while the FWI sensitivity kernel for V_p, and V_s are:

$$K_{V_p}(\mathbf{x}) = 2\left(\frac{\kappa + 3/4\mu}{\kappa} K_\kappa(\mathbf{x})\right) \tag{15}$$

$$K_{V_s}(\mathbf{x}) = 2\left(K_\mu - \frac{4}{3}\frac{\mu}{\kappa} K_\kappa(\mathbf{x})\right) \tag{16}$$

Different parameterizations can be used to perform the full waveform inversion (Köhn, 2011; Pan et al., 2018). The gradient calculations of the parameters in different parameterizations can be calculated by using the Chain's Rule.

Gradient based method

After obtaining the gradient method, we use the following equation to update the elastic parameters:

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \Delta\mathbf{m}, \tag{17}$$

where the $\Delta\mathbf{m}$ is:

$$\Delta\mathbf{m} = -\eta\mathbf{p}, \tag{18}$$

where the $\mathbf{p}$ is the direction of the model update. There are several ways for us to calculate the direction of the model update. For instance, in the steepest decent optimization method, the direction $\mathbf{p}$ is the negative value of the gradient Tarantola (2005). The conjugate gradient method is also one of the most popular optimization methods (Mora, 1987; Tarantola, 1986; Crase et al., 1990), in which the direction $\mathbf{p}$ is the linear combination of the gradient at iteration k. Also, there are several ways of performing this linear combination, for instance, the Fletcher–Reeves formula (Fletcher and Reeves, 1964), Polak–Ribiére formula Polak and Ribiere (1969), Hestenes-Stiefel formula (Hestenes and Stiefel, 1952), the Dai-Yuan formula (Dai and Yuan, 1999). These formulas are equivalent to a quadratic function, but for nonlinear optimization, the preferred formula is a matter of heuristics or taste.

Newton based method

In the newton based optimization-based method, the update of the elastic model can be written as:

$$\Delta\mathbf{m} = -\eta\mathbf{H}^{-1}\mathbf{g}, \tag{19}$$

where $\mathbf{H}^{-1}$ is the inverse Hessian, which is very hard to calculate explicitly. Finite approximations of the Hessian and its inverse can be computed using quasi-Newton methods such as the BFGS algorithm named after its discoverers Broyden, Fletcher, Goldfarb, and Shanno. The main idea of this method is to update the approximation of the Hessian or its inverse Hessian at each iteration of the inversion, taking into account of the additional knowledge provided by the gradient. For large-scale problems such as FWI in which the cost of storing and working with the approximation of the Hessian matrix is prohibitive, a limited-memory variant of the quasi-Newton BFGS method known as the L-BFGS algorithm (Clark, 2010; Nocedal and Wright, 1999). The more accurate, although more computationally intensive, Gauss-Newton and Newton algorithms are described in Akcelik (2002); Askan et al. (2007); Askan and Bielak (2008).

Line search method

In each iteration of FWI, in order to properly update the elastic models in equation 18 and 19, we need to properly estimate the step length η to update the elastic models. According to Nocedal and Wright (1999), the Wolfe condition is one of the most commonly used line search conditions used for the nonlinear optimization method. The Wolfe condition is formulated as follows:

$$\begin{aligned} \phi(\mathbf{m}_k + \eta_k \mathbf{p}_k) &\leq \phi(m_k) + c_1 \eta_k \mathbf{g}_k^\top \mathbf{p}_k \\ \mathbf{g}(\mathbf{m}_k + \eta_k \mathbf{p}_k)^\top \mathbf{p}_k &\geq c_2 \mathbf{g}_k^\top \mathbf{p}_k \end{aligned} \tag{20}$$

with $0 < c_1 < c_2 < 1$. The value of $c_1 = 0.001$ and $c_2 = 0.999$ is suggested for the Newton based method. $c_1 = 0.001$ and $c_2 = 0.6$ is suggested for gradient based method. The first equation in equation 20 is also called the Armijo condition, which makes sure that the objective function must decline at a certain rate in each iteration. The second condition is known as the curvature condition, and together, they are the Wolfe condition. The line search method makes reasonable decisions for the step length, ensuring that the inversion converges at a certain rate for each iteration. However, extra forward modeling (at least one) and backpropagation (at least one) are needed to evaluate the proper step lengths. However, this is usually hard to avoid. For instance, the step length that satisfies the Wolfe condition is the necessary and sufficient condition for the conv

Gradient based method without line search

The optimization method we use in the following numerical tests is Adam's algorithm (Kingma et al., 2020). Adams's algorithm is suitable for full waveform inversion. First is in Adam's algorithm, the direction for each parameter is calculated through their gradient and momentum. When the inversion is trapped in the local minimum, when the update direction is calculated with the gradient only, as the gradient value at the local minimums is zero, the direction would be zero, which makes the inversion unable to escape from the local minimums. If the update direction is calculated with the gradient and the momentum in the local minimums, when the gradient is zeros, the momentum is unusually not zero, which could lead to the inversion out of the local minimums. The second advantage is related to inversion efficiency. In Adam's algorithm, the line search is not required. The step length is given manually before the beginning of the inversion. The step length used to update the model can decrease as the increase of iteration number.

NUMERICAL TESTS

In this section, we will use the Marmousi model to illustrate the validation of our proposed objective function. The grid length of the model is 30 × 30. The grid size of the model is 130×225. The source we use to generate the observed and synthetic data is Ricker's wavelet, with a listening time of 3 seconds, using time interval $dt = 0.002s$. We use the Ocean Bottom Cable (OBC) acquisition system to receive the seismic data, which means that all the sources are on the surface of the model, and the receivers are located on the bottom of the sea floor. In our elastic model, we have a water layer with a depth of $300m$, and ten layers of water are used as the water layer.

Numerical tests

In this section, we will test the performance of auto-adjoint EFWI using good initial models with a sufficient amount of low frequencies in the observed data. In this test, we will use the Gaussian smoothing method to obtain the initial models. We use the Gaussian smooth parameter $\sigma_G = 10$ to smooth the true models to obtain the initial models. The source wavelet used in this test is Ricker's wavelet with a main frequency 7Hz. We will use the different objective functions, WD, ABSWD, EN, GC and $\ell 1$ RI objective functions, to perform the FWI and compare the inversion results.

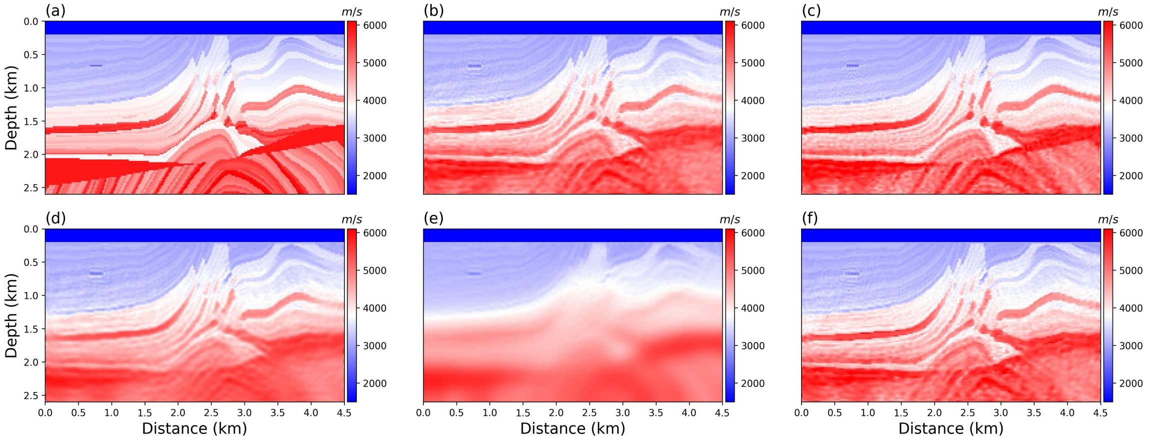

FIG. 5. Inversion results for V_p. (a) true V_p. (b) WD V_p inversion result. (c) ABSWD V_p inversion result. (d) EN V_p inversion result. (e) GC V_p inversion result. (e) $\ell 1$ RI V_p inversion result.

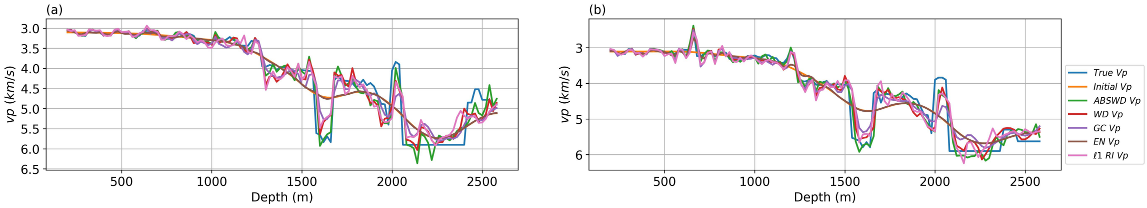

FIG. 6. Vertical inversion profiles for V_p. (a) Vertical inversion for V_p at 2000m in distance of the model. (b) Vertical inversion for V_p at 3000m in distance of the model.

Figure 5, 7, and 9 shows the inversion results for V_p, V_s, and ρ, respectively. In each of the figures above, (a) is the true elastic model. (b),(c),(d),(e) and (f) are the inversion results using WD, ABSWD, EN, GC and $\ell 1$ RI objective functions, respectively. The GC-based objective function failed to give detailed information about the deeper part of the model, but the shallower part of the model is correctly inverted. Comparing with the result in the deeper part of the model, which is (2.0km-2.6km) of the model. The $\ell 1$ norm and $\ell 1$ RI objective function can better recover the deeper part of the model. To further prove the validity of the out method. The vertical profile of the inversion results for V_p, V_s and ρ at 300m and 3500m in a distance of the models are also plotted in Figures 6, 8, and 10. We

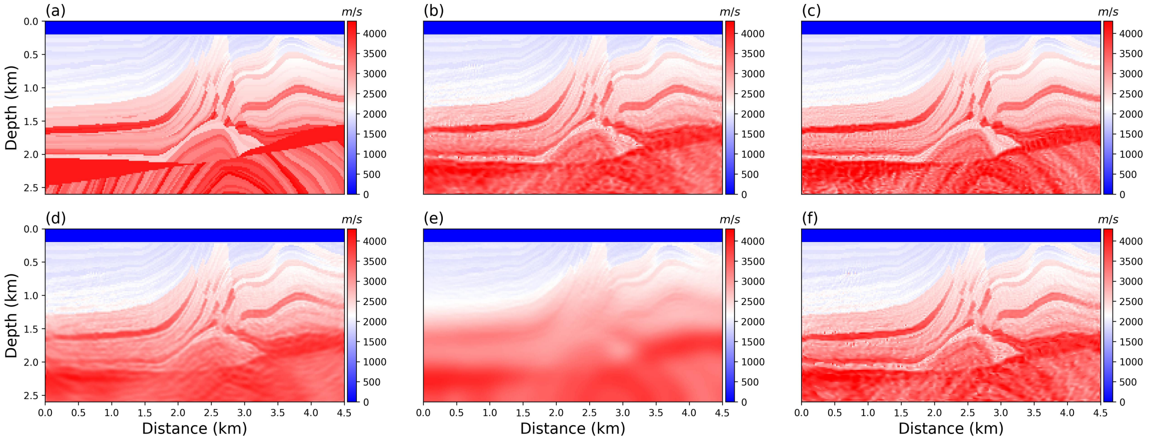

FIG. 7. Inversion results for V_s. (a) true V_s. (b) WD V_s inversion result. (c) ABSWD V_s inversion result. (d) EN V_s inversion result. (e) GC V_s inversion result. (e) $\ell 1$ RI V_s inversion result.

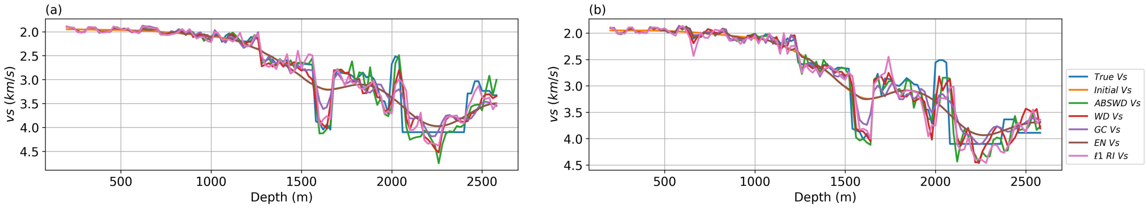

FIG. 8. Vertical inversion profiles for V_s. (a) Vertical inversion for V_s at 2000m in distance of the model. (b) Vertical inversion for V_s at 3000m in distance of the model.

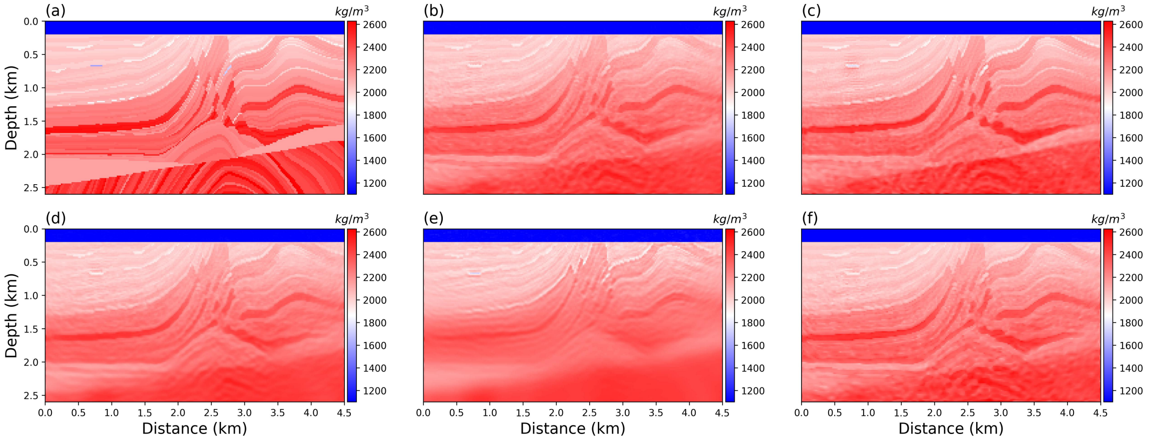

FIG. 9. Inversion results for ρ. (a) true ρ. (b) WD ρ inversion result. (c) ABSWD ρ inversion result. (d) EN ρ inversion result. (e) GC ρ inversion result. (e) $\ell 1$ RI ρ inversion result.

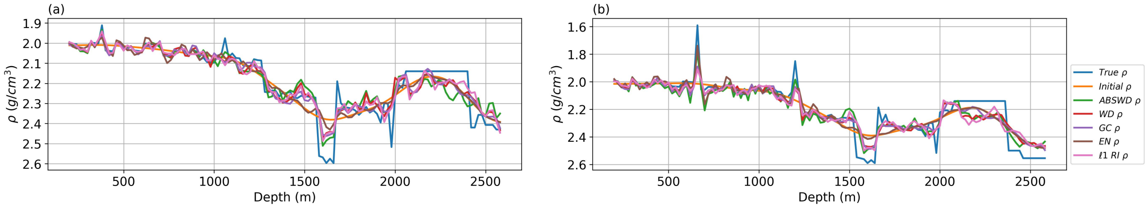

FIG. 10. Vertical inversion profiles for ρ. (a) Vertical inversion for ρ at 2000m in distance of the model. (b) Vertical inversion for ρ at 3000m in distance of the model.

can also see that the predictions for the shallower part of the models are correct, while the $\ell 1$ and the $\ell 1$ RI give a better estimation for the deeper part of the models.

Poor initial model test

In this section, we will use the poor initial models to perform the inversion to see the performance of different objective functions. We will use Gaussian smooth using parameter $\sigma_G = 20$ to smooth the true models to obtain the initial models. The larger σ_G are, the smoother the figure would be, and we obtain poorer initial models.

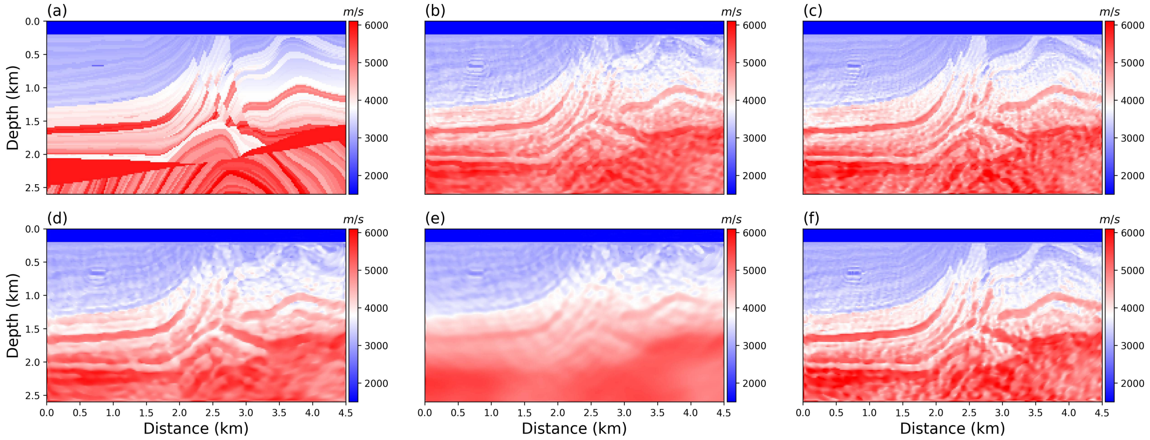

FIG. 11. Stress test inversion results for V_p. (a) true V_p. (b) WD V_p inversion result. (c) ABSWD V_p inversion result. (d) EN V_p inversion result. (e) GC V_p inversion result. (e) $\ell 1$ RI V_p inversion result.

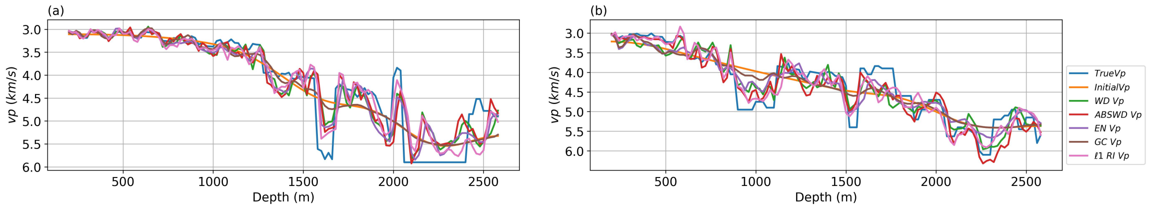

FIG. 12. Stress test inversion profiles for V_p. (a) Vertical inversion for V_p at 2000m in distance of the model. (b) Vertical inversion for V_p at 3000m in distance of the model.

Figures 11, 13, and 15 illustrates the inversion results using the poor initial models. In the above figures, still (a) is the true model, and (b)-(f) are the inversion results for WD, ABSWD, EN, GC, and $\ell 1$ RI-based objective functions. With the poorer initial model. We can clearly see that the inversion results are poorer compared with the results in the previous section. Take the inversion results for V_p as an example. The inversion results using WD are poor in the deeper part of the model, and there are some artifacts in the upper right part of the model. The inversion results for EN only recover the large-scale information of the model and fail to give detailed information about the inversion results. The V_p inverted by using the GC is still poor for the deeper part of the model, and only the upper part of the model is inverted. The inversion results for the $\ell 1$ RI and $\ell 1$ objective functions are very similar, and both the deeper and the shallower parts of the model are inverted. To further prove the advantages of using the $\ell 1$ RI-based objective function for the FWI. Figures , ,

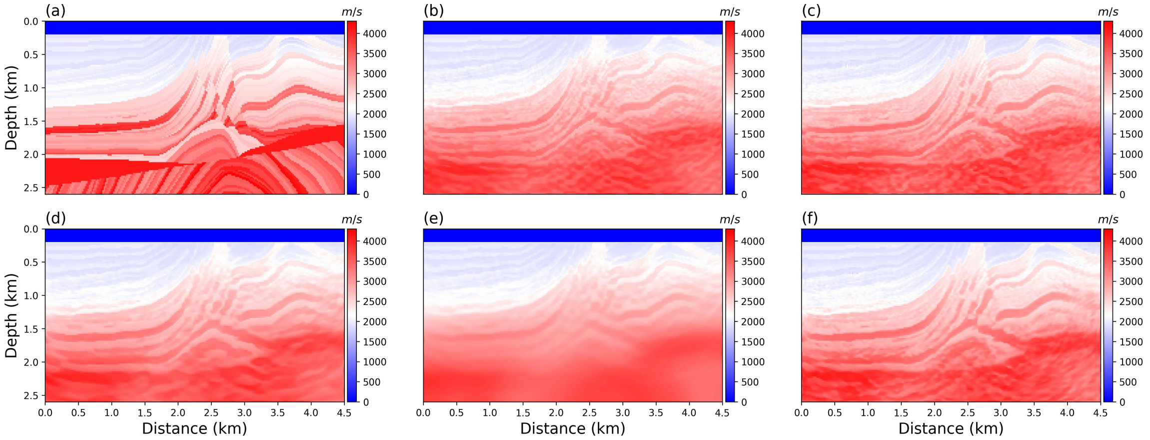

FIG. 13. Stress test inversion results V_s. (a) true V_s. (b) WD V_s inversion result. (c) ABSWD V_s inversion result. (d) EN V_s inversion result. (e) GC V_s inversion result. (e) $\ell 1$ RI V_s inversion result.

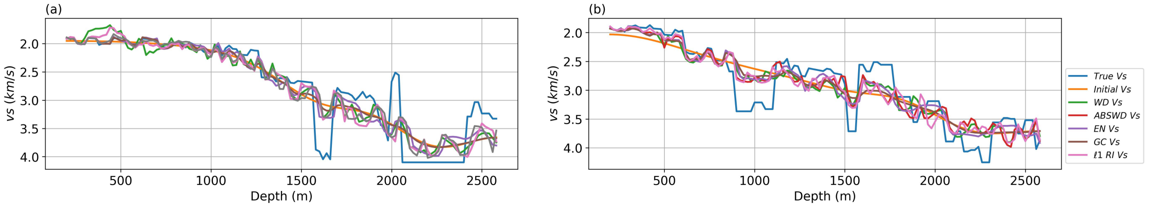

FIG. 14. Stress test inversion profiles for V_s. (a) Vertical inversion for V_s at 2000m in distance of the model. (b) Vertical inversion for V_s at 3000m in distance of the model.

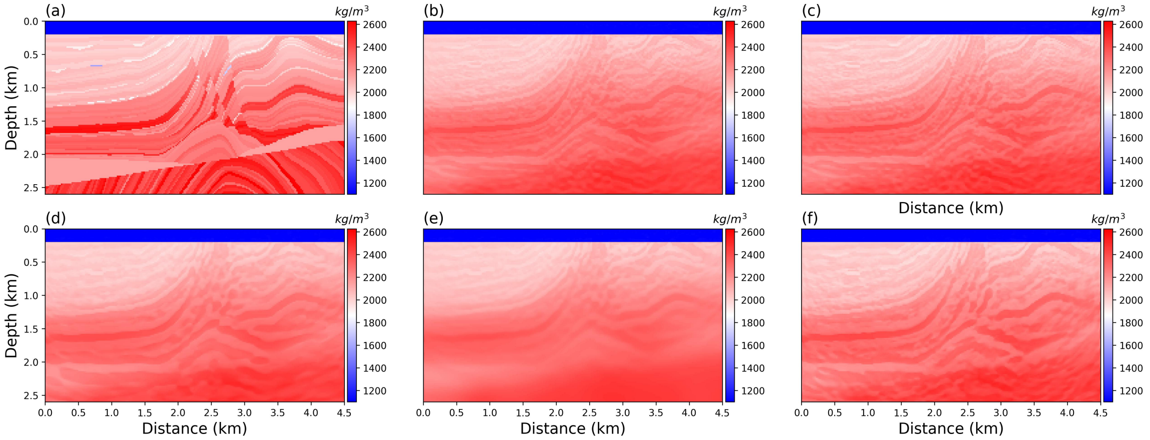

FIG. 15. Stress test inversion results ρ. (a) true ρ. (b) WD ρ inversion result. (c) ABSWD ρ inversion result. (d) EN ρ inversion result. (e) GC ρ inversion result. (e) $\ell 1$ RI ρ inversion result.

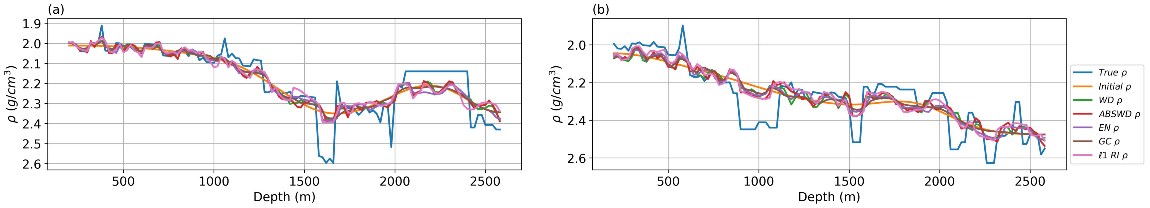

FIG. 16. Stress test inversion profiles for ρ. (a) Vertical inversion for ρ at 2000m in distance of the model. (b) Vertical inversion for ρ at 3000m in distance of the model.

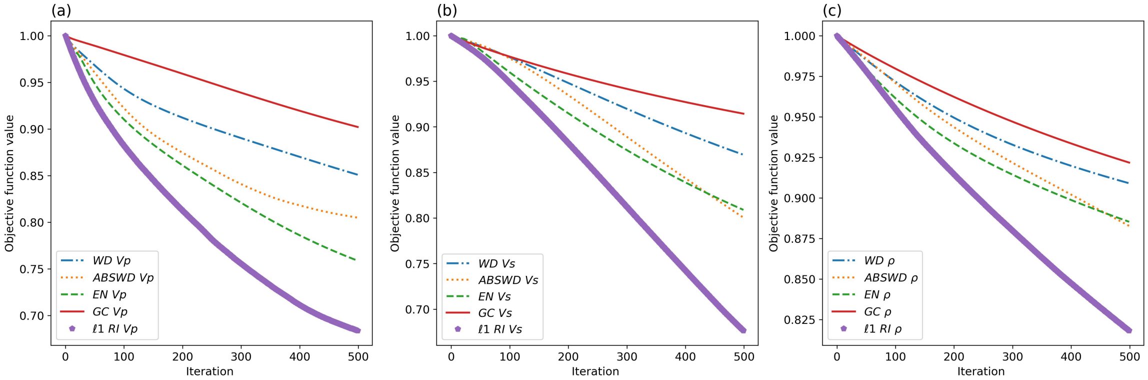

FIG. 17. Stress test inversion profiles for model loss with respect to iteration time. (a) Model misfit for V_p. (b) Model misfit for V_s. (c) Model misfit for ρ.

plot the vertical profile for the V_p, V_s, and ρ inversion results at 200m and 3800m latterly of the model. We can still see that the inversion results given by $\ell1$ RI (purple line) are closer to the true models. To further prove the validity of the $\ell1$ norm objective function. We plot the Mean Squares Error (MSE) between the inversion results and the true models for each parameter at each iteration in Figure 17. In figure 17, we can see that the $\ell1$ RI-based objective function has the best convergence rate, and EN has the slowest convergence rate.

Bad initial model test with noise

In this test, We will add Gaussian noise into the observed data to test the performance of different objective functions. The signal to noise ratio (SNR) we use in this test is $\mathrm{SNR} = 10$, where the SNR is calculated with (Li et al., 2016):

$$\mathrm{SNR} = 10\log_{10}\frac{||s||^2}{||n||^2}, \tag{21}$$

where $||s||^2$ is the energy of the observed data trace and $||n||^2$ is the energy of the noise in the data. The noises are added to both the x and y components of the data to perform the inversion.

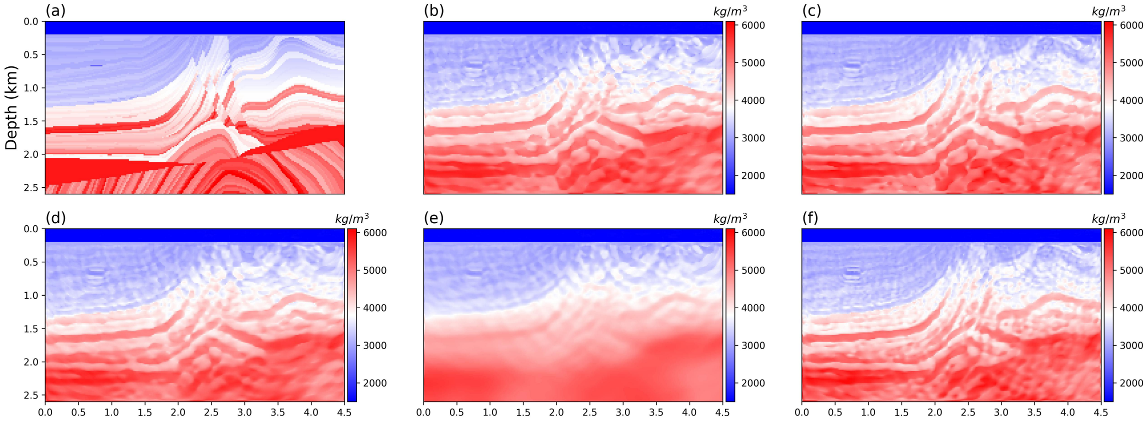

FIG. 18. Stress test inversion results V_p. (a) true V_p. (b) WD V_p inversion result. (c) ABSWD V_p inversion result. (d) EN V_p inversion result. (e) GC V_p inversion result. (e) $\ell1$ RI V_p inversion result.

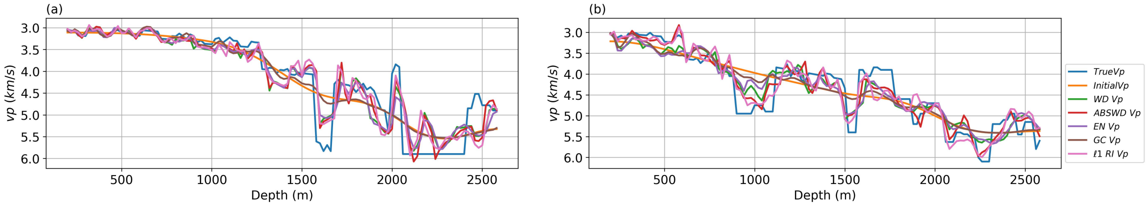

FIG. 19. Stress test inversion profiles for V_p. (a) Vertical inversion for V_p at 2000m in distance of the model. (b) Vertical inversion for V_p at 3000m in distance of the model.

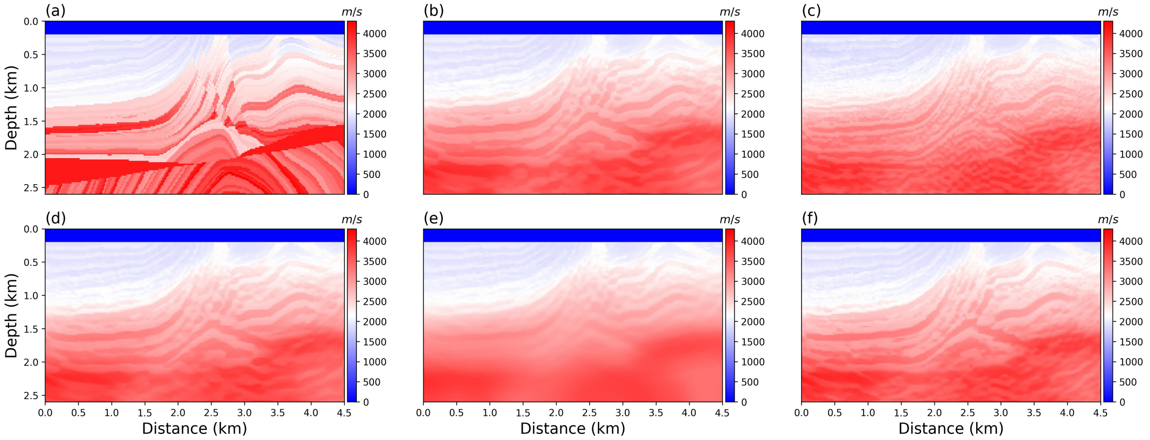

FIG. 20. Stress test inversion results V_s. (a) true V_s. (b) WD V_s inversion result. (c) ABSWD V_s inversion result. (d) EN V_s inversion result. (e) GC V_s inversion result. (e) ℓ1 RI V_s inversion result.

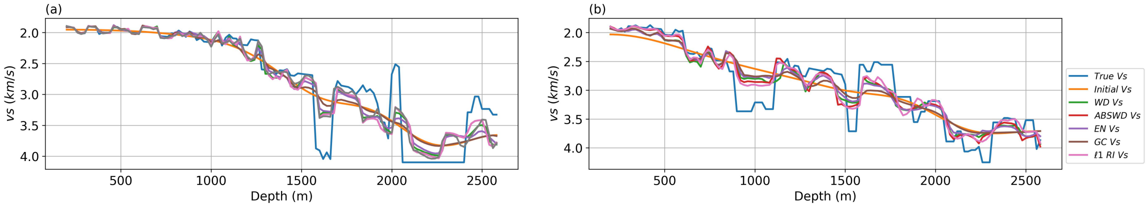

FIG. 21. Stress test inversion profiles for V_s. (a) Vertical inversion for V_s at 2000m in distance of the model. (b) Vertical inversion for V_s at 3000m in distance of the model.

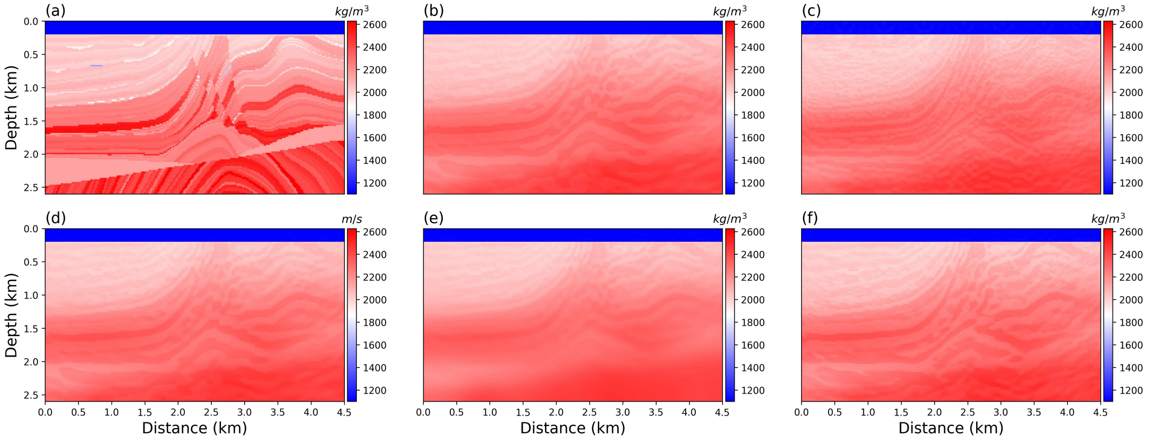

FIG. 22. Stress test inversion results ρ. (a) true ρ. (b) WD ρ inversion result. (c) ABSWD ρ inversion result. (d) EN ρ inversion result. (e) GC ρ inversion result. (e) ℓ1 RI ρ inversion result.

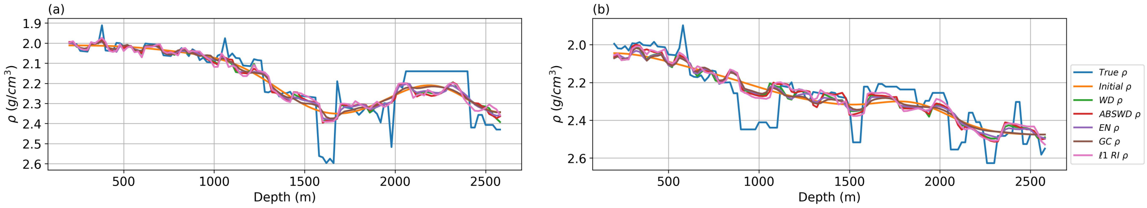

FIG. 23. Stress test inversion profiles for ρ. (a) Vertical inversion for ρ at 2000m in distance of the model. (b) Vertical inversion for ρ at 3000m in distance of the model.

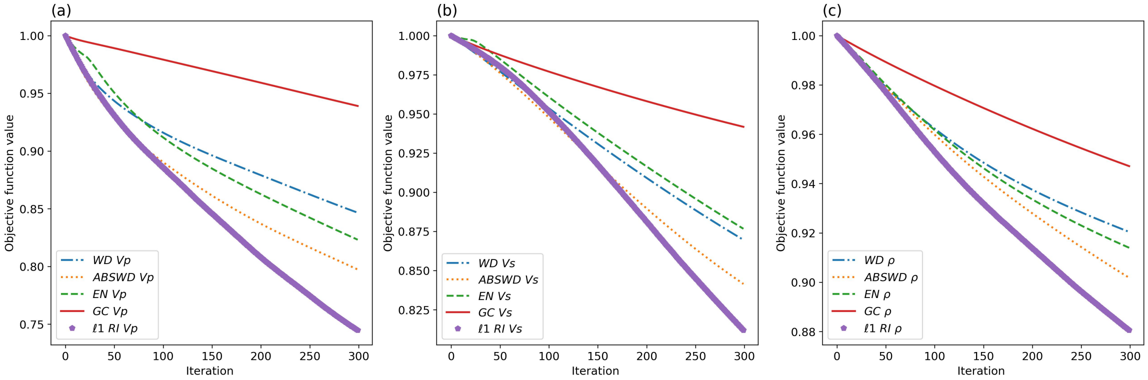

FIG. 24. Stress test inversion profiles for model loss with respect to iteration time. (a) Model misfit for V_p. (b) Model misfit for V_s. (c) Model misfit for ρ.

Figures 18, 20, and 22 shows the inversion results for this stress test. In these figures, (a) is still the true model and (b)-(f) are the inversion results by using the WD, ABSWD, GC, EN and $\ell 1$ RI objective functions. Taking the inversion results for V_p as an example. In the inversion results for WD based objective function, Figure 18b, we can see that, due to the presence of the noise, the upper right part of the model has been poorly inverted, and the deeper part of the model has been badly blurred. The inversion results given by $\ell 1$ norm can better invert the upper right part of the model, while the bottom part of the model also has a low resolution. Similar to the test in the previous section, EN gives large-scale information on inversion results, and the high wavenumber components of the elastic model are rarely updated. GC-based objective function failed to give the apparent update for the deeper structure of the model. Compared with all the inversion results, $\ell 1$ RI objective function gives the best inversion results, we admit that this result is no better than the noise-free case, but we can see that both the shallow and deep structure of the elastic model has been correctly updated compared with the other objective functions. Similar inversion patterns can be observed for V_p, V_s and ρ inversion. To better illustrate the advantages given by the $\ell 1$ RI objective function. We also plot how the MSE model error varies with iterations in Figure 24. We can also see that among all these objective functions. The $\ell 1$ RI-based objective function has the best convergence rate.

CONCLUSIONS

We propose the Auto-adjoint time domain elastic full waveform inversion in this report, which is an FWI framework accelerated with CUDA using adjoint sources calculated with the automatic differential method. In this FWI framework, the forward modeling and the adjoint modeling are accelerated by CUDA, and the adjoint sources are calculated by the automatic differential method. These two features allow us to perform time domain FWI with GPU acceleration and explore how different kinds of objective functions can influence the inversion results effectively. We study the objective function behaviour for the ℓ-2 norm. ℓ1-norm, Global-correlation based, Envelope based, objective function, and ℓ-1 norm between the real and imaginary part of the synthetic data and the observed data (ℓ1 RI objective function). According to the numerical test we did in this paper, the ℓ1 RI objective function has a better ability to tolerate the noise when the poor initial model is used for inversion, compared with all the other objective functions we considered.

ACKNOWLEDGMENT

We thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 543578-19. The first author is also supported by the Chine Scholarship Council (CSC).

REFERENCES

Akcelik, V., 2002, Multiscale newton-krylov methods for inverse acoustic wave propagation: Ph.D. thesis, Carnegie Mellon University.

Askan, A., Akcelik, V., Bielak, J., and Ghattas, O., 2007, Full waveform inversion for seismic velocity and anelastic losses in heterogeneous structures: Bulletin of the Seismological Society of America, **97**, No. 6, 1990–2008.

Askan, A., and Bielak, J., 2008, Full anelastic waveform tomography including model uncertainty: Bulletin of the Seismological Society of America, **98**, No. 6, 2975–2989.

Brossier, R., Operto, S., and Virieux, J., 2010, Which data residual norm for robust elastic frequency-domain full waveform inversion?: Geophysics, **75**, No. 3, R37–R46.

Choi, Y., and Alkhalifah, T., 2012, Application of multi-source waveform inversion to marine streamer data using the global correlation norm: Geophysical Prospecting, **60**, No. 4-Simultaneous Source Methods for Seismic Data, 748–758.

Choi, Y., and Alkhalifah, T., 2015, Unwrapped phase inversion with an exponential damping: Geophysics, **80**, No. 5, R251–R264.

Clark, D., 2010, The best papers in geophysics (by the numbers).

Crase, E., Pica, A., Noble, M., McDonald, J., and Tarantola, A., 1990, Robust elastic nonlinear waveform inversion: Application to real data: Geophysics, **55**, No. 5, 527–538.

Dai, Y.-H., and Yuan, Y., 1999, A nonlinear conjugate gradient method with a strong global convergence property: SIAM Journal on optimization, **10**, No. 1, 177–182.

Djikpéssé, H. A., and Tarantola, A., 1999, Multiparameter l1 norm waveform fitting: Interpretation of gulf of mexico reflection seismograms: Geophysics, **64**, No. 4, 1023–1035.

Dong, S., Han, L., Hu, Y., and Yin, Y., 2020, Full waveform inversion based on a local traveltime correction and zero-mean cross-correlation-based misfit function: Acta Geophysica, **68**, No. 1, 29–50.

Fletcher, R., and Reeves, C. M., 1964, Function minimization by conjugate gradients: The computer journal, **7**, No. 2, 149–154.

Gholami, Y., Brossier, R., Operto, S., Ribodetti, A., and Virieux, J., 2013, Which parameterization is suitable for acoustic vertical transverse isotropic full waveform inversion? part 1: Sensitivity and trade-off analysis: Geophysics, **78**, No. 2, R81–R105.

Hestenes, M. R., and Stiefel, E., 1952, Methods of conjugate gradients for solving: Journal of research of the National Bureau of Standards, **49**, No. 6, 409.

Huber, P. J., 1973, Robust regression: asymptotics, conjectures and monte carlo: The annals of statistics, 799–821.

Kingma, D. P., Ba, J. A., and Adam, J., 2020, A method for stochastic optimization. arxiv 2014: arXiv preprint arXiv:1412.6980, **106**.

Köhn, D., 2011, Time domain 2d elastic full waveform tomography: Ph.D. thesis.

Li, B., Wang, D., Liu, Y., and Liu, C., 2016, Local snr estimation method for seismic data, *in* SEG Technical Program Expanded Abstracts 2016, Society of Exploration Geophysicists, 4720–4724.

Liu, Y., Teng, J., Xu, T., Wang, Y., Liu, Q., and Badal, J., 2017, Robust time-domain full waveform inversion with normalized zero-lag cross-correlation objective function: Geophysical Journal International, **209**, No. 1, 106–122.

Mora, P., 1987, Nonlinear two-dimensional elastic inversion of multioffset seismic data: Geophysics, **52**, No. 9, 1211–1228.

Nocedal, J., and Wright, S. J., 1999, Numerical optimization: Springer.

Pan, W., Innanen, K. A., and Geng, Y., 2018, Elastic full-waveform inversion and parametrization analysis applied to walk-away vertical seismic profile data for unconventional (heavy oil) reservoir characterization: Geophysical Journal International, **213**, No. 3, 1934–1968.

Polak, E., and Ribiere, G., 1969, Note sur la convergence de méthodes de directions conjuguées: Revue française d'informatique et de recherche opérationnelle. Série rouge, **3**, No. 16, 35–43.

Sen, M. K., and Stoffa, P. L., 1991, Nonlinear one-dimensional seismic waveform inversion using simulated annealing: Geophysics, **56**, No. 10, 1624–1638.

Stoffa, P. L., and Sen, M. K., 1991, Nonlinear multiparameter optimization using genetic algorithms: Inversion of plane-wave seismograms: Geophysics, **56**, No. 11, 1794–1810.

Tarantola, A., 1986, A strategy for nonlinear elastic inversion of seismic reflection data: Geophysics, **51**, No. 10, 1893–1903.

Tarantola, A., 2005, Inverse problem theory and methods for model parameter estimation: SIAM.

Wu, R.-S., Luo, J., and Wu, B., 2014, Seismic envelope inversion and modulation signal model: Geophysics, **79**, No. 3, WA13–WA24.