# Efficient pseudo Gauss-Newton full waveform inversion in the time-ray parameter domain

Wenyong Pan, Kris Innanen, Gary Margrave


## ABSTRACT

Full Waveform Inversion (FWI) has been widely studied in recent years but still cannot be practiced in industry effectively. Generally, its failure can be attributed to expensively computational cost, slow convergence rate, cycle skipping problem and so on. For traditional FWI, the gradient is calculated shot by shot based on the adjoint state method. The computational burden rises significantly when considering large 2D velocity model or 3D experiments. A linear phase encoding strategy is employed to construct the gradient in the time-ray parameter domain. The phase encoding approach forms supergathers by summing densely distributed individual shots and can reduce the computational burden considerably. Furthermore, we propose the gradient be calculated using one single ray parameter per FWI iteration, with the ray parameter value varied for different iterations. The computational cost is reduced further within this strategy. The gradient is a poorly scaled image which can be considerably enhanced by multiplying the inverse Hessian. The Hessian matrix serves as a nonstationary deconvolution operator to compensate the geometrical spreading effects and suppress the multiple scattering effects. While explicit calculation of the gradient is also considered to be unfeasible. Under the assumption of high frequency limitation, the diagonal Hessian can work as a good approximation and it can also be constructed by the phase encoding method. In this research, preconditioning the gradient using the diagonal part of the phase encoded Hessian forms one pseudo Gauss-Newton step. Several numerical examples are presented to analyze the gradient contributions and compare different Hessian approximations. Finally, a modified Marmousi model is illustrated for full waveform inversion. we compared the effects for fixed ray parameter and varied ray parameter and analyzed the sensitivity to the ray parameter range, sensitivity to the Gaussian noise and sensitivity to the number of encoded sources. And the inversion results with different scaling methods are also provided for comparison.


## INTRODUCTION

Full Waveform Inversion (FWI), formulated as a least-squares inverse problem, seeks to minimize the difference between the observed data and synthetic data (Lailly, 1983; Tarantola, 1984; Pratt et al., 1998; Virieux and Operto, 2009; Margrave et al., 2011). It has drawn people's attention in the past several decades for its great powerfulness in estimating the subsurface parameters with high resolution. While FWI suffers from prohibitively computational burden, slow convergence rate, cycle skipping problem and some other problems which make it unfeasible in industry practice. Aiming at the problems existed in FWI, we proposed a pseudo Gauss-Newton (PGN) method in time-ray parameter domain which employs phase encoding technique to construct the gradient and the Hessian matrix.

For conventional FWI, simulations from each source and receiver should be performed to construct the gradient which is extensively expensive. And an efficient gradient calculation strategy proposed by Lailly (1983) and Tarantola (1984) is the adjoint state method, which allows us to construct the gradient by applying a zero-lag cross-correlation between the forward modeling wavefields from each source and back-propagated wavefields from all receivers simultaneously. Thus, we only need $2Ns$ simulations to calculate the gradient, where $Ns$ indicates the number of sources. However, for large 2D velocity model, especially for 3D inverse problem, the computational cost arises considerably.

In this research, the phase encoding strategy, which was firstly introduced in prestack migration (Morton and Ober, 1998; Romero et al., 2000; Zhang et al., 2005; Liu et al., 2006; Dai and Schuster, 2013), is employed to construct the gradient. The phase encoding technique forms supergathers by summing densely distributed individual shots and can reduce the computational burden considerably for FWI problem (Vigh and Starr, 2008; Krebs et al., 2009; Ben-Hadj-Ali et al., 2011; Gao et al., 2010; Tao and Sen, 2013) but unfortunately introduces coherent crosstalk artifacts resulting from the undesired interactions between unrelated source and receiver wavefields. The crosstalk artifacts introduced can be suppressed reasonably well with a sufficient number of ray parameters (Liu et al., 2006), which are controlled by the take-off angle at the surface location and the top surface velocity (Tao and Sen, 2013; Dai and Schuster, 2013). Thus, the number of simulations for the gradient calculation in one FWI iteration becomes $2Np$, where $Np$ means the number of ray parameters. Generally, the ray parameter range can be determined by the geological structures of the target area and different ray parameters account for the subsurface layers with different dip angles. Zhang et al. (2005) and Stork and Kapoor (2004) used ray parameter component sampling theory to determine the number of ray parameters $N_p$ and ray parameter interval $\triangle p$. In this research, we propose the gradient be calculated using one single ray parameter per FWI iteration, with the ray parameter value varied for different iterations (Pan et al., 2014). This means that we only need 2 simulations to calculate the gradient for one FWI iteration. The computational cost is reduced further within this strategy. And comparing with conditions using fixed ray parameter, varying the ray parameters can correct the biased gradient direction and improve the convergence rate. While this strategy is sensitive to the assembled sources. When the encoded sources distribution are not dense enough, the footprints or crosstalk artifacts will become obvious especially for the shallow layers and the convergence rate will also be decreased.

Another big problem for FWI is its slow convergence rate. It is known that the gradient is equivalent to a reverse time migration image based on cross-correlation imaging condition. And the image is a poorly scaled for the energy loss during forward modeling and backward propagating the data residual (Shin et al., 2001a,b). So, the steepest decent method is considered to be a crude scaling method for simplifying the Hessian matrix to an identity matrix. The poorly scaled image can be enhanced considerably by multiplying the inverse Hessian (Pratt et al., 1998). Hessian matrix works as a nonstationary deconvolution operator to compensate the geometrical spreading effects saptially at each image point and recover the amplitude at deep reflectors. Hessian matrix also predicts the multiple scattered wavefields (Pratt et al., 1998) and can suppress the spurious part of the gradient caused by the multiple scattering effects. Then the resolution of the gradient can be enhanced. Under the assumption of high frequency limit, the Hessian matrix is diagonally dominant and the diagonal part of the Hessian can serve as a good approximation to recover and balance the amplitudes. However, even explicit calculation of the diagonal Hessian matrix is prohibitively expensive in practical application (Pratt et al., 1998; Tang, 2009; Tao and Sen, 2013).

The nonlinear term in full Hessian matrix accounting for the multiple scattering is always neglected and the linear part forms the approximate Hessian used in Gauss-Newton method. Under the assumption of infinite receiver coverage, Shin et al. (2001a) proposed the pseudo-Hessian, which can be constructed by the forward modeling wavefields, used as the virtual sources in reverse time migration. It assumes the receiver-side Green's functions in the approximate Hessian as an constant (Plessix and Mulder, 2006) and preconditioning the gradient using the diagonal part of the pseudo-Hessian is equivalent to a deconvolution imaging condition in reverse time migration (Pan et al., 2013a,b). And the diagonal part of the pseudo-Hessian overestimates the illumination energy and is limited to compensate the geometrical spreading effects. Tang (2009) introduced a receiver-side phase encoded Hessian by construct the receiver-side Green's functions using a random phase encoding strategy. Tao and Sen (2013) calculated the diagonal part of the approximate Hessian using a linear phase encoding strategy with plane wave data. In this research, we introduced a chirp phase encoding method to calculate the diagonal part of the approximate Hessian. This encoding strategy can be regarded as a combination of linear and random phase encoding strategies. Compared to the linear phase encoding method, it can disperse the crosstalk noise better and get a more close approximation to the exact diagonal Hessian with the same computational cost. And preconditioning the gradient using the diagonal phase encoded Hessian forms one pseudo Gauss-Newton step in this research.

FWI also suffers from cycle skipping problem. The poor starting velocity model increases the nonlinearity of the least-squares inverse problem. The lack of low frequency information results in the miss of global minimum. Low frequency is essential to catch the long wavenumber component of the velocity model. And the high frequency is responsible to add detailed information. A multiscale approach, performed with increasing low frequency to high frequency groups, can improve the chances that the global minimum is reached and overcome the cycle skipping difficulty (Vigh and Starr, 2008). Furthermore, a multiscale method is computationally efficient and converges faster than a conventional, single-scale method (Boonyasiriwat et al., 2009). In frequency domain, the multiscale approach can be implemented by selecting a few frequencies. In this research, a time domain multiscale approach is implemented by applying a low-pass filtering to the data residual. And the frequency bands are broaden with increasing the number of iterations.

This paper is organized as follows. Firstly, the basic theory for least squares inverse problem is reviewed and the similarity of gradient calculation in FWI and reverse time migration is explained. Then we discussed different Hessian approximations and explained phase encoded Hessian combining with several numerical examples. Finally, we presented a numerical example for full waveform inversion based on a modified Marmousi model. And we discussed the effects for fixed ray parameter and varied ray parameter and analyzed the sensitivities to ray parameter range, the number of encoded sources and Gaussian noise. We compared the inversion results for different scaling methods and the imaging results using initial velocity model, true velocity model and inverted velocity model are provided for comparison.

## THEORY AND METHODS

In this section, firstly we reviewed the basic principle for least-squares inverse problem and illustrated the equivalence between the gradient in FWI and reverse time migration. And then we compared the full Hessian, approximate Hessian and pseudo-Hessian and presented how to construct the approximate Hessian using phase encoding method.

### Gradient Calculation in FWI and Reverse Time Migration

As a least-squares local optimization, full waveform inversion seeks to minimize the difference between the synthetic data and observed data (Lailly, 1983; Tarantola, 1984) and update the model iteratively. The misfit function $\phi$ is given in a least-squares norm:

$$\phi\left(s_0^{(n)}(\mathbf{r})\right) = \frac{1}{2}\int d\omega \left(\sum_{\mathbf{r}_s}\sum_{\mathbf{r}_g} \|\delta P\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}(\mathbf{r})\right)\|_2\right), \tag{1}$$

where $s_0^{(n)}(\mathbf{r}) = \frac{1}{\left(c_0^{(n)}(\mathbf{r})\right)^2}$ are the model parameters, the square of the slowness in the $n$th iteration, and $c_0^{(n)}(\mathbf{r})$ is the velocity. $\delta P\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}(\mathbf{r})\right)$ mean the data residuals, the difference between the observed data $P\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)$ and synthetic data $G\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)$, $\|\cdot\|_2$ indicates the $\ell - 2$ norm.

The minimum value of the misfit function is sought in the vicinity of the starting model $s_0(\mathbf{r})$ and the updated model can be written as the sum of the starting model and a model perturbation $\delta s_0^{(n)}(\mathbf{r})$ (Virieux and Operto, 2009).

$$s^{(n)}(\mathbf{r}) = s_0^{(n)}(\mathbf{r}) + \mu^{(n)}\delta s_0^{(n)}(\mathbf{r}), \tag{2}$$

where $\mu^{(n)}$ is the step length in $n$th iteration, which is a scalar constant used to scale the model perturbation and can be obtained through a line search method (Gauthier et al., 1986; Pica et al., 1990).

Applying a second order Taylor-Lagrange development of the misfit function and then taking partial derivative to the model parameter give equation (3) and (4):

$$\phi\left(s_0^{(n)} + \delta s_0^{(n)}\right) \simeq \phi\left(s_0^{(n)}\right) + \int d\mathbf{r}' \frac{\partial \phi}{\partial s_0^{(n)}(\mathbf{r}')}\delta s_0^{(n)}(\mathbf{r}'), \tag{3}$$

$$\begin{aligned} \frac{\partial\phi\left(s_0^{(n)}+\delta s_0^{(n)}\right)}{\partial s_0^{(n)}(\mathbf{r}')} &\simeq \frac{\partial\phi\left(s_0^{(n)}\right)}{\partial s_0^{(n)}(\mathbf{r}')} + \frac{\partial}{\partial s_0^{(n)}(\mathbf{r}')}\int d\mathbf{r}\frac{\partial\phi\left(s_0^{(n)}\right)}{\partial s_0^{(n)}(\mathbf{r})}\delta s_0^{(n)}(\mathbf{r}), \\ &= g^{(n)}(\mathbf{r}') + \int d\mathbf{r} H^{(n)}(\mathbf{r}', \mathbf{r})\delta s_0^{(n)}(\mathbf{r}). \end{aligned} \tag{4}$$

where $g^{(n)}(\mathbf{r}')$ is the gradient and $H^{(n)}(\mathbf{r}', \mathbf{r})$ is the Hessian matrix, When equation (4) equals to zero, the misfit function towards to the minimum and the model perturbation can be expressed as:

$$\delta s_0^{(n)} = -\int d\mathbf{r}' H^{(n)-}(\mathbf{r}, \mathbf{r}') g^{(n)}(\mathbf{r}), \tag{5}$$

Then inserting equation (5) into equation (2), the model can be updated iteratively using the following equation:

$$s^{(n)}(\mathbf{r}) = s_0^{(n)}(\mathbf{r}) - \mu^{(n)} \left( - \int d\mathbf{r}' H^{(n)-}(\mathbf{r}, \mathbf{r}') g^{(n)}(\mathbf{r}) \right), \tag{6}$$

For gradient, the first order derivative of the misfit function $\phi$ with respect to the model parameters, can be obtained by a zero-lag correlation between the complex conjugate of data residuals and the first order partial derivative wavefields:

$$g^{(n)}(\mathbf{r}) = -\sum_{\mathbf{r}_s} \sum_{\mathbf{r}_g} \int d\omega \Re \left( \frac{\delta G\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)}{\delta s_0^{(n)}(\mathbf{r})} \delta P^* \left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right) \right), \tag{7}$$

where $\frac{\delta G\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)}{\delta s_0^{(n)}(\mathbf{r})}$ is the sensitive matrix (it is also called Fr$\acute{e}$chet derivative matrix or Jacobian matrix). Then apply a perturbative derivation and truncate the Born series based on the assumption of small model perturbation (as shown by APPENDIX A), the sensitive matrix can be written as:

$$\frac{\delta G\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)}{\delta s_0^{(n)}(\mathbf{r})} = -\omega^2 G\left(\mathbf{r}_g, \mathbf{r}, \omega | s_0^{(n)}\right) G\left(\mathbf{r}, \mathbf{r}_s, \omega | s_0^{(n)}\right), \tag{8}$$

Then insert equation (8) into equation (7), the gradient becomes:

$$g^{(n)}(\mathbf{r}) = \sum_{\mathbf{r}_s} \sum_{\mathbf{r}_g} \int d\omega \Re \left( \omega^2 \mathcal{F}_s(\omega) G(\mathbf{r}, \mathbf{r}_s, \omega) G(\mathbf{r}_g, \mathbf{r}, \omega) \delta P^* \left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right) \right), \tag{9}$$

where $G(\mathbf{r}, \mathbf{r}_s, \omega)$ and $G(\mathbf{r}_g, \mathbf{r}, \omega)$ are the source-side and receiver-side Green's functions, respectively. Then the gradient can be calculated using the adjoint state method by applying a zero-lag convolution between the forward modeling wavefields and back-propagated data residuals, which avoids the direct computation of the partial derivative wavefields.

Recall reverse time migration, we can recognize the similarity between the gradient calculation in FWI and reverse time migration image. Both of them employ the adjoint state technique to avoid the direct calculation of the partial derivative wavefields. The only difference is that FWI backpropagates data residuals while RTM backpropagates the observed data. So, if the initial model in FWI is considered as a smooth background velocity model, the data residuals in FWI become measured data (Lailly, 1983; Shin et al., 2001a). Thus, the gradient calculation in the first iteration of FWI is formally identical to the prestack reverse time migration with a crosscorrelation imaging condition.

### Full Hessian Matrix and Hessian Approximations

The image or gradient formed by crosscorrelation imaging condition suffers from geometrical spreading effects, wave attenuation and transmission loss, which result in poor amplitudes for deep reflectors. Multiplying the inverse Hessian matrix can remove the geometrical amplitude decay of the Green's functions (Tao and Sen, 2013). While it is considered as unfeasible to calculate the Hessian matrix directly for its great computational requirements.

*Full Hessian*

The full Hessian matrix is the second order partial derivative of the misfit function with respect to the model parameters. And it can be written as:

$$H^{(n)}(\mathbf{r}',\mathbf{r}) = \frac{\partial^2\phi\left(s_0^{(n)}\right)}{\partial s_0^{(n)}(\mathbf{r})\partial s_0^{(n)}(\mathbf{r}')}, \tag{10}$$

After a series of derivations, the Hessian matrix can be written as the summation of two terms:

$$H^{(n)}(\mathbf{r}',\mathbf{r}) = H_1^{(n)} + H_2^{(n)}, \tag{11}$$

where

$$\begin{cases} H_1^{(n)} = -\sum\limits_{\mathbf{r}_s,\mathbf{r}_g}\int d\omega \Re\left\{\omega^2 \dfrac{\partial\left[G\left(\mathbf{r}',\mathbf{r}_s,\omega|s_0^{(n)}\right)G\left(\mathbf{r}_g,\mathbf{r},\omega|s_0^{(n)}\right)\right]}{\partial s_0^{(n)}}\delta P^*\left(\mathbf{r}_g,\mathbf{r}_s,\omega|s_0^{(n)}\right)\right\} \\ H_2^{(n)} = \sum\limits_{\mathbf{r}_s,\mathbf{r}_g}\int d\omega \Re\left\{\omega^4 G\left(\mathbf{r}_g,\mathbf{r}',\omega|s_0^{(n)}\right)G\left(\mathbf{r}',\mathbf{r}_s,\omega|s_0^{(n)}\right)G^*\left(\mathbf{r}_g,\mathbf{r}'',\omega|s_0^{(n)}\right)G^*\left(\mathbf{r}'',\mathbf{r}_s,\omega|s_0^{(n)}\right)\right\} \end{cases}, \tag{12}$$

The first term of full Hessian matrix $H_1^{(n)}$ can be computed by multiplying the data residuals using the second order partial derivative wavefields, which is defined as the variation of the first order partial derivative wavefields corresponding to perturbation in the model parameters. If we apply a further derivation to the second order partial derivative wavefields, we can get:

$$\begin{aligned} \frac{\partial\left[G\left(\mathbf{r}',\mathbf{r}_s,\omega|s_0^{(n)}\right)G\left(\mathbf{r}_g,\mathbf{r},\omega|s_0^{(n)}\right)\right]}{\partial s_0^{(n)}} = & -\left\{G\left(\mathbf{r}',\mathbf{r}_s,\omega|s_0^{(n)}\right)G\left(\mathbf{r},\mathbf{r}',\omega|s_0^{(n)}\right)G\left(\mathbf{r}_g,\mathbf{r},\omega|s_0^{(n)}\right)\right\} \\ & -\left\{G\left(\mathbf{r},\mathbf{r}_s,\omega|s_0^{(n)}\right)G\left(\mathbf{r}',\mathbf{r},\omega|s_0^{(n)}\right)G\left(\mathbf{r}_g,\mathbf{r}',\omega|s_0^{(n)}\right)\right\} \end{aligned}, \tag{13}$$

So, the second order partial derivative wavefields actually is the summation of the second order multiple scatterings at positions of $\mathbf{r}'$ and $\mathbf{r}$ excited by the first order multiple scatterings at the positions of $\mathbf{r}$ and $\mathbf{r}'$, respectively. It indicates the nonlinearity of the least-squares inverse problem. The crosscorrelation method produces false anomalies (low frequency artifacts in reverse time migration) caused by correlating the multiple scattering in data residuals with the partial derivative wavefields (Pratt et al., 1998). These anomalies are spurious or false parts in the gradient. The first term $H_1^{(n)}$ predicts these high-order multiple scatterings and works as a de-multiple operator to suppress these artifacts in the gradient.

It is always expensive to compute the first term of the Hessian matrix directly. Pratt et al. (1998) proposed to multiply the second order virtual sources (calculated from the first order scattering wavefields) by the back-propagated wavefields, which also needs $m$ forward modeling simulations, where $m$ is the number of the model parameters. Even though, the computation cost can be reduced considerably, it is still unfeasible to be practiced in industrial application.

*Approximate Hessian*

When the initial model is close to the real model, the misfit function is near to the global minimum and the data residual becomes very small, just as what we have assumed in truncating the Born series. So, the first term $H_1^{(n)}$ in the full Hessian matrix is always neglected for computation convenience, which makes the inverse problem approximately linear. Thus the full Hessian matrix can be substituted by the approximate Hessian (the second term in equation (11), which is used as a preconditioner for the gradient in the Gauss-Newton method:

$$H_a^{(n)} = H_2^{(n)} \simeq \sum_{\mathbf{r}_s,\mathbf{r}_g} \int d\omega \Re \left\{ \omega^4 G\left(\mathbf{r}_g, \mathbf{r}', \omega | s_0^{(n)}\right) G\left(\mathbf{r}', \mathbf{r}_s, \omega | s_0^{(n)}\right) G^*\left(\mathbf{r}_g, \mathbf{r}'', \omega | s_0^{(n)}\right) G^*\left(\mathbf{r}'', \mathbf{r}_s, \omega | s_0^{(n)}\right) \right\}, \tag{14}$$

Comparing equation (14) with equation (8), each element in the approximate Hessian $H_a^{(n)}$ can be interpreted as the scalar product of two partial derivative wavefields (Pratt et al., 1998), which is equivalent to the zero-lag convolution computation in time domain. Pratt et al. (1998) used the virtual sources to multiply the back-propagated wavefields for constructing the partial derivative wavefields. The gradient can be preconditioned considerably by multiplying the inverse approximate Hessian $H_a^{(n)-}$.

The diagonal elements and off-diagonal elements of the approximate Hessian can be interpreted as the zero-lag auto-correlation and cross-correlation of the partial derivative wavefields respectively (Shin et al., 2001a), as indicated by the first term $\bar{H}_a^{(n)}$ and second term $\tilde{H}_a^{(n)}$ in equation (15).

$$H_a^{(n)} = \bar{H}_a^{(n)} + \tilde{H}_a^{(n)}, \tag{15}$$

where

$$\begin{cases} \bar{H}_a^{(n)} = \sum_{\mathbf{r}_s,\mathbf{r}_g} \int d\omega \Re \left\{ \omega^4 G\left(\mathbf{r}_g, \mathbf{r}', \omega\right) G\left(\mathbf{r}', \mathbf{r}_s, \omega\right) G^*\left(\mathbf{r}_g, \mathbf{r}', \omega\right) G^*\left(\mathbf{r}', \mathbf{r}_s, \omega\right) \right\} \\ \tilde{H}_a^{(n)} = \sum_{\mathbf{r}_s,\mathbf{r}_g} \int d\omega \Re \left\{ \omega^4 G\left(\mathbf{r}_g, \mathbf{r}', \omega\right) G\left(\mathbf{r}', \mathbf{r}_s, \omega\right) G^*\left(\mathbf{r}_g, \mathbf{r}'', \omega\right) G^*\left(\mathbf{r}'', \mathbf{r}_s, \omega\right) \right\}, \mathbf{r}'' \neq \mathbf{r}' \end{cases},$$

In the high-frequency asymptotics and for the reference model with relatively smooth impedance variations, the two partial derivative wavefields are largely uncorrelated with each other but perfectly self-correlated (Pratt et al., 1998; Shin et al., 2001b; Tang, 2009), which means that the approximate Hessian is diagonally dominated. In this case, the off-diagonal elements of the approximate Hessian, the second term $\tilde{H}_a^{(n)}$ in equation (15), can be neglected. And the diagonal elements of the approximate Hessian, the auto-correlation between the source-side Green's functions and receiver-side Green's functions, can serve as a good preconditioner for the gradient to deblur the image. Furthermore, the inverse approximate Hessian can be approximated by the reciprocals of the approximate Hessian $\frac{1}{\bar{H}_a^{(n)}}$ (Ben-Hadj-Ali et al., 1989), which is called as the pseudo-inverse of the approximate Hessian. Thus, the approximate Hessian works like a deconvolution operator to compensate the geometrical spreading effects (Margrave et al., 2011). And equation (5) can be expressible as:

$$\delta s_0^{(n)} = -\mu^{(n)} H^{(n)-} g^{(n)} \simeq -\mu^{(n)} \frac{g^{(n)}}{\bar{H}_a^{(n)}}, \tag{16}$$

And preconditioning the image using the Hessian matrix is also considered to be equivalent to applying a poststack deconvolution to the image and the resolution of the image can be improved after a number of iterations (Hu et al., 2001; Dai et al., 2012). If we examine the diagonal elements of the approximate Hessian further, as shown in Fig.5 of Pratt et al. (1998), we can find that the values of elements along the diagonal line of the Hessian matrix vanish from top left corner to bottom right corner, which indicates the nonstationarity of the deconvolution operator.

For a finite range frequency, the structure of the approximate Hessian matrix is banded, for that the partial derivative wavefields from adjacent nodes are correlated to some extent. The off-diagonal elements account for the limited-bandwidth effects and can improve the resolution of the image. In this condition, the diagonal elements of the approximate Hessian have limited effects in preconditioning the gradient, especially for the areas with poor illumination (Tang, 2009). So, some researchers (Albertin et al., 2004; Valenciano, 2008) computed a limited number of off-diagonal elements of the approximate Hessian for better migration or inversion results.

*Pseudo-Hessian*

Under the assumption of the infinite receiver coverage, the receiver-side Green's functions can be approximated as $d^2(\mathbf{r}_g, \mathbf{r}')$, where $d(\mathbf{r}_g, \mathbf{r}')$ means the distance from position $\mathbf{r}'$ to the receiver position $\mathbf{r}_g$. When the depth in vertical is quite smaller than the distance in horizontal, $d^2(\mathbf{r}_g, \mathbf{r}')$ can be approximated as a constant scalar $L$. This approximation to Hessian matrix is equivalent to the pseudo-Hessian proposed by Shin et al. (2001b). The pseudo-Hessian is constructed using the forward modeling wavefields, used as the virtual sources in reverse rime migration.

$$H_{p\_a}^{(n)} = L \sum_{\mathbf{r}_s} \int d\omega \omega^4 \Re \left\{ G\left(\mathbf{r}'', \mathbf{r}_s, \omega\right) G^*\left(\mathbf{r}', \mathbf{r}_s, \omega\right) \right\}, \tag{17}$$

The faint image can be enhanced effectively by the multiplying the inverse diagonal pseudo-Hessian (when $\mathbf{r}' = \mathbf{r}''$), which can be written as:

$$\bar{H}_{p\_a}^{(n)} = L \sum_{\mathbf{r}_s} \int d\omega \omega^4 \Re \left\{ G\left(\mathbf{r}', \mathbf{r}_s, \omega\right) G^*\left(\mathbf{r}', \mathbf{r}_s, \omega\right) \right\}, \tag{18}$$

If a source function is added into the diagonal pseudo-Hessian, we can recognize that the diagonal pseudo-Hessian is actually equivalent to the source illumination (as shown in APPENDIX B) and the gradient divided by the source illumination is equivalent to a standard deconvolution imaging condition in prestack reverse time migration. Thus, the model perturbation in equation (5) becomes (Jang et al., 2009):

$$\begin{aligned} \delta s_0^{(n)} &\simeq -\mu^{(n)} \left( \bar{H}_{p\_a}^{(n)} + \lambda I \right)^{-1} g^{(n)} \\ &\simeq -\mu^{(n)} \frac{\sum_{\mathbf{r}_s} \sum_{\mathbf{r}_g} \int d\omega \Re\left\{ \omega^2 \mathcal{F}_s(\omega) G(\mathbf{r}, \mathbf{r}_s, \omega) G(\mathbf{r}_g, \mathbf{r}, \omega) \delta P^*(\mathbf{r}_g, \mathbf{r}_s, \omega) \right\}}{L \sum_{\mathbf{r}_s} \int d\omega \omega^4 \Re\{ |\mathcal{F}_s(\omega)|^2 G(\mathbf{r}', \mathbf{r}_s, \omega) G^*(\mathbf{r}', \mathbf{r}_s, \omega)\} + \lambda I} \end{aligned}, \tag{19}$$

where $\lambda I$ in the denominator is the stable term which makes the deconvolution imaging condition stable, $I$ is an identity matrix. And a source wavelet function $\mathcal{F}_s(\omega)$ is added

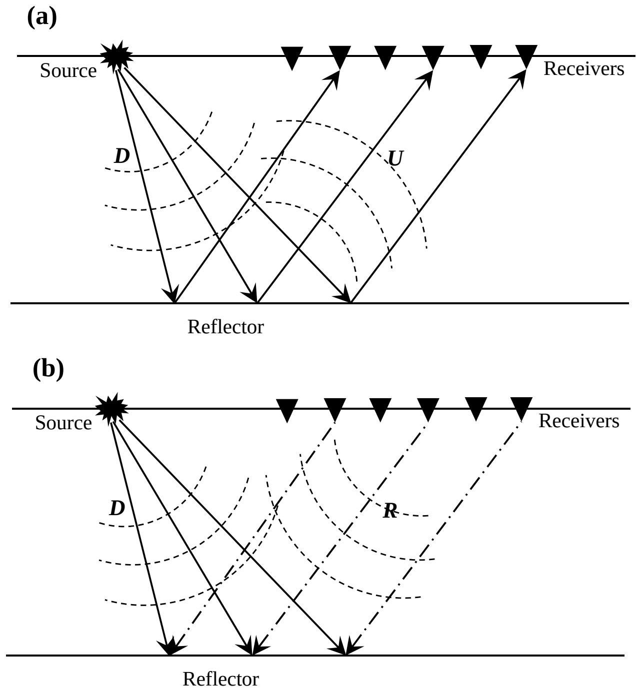


FIG. 1. Schematic diagrams for Reciprocity theory. $D$ indicate the downgoing wavefields from the source. $U$ indicate the upgoing wavefields, which can be interpreted as first order partial derivative wavefields (or first order scattered wavefields) caused by the virtual source. And $R$ is the reciprocal wavefields from the receivers. Based on reciprocal theory, $R$ is considered to be identical to $U$.

into equation (19) to make sure that it is the bandwidth limited. Equation (19) becomes the standard deconvolution imaging condition in reverse time migration.

For finite receiver coverage, this approximation is limited to compensate the geometrical spreading effects and balance the amplitudes, for missing the receiver-side Green's functions. By introducing the reciprocity theory, we can get a better approximation (Shin et al., 2001b). The reciprocity principle proved by Aki and Richards (2002) in an elastic media, states that the recorded wavefields are identical when interchanging the locations of sources and receivers, as shown in Fig. 2.

In this condition, the computation of the partial derivative wavefields doesn't depend on the number of model parameters but depends on the number of sources and receivers (Sheen et al., 2006). Then we can get an improved pseudo-Hessian matrix:

$$H_{im\_pseudo\_a} = \sum_{\mathbf{r}_s} \int d\omega \Re \left\{ \omega^4 G(\mathbf{r}', \mathbf{r}_s, \omega) G^*(\mathbf{r}'', \mathbf{r}_s, \omega) \right\} \sum_{\mathbf{r}_g} \int d\omega \Re \left\{ \tilde{G}(\mathbf{r}', \mathbf{r}_g, \omega) \tilde{G}^*(\mathbf{r}'', \mathbf{r}_s, \omega) \right\}, \tag{20}$$

where $\tilde{G}$ mean the reciprocal wavefields from the receivers.We can get the diagonal part of the improved pseudo-Hessian, when $\mathbf{r}' = \mathbf{r}''$:

$$\bar{H}_{im\_pseudo\_a} = \sum_{\mathbf{r}_s} \int d\omega \Re \left\{ \omega^4 G(\mathbf{r}', \mathbf{r}_s, \omega) G^*(\mathbf{r}', \mathbf{r}_s, \omega) \right\} \sum_{\mathbf{r}_g} \int d\omega \Re \left\{ \tilde{G}(\mathbf{r}', \mathbf{r}_g, \omega) \tilde{G}^*(\mathbf{r}', \mathbf{r}_g, \omega) \right\}, \tag{21}$$

In equation (21) the Green's function $G(\mathbf{r}', \mathbf{r}_s, \omega)G^*(\mathbf{r}', \mathbf{r}_s, \omega)$ serve as the source-side illumination to compensate the geometrical spreading effects on the source side and the reciprocal wavefields and the reciprocal wavefields $\widetilde{G}(\mathbf{r}', \mathbf{r}_g, \omega)\widetilde{G}^*(\mathbf{r}', \mathbf{r}_g, \omega)$ server as the receiver-side illumination to compensate the geometrical spreading effects on the receiver side. Comparing with the pseudo-Hessian matrix, this deconvolution operator takes the receiver-side illumination into consideration and can balance the amplitude better for that its nonstationarity becomes stronger. The source illumination can be obtained in the step of gradient calculation. While we still need $n$ (the number of the receivers) forward simulations to calculate the reciprocal wavefields using an impulse response source in frequency domain (or a band limited source function in time domain) (Sheen et al., 2006), which is also a computationally demanding task.

In this research, we assume that if the lateral velocity variation is small, we can select some of the receivers regularly in the whole acquisition geometry to calculate the reciprocal wavefields, which can reduce the computation cost greatly. A further assumption is that the selected receivers and the sources are collocated. In this condition, the reciprocal wavefields can be substituted by the forward modeling wavefields and we don't need any additional forward simulations to calculate the Hessian matrix. The improved pseudo-Hessian and its diagonal elements are shown as equation (22) and (23) respectively:

$$\begin{aligned} H_{im\_pseudo\_a} &= \sum_{\mathbf{r}_s} \int d\omega \Re \left\{ \omega^4 G(\mathbf{r}', \mathbf{r}_s, \omega) G^*(\mathbf{r}', \mathbf{r}_s, \omega) G(\mathbf{r}'', \ddot{\mathbf{r}}_g, \omega) G^*(\mathbf{r}'', \ddot{\mathbf{r}}_g, \omega) \right\} \\ &\simeq \sum_{\mathbf{r}_s} \int d\omega \Re \left\{ \omega^4 G(\mathbf{r}', \mathbf{r}_s, \omega) G^*(\mathbf{r}', \mathbf{r}_s, \omega) G(\mathbf{r}'', \mathbf{r}_s, \omega) G^*(\mathbf{r}'', \mathbf{r}_s, \omega) \right\}, \end{aligned} \tag{22}$$

$$\bar{H}_{im\_pseudo\_a} = \sum_{\mathbf{r}_s} \int d\omega \Re \left\{ \omega^4 G(\mathbf{r}, \mathbf{r}_s, \omega) G^*(\mathbf{r}, \mathbf{r}_s, \omega) G(\mathbf{r}, \mathbf{r}_s, \omega) G^*(\mathbf{r}, \mathbf{r}_s, \omega) \right\}, \tag{23}$$

where $\ddot{\mathbf{r}}_g$ in equation (22) are the selected receiver locations. So, using equation (22) to substitute the deconvolution operator term in the denominator of equation (19) forms improved deconvolutional gradient:

$$g(\mathbf{r}) = \frac{\sum_{\mathbf{r}_s} \int d\omega \Re \left\{ \omega^2 \mathcal{F}_s(\omega) G(\mathbf{r}, \mathbf{r}_s, \omega) G(\mathbf{r}_g, \mathbf{r}, \omega) \delta P^*(\omega) \right\}}{\bar{H}_{im\_pseudo\_a} + \lambda I}, \tag{24}$$

Because the deconvolution operator $\bar{H}_{im\_pseudo\_a}$ in the denominator is actually equivalent to double source illumination (Pan et al., 2013c), this strategy for improving the pseudo-Hessian is defined as double source illumination method.

*Phase Encoded Hessian*

Another strategy to improve the pseudo-Hessian is to construct the receiver-side Green's functions using phase encoding technique, which can reduce the computational cost considerably but unfortunately involve strong crosstalk artifacts. Tang (2009) calculated the phase encoded Hessian and compared effects of linear phase encoding and random phase encoding approaches in attenuating the crosstalk artifacts. Furthermore, the phase encoding technique can be used for imaging (Liu et al., 2006; Perrone and Sava, 2012) and constructing the gradient in FWI (Krebs et al., 2009; Vigh and Starr, 2008; Ben-Hadj-Ali et al., 2011). Tao and Sen (2013) used the plane wave linear phase encoding approach

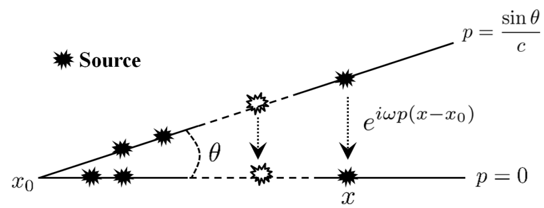


FIG. 2. Linear phase encoding strategy. The phase shift $e^{i\omega p(x-x_0)}$ is controlled by ray parameter $p$ and source position $x$. And the ray parameter $p$ is a function of take-off angle $\theta$ and top surface velocity $c$. This diagram is after Zhang et al. (2005) and Dai and Schuster (2013)

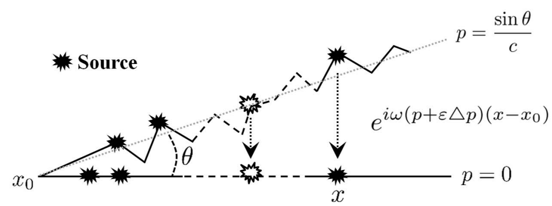


FIG. 3. Chirp phase encoding strategy. Comparing with linear phase encoding strategy, there is one random term $\varepsilon\triangle p$ added into the phase shift. And $\varepsilon$ is random scalar to control the ray parameter perturbation $\triangle p$.

to calculate the gradient for efficient full waveform inversion in frequency-ray parameter domain. The crosstalk artifacts can be reduced effectively with sufficient source and receiver ray parameters. In this research, we used a chirp phase encoding strategy (Perrone and Sava, 2012), which is a combination of the linear phase encoding and random phase encoding methods, to calculate the receiver-side Green's functions. Because, the linear time shift in time domain corresponds to linear phase delay in frequency domain, the linear phase encoded Hessian can be written as:

$$\begin{aligned} H_{linear\_encoded} = \; & \sum_{\mathbf{r}_s} \int d\omega \Re \left\{ \omega^4 G(\mathbf{r}', \mathbf{r}_s, \omega) G^*(\mathbf{r}'', \mathbf{r}_s, \omega) \right\} \\ & \times \sum_{\mathbf{p}_g} \int d\omega \Re \left\{ G(\mathbf{r}'', \mathbf{r}'_g, \omega) e^{i\omega p_g (x'_g - x_{initial})} G^*(\mathbf{r}'', \mathbf{r}_g, \omega) e^{-i\omega p_g (x_g - x_{initial})} \right\}, \end{aligned} \tag{25}$$

where $\mathbf{p}_g$ are the receiver-side ray parameters vector, and:

$$x_{initial} = \begin{cases} x_0, p_g^i \geq 0 \\ x_{max}, p_g^i < 0 \end{cases}, \tag{26}$$

where $x_0$ and $x_{max}$ are the positions of the start and the end receivers, $i$ indicates ray parameter index.

In the second term of equation (25), when $x'_g = x_g$ , the phase encoded Hessian becomes $H_{exact}$ which is the exact approximate without crosstalk noise. And when $x'_g \neq x_g$, only crosstalk artifacts term $H_{crosstalk}$ left. So, equation (25) can be written as the summation of the exact Hessian and crosstalk artifacts (Liu et al., 2006; Tang, 2009; Tao and Sen, 2013):

$$H_{linear\_encoded} = H_{exact} + H_{crosstalk}, \tag{27}$$

And if we apply integration over ray parameter from $-\infty$ to $+\infty$, the crosstalk noise in equation (27) can be dispersed completely (Liu et al., 2006; Tao and Sen, 2013).

The chirp phase encoding strategy used in this research is a combination of linear phase encoding and random phase encoding methods. And it is similar to Perrone and Sava's dithered phasing encoding strategy by Perrone and Sava (2012). A random factor is added into the phase delay term of equation (25), which gives:

$$\begin{aligned} H_{chirp\_encoded} = & \sum_{\mathbf{r}_s} \int d\omega \Re \left\{ \omega^4 G(\mathbf{r}', \mathbf{r}_s, \omega) G^*(\mathbf{r}', \mathbf{r}_s, \omega \right\} \\ & \times \sum_{\mathbf{p}_g} \int d\omega \Re \left\{ G(\mathbf{r}'', \mathbf{r}'_g, \omega) G^*(\mathbf{r}'', \mathbf{r}_g, \omega) e^{i\omega(p_g + \varepsilon \triangle p)(x'_g - x_g)} \right\} \end{aligned}, \tag{28}$$

where $\triangle p = rand * p_g^i, rand \in (0, 1)$ is the random coefficient and i means the number of the ray parameter, $\varepsilon$ is the coefficient used to control the amount of dithering. Chirp phase encoding strategy is expected to reduce the crosstalk noise better than the linear phase encoding strategy with the same number of simulations.

## Multiscale approach

FWI suffers from cycle skipping problem because of the poor starting model and the lack of low frequency data in piratical application. The cycle skipping problem resulted from the high nonlinearity of the least-squares inverse problem may results in the miss of the global minimum, as indicated by Fig.4. Inversion proceeds in a multiscale ap-

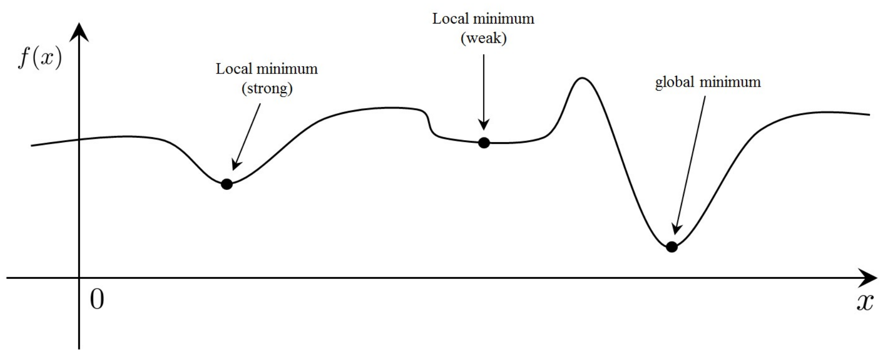


FIG. 4. Diagram indicating local minimum and global minimum.

proach from lower to higher frequencies to improve the chances that the global minimum is reached and avoid the local minimum (Vigh and Starr, 2008). Multiscale approach can be implemented in frequency domain by selecting a few frequencies. A time domain multiscale approach by applying a low-pass filtering to the data residual is employed in this research.It has been demonstrated that a multiscale method is computationally efficient and converges faster than a conventional, single-scale method (Boonyasiriwat et al., 2009). The low frequency is responsible to catch $s(\mathbf{r})^{low}$ the low wavenumber component of the velocity model. And the high frequency is used to add $s(\mathbf{r})^{high}$ the detailed information to the velocity model.

$$s(\mathbf{r}) = s(\mathbf{r})^{low} + s(\mathbf{r})^{high} \tag{29}$$

## Pseudo Gauss-Newton Step

To reduce the computational cost further, we proposed to use one ray parameter in one FWI iteration but change the ray parameter regularly for different iterations. And

Table1. Pseudo code for PGN method

---

**BEGIN** $\leftarrow s_0$, initial model;
**WHILE** $\epsilon \leq \epsilon_{min}$ or $n \leq n_{max}$
  Identify the ray parameter $p_s^{(n)}$
  Identify the frequency band $f^{(n)} = f_0 \rightarrow f_{max}$, $f_{interval}$, every $k$ iterations
  Generate the data residual $\delta P$ and apply low-pass filtering $\delta \tilde{P} = \textbf{low\_pass}\left(\delta P, f^{(n)}\right)$
  Generate the linear phase encoded gradient $g^{(n)}\left(p_s^{(n)}\right)$
  **FOR** $i = 1$ to $\mathbf{p}_s^H, \mathbf{p}_r^H$, every 1 or $m$ iterations
    Construct the diagonal part of the hybrid phase encoded Hessian $diag\left(H_{chirp\_encoded}^{(n)}\right)$
  **END FOR**
  Calculate the step length $\mu^{(n)}$ using the line search method
  update the velocity model:
    $s^{(n+1)}(\mathbf{r}) = s^{(n)}(\mathbf{r}) - \mu^{(n)}\left\{diag\left(H_{chirp\_encoded}^{(n)}\right) + \lambda I\right\}^{-1} g^{(n)}\left(p_s^{(n)}\right)$
  Calculate the relative least-squares error:
    $\epsilon = \frac{\|s^{(n)}(\mathbf{r}) - s^{true}(\mathbf{r})\|_2}{\|s^{true}(\mathbf{r})\|_2}$
**END WHILE**

---

preconditioning the gradient using the diagonal part of the chirp phase encoded Hessian forms one pseudo Gauss-Newton step in PGN method, as indicated by equation (30):

$$\delta s\left(\mathbf{r}\right) = \frac{\int d\omega \Re\left\{\omega^2 \mathcal{F}_s(\omega)\tilde{G}(\mathbf{r}, \mathbf{r}_s, \omega)G(\mathbf{r}_g, \mathbf{r}, \omega)\delta P^*\right\}}{diag\left(H_{chirp\_encoded}\right) + \lambda I} \tag{30}$$

The pseudo code for the PGN method is presented in the Table. 1. For each PGN step, we need to identify the ray parameter and frequency band. And then construct the linear phase encoded gradient and diagonal part of the chirp phase encoded Hessian. And then calculate the step length and update the velocity model. And we use the relative least-squares error $\epsilon$ to evaluate the quality of the inverted model:

$$\epsilon = \frac{\parallel s^{(n)}(\mathbf{r}) - s^{true}(\mathbf{r}) \parallel_2}{\parallel s^{true}(\mathbf{r}) \parallel_2}, \tag{31}$$

where $s^{(n)}(\mathbf{r})$ and $s^{true}(\mathbf{r})$ indicate the inverted slowness and true slowness respectively.

### Computational cost comparison for different strategies

We compared the computation costs for traditional Gauss-Newton(TGN) method, source encoding Gauss-Newton method(SEGN) and pseudo Gauss-Newton(PGN) method, as presented in Table. 2. We can notice that the PGN method is more efficient than TGN and SEGN methods. PGN method only needs 2 simulations to calculate the gradient in one iteration. While the number of simulations needed to construct the gradient in TGN and

Table2. Computational cost comparison for different strategies

| Methods | Gradient | $H_a$ | $diag\,(H_{encoded})$ | Step length | Cost for one iteration |
|---|---|---|---|---|---|
| TGN Method | $2N_s$ | $N_s \times N_r$ | \ | 1 | $2N_s + N_s \times N_r + 1$ |
| SEGN Method | $2N_p^g$ | \ | $N_{ps}^H + N_{pr}^H$ | 1 | $2N_p^g + N_{ps}^H + N_{pr}^H + 1$ |
| PGN Method | 2 | \ | $N_{ps}^H + N_{pr}^H$ | 1 | $N_{ps}^H + N_{pr}^H + 3$ |

SEGN methods are $2N_s$ and $2N_p^g$ respectively, where $N_s$ and $N_p^g$ are the number of sources and ray parameters. And in TGN method, we need $N_s \times N_g$ simulations to calculate the approximate Hessian. While in SEGN and PGN method,we only need $N_{ps}^H + N_{pr}^H$ simulations to calculate the diagonal part of the phase encoded Hessian. $N_{ps}^H$ and $N_{pr}^H$ mean the numbers of the source-side ray parameters and receiver-side ray parameters respectively.

## NUMERICAL EXPERIMENTS

In this section, several numerical examples are illustrated for analysis and discussion. The first numerical example is a single interface model with one source-receiver pair. An analysis of the contributions to the gradient or RTM image is performed based on this simple model. The second numerical example is a homogeneous model with a constant background velocity. The exact Hessian, Hessian contaminated by the crosstalk artifacts, pseudo-Hessian and phase encoded Hessian are presented for comparison. The final numerical example for FWI is performed based on the modified Marmousi model. The effects for fixed ray parameter and varied ray parameters are analyzed. We also analyzed the sensitivities to the ray parameter range, the number of encoded sources, and Gaussian noise. Finally, we compared the FWI results for different scaling methods.

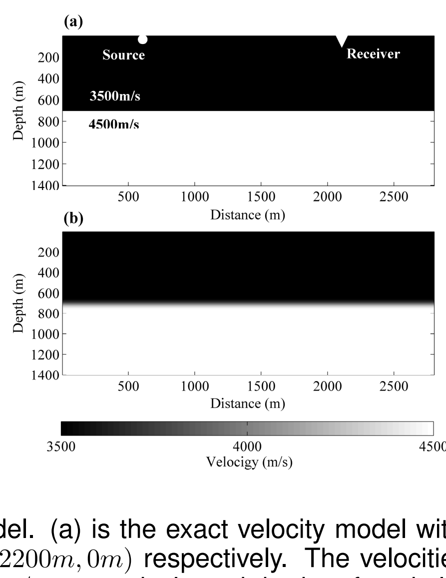


FIG. 5. Single interface model. (a) is the exact velocity model with one source and receiver pair located at $(700m, 0m)$ and $(2200m, 0m)$ respectively. The velocities of the first layer and second layer are $3500m/s$ and $4500m/s$ respectively and the interface is located at $700m$ in depth. (b) is the reference velocity model smoothed by a Gaussian function.

## Gradient Contribution Analysis

Fig. 5a and Fig. 5b show the true single interface model and slightly smoothed velocity model using Gaussian function. The source and receiver are located at $700m$ and $2200m$ on the top surface. A Ricker wavelet with dominant frequency of $15Hz$ is used as the source function.

Fig. 6a is the RTM image by migrating the measured data using the exact velocity model. It can be seen that besides the exact image part what we want, the image is contaminated by the artifacts caused by directive wave and scattered wave. Fig. 6b is the gradient formed by migrating the data residuals using the smoothed velocity model, which is equivalent to the first iteration in FWI. The main difference between Fig. 6a and Fig. 6b is the direct wave contribution, as shown by Fig. 6c. The low frequency artifacts in Fig. 6a are equivalent to the spurious parts of the gradient in Fig. 6b.

Fig. 6d shows the contribution of the unperturbed wavefields. It is also the gradient when the reference velocity model is simplified to a homogeneous background model (Xu et al., 2012). Fig. 6e and f are the scattered wavefields contributions to the gradient and they formed by convolving the backpropagated data residuals and the scattered wavefields caused by the reflector. The gradients in Fig. 6b preconditioned by the source-side Green's functions and receiver-side Green's functions are shown in Fig. 6g and Fig. 6h respectively. Fig. 6i shows the valuable part of the gradient in Fig. 5b. The pseudo-Hessian only contains the source-side Green's functions, which is not enough to precondition the gradient. Hence, the receiver-side Green's functions should also be taken into consideration, as we discussed above. Fig 6j shows the gradient preconditioned by the diagonal part of the approximate Hessian. We can see that the spurious part of the gradient is suppressed. And we can extract vertical lines from Fig. 6b and j for comparison as indicated by the blue and red lines in Fig 7. It can be seen that the gradient preconditioned by the diagonal Hessian is more sharp. This means that the Hessian matrix can suppress the multiple scattering effects and improve the resolution of the gradient in least-squares inverse problem.

## Comparison of Hessian Approximations

Fig.8 shows the homogeneous model with a constant velocity of $2500m/s$ used to calculate the Hessian matrix. The model consists of $50 \times 50$ grid cells with $5m$ horizontal and vertical grid intervals, which means that the total number of the parameters is $2500$. The sources and receivers are co-located at the top surface of the model and several acquisition geometries are designed to construct the Hessian. The source function is a Ricker wavelet with a $25Hz$ dominant frequency banlimited between $0$ and $30Hz$. The first acquisition geometry is designed with one source located at $(125m, 0m)$ and two receivers located at $(75m, 0m)$ and $(175m, 0m)$ respectively. Fig. 9a shows the exact approximate Hessian with the parameters from $1000$ to $2500$. A detailed view of the area delineated by the black square is given in the right corner. We can see that the approximate Hessian is dominated by the band along the diagonal line. Furthermore, the energy of the band vanishes from the top-left corner to the bottom-right corner, which indicates the nonstationarity of the diagonal Hessian. So, multiplying the gradient with the inverse Hessian is equivalent to applying an inverse Q filtering to the gradient and the amplitude of the deep parts of the gradient can

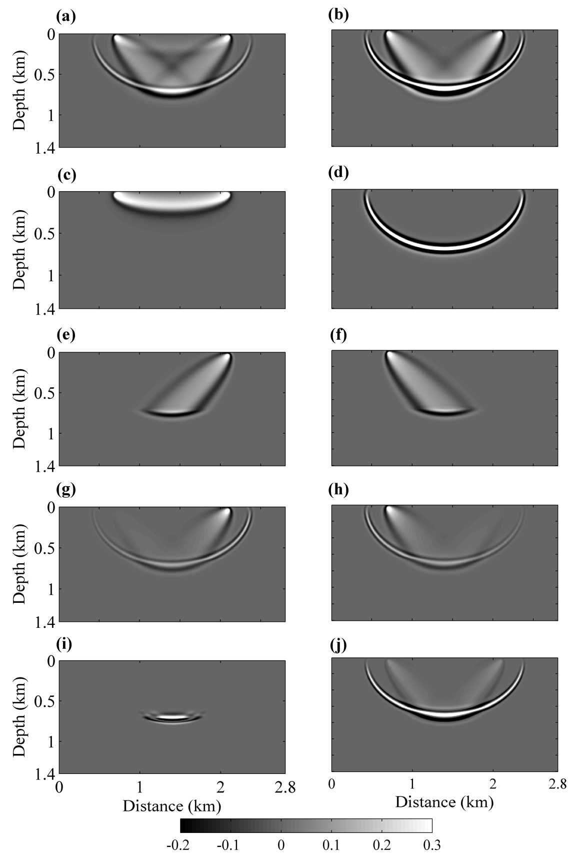


FIG. 6. Gradient contribution analysis.(a) and (b) are the RTM image and gradient respectively. (c) is the directive wave contribution. (d) is the unperturbed wavefield contribution. (e) and (f) are the scattered wavefield contribution. (g) and (h) are the gradients preconditioned by the source-side and receiver-side Green's functions respectively. (i) shows the valuable part of the gradient. (j) is the gradient preconditioned by the diagonal approximate Hessian.

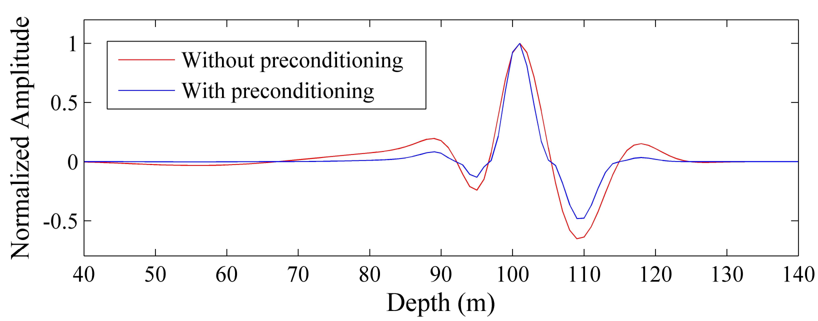


FIG. 7. Comparison of the gradients with and without precondition. The red line and blue line indicate the gradient with and without precondition respectively.

be enhanced.

The approximate Hessian in Fig. 9b is formed by constructing the receiver-side Green's function with the ray parameter $p = 0$. This approximate Hessian is contaminated by the crosstalk artifacts. Fig. 9c shows the approximate Hessian formed using the linear phase encoding method and the ray parameter range is $[-0.3s/km, 0.3s/km]$ with a step of $0.1s/km$. It can be seen that the linear phase encoding method can effectively eliminate

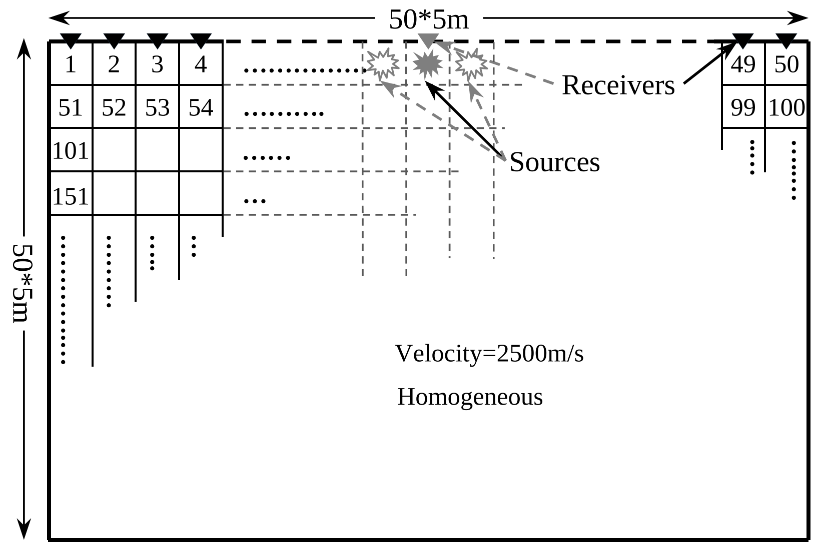


FIG. 8. The homogeneous model used to calculate the Hessian matrix. This model consists of $50 * 50 = 2500$ grid cells with a grid interval of $5m$. The sources and receivers are located at the top surface.

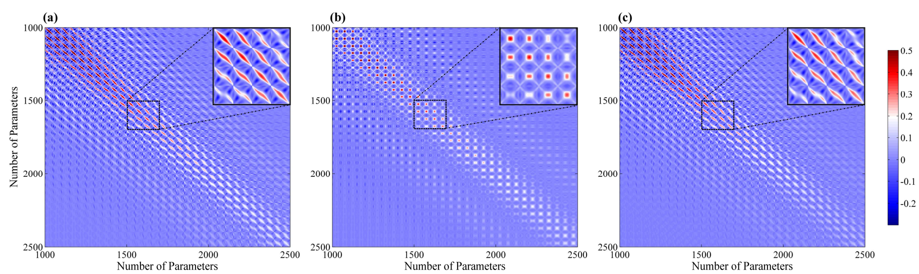


FIG. 9. A Comparison of the approximate Hessian. (a) The exact approximate Hessian; (b) The Hessian contaminated by crosstalk artifacts; (c) The Hessian constructed using the linear phase encoding method and the ray parameter range is $[-0.3s/km, 0.3s/km]$ with a step of $0.1s/km$.

the crosstalk artifacts and the phase encoded Hessian is very close to the exact approximate Hessian, as shown by Fig. 9a.

Then the acquisition geometries are designed with multi-sources and receivers. Fig. 10a shows the exact approximate Hessian. The single source location is $(125m, 0m)$. While we have $50$ receivers arranged from $0$ to $2500m$ with a $25m$ spacing. Fig. 10b shows the exact approximate Hessian with the same receiver arrangement but $9$ sources arranged from $25m$ to $225m$ with a $25m$ spacing.

Even though the exact approximate Hessian with one source looks similar to the exact approximate Hessian with multi-sources, we can still recognize the slight difference through the detailed view. It is noted that the energy of the diagonal part of the Hessian with multi-sources is more concentrated than that of the Hessian with single source.

Fig. 11a and Fig. 11b show the pseudo-Hessian matrices of single source and multi-

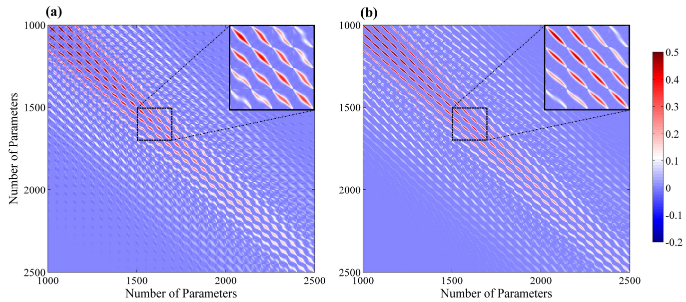


FIG. 10. The exact approximate Hessian with different sources arrangements. (a) is exact approximate Hessian formed by one source and $50$ receivers. (b) is exact approximate Hessian formed by $9$ sources and $50$ receivers.

sources respectively. Comparing them with Fig. 10a and Fig. 10b, it can be recognized that the energy of the pseudo-Hessian is less concentrated on the band parts and the energy decrease slower along diagonal line. The pseudo-Hessian is a crude approximation to the exact diagonal of the Hessian because it assumes receiver-side Green's functions as an constant and ignores the effects of the limited receiver aperture (Tang, 2009).

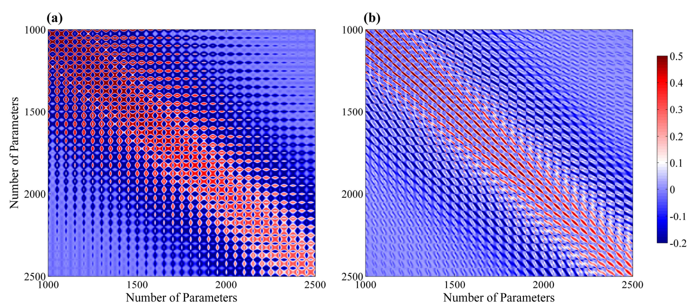


FIG. 11. The pseudo-Hessian with different sources arrangements. (a) is the pseudo-Hessian formed by $1$ source and $50$ receivers. (b) is the pseudo-Hessian formed by $9$ sources and $50$ receivers.

We constructed the Hessian approximations using the double-illumination method by Plessix and Mulder (2006) (Fig. 12a and e), linear phase encoding method (Fig. 12b, c, f and g) and chirp phase encoding method (Fig. 12d and h). For Fig. 10b and f, the approximate Hessian matrices were constructed using the linear phase encoding technique with the ray parameter $p = 0$. The crosstalk artifacts contaminate the approximate Hessian obviously. Fig. 12c and g are the improved pseudo-Hessian matrices using the linear phase encoding strategy and the ray parameters range from $-0.3s/km$ to $0.3s/km$ with a step of $0.1s/km$. Fig. 12d and h are calculated by stacking the same range of ray parameters but using the chirp phase encoding strategy. Both of these two phase encoding methods can disperse the crosstalk noise effectively and reduce the computation burden.

We also calculate the error matrices of the Hessian contaminated by crosstalk artifacts, linear phase encoded Hessian and chirp phase encoded Hessian, as indicated by Fig. 13a, b and c respectively. And we extracted the diagonal lines of these error matrices, as shown in Fig. 14. We can notice that the red line is more close to $0$, compared to the blue line. So, the chirp phase encoding strategy can reduce the crosstalk noise better than the linear

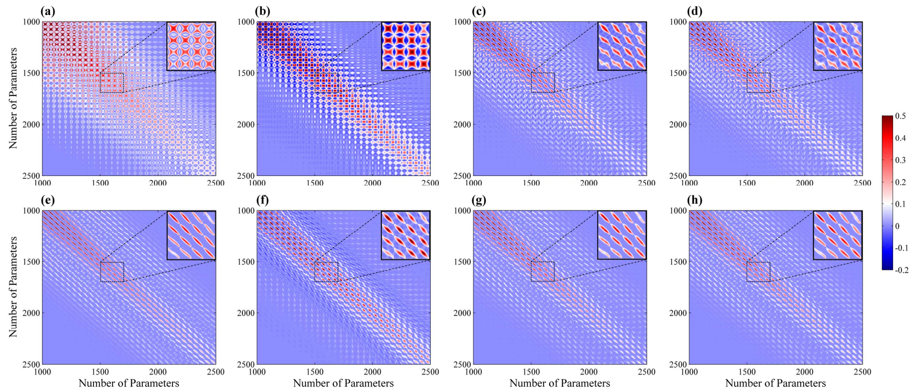


FIG. 12. The Hessian matrices with different strategies and sources arrangements. (a), (b), (c) and (d) are the Hessian matrices for one source and $50$ receivers. (e), (f), (g) and (h) are the Hessian matrices for $9$ source and $50$ receivers. (a) and (e) are the Hessian matrices based on the double source illumination method. (b) and (f) are the Hessian matrices using linear phase encoding strategy with ray parameter $p = 0$. (c) and (g) are the Hessian matrices using the linear phase encoding strategy and the ray parameters range from $-0.3s/km$ to $0.3s/km$ with a step of $0.1s/km$. (d) and (h) are the Hessian matrices using the chirp phase encoding strategy and the ray parameters range from $-0.3s/km$ to $0.3s/km$ with a step of $0.1s/km$.

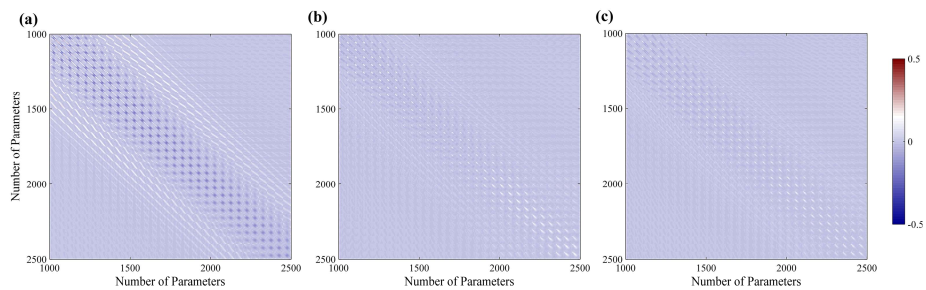


FIG. 13. Error matrices for different Hessian approximations. (a) shows the error matrix of the Hessian contaminated by crosstalk noise. (b) is the error matrix of the linear phase encoded Hessian. (c) is the error matrix of the chirp phase encoded Hessian.

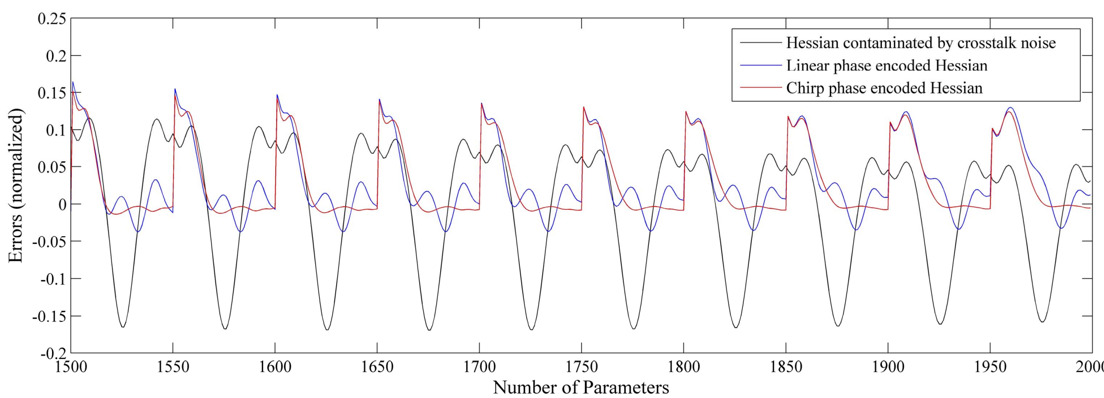


FIG. 14. Diagonal part of the error matrices. The black line is diagonal error matrix of the Hessian contaminated by crosstalk noise. The blue line is diagonal error matrix of the linear phase encoded Hessian. The red line is diagonal error matrix of the linear phase encoded Hessian.

phase encoding method with the same number of simulations.

Here, we present another numerical example to show how the phase encoding method can disperse the crosstalk artifacts. The velocity model is homogeneous with a velocity of $2500m/s$. One source is arranged at $(1km, 0km)$ and four receivers are arranged on both sides of the source. Fig. 15a is the exact diagonal Hessian. Fig. 15b is the diagonal part of phase encoded Hessian when ray parameter $p = 0$. The crosstalk artifacts are very obvious. Fig. 15c shows the diagonal part of the phase encoded Hessian with 7 simulations when integrating the ray parameter from $-0.3s/km$ to $0.3s/km$ with a step of $0.1s/km$. Fig. 15d shows the diagonal part of the phase encoded Hessian with 14 simulations when integrating the ray parameter from $-0.3s/km$ to $0.3s/km$ with a step of $0.05s/km$. It can be seen that with increasing the number of simulations, the crosstalk artifacts are dispersed effectively.

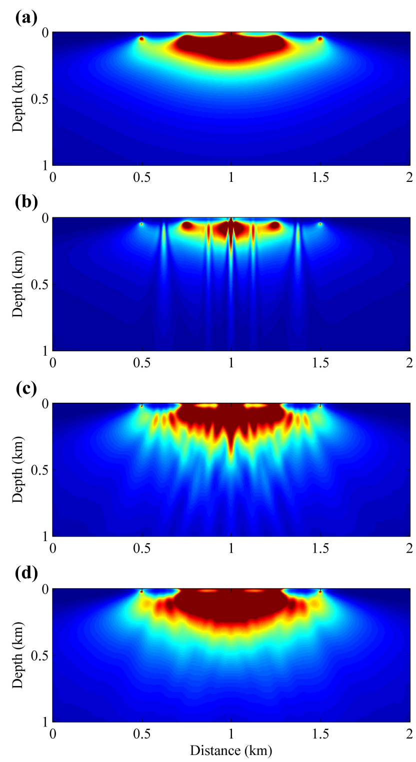


FIG. 15. (a) is the exact diagonal Hessian. (b) shows the diagonal Hessian contaminated by the crosstalk noise. (c) is the diagonal phase encoded Hessian with $7$ simulations. (d) is the diagonal phase encoded Hessian with $14$ simulations.

## Numerical Example for Full Waveform Inversion

The strategies proposed above are practiced and applied on a modified Marmousi Model. The Marmousi model is modified by introducing one water layer with a thickness of $70m$ and a velocity of $1500m/s$. And the model has $180 \times 767$ grid cells with the same grid interval of $5m$ in horizontal and vertical. $380$ point sources are distributed on the surface with a source interval of $10m$ from 0 to $3800m$ and $767$ receivers are deployed on the surface with a receiver interval of $5m$ from $5m$ to $5750m$. The source function is a Ricker wavelet with a dominant frequency of $30Hz$. The ray parameter range used for linear phase encoding method is $[-0.3s/km, 0.3s/km]$ with a step of $0.1s/km$ are tested for comparison. The lowest frequency band used is $[0Hz, 5Hz]$ and the frequency band increases by $5Hz$ for every 10 iterations. Fig.16a shows the exact P-wave velocity model and Fig.16b shows the smoothed P-wave velocity model used as the reference model in this research.

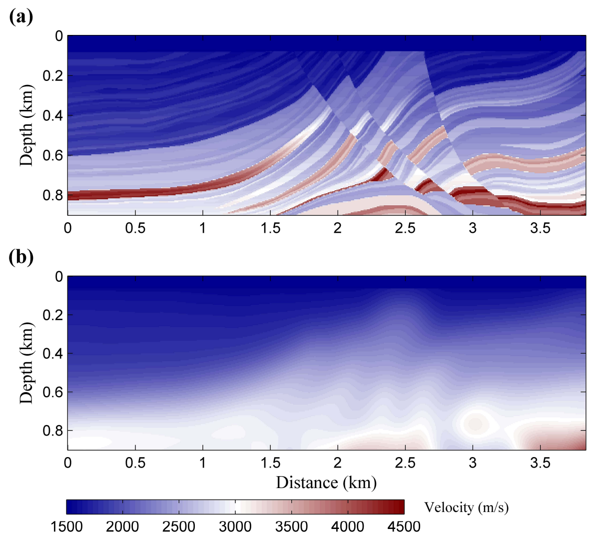


FIG. 16. (a) is true velocity model; (b) is the initial velocity model.

*Fixed and varied ray parameter effects*

Fig.17 shows the FWI results after 50 iterations by preconditioned method with fixed ray-parameter and varied ray parameter in each iteration. Fig.16a shows the result when ray parameter was fixed at $p = 0s/km$. Fig.16b shows the result when ray parameter was fixed at $p = -0.2s/km$. Fig.16c shows the result when ray parameter was fixed at $p = 0.2s/km$. It can be seen that the updated model is not balanced with just one fixed ray-parameter during iterations. Some noise or anomalies are obvious. Fig.16d shows the result when we changed the ray parameters from $-0.3s/km$ to $0.3s/km$ with a step of $0.1s/km$ during the iterations. We can see that the update is balanced and the updated model is much more better. Fig. 18 shows the relative least-squares errors for different ray parameter arrangements. It can seen that varying ray parameter during iterations can provide a higher quality inversion result after 50 iterations, as indicated by the black cycle line.

*Sensitivity to the ray parameter range*

Because the ray parameter is the function of the take-off angle and top surface velocity, different ray parameters are responsible to update the subsurface layers with different dip angles. We compared the inversion results after 50 iterations with different ray parameter ranges as shown in Fig. 19. Fig. 19a shows the inversion result when the ray parameter ranges from $-0.1s/km$ to $0.1s/km$ with a step of $0.05s/km$. We can see that because the ray parameter range is too small, the subsurface layers with dip angles cannot be up-

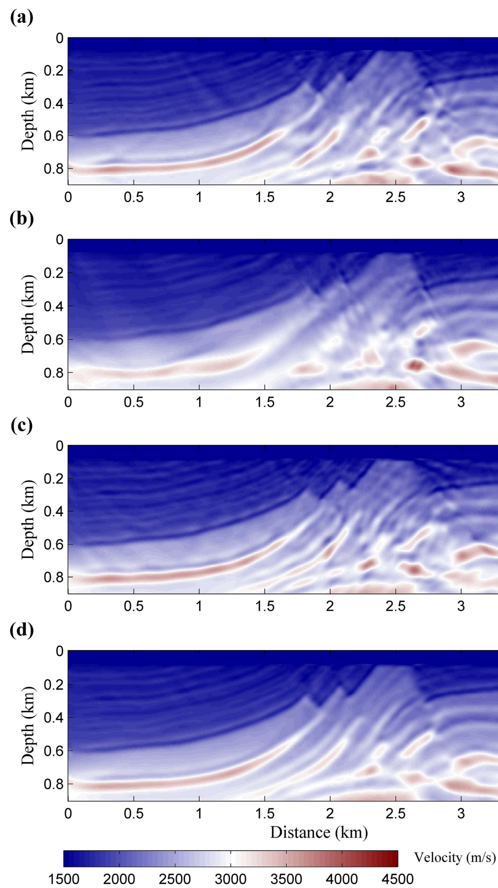


FIG. 17. FWI results after $50$ iterations with single ray parameter in each FWI iteration. (a) ray parameter is fixed at $p = 0s/km$; (b) ray parameter is fixed at $p = -0.2s/km$; (c) ray parameter is fixed at $p = 0.2s/km$; (d) ray parameter varies from $-0.3s/km$ to $0.3s/km$ with a step of $0.1s/km$.

dated in balance. Fig. 19b and c show the inverted results when the ray parameter ranges are $[-0.4s/km, 0.4s/km]$ with a step of $0.1s/km$ and $[-0.6s/km, 0.6s/km]$ with a step of $0.2s/km$ respectively. It can be seen that if the ray parameter range is too large, the convergence rate can be decreased.

*Sensitivity to the number of encoded sources*

The phase encoding technique can reduce the computational cost effectively but unfortunately introduce unwanted crossterms between unrelated shot and receiver wavefields. Generally, slant stacking over sufficient ray parameters can disperse or shift these crossterms reasonably. In theory, when the shot coordinates are sparsely distributed, the crosstalk noise becomes more serious with increasing the number of the encoded sources (Romero et al., 2000). While when the shot locations densely cover the imaging area, each plane-wave encoded gradient exhibits a limited amount of noise (Liu et al., 2006).

We also analyzed the sensitivity to the number of encoded sources. Fig. 20 shows the inversion results when the source number $N_s$ is $38$, $76$ and $350$, as indicated by Fig. 20a, c and d respectively. Fig. 20b is the inversion result when the source number is also $38$. But

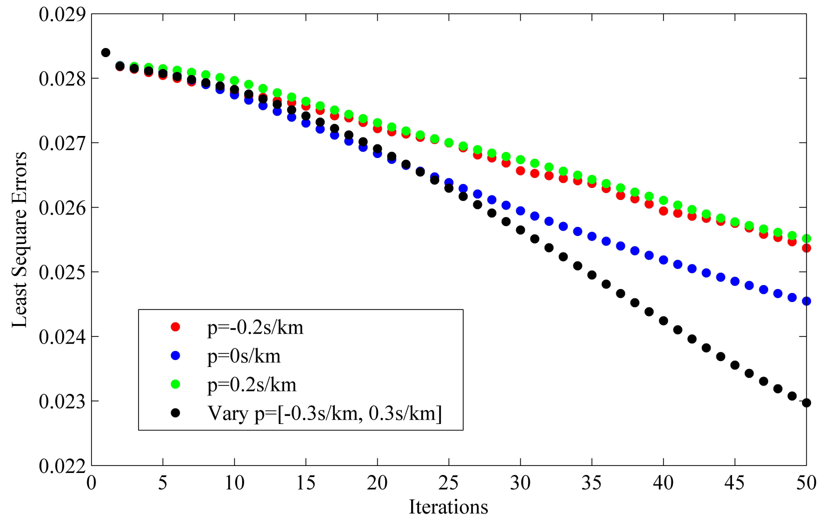


FIG. 18. Relative least-squares errors comparison for different ray parameter settings. The red cycle line, blue cycle line and green cycle line indicate the least squares errors when the ray parameter $p$ is fixed at $-0.2s/km$, $0s/km$ and $0.2s/km$ respectively. And the black cycle line is the least squares errors when varying the ray parameter from $-0.3s/km$ to $0.3s/km$ with a spacing of $0.1s/km$.

we applied an integration over ray parameter from $-0.2s/km$ to $0.2s/km$ with a step of $0.1s/km$. For Fig. 20a, we can note that the crosstalk artifacts are very obvious especially for the shallow layers. This is because the interferences between unrelated source and receiver wavefields. And if we apply integration over ray parameter in each FWI iteration, the crosstalk artifacts can be reduced effectively but more expensive. With increasing the number of sources, as shown in Fig.20c and d, the crosstalk noise becomes weaker and the convergence rate can also be increased. Fig. 21 shows the vertical lines extracted from Fig. 20a, c and d at $0.5km$ and $3km$ for comparison. It is easy for us to observe that the inversion results with $350$ sources (the green lines) approach the true velocity model better (the bold black lines).

*Sensitivity to Gaussian noise*

We also considered the influence of the Gaussian noise to the inversion results. Fig. 22a, b, c and d show the data residuals with Gaussian noise. The signal to noise ratios (SNR) are $5$, $4$, $3$, and $2$ respectively. Fig. 23a, b, c and d show the corresponding inverted results. It can be seen that with increasing the strength of the random noise, the inverted results become worse and the convergence rate can also be decrease. Some small layers in the shallow parts of the velocity model cannot be inverted. But we can still recognize the bigger structures in the deep parts of the velocity model.

*Comparison of different scaling methods*

We compared the inversion results of the gradient based method, preconditioned by the diagonal pseudo-Hessian, preconditioned by the diagonal chirp phase encoded Hessian. Fig. 24a shows the exact diagonal Hessian using the true velocity model. Fig. 24b shows

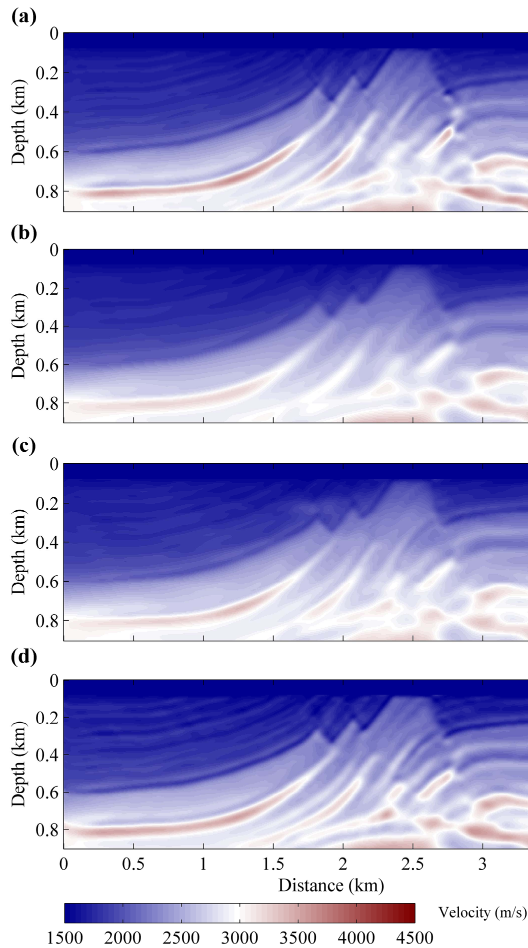


FIG. 19. Sensitivity to the ray parameter range. (a) ray parameter range is $[-0.1s/km, 0.1s/km]$ with s step of $0.05s/km$; (b) ray parameter range is $[-0.4s/km, 0.4s/km]$ with a step of $0.1s/km$; (c) ray parameter range is $[-0.6s/km, 0.6s/km]$ with a step of $0.2s/km$; (d) ray parameter range is $[-0.3s/km, 0.3s/km]$ with a step of $0.1s/km$.

the diagonal part of the pseudo-Hessian. We can notice that the pseudo-Hessian overestimates the energy. Fig. 24c and d show the diagonal parts of linear phase encoded Hessian and chirp phase encoded Hessian. We can see that the linear phase encoded Hessian and chirp phase encoded Hessian can approach the exact Hessian very well but introduce slight noise. Fig. 25 shows the diagonal parts of the Hessian using the initial velocity model. Fig. 25a is the exact diagonal Hessian. Fig. 25b shows the diagonal part of pseudo-Hessian. Fig. 25c and d are the diagonal parts of linear phase encoded Hessian constructed by $14$ and $28$ simulations respectively. Fig. 25e and f are the diagonal parts of chirp phase encoded Hessian constructed by $14$ and $28$ simulations respectively. We can see that from Fig. 25c to Fig. 25d, the crosstalk noise are reduced with increasing the number of simulations. While the crosstalk artifacts are still very obvious. While for the diagonal parts of the chirp phase encoded Hessian, the crosstalk artifacts are reduced more effectively in Fig. 25e and f.

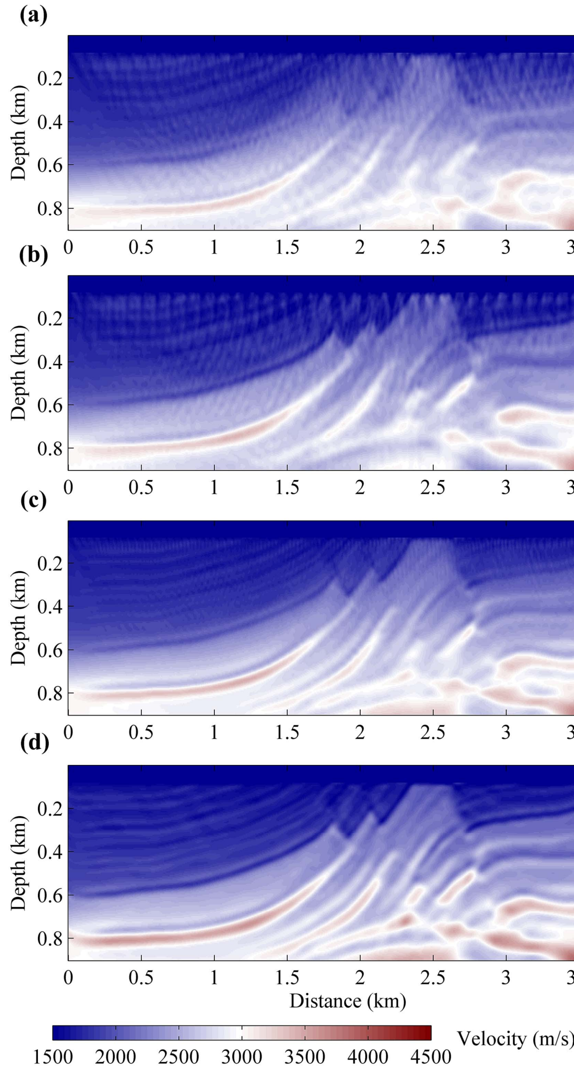


FIG. 20. Sensitivity to the number of encoded sources. (a) source number $N_s = 38$; (b) source number $N_s = 38$ with integration over ray parameter from $-0.2s/km$ to $0.2s/km$ in each FWI iteration; (c) source number $N_s = 76$; (d) source number $N_s = 350$.

And then we performed full waveform inversion for $200$ iterations using gradient based method, with diagonal pseudo-Hessian precondition and with diagonal chirp phase encoded Hessian respectively. Fig. 26 shows the inversion results for different scaling strategies. Fig. 26a is the true velocity model. Fig. 26b shows the inversion result of the gradient based method. Fig. 26c shows the inversion result with the diagonal pseudo Hessian precondition. And Fig. 26d is the inversion result with diagonal chirp phase encoded Hessian precondition. We can see that the gradient based method can not invert the velocity model very well especially for the deep reflectors. The inversion results with diagonal pseudo-Hessian precondition and diagonal phase encoded Hessian precondition are much better. We extract the vertical lines from Fig. 26b, c and d at $0.5km$ and $3.25km$ respectively, as indicated by the Fig. 27.

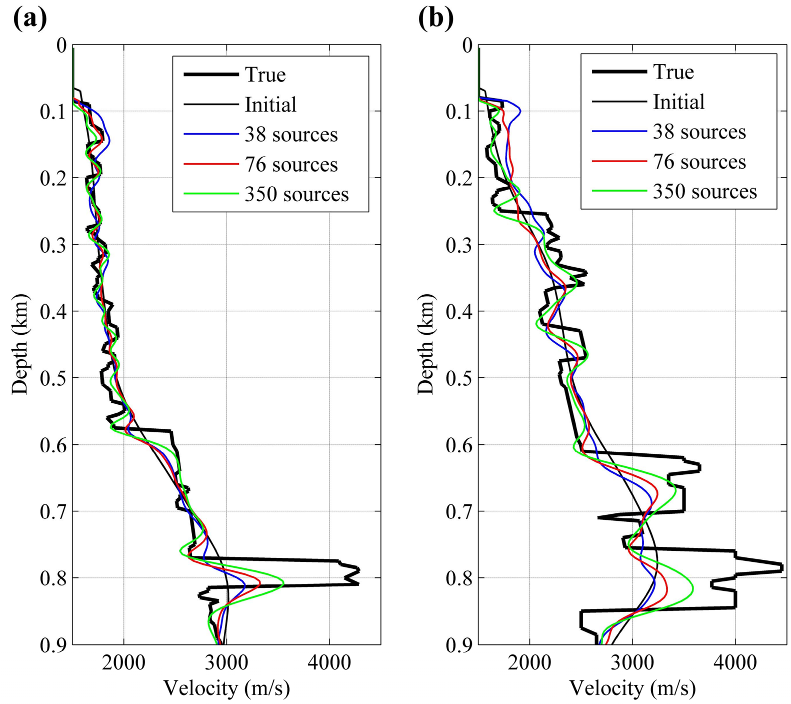


FIG. 21. Inversion results comparison for different number of encoded sources at $0.5km$ (a) and $3km$ (b) respectively. The bold black lines are the true velocity model. The thin black lines are the initial velocity model. The blue, red and green lines are the inversion results when the source number is $38$, $76$ and $350$ respectively.

### Imaging Results Comparison

We produced the reverse time images using the initial velocity model, true velocity model and inverted velocity model respectively for comparison. Fig. 28a shows the migrated image using the initial velocity model. Fig. 28b shows the migrated image using the true velocity model. We can see that the image using the initial velocity model is seriously distorted and the positions of the subsurface reflectors are shifted. Fig. 29a shows the image by the inverted velocity model in Fig. 26d. We can see that the image using the inverted velocity model is very close the image using the true velocity model.

## CONCLUSION

From what we have discussed above, several conclusions can be achieved: (1) Hessian matrix server as a nonstationary deconvolution operator to improve the convergence rate of least-squares inverse problem; (2) Varying ray-parameter during iterations can reduce the computational cost further and balance the model update; (3) If the ray-parameter range is too small, the layers with dip angles cannot be inverted in balance, if the ray-parameter range is too large, the convergence rate will be decreased; (4) If the assembled sources are not dense enough, the crosstalk noise will be very obvious, especially for shallow layers; (5) Chirp phase encoding strategy can reduce the crosstalk noise better than linear phase encoding strategy with the same number of simulations; (6) Diagonal part of the phase encoded Hessian can server as a good approximation of the Hessian to precondition

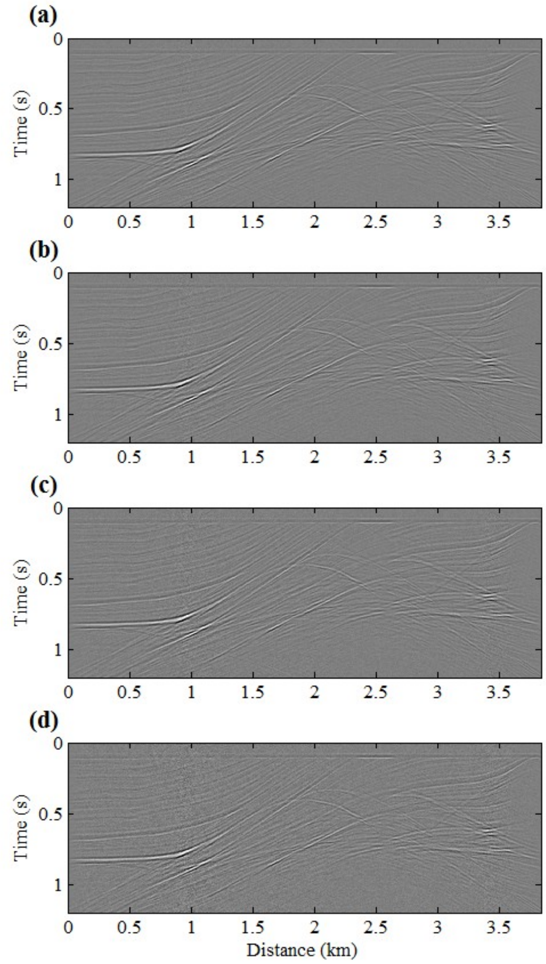


FIG. 22. The data residuals with Gaussian noise when SNRs are 5(a),4(b),3(c),2(d) respectively.

the gradient and increase the convergence rate; (7) Full waveform inversion with source encoding is efficient for numerical modeling but will increase difficulties in seismic data acquisition and preprocessing.

## ACKNOWLEDGMENTS

This research was supported by the Consortium for Research in Elastic Wave Exploration Seismology (CREWES). And I wish to thank Danping Cao,Joe Wang, Vladimir Zubov, Raul Cova, Babatunde Arenrin's discussions with me.

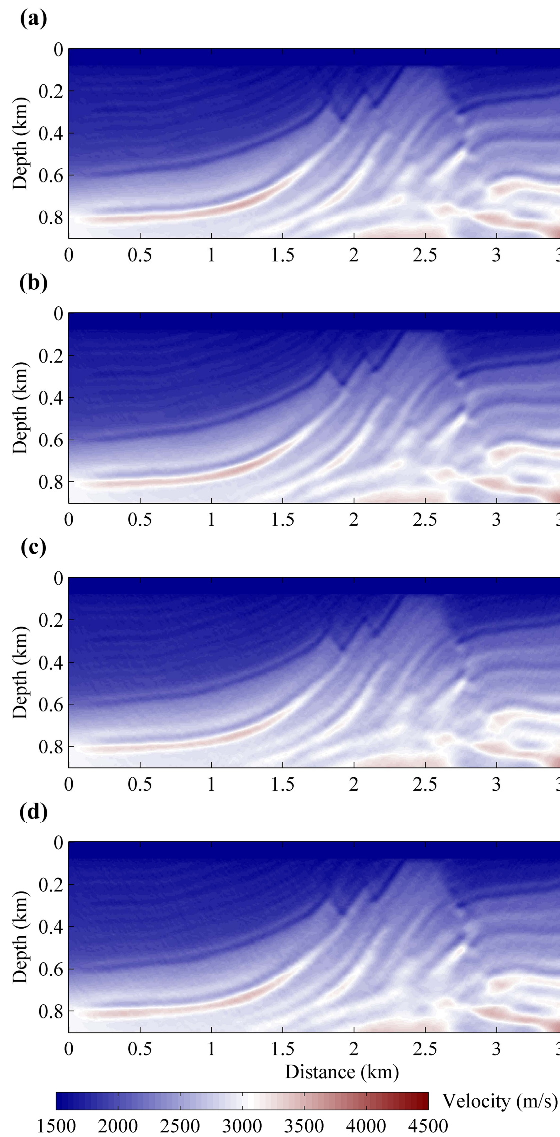


FIG. 23. The inverted results when $SNR = 5$ (a), $SNR = 4$ (b), $SNR = 3$ (c) and $SNR = 2$ (d) respectively.

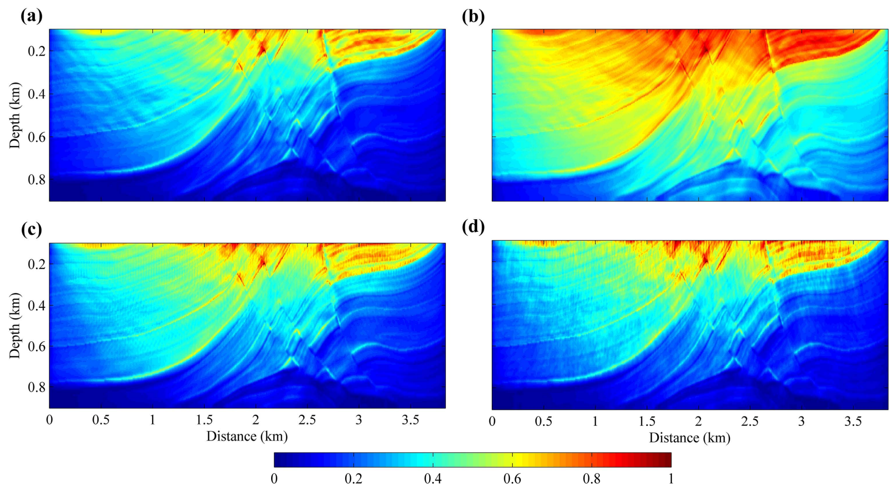


FIG. 24. Diagonal Hessian comparison by the true velocity model. (a) the exact diagonal Hessian; (b) the diagonal part of the pseudo-Hessian; (c) the diagonal part of the linear phase encoded Hessian with $35$ simulations; (d) the diagonal part of the chirp phase encoded Hessian with $35$ simulations.

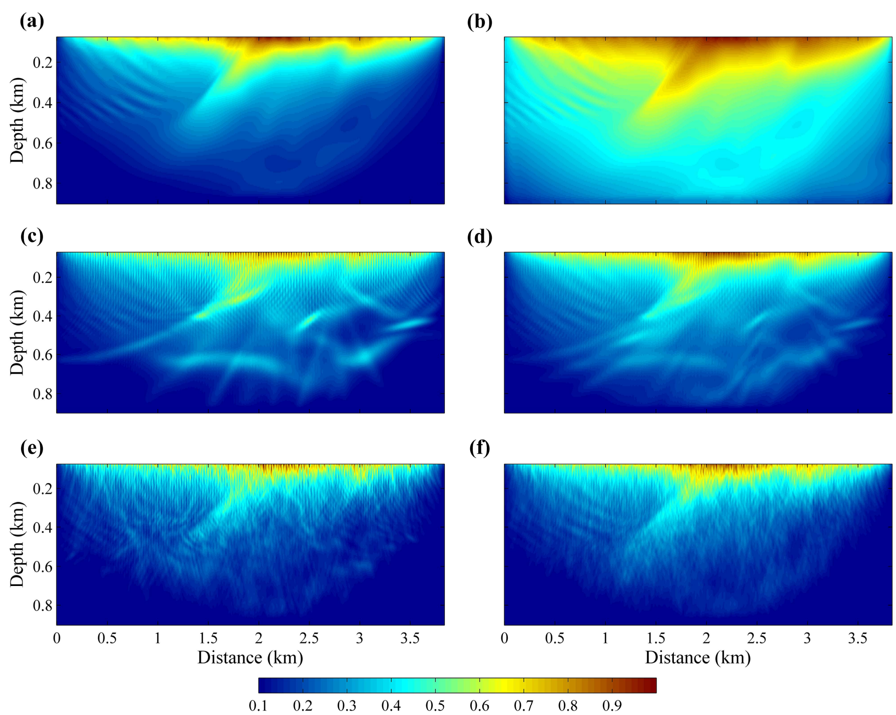


FIG. 25. Diagonal Hessian comparison by the initial velocity model. (a) the exact diagonal Hessian; (b) the diagonal part of the pseudo-Hessian; (c) and (d) are the diagonal parts of the linear phase encoded Hessian with $14$ and $28$ simulations respectively. (e) and (f) are the diagonal parts of the chirp phase encoded Hessian with $14$ and $28$ simulations respectively.

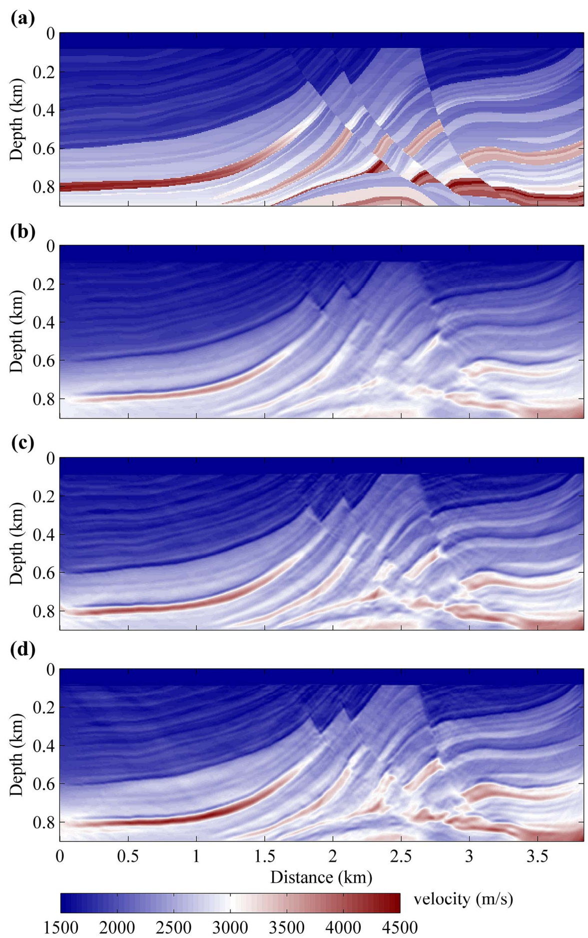


FIG. 26. FWI results after 200 iterations for different scaling methods. (a)True velocity model;(b) Inversion result gradient based method; (c) Inversion result with diagonal pseudo-Hessian precondition; (d) Inversion result with diagonal chirp phase encoded Hessian precondition.

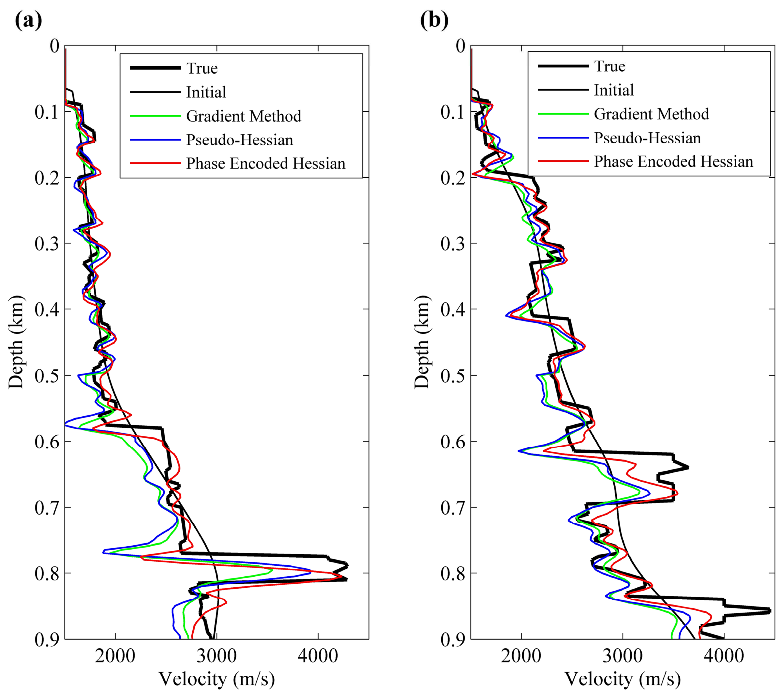


FIG. 27. Inversion results comparison for different scaling methods at $0.5km$ (a) and $3.25km$ (b) respectively. The bold black lines are the true velocity model. The thin black lines are the initial velocity model. The green, blue and red lines indicate the inversion results by gradient based method, with diagonal pseudo Hessian precondition and with diagonal chirp phase encoded Hessian precondition respectively.

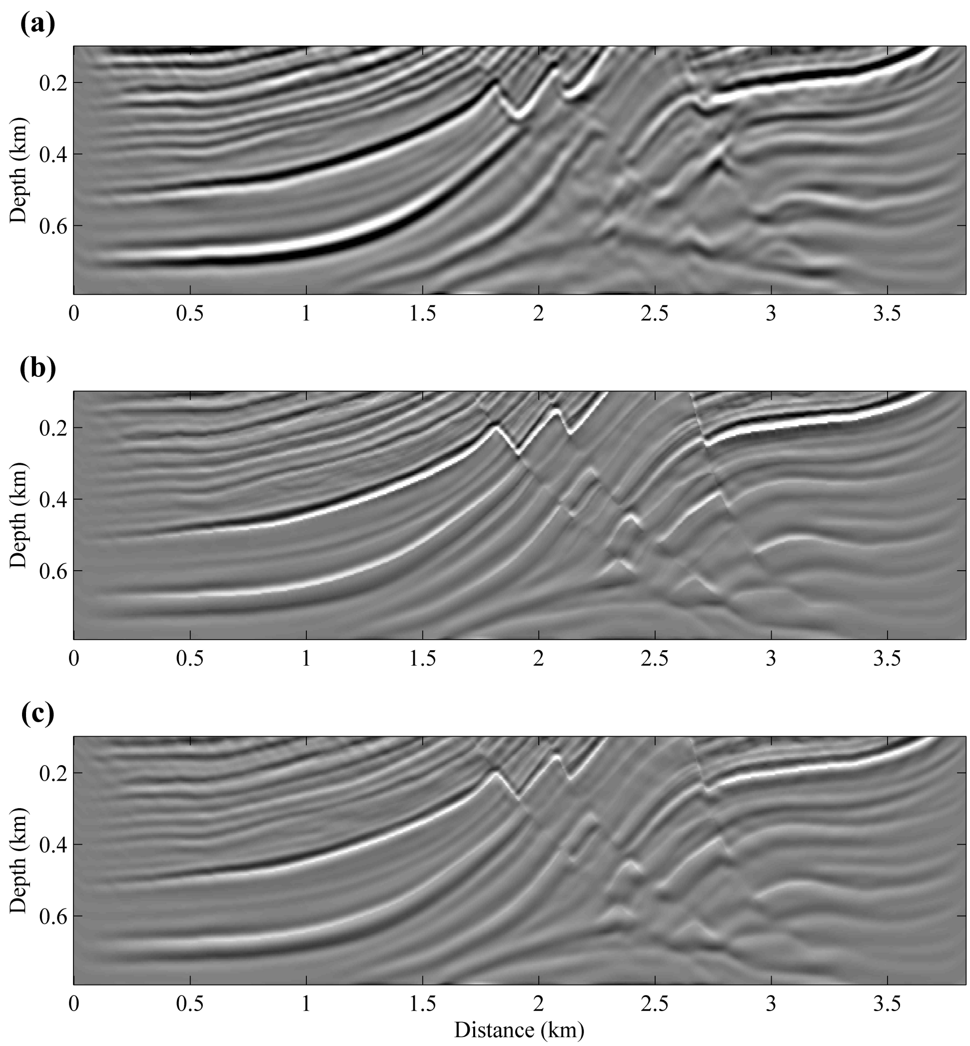


FIG. 28. Reverse time migration images comparison.(a) is the reverse time migration image using the initial velocity model; (b) is the reverse time migration image by the true velocity model. (c) is the reverse time migration image by the inverted velocity model.

## APPENDIX A - THE DERIVATION OF PARTIAL DERIVATIVE WAVEFIELDS BASED ON BORN SERIES TRUNCATION

For the partial derivative wavefields with respect to the model parameters:

$$\frac{\delta G\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)}{\delta s_0^{(n)}(\mathbf{r})} = \frac{G_1\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right) - G_0\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)}{\delta s_0^{(n)}(\mathbf{r})}, \tag{32}$$

where $G_1\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)$ and $G_0\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)$ are the perturbed and unperturbed wavefields respectively. And $G_1$ and $G_0$ satisfy the following wave equations (Innanen, 2009):

$$\left[\nabla^2 + \omega^2 s_0^{(n)}(\mathbf{r})\right] G_0\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right) = \delta(\mathbf{r}_g - \mathbf{r}_s), \tag{33}$$

$$\left[\nabla^2 + \omega^2 s_0^{(n)}(\mathbf{r}) + \omega^2 \delta s_0^{(n)}(\mathbf{r})\right] G_1\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right) = \delta(\mathbf{r}_g - \mathbf{r}_s), \tag{34}$$

Combing equations (A-2) and (A-3), we can isolate $G_1\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)$ on the right hand side and get the Lippmann-Schwinger equation (Newton, 1966; Taylor, 1972; Stolt and Weglein, 2012):

$$G_1\left(\mathbf{r}, \mathbf{r}_s, \omega | s_0^{(n)}\right) = G_0\left(\mathbf{r}, \mathbf{r}_s, \omega | s_0^{(n)}\right) - \int d\mathbf{r}' G_0\left(\mathbf{r}', \mathbf{r}_s, \omega | s_0^{(n)}\right) V G_1\left(\mathbf{r}', \mathbf{r}_s, \omega | s_0^{(n)}\right), \tag{35}$$

where $V = \omega^2 \delta s_0^{(n)}(\mathbf{r}')$ is the scattering potential (Innanen and Weglein, 2007), the difference between the exact and approximate wave modeling operators. Equation (A-4) can be expanded as a power series in the quantity of $G_0 V$:

$$G_1 = (G_0 V)^0 G_0 - (G_0 V)^1 G_0 + (G_0 V)^2 G_0 - (G_0 V)^3 G_0 + \ldots, \tag{36}$$

If the scattering potential is small or more precisely, the norm of operator $G_0 V$ is smaller than 1, the Born series can be truncated by neglecting the high order terms, which gives the Born approximation to the Lippmann-Schwinger equation (Stolt and Weglein, 2012):

$$G_1 \simeq G_0 - G_0 V G_0, \tag{37}$$

This Born approximation can be viewed as a linear term in Lippmann-Schwinger equation and the error is acceptable when model perturbation $\delta s_0^{(n)}(\mathbf{r}')$ or $G_0 V$ is small (Stolt and Weglein, 2012). If we move $G_0$ in the right hand side of equation (A-6) to the left hand side, the wavefields perturbation can be written as:

$$\delta G = G_1 - G_0 = -G_0 V G_0 = -\omega^2 G_0 \delta s_0 G_0, \tag{38}$$

Then the partial derivative wavefields in equation (A-1) becomes:

$$\frac{\delta G\left(\mathbf{r}_g, \mathbf{r}_s, \omega | s_0^{(n)}\right)}{\delta s_0^{(n)}(\mathbf{r})} = -\omega^2 G_0\left(\mathbf{r}_g, \mathbf{r}, \omega | s_0^{(n)}\right) G_0\left(\mathbf{r}, \mathbf{r}_s, \omega | s_0^{(n)}\right), \tag{39}$$

## APPENDIX B - PROOF OF THE EQUIVALENCE BETWEEN SOURCE ILLUMINATION AND DIAGONAL PART OF THE PSEUDO-HESSIAN

Firstly, we can re-examine the acoustic wave equation:

$$L_0^{(n)}\left(\mathbf{r}_g,\mathbf{r},\omega|s_0^{(n)}\right)G\left(\mathbf{r}_g,\mathbf{r}_s,\omega|s_0^{(n)}\right)=\mathcal{F}_s(\omega)\delta(\mathbf{r}_g-\mathbf{r}_s), \tag{40}$$

where $L_0^{(n)}\left(\mathbf{r}_g,\mathbf{r},\omega|s_0^{(n)}\right)=\left(\nabla^2+\omega^2 s_0^{(n)}(\mathbf{r}_g)\right)$ is the wave modeling operator, $\nabla^2$ is the Laplace operator, $\mathcal{F}_s(\omega)\delta(\mathbf{r}_g-\mathbf{r}_s)$ is the source term. And then take partial derivative with respect to model parameters on both sides of the equation (B-1):

$$\begin{aligned} L_0^{(n)}\left(\mathbf{r}_g,\mathbf{r},\omega|s_0^{(n)}\right)\frac{\partial G\left(\mathbf{r}_g,\mathbf{r}_s,\omega|s_0^{(n)}\right)}{\partial s_0^{(n)}(\mathbf{r})} &= -\frac{\partial L_0^{(n)}\left(\mathbf{r}_g,\mathbf{r}_s,\omega|s_0^{(n)}\right)}{\partial s_0^{(n)}(\mathbf{r})}G\left(\mathbf{r}_g,\mathbf{r}_s,\omega|s_0^{(n)}\right) \\ &= -\omega^2\mathcal{F}_s(\omega)\delta(\mathbf{r}-\mathbf{r}_g)G\left(\mathbf{r}_g,\mathbf{r}_s,\omega|s_0^{(n)}\right) \\ &= -\int d\mathbf{r}'\mathcal{F}_s(\omega)\delta(\mathbf{r}-\mathbf{r}')G\left(\mathbf{r}',\mathbf{r}_s,\omega|s_0^{(n)}\right) \end{aligned}, \tag{41}$$

According to the sifting property of delta function:

$$\begin{aligned} f_{virtual} &= L_0^{(n)}\left(\mathbf{r}_g,\mathbf{r},\omega|s_0^{(n)}\right)\frac{\partial G\left(\mathbf{r}_g,\mathbf{r}_s,\omega|s_0^{(n)}\right)}{\partial s_0^{(n)}(\mathbf{r})} \\ &= -\omega^2\mathcal{F}_s(\omega)G\left(\mathbf{r},\mathbf{r}_s,\omega|s_0^{(n)(\mathbf{r})}\right) \end{aligned}, \tag{42}$$

where $f_{virtual}$ is the virtual source (Pratt et al., 1998). The pseudo-Hessian proposed by Shin et al. (2001b) becomes:

$$H_{p\_a}^{(n)}=f_{virtual}f_{virtual}^*=\sum_{\mathbf{r}_s}\int d\omega\Re\{\omega^4|\mathcal{F}_s(\omega)|^2G(\mathbf{r}',\mathbf{r}_s,\omega)G^*(\mathbf{r}'',\mathbf{r}_s,\omega)\}, \tag{43}$$

And when $\mathbf{r}'=\mathbf{r}''$, the diagonal of the pseudo-Hessian is expressible as:

$$\bar{H}_{p\_a}^{(n)}=diag\left(f_{virtual}f_{virtual}^*\right)=\sum_{\mathbf{r}_s}\int d\omega\Re\{\omega^4|\mathcal{F}_s(\omega)|^2G(\mathbf{r},\mathbf{r}_s,\omega)G^*(\mathbf{r},\mathbf{r}_s,\omega)\}, \tag{44}$$

So, it can be seen that the diagonal part of pseudo-Hessian is actually equivalent to the source illumination, as shown in the denominator of equation (24).

## REFERENCES


Aki, K., and Richards, P. G., 2002, Quantitative Seismology: University Science Books, 2nd edn.

Albertin, U., Yingst, D., Kitchenside, P., and Tcheverda, V., 2004, True-amplitude beam migration, 949–952.

Ben-Hadj-Ali, H. S., Operto, S., and Virieux, J., 1989, Elastic ray-born l2-migration/inversion: Geophysical Journal International, **97**, 151–160.

Ben-Hadj-Ali, H. S., Operto, S., and Virieux, J., 2011, An efficient frequency-domain full waveform inversion method using simultaneous encoded sources: Geophysics, **76**, WCC177–WCC188.

Boonyasiriwat, C., Valasek, P., Routh, P., Cao, W., Schuster, G. T., and Macy, B., 2009, An efficient multiscale method for time-domain waveform tomography: Geophysics, **74**, WCC59–WCC68.

Dai, W., Fowler, P., and Schuster, G. T., 2012, Multi-source least-squares reverse time migration: Geophysical Prospecting, **60**, 681–695.

Dai, W., and Schuster, G. T., 2013, Plane-wave least-squares reverse-time migration: Geophysics, **78**, S165–S177.

Gao, F., Atle, A., and Williamson, P., 2010, Full waveform inversion using deterministic source encoding, 1013–1017.

Gauthier, O., Virieux, J., and Tarantola, A., 1986, Two-dimensional nonlinear inversion of seismic waveforms: numerical results: Geophysics, **51**, 1387–1403.

Hu, J., Schuster, G. T., and A.Valasek, P., 2001, Poststack migration deconvolution: Geophysics, **66**, 939–952.

Innanen, K. A., 2009, Born series forward modeling of seismic primary and multiple reflections: an inverse scattering shortcut: Geophysical Journal International, **177**, 1197–1204.

Innanen, K. A., and Weglein, A. B., 2007, On the construction of an absorptive-dispersive medium model via direct linear inversion of reflected seismic primaries: Inverse Problem, **23**, 2289–2310.

Jang, U., Min, D., and Shin, C., 2009, Comparison of scaling methods for waveform inversion: Geophysical Prospecting, **57**, 49–59.

Krebs, J. R., Anderson, J. E., Henkley, D., Neelamani, R., Lee, S., Baumstein, A., and Lacasse, M., 2009, Fast full-wavefield seismic inversion using encoded sources: Geophysical Prospecting, **74**, WCC177–WCC188.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migration, 206–220.

Liu, F., Whitemore, D. W., Day, R. S., and Stolt, R. H., 2006, Toward a unified analysis for source plane-wave migration: Geophysics, **71**, S129–S139.

Margrave, G. F., Ferguson, R. J., and Hogan, C. M., 2011, Full waveform inversion using wave-equation depth migration with tying to wells, 2454–2458.

Morton, S. A., and Ober, C. C., 1998, Faster shot-record depth migrations using phase encodingtest, 1131–1134.

Newton, R., 1966, Scattering Theory of Waves and Particles: New York:McGraw-Hill.

Pan, W., Innanen, K. A., and Margrave, G. F., 2013a, Efficient full waveform inversion with phase encoded pseudo-hessian, 1–15.

Pan, W., Innanen, K. A., and Margrave, G. F., 2013c, A comparison of different scaling methods for least-squares inversion/migration, 1–15.

Pan, W., Innanen, K. A., and Margrave, G. F., 2014, Efficient full waveform inversion in the time-ray parameter domain using iteration-dependent sets of ray parameters, 1–5.

Pan, W., Margrave, G. F., and Innanen, K. A., 2013b, On the role of the deconvolution imaging condition in full waveform inversion, 1–14.

Perrone, F., and Sava, P., 2012, Wave-equation migration with dithered plane waves: Geophysical Prospecting, **60**, 444–456.

Pica, A., Diet, J. P., and Tarantola, A., 1990, Nonlinear inversion of seismic reflection data in a laterally invariant medium: Geophysics, **55**, 284–292.

Plessix, R. E., and Mulder, W. A., 2006, Frequency-domain finite-difference amplitude-preserving migration: Geophysical Journal International, **157**, 975–987.

Pratt, R. G., Shin, C., and Hicks, G. J., 1998, Gauss-newton and full newton methods in frequency-space seismic waveform inversion: Geophysical Journal International, **133**, 341–362.

Romero, L. A., Ghiglia, D. C., Ober, C. C., and Morton, S. A., 2000, Phase encoding of shot records in prestack migration: Geophysics, **65**, 426–436.

Sheen, D. H., Tuncay, K., Baag, C. E., and Ortoleva, P. J., 2006, Time domain gauss-newton seismic waveform inversion in elastic media: Geophysical Journal International, **167**, 1373–1384.

Shin, C., Jang, S., and Min, M., 2001a, Improved amplitude preservation for prestack depth migration by inverse scattering theory: Geophysical Prospecting, **49**, 592–606.

Shin, C., Yoon, K., Marfurt, K. J., Park, K., Yang, D. W., Chung, S., and Shin, S., 2001b, Efficient calculation of a partial-derivative wavefield using reciprocity for seismic imaging and inversion: Geophysical Prospecting, **66**, 1856–1863.

Stolt, R. H., and Weglein, A. B., 2012, Seismic Imaging and Inversion: Application of Linear Inverse Theory: Cambriage University Press.

Stork, N., and Kapoor, J., 2004, How many p values do you want to migrate for delayed shot wave equation migration?, 1041–1044.

Tang, Y., 2009, Target-oriented wave-equation least-squares migration/inversion with phase-encoded hessian: Geophysics, **74**, WCA95–WCA107.

Tao, Y., and Sen, M. K., 2013, Frequency-domain full waveform inversion with plane-wave data: Geophysics, **78**, R13–R23.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266.

Taylor, J., 1972, Scattering Theory: The Quantum Theory of Nonrelativistic Collisions: New York: John Wiley and Sons.

Valenciano, A., 2008, Imaging by wave-equation inversion: Ph.D. thesis, Stanford University.

Vigh, D., and Starr, E. W., 2008, 3d prestack plane-wave, full-waveform inversion: Geophysics, **73**, VE135–VE144.

Virieux, A., and Operto, S., 2009, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **74**, WCC1–WCC26.

Xu, S., Wang, D., Chen, F., Zhang, Y., and Lambare, G., 2012, Full waveform inversion for reflected seismic data.

Zhang, Y., Sun, J., Notfors, C., Gray, S. H., Chernis, L., and Young, J., 2005, Delayed-shot 3d depth migration: Geophysics, **70**, E21–E28.