# Preconditioning for the Hessian-free Gauss-Newton full-waveform inversion

Wenyong Pan, Kris Innanen, Wenyuan Liao


## ABSTRACT

Full-waveform inversion (FWI) has emerged as a powerful strategy for estimating the subsurface model parameters by iteratively minimizing the difference between the synthetic data and observed data. The gradient-based methods promise to converge globally but suffer from slow convergence rate. The Newton-type methods provide a quadratic convergence, but the computation, storage and inversion of the Hessian are beyond the current computation ability for large-scale inverse problem. The Hessian-free (HF) optimization method represents an attractive alternative to these above-mentioned optimization methods. At each iteration, it obtains the search direction by approximately solving the Newton linear system using a conjugate-gradient (CG) algorithm with a matrix-free fashion. One problem of the HF optimization method is that the CG algorithm requires many iterations. The main goal of this paper is to accelerate the HF FWI by preconditioning the CG algorithm. In this research, different preconditioning schemes for the HF Gauss-Newton optimization method are developed. The preconditioners are designed as Hessian approximations (e.g., diagonal pseudo-Hessian and diagonal Gauss-Newton Hessian) or its inverse approximations. We also developed a new pseudo diagonal Gauss-Newton Hessian approximation for preconditioning based on the reciprocal property of the Green's function. Furthermore, a quasi-Newton $l$-BFGS inverse Hessian approximation preconditioner with the diagonal Hessian approximation as initial guess is proposed and developed. Several numerical examples are solved to demonstrate the effectiveness of the preconditioning schemes. It is concluded that the quasi-Newton $l$-BFGS preconditioning scheme with the pseudo diagonal Gauss-Newton Hessian as initial guess shows the best performances in speeding up the HF Gauss-Newton FWI, improving the convergence rate and reducing the computation burden.


## INTRODUCTION

Full-waveform inversion (FWI) promises to provide high-resolution estimates of the subsurface parameters, such as compressional and shear wave velocities, density, attenuation and anisotropic properties. FWI iteratively reconstructs the model parameters by minimizing a L2 norm misfit function, which measures the difference between the observed data and synthetic data (Lailly, 1983; Tarantola, 1984; Virieux and Operto, 2009).

The traditional optimization methods for FWI in exploration geophysics are gradient-based methods (e.g., steepest-descent (SD) and non-linear conjugate-gradient (NCG) methods). In SD method, the search direction is simply the negative of the gradient and in NCG method, the search direction is just the linear combination of the current gradient and previous search direction (Fletcher and Reeves, 1964; Nocedal and Wright, 2006; Hu et al., 2011). Within the adjoint-state method, the gradient of the misfit function can be calculated efficiently by applying a zero-lag cross-correlation between the forward modelled wavefield and back-propagated data residual wavefield (Pratt et al., 1998; Tromp et al.,

2005; Fichtner and Trampert, 2011). Thus, the gradient-based methods are computationally attractive for large-scale inverse problem but also suffer from slow convergence rate.

The second-order partial derivative of the misfit function (namely Hessian) carries crucial information in the reconstruction process (Santosa and Symes, 1988; Fichtner and Trampert, 2011). The search direction can be significantly enhanced by multiplying the gradient with the inverse Hessian matrix, which serves as a deconvolution operator for compensating the geometrical spreading effects and de-blurring the gradient (Pratt et al., 1998). Furthermore, the second-order term in the Hessian matrix that accounts for the non-linear scattering effects, can help to remove the second-order scattering artifacts in the gradient (Pratt et al., 1998; Pan et al., 2015b,c). However, explicit calculation, storage and inversion of the Hessian at each iteration are beyond the current computation capability for large-scale inverse problem. Hence, various approaches have been proposed for approximating the Hessian (Shin et al., 2001a; Plessix and Mulder, 2004; Tang, 2009; Jun et al., 2015) or inverse Hessian (Nocedal and Wright, 2006; Nammour and Symes, 2009; Demanet et al., 2012; Letourneau et al., 2012). In Gauss-Newton method, an approximate Hessian is introduced by only considering the first-order term and ignoring the second-order term in Hessian (Pratt et al., 1998). Operto et al. (2006) calculated the diagonal elements of the Gauss-Newton approximate Hessian for preconditioning. Tang (2009) and Pan et al. (2015a) used phase-encoding technology to construct the diagonal Gauss-Newton Hessian efficiently. Shin et al. (2001a) proposed a pseudo-Hessian approximation by replacing the Fréchet derivative wavefield with the virtual source during the auto-correlation process. Preconditioning the gradient with the diagonal pseudo-Hessian resembles a deconvolution imaging condition (Pan et al., 2014b, 2015a).

Instead of constructing the Hessian explicitly, the quasi-Newton methods approximate the inverse Hessian iteratively by storing the model and gradient changes from previous iterations (Nocedal and Wright, 2006). One popular quasi-Newton method is the BFGS method, named after its four inventors Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970). While, the storage requirement of the inverse Hessian approximation and computation cost of preconditioning is still very high. To mitigate this difficulty, a limited-memory BFGS ($l$-BFGS) method is developed by storing information from a limited number $M$ ($M < 10$) of previous iterations (Nocedal, 1980; Byrd et al., 1995; Nocedal and Wright, 2006). Compared to gradient-based methods, $l$-BFGS methods provide a faster convergence rate for large-scale inverse problem (Brossier et al., 2010; Ma and Hale, 2012). Demanet et al. (2012) and Letourneau et al. (2012) proposed to inverse the wave-equation Hessian approximately via matrix probing with a randomized preconditioner.

The Hessian-free optimization method (truncated-Newton or inexact-Newton method) represents an attractive alternative to the above-described optimization methods (Nash, 1985; Santosa and Symes, 1988; Nash, 2000; Métivier et al., 2012, 2014). At each iteration, the search direction is computed by approximately solving the Newton equations through a matrix-free fashion of the conjugate-gradient (CG) algorithm, which is an optimal method for solving a positive definite system (Nash, 1985; Hu et al., 2009). This linear iterative solver only requires computation of the Hessian-vector products instead of forming the Hessian operator explicitly (Fichtner and Trampert, 2011; Métivier et al., 2014). In this paper, the full Hessian is replaced with the Gauss-Newton Hessian, which is always

symmetric positive definite. One problem of the HF optimization method is that obtaining the search direction approximately requires a large number of CG iterations. Our main goal in this paper is to precondition the CG algorithm for reducing the CG iterations and accelerating the HF Gauss-Newton full-waveform inversion (Nash, 2000; Sainath et al., 2013).

The preconditioner for the CG algorithm is designed by approximating the Hessian or its inverse (Nash, 2000). Different preconditioning schemes are developed for comparison in this paper. The traditional Hessian approximations, diagonal pseudo-Hessian and diagonal Gauss-Newton Hessian, are first considered as preconditioners for the CG algorithm. In this paper, based on the reciprocal property of the Green's function, a new pseudo diagonal Gauss-Newton Hessian approximation is proposed as the preconditioner for the CG solver. Furthermore, an inverse Hessian approximation based on quasi-Newton $l$-BFGS method is developed as a preconditioner for the CG solver. The initial guess of the inverse Hessian approximation is important to the performance of the $l$-BFGS method. Thus, we propose to use the diagonal pseudo-Hessian, diagonal Gauss-Newton Hessian and the new pseudo diagonal Gauss-Newton Hessian as the initial guess for constructing the $l$-BFGS preconditioner. Preconditioning makes the CG problem well-conditioned, hence is easier to solve and reduces the number of CG iterations. We demonstrate that the inverse Hessian approximation based on $l$-BFGS method serves as an effective preconditioner for CG-based HF optimization method, which is appropriate for large-scale inverse problem in exploration geophysics.

The paper is organized as follows. We first review the forward modelling problem in the frequency domain and the basic principle of least-squares inverse problem. Different optimization methods for full-waveform inversion are then described and the CG-based Hessian-free optimization method is introduced in detail. We discuss different preconditioning schemes for the inner CG iteration in HF method. In the numerical modelling section, two numerical examples are presented to show the effectiveness and the efficiency of the preconditioning schemes. It is demonstrated that the quasi-Newton $l$-BFGS preconditioner with the pseudo diagonal Gauss-Newton Hessian as initial guess can accelerate the HF Gauss-Newton FWI most efficiently.

## METHODOLOGY

### Frequency domain forward modelling problem

In frequency domain, the forward modelling problem in acoustic medium is governed by the following equation (Helmholtz equation) (Marfurt, 1984):

$$\nabla^2 u(\mathbf{x}, \mathbf{x}_s, \omega) + m(\mathbf{x})\,\omega^2 u(\mathbf{x}, \mathbf{x}_s, \omega) = f_s(\omega)\delta(\mathbf{x} - \mathbf{x}_s), \tag{1}$$

where $\omega$ is the angular frequency, $\mathbf{x} = (x, y, z)$ denotes the subsurface location with Cartesian coordinates, $\nabla^2$ is the Laplacian operator, $m(\mathbf{x})$ is the model parameter (square of slowness), $u(\mathbf{x}, \mathbf{x}_s, \omega)$ denotes the pressure wavefield at position $\mathbf{x}$, $\delta(\mathbf{x} - \mathbf{x}_s)$ is the Dirac delta function, and $f_s(\omega)$ means the source signature at position $\mathbf{x}_s$. The solution of equation (1) can be written as the convolution of source $f_s(\omega)$ with Green's function $G(\mathbf{x}, \mathbf{x}_s, \omega)$:

$$u(\mathbf{x}, \mathbf{x}_s, \omega) = f_s(\omega)\, G(\mathbf{x}, \mathbf{x}_s, \omega), \tag{2}$$

where the Green's function is defined as the solution of wave equation due to an impulse source. A 9-point finite difference scheme is employed to discretize the model (Jo et al., 1996) and a Engquist-Majda boundary condition (first-order) is applied on all of the boundaries of the model (Engquist and Majda, 1977). Equation (1) can be rewritten in a matrix form after discretization:

$$\mathbf{A}\left(\mathbf{m},\omega\right)\mathbf{u}(\mathbf{x},\mathbf{x}_s,\omega)=\mathbf{f}_s(\omega)\delta\left(\mathbf{x}-\mathbf{x}_s\right), \tag{3}$$

where $\mathbf{m}$ is the square of slowness vector, $\mathbf{u}(\mathbf{x},\mathbf{x}_s,\omega)$ and $\mathbf{f}_s(\omega)$ are the discrete pressure wavefield and source vectors. $\mathbf{A}\left(\mathbf{m},\omega\right)$ is the discretized impedance matrix, which is typically sparse and symmetric. In this research, the linear equation (equation (3)) is solved with a direct solver based on multi-frontal Lower Upper (LU) decomposition (Davis and Duff, 1997), which is efficient for a multi-source problem with forward and backward substitutions (Hu et al., 2011).

**The non-linear least-squares inverse problem**

As a non-linear least-squares optimization problem, FWI seeks to estimate the subsurface parameters by iteratively minimizing the difference between the synthetic data $\mathbf{d}_{syn}$ and observed data $\mathbf{d}_{obs}$ (Lailly, 1983; Tarantola, 1984; Virieux and Operto, 2009). The misfit function $\Phi$ is formulated in a least-squares form:

$$\Phi\left(\mathbf{m}\right)=\frac{1}{2}\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\|\Delta\mathbf{d}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)\|^2, \tag{4}$$

where $\Delta\mathbf{d}=\mathbf{d}_{obs}-\mathbf{d}_{syn}$ is the data residual vector, and $\|\cdot\|$ means the L2 norm. Here, the synthetic data $\mathbf{d}_{syn}$ is related to the seismic wavefield $\mathbf{u}$ by a detection operator $\mathcal{P}$, which samples the wavefield at the receiver locations: $\mathbf{d}_{syn}=\mathcal{P}\mathbf{u}$. The Newton optimization approach is developed based on the second-order Taylor-Lagrange expansion of the misfit function $\Phi$:

$$\Phi\left(\mathbf{m}+\Delta\mathbf{m}\right)\approx\Phi\left(\mathbf{m}\right)+\mathbf{g}^{\dagger}\Delta\mathbf{m}+\frac{1}{2}\Delta\mathbf{m}^{\dagger}\mathbf{H}\Delta\mathbf{m}, \tag{5}$$

where the symbol "†" means transpose, $\Delta\mathbf{m}$ is the search direction, $\mathbf{g}=\nabla_{\mathbf{m}}\Phi\left(\mathbf{m}\right)$ and $\mathbf{H}=\nabla_{\mathbf{m}}\nabla_{\mathbf{m}}\Phi\left(\mathbf{m}\right)$ indicate gradient and Hessian respectively.

To minimize the quadratic approximation of the misfit function, the updated model at the $(k+1)$th iteration can be written as the sum of the model at the $k$th iteration and the search direction $\Delta\mathbf{m}_k$:

$$\mathbf{m}_{k+1}=\mathbf{m}_k+\mu_k\Delta\mathbf{m}_k, \tag{6}$$

where $\mu_k$ is the step length, a scalar constant calculated through a line search method satisfying the weak Wolfe condition (Gauthier et al., 1986; Pica et al., 1990; Nocedal and Wright, 2006). Within a Newton optimization framework, the search direction $\Delta\mathbf{m}_k$ is the solution of the Newton linear system:

$$\mathbf{H}_k\Delta\mathbf{m}_k=-\mathbf{g}_k. \tag{7}$$

The gradient is the first-order partial derivative of the misfit function with respect to the model parameter and it indicates the direction in which the misfit function is increasing

most rapidly (Pratt et al., 1998). It can be constructed by zero-lag correlation between the Fréchet derivative wavefield with complex conjugate of the data residuals $\Delta\mathbf{d}$:

$$\mathbf{g}\left(\mathbf{x}\right)=\nabla_{\mathbf{m}(\mathbf{x})}\Phi\left(\mathbf{m}\right)=-\sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega}\Re\left(\frac{\partial\mathbf{d}_{syn}^{\dagger}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}\right)}\Delta\mathbf{d}^{*}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)\right), \tag{8}$$

where the symbol "$*$" means complex conjugate, $\frac{\partial\mathbf{d}_{syn}(\mathbf{x}_g,\mathbf{x}_s,\omega)}{\partial\mathbf{m}(\mathbf{x})}$ indicates the Fréchet derivative wavefield (or Jacobian matrix) recorded at the receiver $\mathbf{x}_g$ due to model perturbation at position $\mathbf{x}$ and $\Re\left(\cdot\right)$ denotes the real part. Within the adjoint-state formalism (Plessix, 2006), the gradient can be expressed as (Sirgue and Pratt, 2004; Plessix and Mulder, 2004; Tao and Sen, 2013; Pan et al., 2015a):

$$\mathbf{g}\left(\mathbf{x}\right)=\sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega}\Re\left(\omega^{2}f_s\left(\omega\right)G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G\left(\mathbf{x}_g,\mathbf{x},\omega\right)\Delta\mathbf{d}^{*}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)\right), \tag{9}$$

where $G\left(\mathbf{x},\mathbf{x}_s,\omega\right)$ and $G\left(\mathbf{x}_g,\mathbf{x},\omega\right)$ indicate source-side and receiver-side Green's functions respectively. Following equation (9), the gradient can be constructed efficiently by cross-correlating the forward modelled wavefield with the back-propagated data residual wavefield (Virieux and Operto, 2009; Tromp et al., 2005). The gradient is poorly-scaled due to geometrical spreading, and it is also contaminated by spurious correlations because of finite-frequency effects and doubly-scattered energy (Pratt et al., 1998). The Hessian operator is the second-order partial derivative of the misfit function with respect to the model parameter (Pratt et al., 1998; Plessix and Mulder, 2004):

$$\begin{aligned}\mathbf{H}\left(\mathbf{x},\mathbf{x}'\right)&=\nabla_{\mathbf{m}(\mathbf{x})}\nabla_{\mathbf{m}(\mathbf{x}')}\Phi\left(\mathbf{m}\right)\\&=\sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega}\Re\left(\frac{\partial\mathbf{d}_{syn}^{\dagger}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}\right)}\frac{\partial\mathbf{d}_{syn}^{*}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}'\right)}+\frac{\partial^{2}\mathbf{d}_{syn}^{\dagger}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}\right)\partial\mathbf{m}\left(\mathbf{x}'\right)}\Delta\mathbf{d}^{*}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)\right),\end{aligned} \tag{10}$$

where $\mathbf{x}'$ is the neighboring position around the position $\mathbf{x}$ (Valenciano, 2008; Pan et al., 2015a) and $\frac{\partial^2\mathbf{d}_{syn}(\mathbf{x}_g,\mathbf{x}_s,\omega)}{\partial\mathbf{m}(\mathbf{x})\partial\mathbf{m}(\mathbf{x}')}$ means the second-order partial derivative wavefield due to model perturbations at positions $\mathbf{x}$ and $\mathbf{x}'$. Multiplying the gradient with the inverse Hessian can greatly enhance the model update, which provides a quadratic convergence rate.

### Full Newton and Gauss-Newton methods

The Newton-type optimization methods (e.g., full Newton (FN) and Gauss-Newton (GN) methods) use the quadratic search direction and converge fast for inverting a limited number of unknown parameters. The search direction in FN method is formed by preconditioning the gradient with the full Hessian $\mathbf{H}$ (equation (10)):

$$\Delta\mathbf{m}_k=-\mathbf{H}_k^{-1}\mathbf{g}_k. \tag{11}$$

The approximate Hessian $\tilde{\mathbf{H}}$ used in Gauss-Newton method is constructed by correlating two Fréchet derivative wavefields, which only accounts for the first-order scattering effects,

as indicated by the first term of equation (10):

$$\tilde{\mathbf{H}}\left(\mathbf{x},\mathbf{x}'\right)=\sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega}\Re\left(\frac{\partial\mathbf{d}^{\dagger}_{syn}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}\right)}\frac{\partial\mathbf{d}^{*}_{syn}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}'\right)}\right). \tag{12}$$

With Born approximation, the two Fréchet derivative wavefields in equation (12) can be expressed using Green's functions (Tang, 2009; Tao and Sen, 2013):

$$\tilde{\mathbf{H}}\left(\mathbf{x},\mathbf{x}'\right)\approx\sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega}\Re\left(\omega^4|f_s(\omega)|^2G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G\left(\mathbf{x}_g,\mathbf{x},\omega\right)G^{*}\left(\mathbf{x}',\mathbf{x}_s,\omega\right)G^{*}\left(\mathbf{x}_g,\mathbf{x}',\omega\right)\right), \tag{13}$$

where the element $\tilde{\mathbf{H}}\left(\mathbf{x},\mathbf{x}'\right)$ in the Gauss-Newton Hessian is formed by correlating the two Fréchet derivative wavefields at the receivers' locations due to model perturbations at positions $\mathbf{x}$ and $\mathbf{x}'$. When $\mathbf{x}\neq\mathbf{x}'$, it indicates the off-diagonal element of the Gauss-Newton Hessian. When $\mathbf{x}=\mathbf{x}'$, we can obtain its diagonal part $\tilde{\mathbf{H}}_{diag}$:

$$\tilde{\mathbf{H}}_{diag}\left(\mathbf{x}\right)=\sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega}\Re\left(\omega^4|f_s(\omega)|^2G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G\left(\mathbf{x}_g,\mathbf{x},\omega\right)G^{*}\left(\mathbf{x},\mathbf{x}_s,\omega\right)G^{*}\left(\mathbf{x}_g,\mathbf{x},\omega\right)\right). \tag{14}$$

Because of the band-limited signature of the seismic data, the Fréchet derivative wavefields are often uncorrelated to a great degree, which means that Gauss-Newton Hessian $\tilde{\mathbf{H}}$ is diagonally dominant and banded (Pratt et al., 1998; Valenciano, 2008; Tang, 2009; Pan et al., 2014a). The diagonal elements account for compensating the geometrical spreading effects and the off-diagonal elements are responsible for removing the finite-frequency effects and de-blurring the gradient (Pratt et al., 1998). In multi-parameter FWI, the Gauss-Newton Hessian can mitigate the parameter cross-talk problem (Operto et al., 2013; Innanen, 2014; Pan et al., 2015b,c). Furthermore, the doubly scattered energy in the data residuals can result in second-order scattering artifacts in the gradient. The second term in equation (10) predicts the second-order non-linear effects by correlating the second-order partial derivative wavefields with the data residuals. With this second-order preconditioner, the second-order scattering artifacts in the gradient can be suppressed effectively.

For these Newton-type methods, explicit evaluation and inversion the Hessian matrix $\mathbf{H}$ and Gauss-Newton Hessian $\tilde{\mathbf{H}}$ at each iteration are required. Considering the subsurface model with a number of $N_m$ model parameters, the Hessian is an $N_m\times N_m$ square and symmetric matrix. Even though, the Newton-type methods benefit from fast convergence rate, the computation, storage and inversion of Hessian at each iteration are prohibitively expensive, which limits their applications for large-scale inverse problems in exploration geophysics.

## Gradient-based methods

The gradient-based methods (e.g., steepest-descent (SD) and non-linear conjugate-gradient (NCG) methods) assume the Hessian matrix $\mathbf{H}$ as an identity matrix $\mathbf{I}$ and they are more attractive than the Newton-type ones when inversing a large number of unknown model parameters. The SD method simply determines the search direction to be the negative of the gradient. In mathematics, conjugate-gradient (CG) method seeks the solution of a linear system. The NCG method generalizes the CG method to non-linear optimization and

**Table 1** Pseudo-code of the preconditioned conjugate-gradient (PCG) method.

**Notations**: $\mathbf{W}$ is a symmetric and positive definite matrix;
$\mathbf{q}$ is the residual; $\tilde{k}$ is the iteration index; $\mathcal{M}$ is the preconditioner;
$\gamma_{min}$ is the relative residual tolerance; $\tilde{k}_{max}$ is the maximum iteration.

**Input**: $\mathbf{W}$, $\mathbf{b}$, $\tilde{k}_{max}$, $\mathcal{M}$, $\gamma_{min}$

**Output**: $\mathbf{x}$, $\gamma$

**Initialization**: $\mathbf{q}_0 = \mathbf{b} - \mathbf{W}\mathbf{x}_0$, $\mathbf{z}_0 = \mathcal{M}^{-1}\mathbf{q}_0$, $\mathbf{p}_0 = \mathbf{z}_0$, $\tilde{k} = 0$;

**While** $\gamma_{\tilde{k}} > \gamma_{min}$ & $\tilde{k} < \tilde{k}_{max}$

$\alpha_{\tilde{k}} = \frac{\mathbf{q}_{\tilde{k}}^\dagger \mathbf{z}_{\tilde{k}}}{\mathbf{p}_{\tilde{k}}^\dagger \mathbf{W}\mathbf{p}_{\tilde{k}}}$;

$\mathbf{x}_{\tilde{k}+1} = \mathbf{x}_{\tilde{k}} + \alpha_{\tilde{k}}\mathbf{p}_{\tilde{k}}$;

$\mathbf{q}_{\tilde{k}+1} = \mathbf{q}_{\tilde{k}} - \alpha_{\tilde{k}}\mathbf{W}\mathbf{p}_{\tilde{k}}$;

$\mathbf{z}_{\tilde{k}+1} = \mathcal{M}^{-1}\mathbf{q}_{\tilde{k}+1}$;

$\beta_{\tilde{k}+1} = \frac{\mathbf{z}_{\tilde{k}+1}^\dagger \mathbf{q}_{\tilde{k}+1}}{\mathbf{z}_{\tilde{k}}^\dagger \mathbf{q}_{\tilde{k}}}$;

$\mathbf{p}_{\tilde{k}+1} = \mathbf{z}_{\tilde{k}+1} + \beta_{\tilde{k}}\mathbf{p}_{\tilde{k}}$;

$\gamma_{\tilde{k}} = \frac{\|\mathbf{b} - \mathbf{W}\mathbf{x}_{\tilde{k}+1}\|}{\|\mathbf{b}\|}$;

$\tilde{k} = \tilde{k} + 1$;

**End**

obtain the local minimum of a non-linear function using its gradient alone (Nocedal and Wright, 2006). The search direction is just a linear combination of current gradient and previous search direction. The gradient-based methods are known to converge globally, but possibly very slowly. In most cases, preconditioning is necessary to ensure the fast convergence of the CG method (Hu et al., 2011).

## Quasi-Newton methods

The quasi-Newton methods provide an attractive alternative to Newton-type and gradient-based methods by approximating the inversion Hessian iteratively instead of constructing the Hessian matrix explicitly (Brossier et al., 2009; Ma and Hale, 2012). BFGS method, named after Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970), is one popular quasi-Newton strategy to approximate the inverse Hessian iteratively using the changes of the model and gradient (Nocedal and Wright, 2006).

In the BFGS updating formula, we are given a symmetric and positive definite matrix $\mathcal{H}_k$ that approximates the inverse of the Hessian, and a pair of vectors $\mathbf{s}_k = \mathbf{m}_{k+1} - \mathbf{m}_k$, and $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$ that indicates the model and gradient changes and satisfies the condition $\mathbf{s}_k^\dagger \mathbf{y}_k > 0$. Using these vectors, we compute the inverse Hessian approximation $\mathcal{H}_{k+1}$ by the following formula:

$$\mathcal{H}_{k+1} = \mathbf{v}_k^\dagger \mathcal{H}_k \mathbf{v}_k + \mathbf{w}_k \mathbf{s}_k \mathbf{s}_k^\dagger, \tag{15}$$

where $\mathbf{w}_k = 1/\mathbf{y}_k^\dagger \mathbf{s}_k$, $\mathbf{v}_k = \mathbf{I} - \mathbf{w}_k \mathbf{y}_k \mathbf{s}_k^\dagger$ and $\mathbf{I}$ is the identity matrix. The initial inverse Hessian approximation $\mathcal{H}_0$ is important to BFGS method and it is usually set as an identity matrix to make sure that the updated matrix maintains positive definiteness (Wu et al., 2015).

Although, BFGS method can reduce the computation cost to approximate the Hessian, the memory requirement for storage and computation cost for preconditioning remain to be challenging, especially for large-scale optimization problem. To overcome this difficulty, a limited-memory BFGS ($l$-BFGS) method was developed by storing the model and gradient changes from a limited number $M$ of previous iterations (typically $M < 10$) (Nocedal, 1980). The stored information is then used to construct an approximated inverse Hessian. A "two-loop recursion" scheme is implemented in this research to obtain the search direction using the information of previous updates (Nocedal and Wright, 2006).

**Hessian-free optimization method**

The FN and GN methods require the computation, storage and inversion of the Hessian at each iteration in order to determine a search direction of the form: $\Delta\mathbf{m} = -\mathbf{H}^{-1}\mathbf{g}$. Instead of constructing Hessian approximations or approximating inverse Hessian, the Hessian-free (HF) optimization method, also known as truncated-Newton or inexact-Newton method, obtains the search direction by solving the Newton linear system (equation (7)) approximately using a conjugate-gradient (CG) method with matrix-free fashion (Saad, 2003; Anagaw and Sacchi, 2012; Métivier et al., 2014). The CG method is an optimal algorithm for solving a symmetric positive definite system **Wx=b** and it only requires computing the Hessian-vector products $\mathbf{H}v$ instead of forming the Hessian matrix explicitly, where $v$ is an arbitrary vector in model space. The Hessian-vector products can be calculated via finite-difference method (Nocedal and Wright, 2006) or the second-order adjoint-state method (Fichtner and Trampert, 2011; Métivier et al., 2014). In this paper, the second method is employed for Hessian-vector products calculation.

The full Hessian $\mathbf{H}$ (equation (10)) arising from the second-order partial derivative might not be positive definite (Nash, 2000). Thus, CG method is no longer appropriate for solving an indefinite linear system. In the context of HF optimization method, the Hessian $\mathbf{H}$ (equation (10)) is always replaced with Gauss-Newton Hessian $\tilde{\mathbf{H}}$ (equation (12)), which is always symmetric and positive definite:

$$\left(\tilde{\mathbf{H}}_k + \epsilon\mathcal{A}\right)\Delta\mathbf{m}_k = -\mathbf{g}_k, \tag{16}$$

where $\epsilon\mathcal{A}$ is the damping term, $\epsilon$ is a small constant value and $\mathcal{A}$ indicates the maximum value of the Gauss-Newton Hessian. In this paper, the HF Gauss-Newton FWI is implemented in a double-iterative scheme: the outer loop is to iteratively update the model parameters for the non-linear optimization problem, and the inner loop is to solve the linear system (equation(16)) iteratively with the CG algorithm. The inner iteration is typically stopped or "truncated" before the solution of the Newton equation is obtained. The convergence rate of the outer iteration is proven to be related to the accuracy with which equation (16) is solved (Dembo and Eisenstat, 1982).

A Hessian-free optimization method will be more competitive if further enhancements

are used, for example, an effective preconditioner for the linear system and appropriate stopping criteria for the inner iterative algorithm. With these enhancements, Hessian-free optimization method is a powerful tool for large-scale inverse problem.

*Preconditioning*

One problem of the CG iterative algorithm is that it requires many iterations when obtaining the approximate solution of a linear system $\mathbf{Wx} = \mathbf{b}$. The convergence rate of the CG method depends on the spectral properties (e.g., its eigenvalues) of the coefficient matrix $\mathbf{W}$ (Nash, 2000). It is often convenient to transform the equation system into a system that has the same solution but has more favorable spectral properties. This can be achieved by applying a suitable preconditioner $\mathcal{M}$ on the linear system: $\mathcal{M}^{-1}\mathbf{Wx} = \mathcal{M}^{-1}\mathbf{b}$. Thus, the preconditioned Newton system for the HF Gauss-Newton FWI is given by:

$$\mathcal{M}_k^{-1}\left(\tilde{\mathbf{H}}_k + \epsilon\mathcal{A}\right)\Delta\mathbf{m}_k = -\mathcal{M}_k^{-1}\mathbf{g}_k. \tag{17}$$

The solution of equation (17) can be obtained by the preconditioned conjugate-gradient (PCG) method. The pseudo-code of the PCG method is described in Table 1. The PCG method is expected to reduce the number of inner iterations, improve the convergence rate and accelerate the HF Gauss-Newton FWI. In this research, we aim at developing efficient and stable preconditioning schemes for inner iterative algorithm.

The preconditioner for the CG method is always devised to approximate the Hessian or the inverse Hessian. We first consider the traditional Hessian approximations (e.g., diagonal pseudo-Hessian and diagonal Gauss-Newton Hessian (equation (14))) as the preconditioners for the CG inner iteration. The pseudo-Hessian $\mathbb{H}$ is constructed by replacing the Fréchet derivative wavefield with the virtual source $\tilde{f}_s(\omega)$ in the correlation process (Shin et al., 2001b). Taking partial derivative with respect to model parameter on both sides of equation (1) gives:

$$\left(\nabla^2 + m(\mathbf{x})\,\omega^2\right)\frac{\partial u(\mathbf{x},\mathbf{x}_s,\omega)}{\partial m(\mathbf{x})} = \tilde{f}_s(\omega), \tag{18}$$

where $\tilde{f}_s(\mathbf{x},\omega) = -\omega^2 u(\mathbf{x},\mathbf{x}_s,\omega)$ is the virtual source. Considering equation (2), the virtual source can be expressed with Green's function:

$$\tilde{f}_s(\mathbf{x},\omega) = -\omega^2 f_s(\omega)\, G(\mathbf{x},\mathbf{x}_s,\omega). \tag{19}$$

The pseudo-Hessian is obtained by correlating two virtual sources:

$$\mathbb{H}(\mathbf{x},\mathbf{x}') = \tilde{f}_s(\mathbf{x},\omega)\,\tilde{f}_s^*(\mathbf{x}',\omega) = \sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\Re\left(\omega^4 \mid f_s(\omega)\mid^2 G(\mathbf{x},\mathbf{x}_s,\omega)\,G^*(\mathbf{x}',\mathbf{x}_s,\omega)\right), \tag{20}$$

where when $\mathbf{x} = \mathbf{x}'$, we can obtain the diagonal pseudo-Hessian $\mathbb{H}_{diag}$, the auto-correlation of the two virtual sources:

$$\mathbb{H}_{diag}(\mathbf{x}) = \sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\Re\left(\omega^4 \mid f_s(\omega)\mid^2 G(\mathbf{x},\mathbf{x}_s,\omega)\,G^*(\mathbf{x},\mathbf{x}_s,\omega)\right). \tag{21}$$

Comparing equation (21) with equation (14), we observe that the diagonal pseudo-Hessian $\mathbb{H}_{diag}$ can be obtained by ignoring the receiver-side Green's functions in the diagonal Gauss-Newton Hessian $\tilde{\mathbf{H}}_{diag}$, which means that forming the diagonal pseudo-Hessian at each iteration does not need additional cost. While when employing the diagonal Gauss-Newton Hessian $\tilde{\mathbf{H}}_{diag}$ as a preconditioner for the CG method, more computation cost is required for constructing the receiver-side Green's functions (Tao and Sen, 2013; Pan et al., 2015a).

In this paper, we propose a pseudo diagonal Gauss-Newton Hessian approximation $\tilde{\tilde{\mathbf{H}}}_{diag}$ as the preconditioner for the CG algorithm in the inner loop. When considering the reciprocal property of the Green's function $G\left(\mathbf{x}_g, \mathbf{x}, \omega\right) = G\left(\mathbf{x}, \mathbf{x}_g, \omega\right)$, equation (14) can be rewritten as:

$$\tilde{\mathbf{H}}_{diag}\left(\mathbf{x}\right) = \sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega} \Re\left(\omega^4 |f_s(\omega)|^2 G\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G\left(\mathbf{x}, \mathbf{x}_g, \omega\right) G^*\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G^*\left(\mathbf{x}, \mathbf{x}_g, \omega\right)\right). \tag{22}$$

When considering reflection survey and assuming that the sources and receivers are co-located, the receiver-side Green's function $G\left(\mathbf{x}, \mathbf{x}_g, \omega\right)$ can be replaced by the source-side Green's function $G\left(\mathbf{x}, \mathbf{x}_s, \omega\right)$, which gives the new diagonal Hessian approximation $\tilde{\tilde{\mathbf{H}}}_{diag}$:

$$\tilde{\tilde{\mathbf{H}}}_{diag}\left(\mathbf{x}\right) = \sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega} \Re\left(\omega^4 |f_s(\omega)|^2 G\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G^*\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G^*\left(\mathbf{x}, \mathbf{x}_s, \omega\right)\right). \tag{23}$$

This diagonal Hessian approximation is named as pseudo diagonal Gauss-Newton Hessian and no additional cost is required for constructing it. The diagonal pseudo-Hessian, diagonal Gauss-Newton Hessian and pseudo diagonal Gauss-Newton Hessian preconditioners are given by:

$$\mathcal{M}_k^{DPH} = \mathbb{H}_{diag}^k + \lambda\mathcal{B}_k, \tag{24}$$

$$\mathcal{M}_k^{DGH} = \tilde{\mathbf{H}}_{diag}^k + \lambda\mathcal{C}_k, \tag{25}$$

$$\mathcal{M}_k^{PDGH} = \tilde{\tilde{\mathbf{H}}}_{diag}^k + \lambda\mathcal{D}_k, \tag{26}$$

where $\lambda\mathcal{B}_k$, $\lambda\mathcal{C}_k$ and $\lambda\mathcal{D}_k$ are the stabilization terms, $\lambda$ is a small constant value, $\mathcal{B}_k$, $\mathcal{C}_k$ and $\mathcal{D}_k$ indicate the maximum value of the $\mathbb{H}_{diag}^k$, $\tilde{\mathbf{H}}_{diag}^k$ and $\tilde{\tilde{\mathbf{H}}}_{diag}^k$. These three different preconditioning strategies are named as DPH-GN, DGH-GN and PDGH-GN methods in this paper. When the parameter $\lambda$ is very large, these preconditioning methods approach the non-preconditioned HF Gauss-Newton method (CG-GN method).

Furthermore, we develop an $l$-BFGS preconditioning scheme for the HF optimization method (Nash, 1985, 2000; Métivier et al., 2012; Sainath et al., 2013), namely $l$-BFGS-GN method for $\mathcal{H}_0 = \mathbf{I}$. The $l$-BFGS method approximates the inverse Hessian $\mathcal{H}$ (equation (15)) iteratively using the model and gradient changes from a limited number $M$ of previous iterations. The approximated inverse Hessian $\mathcal{H}$ can also be used as a preconditioner for the CG iterative method:

$$\mathcal{H}_k\left(\tilde{\mathbf{H}}_k + \epsilon\mathcal{A}\right)\Delta\mathbf{m}_k = -\mathcal{H}_k\mathbf{g}_k. \tag{27}$$

An identity matrix $\mathbf{I}$ is usually set as the initial guess $\mathcal{H}_0$, which is actually important for the $l$-BFGS method. To improve the performance of the $l$-BFGS preconditioning scheme, we consider using the stabilized diagonal pseudo-Hessian (equation (24)), diagonal Gauss-Newton Hessian (equation (25)) and pseudo diagonal Gauss-Newton Hessian (equation (26)) as the initial guess for constructing the $l$-BFGS preconditioner, which are named as $l$-BFGS-GN-DPH, $l$-BFGS-GN-DGH and $l$-BFGS-GN-PDGH respectively. Similarly, when parameter $\lambda$ is very large, these three methods approach $l$-BFGS-GN method ($\mathcal{H}_0 = \mathbf{I}$).

*Stopping criteria*

Newton's method is based on the Taylor series approximation (equation (5)). If this approximation is inaccurate then it may not be suitable to solve the Newton equations accurately and "over-solving" the Newton equation will not produce a better search direction (Nash, 2000). The CG algorithm should be terminated with an appropriate stopping criteria. We can define the maximum inner iteration number $\tilde{k}_{max}$ and relative residual $\gamma_{\tilde{k}}$:

$$\gamma_{\tilde{k}} = \frac{\|\tilde{\mathbf{H}}_{\tilde{k}}\Delta\mathbf{m}_{\tilde{k}} + \mathbf{g}_{\tilde{k}}\|}{\|\mathbf{g}_{\tilde{k}}\|}, \tag{28}$$

where $\tilde{k}$ indicates the CG inner iteration index. The inner iteration is stopped when $\gamma_{\tilde{k}} < \gamma_{min}$, where $\gamma_{min}$ indicates the relative residual tolerance.

**Line search with weak Wolfe condition**

The step length $\mu_k$ in equation (6) is obtained by a line search method satisfying the Wolfe condition, as measured by the following inequalities (Nocedal and Wright, 2006):

$$\Phi\left(\mathbf{m}_k + \mu_k\Delta\mathbf{m}_k\right) \leq \Phi\left(\mathbf{m}_k\right) + \mu_k c_1 \Delta\mathbf{m}_k^{\dagger}\nabla\Phi\left(\mathbf{m}_k\right), \tag{29}$$

$$\nabla\Phi\left(\mathbf{m}_k + \mu_k\Delta\mathbf{m}_k\right) \geq c_2\nabla\Phi\left(\mathbf{m}_k\right), \tag{30}$$

where the constant parameters $c_1$ and $c_2$ satisfy $0 < c_1 < c_2 < 1$. In practice, parameter $c_1$ can be chosen as $c_1 = 10^{-4}$ and parameter $c_2$ should be much larger $c_2 = 0.9$ (Nocedal and Wright, 2006). Equation (29) is the Armijo condition, which ensures that the step length $\mu_n$ decreases the misfit function sufficiently (Armijo, 1966). Equation (30) is known as the weak Wolfe condition, which ensures that the slope has been reduced sufficiently (Nocedal and Wright, 2006). The initial step length $\mu_0$ is always chosen as 1 and then after a set of trial step lengths, the optimal one will be accepted to satisfy the above conditions. The pseudo-code of the HF Gauss-Newton method with a PCG algorithm is illustrated in Table 2.

## NUMERICAL EXPERIMENTS

In this section, we first apply the Hessian-free Gauss-Newton FWI on a Gaussian-anomaly model in comparison with steepest-descent (SD) and $l$-BFGS methods. The above mentioned preconditioning schemes for the HF Gauss-Newton FWI are also examined. These different methods are then enacted on a more complex modified Marmousi model. The inversion results verify the effectiveness of the preconditioning methods in accelerating the HF Gauss-Newton FWI, improving the convergence rate and reducing the computation cost.

**Table 2** Pseudo-code of the HF Gauss-Newton FWI with PCG method.

**Notations**: $k_{max}$ is the maximum outer iteration; $\phi$ is the normalized misfit; $\phi_{min}$ is the minimum normalized misfit; $\mathbf{N}_\omega$ is the multi-frequency group.

**Input**: $\mathbf{m}_0$, $\phi_{min}$, $\epsilon$, $k_{max}$, $M$, $\mathbf{N}_\omega$, $\lambda$, $\mathbf{d}_{obs}$, $\tilde{k}_{max}$, $\gamma_{min}$

**Output**: $\mathbf{m}$, $\phi$

**Initialization**: $k = 0$;

**For** $k < k_{max}$ & $\phi_k > \phi_{min}$

1. Calculate $\mathbf{d}_{syn}\left(\mathbf{m}_k, \mathbf{N}_\omega^k\right)$ and $\Delta\mathbf{d}\left(\mathbf{m}_k, \mathbf{N}_\omega^k\right) = \mathbf{d}_{obs}\left(\mathbf{N}_\omega^k\right) - \mathbf{d}_{syn}\left(\mathbf{m}_k, \mathbf{N}_\omega^k\right)$;
2. Construct $\mathbf{g}_k\left(\mathbf{m}_k, \mathbf{N}_\omega^k\right) \leftarrow \Delta\mathbf{d}\left(\mathbf{m}_k, \mathbf{N}_\omega^k\right)$ with the adjoint-state method;
3. Construct Hessian-vector products $\tilde{\mathbf{H}}_k\upsilon$;
4. Construct the preconditioner $\mathcal{M}_k$:
   $\mathcal{M}_k = \mathcal{M}_k^{DPH}, \mathcal{M}_k^{DGH}, \mathcal{M}_k^{PDGH}$;
   or $\mathcal{M}_k^{-1} = \mathcal{H}_k$;
5. Call PCG method (Table 1) $\leftarrow \tilde{k}_{max}, \gamma_{min}, \mathcal{M}_k, \mathbf{g}_k, \tilde{\mathbf{H}}_k\upsilon$ and calculate $\Delta\mathbf{m}_k$:
   $\mathcal{M}_k^{-1}\left(\tilde{\mathbf{H}}_k + \epsilon\mathcal{A}\right)\Delta\mathbf{m}_k = -\mathcal{M}_k^{-1}\mathbf{g}_k$;
6. Get $\mu_k$ with a line search method satisfying weak Wolfe condition;
7. $\mathbf{m}_{k+1} = \mathbf{m}_k + \mu_k\Delta\mathbf{m}_k$;
8. Calculate the normalized misfit $\phi_k$;
9. $k = k + 1$;

**End**

### The Gaussian-anomaly model

In this numerical example, the SD, $l$-BFGS and HF Gauss-Newton methods are applied on a Gaussian-anomaly model. We examine the quadratic convergence rate of the HF Gauss-Newton FWI in reconstructing the velocity model compared to SD and $l$-BFGS methods. The effectiveness of the preconditioning schemes in speeding up the HF Gauss-Newton FWI is also verified.

The Gaussian-anomaly model consists of $50 \times 100$ grid cells with a grid interval of $10$ m in both horizontal and vertical directions. A total of $49$ sources are deployed from $20$ m to $980$ m with a source interval of $20$ m and a depth of $20$ m. A total of $100$ receivers are distributed from $10$ m to $1000$ m with a receiver interval of $10$ m and a depth of $20$ m. A Ricker wavelet with a $30$ Hz dominant frequency is used as the source function. Figure 1 shows the true Gaussian-anomaly P-wave velocity model. The initial velocity model is a homogeneous model with a constant velocity of $2$ km/s.

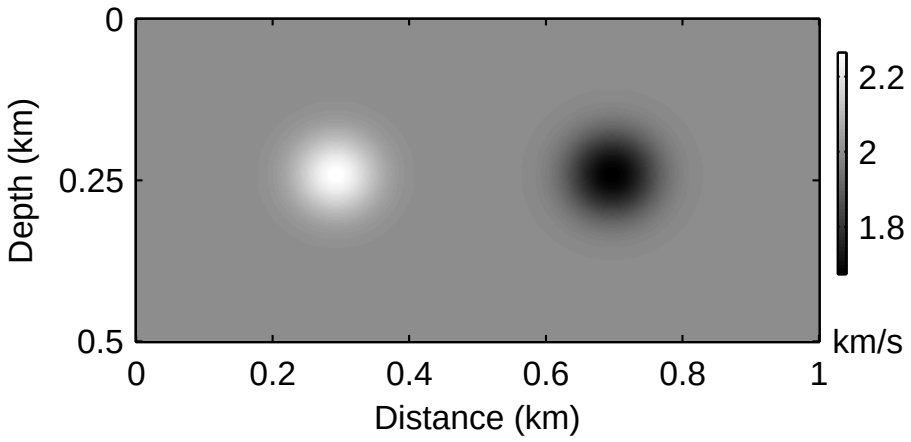


FIG. 1. The true Gaussian-anomaly model.

For comparison of different methods, we set the stopping criteria for inversion as: the maximum outer iteration $k_{max} = 101$ and the minimum normalized misfit $\phi_{min} = 2.0e\text{-}5$. To illustrate the performances of different preconditioning strategies, the stopping criteria for the inner CG algorithm is also defined: the maximum inner iteration $\tilde{k}_{max} = 100$ and the minimum relative residual $\gamma_{min} = 1.0e\text{-}2$ (equation (4)). The parameters used for stabilization are $\epsilon = 1.0e\text{-}2$ (equation (16)) and $\lambda = 1.0e\text{-}2$. Figure 2 illustrates the inversion results obtained with SD, $l$-BFGS and HF Gauss-Newton methods with different preconditioning strategies. Figures 2a and 2b show the inversion results by SD and $l$-BFGS methods. Figure 2c shows the inversion result by HF Gauss-Newton method without preconditioning (CG-GN method). Figures 2d, 2e and 2f are the inversion results by DPH-GN, DGH-GN and PDGH-GN methods respectively. Figures 2g, 2h and 2i are the inversion results by $l$-BFGS-GN method ($\mathcal{H}_0 = \mathbf{I}$) when $M = 3$, $M = 5$ and $M = 8$ respectively. Figures 2j, 2k and 2l are the inversion results by $l$-BFGS-GN-DPH, $l$-BFGS-GN-DGH and $l$-BFGS-GN-PDGH methods with $M = 5$. Clearly, the HF Gauss-Newton methods can obtain better inversion results in comparison with SD and $l$-BFGS methods.

Table 3 illustrates the maximum iterations, normalized minimum misfit ($\phi$), number of

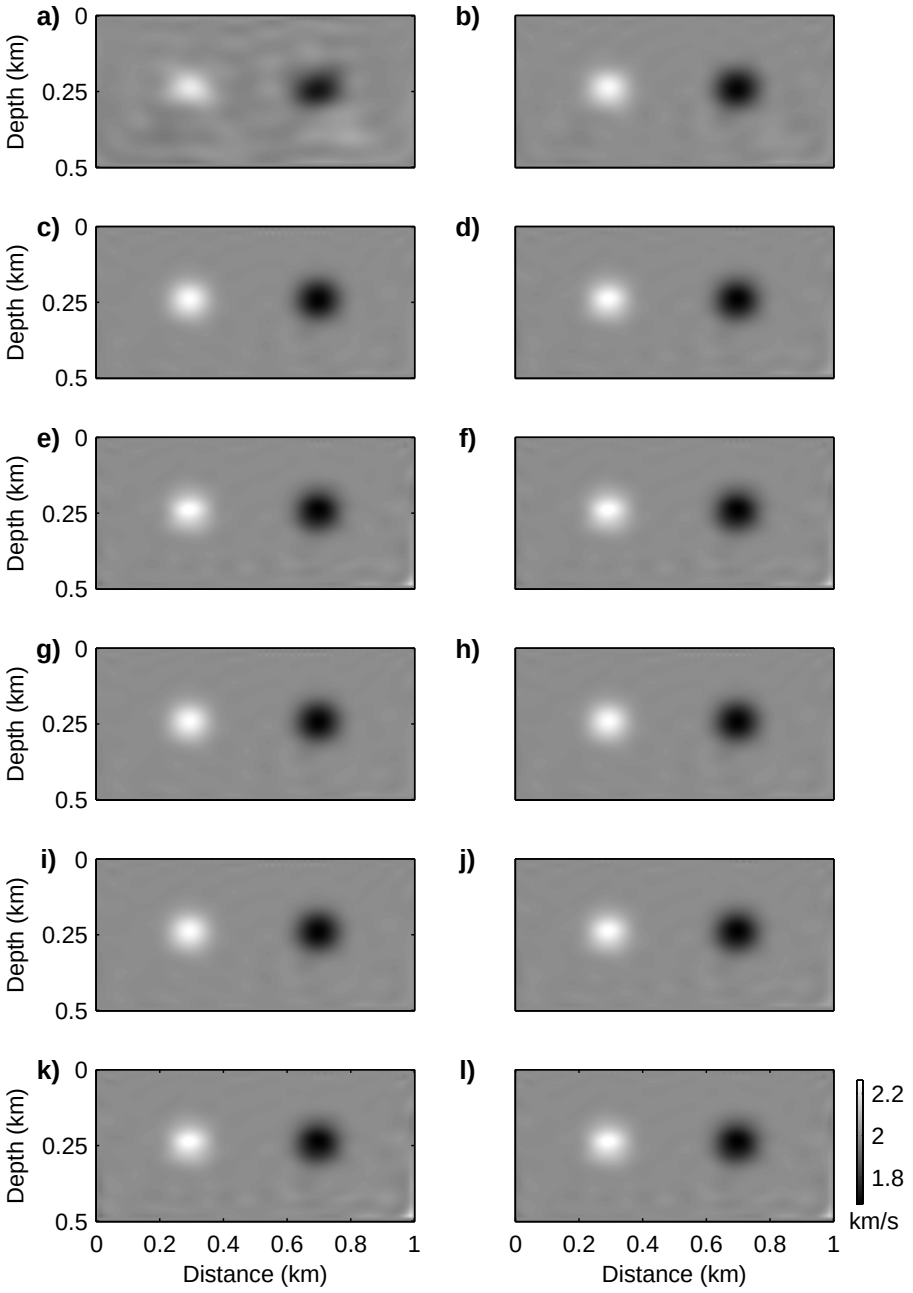


FIG. 2. (a) SD method; (b) $l$-BFGS method; (c) CG-GN method; (d) DPH-GN method; (e) DGH-GN method; (f) PDGH-GN method; (g) $l$-BFGS-GN method ($M = 3$ and $\mathcal{H}_0 = \mathbf{I}$); (h) $l$-BFGS-GN method ($M = 5$ and $\mathcal{H}_0 = \mathbf{I}$); (i) $l$-BFGS-GN method ($M = 8$ and $\mathcal{H}_0 = \mathbf{I}$); (j) $l$-BFGS-GN-DPH method ($M = 5$); (k) $l$-BFGS-GN-DGH method ($M = 5$); (l) $l$-BFGS-GN-PDGH method ($M = 5$).

forward problems solved and computation time of these different methods. As it can be seen that, the SD and $l$-BFGS methods undergo $k_{max} = 101$ iterations without reaching the minimum misfit requirement of $\phi_{min} = 2.0e\text{-}5$. While the HF Gauss-Newton methods can converge towards the minimum misfit requirement quadratically within several iterations. The methods of DGH-GN, PDGH-GN, $l$-BFGS-GN-DPH, $l$-BFGS-GN-DGH and $l$-BFGS-GN-PDGH achieve the minimum misfit most quickly with $6$ outer iterations. The performances of $l$-BFGS-GN methods with different values of $M$ are similar. The DGH-

GN method gives the best inversion result with the minimum normalized misfit of $1.59e$-5. However, this method needs to solve more forward modelling problems for constructing the receiver-side Green's functions, which takes more computation time. The $l$-BFGS-GN-DPH method show the best performance in computational efficiency with number of $488$ problems solved and computation time of $551.20$ s. While the quality of final inversion result by $l$-BFGS-GN-DPH method is not satisfactory. The PDGH-GN and $l$-BFGS-GN-PDGH methods with the proposed pseudo diagonal Gauss-Newton Hessian approximation show good performances in the convergence rate, quality of the inversion results and computation time.

From Figure 3 to Figure 5, we use the same line styles to indicate different methods. Figures 3a and 3b show the convergence history of different methods. We can observe that the HF Gauss-Newton methods converge much faster than SD and $l$-BFGS methods. Figure 3b is a detailed view of Figure 3a. Generally, the HF Gauss-Newton methods with preconditioning converge faster than the one without preconditioning. Even though, the convergence rate of the $l$-BFGS-GN method ($M = 5$) is close to that of the CG-GN method. The computation time is reduced with $l$-BFGS preconditioning, as illustrated in Table 3.

Figure 4 shows the relative residual $\gamma$ at each outer iteration of the HF Gauss-Newton methods. We can see that, the relative residuals of the preconditioned HF Gauss-Newton methods are generally smaller than those of non-preconditioned HF Gauss-Newton method (CG-GN method). This means that the PCG method can obtain search directions more accurately than the non-preconditioned CG method, which gives faster convergence rate for outer iteration. Figure 5a shows the number of forward problems solved as the iteration proceeds by the preconditioned HF Gauss-Newton methods. DGH-GN and $l$-BFGS-GN-DGH methods need to solve more forward modelling problems for constructing the receiver-side Green's functions. Figures 5c and 5d show the Normalized misfit vs. Number of forward problems solved and Normalized misfit vs. Computation time (s) for the HF Gauss-Newton methods respectively. As we can see, to obtain the same quality inversion result, most of the preconditioning strategies except for $l$-BFGS-GN-DGH method can reduce the number of forward problems solved and the computation time.

**The modified Marmousi model**

A more complex modified Marmousi model is used to examine the efficiency of different preconditioning schemes for the HF Gauss-Newton full-waveform inversion. The truncated Marmousi model has $100 \times 100$ grid cells with a grid interval of $10$ m in both horizontal and vertical directions. We deploy $38$ sources from $50$ m to $475$ m at depth of $20$ m with a regular source spacing of $25$ m. Fifty receivers are arranged from $10$ m to $1000$ m every $20$ m at the depth of $20$ m. The source function is a Ricker wavelet with a dominant frequency of $30$ Hz. Figures 6a and 6b show the true P-wave velocity model and initial P-wave velocity model. The initial velocity model is obtained by smoothing the true model with a Gaussian function. The inversion process is carried out with a multi-scale approach for mitigating the cycle-skipping problem (Pratt and Chapman, 1992; Bunks et al., 1995; Sirgue and Pratt, 2004). The frequency is increased from $5$ Hz to $30$ Hz with a partial overlap-frequency selection strategy, in which a group of $3$ frequencies are used for inversion simultaneously. The frequency group increases from low to high with $2$ frequencies

**Table 3** Efficiency comparison for different methods with the Gaussian-anomaly model.

| Methods | Maximum Iteration | Minimum Misfit | Problems Solved | Computation Time (s) |
|---|---|---|---|---|
| SD | 101 | 2.40e-3 | 1350 | 1.73e3 |
| $l$-BFGS | 101 | 2.73e-5 | **210** | **267.02** |
| CG-GN | 8 | 1.78e-5 | 621 | 692.95 |
| DPH-GN | 7 | 1.73e-5 | 601 | 679.16 |
| DGH-GN | **6** | **1.59e-5** | 633 | 707.43 |
| PDGH-GN | **6** | 1.62e-5 | 598 | 676.16 |
| $l$-BFGS-GN ($M = 3$) | 8 | 1.85e-5 | 555 | 626.28 |
| $l$-BFGS-GN ($M = 5$) | 8 | 1.84e-5 | 561 | 632.10 |
| $l$-BFGS-GN ($M = 8$) | 8 | 1.83e-5 | 561 | 631.50 |
| $l$-BFGS-GN-DPH | **6** | 1.90e-5 | **488** | **551.20** |
| $l$-BFGS-GN-DGH | **6** | 1.94e-5 | 876 | 969.63 |
| $l$-BFGS-GN-PDGH | **6** | 1.61e-5 | 554 | 651.20 |

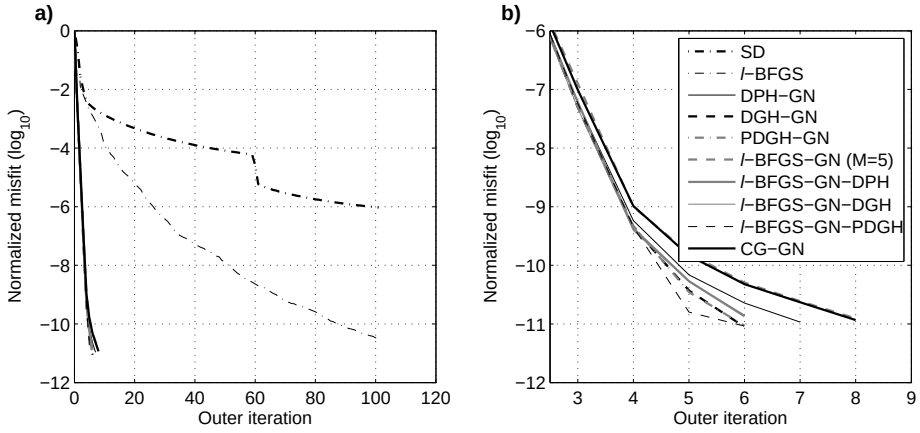


FIG. 3. Comparison of the convergence history for different methods. (b) is the detailed view of (a). The black-bold-dash-dot and black-thin-dash-dot lines are convergence history of SD and $l$-BFGS methods. The thin-black-solid, black-bold-dash, gray-bold-dash-dot, black-bold-dash, gray-bold-solid, gray-thin-solid, gray-thin-dash and black-bold-solid lines show the convergence history of DPH-GN, DGH-GN, PDGH-GN, $l$-BFGS-GN ($M = 5$),$l$-BFGS-GN-DPH, $l$-BFGS-GN-DGH and $l$-BFGS-GN-PDGH methods respectively.

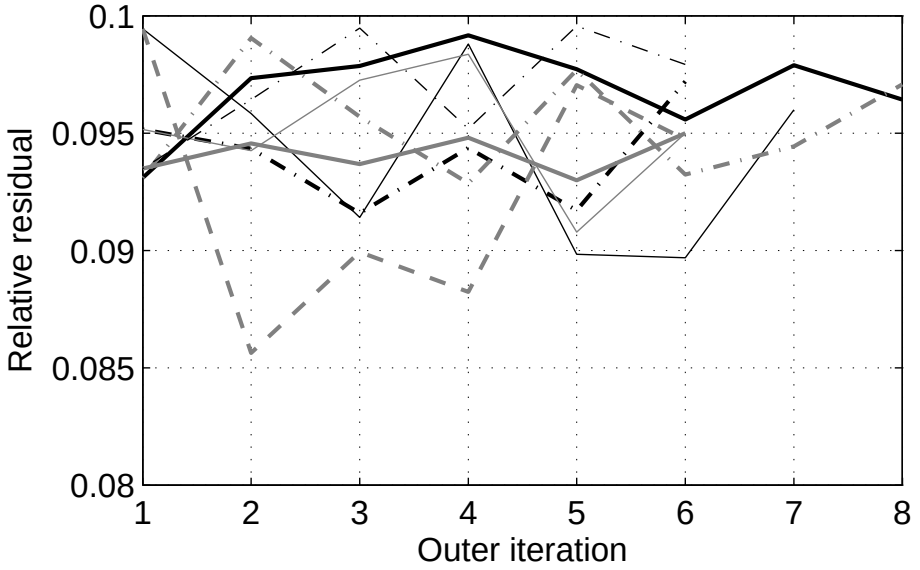


FIG. 4. Relative residual $\gamma$ at each outer iteration of the HF Gauss-Newton methods with and without preconditioning.

overlapped and for each frequency band, a number of $5$ outer iterations are performed. The stopping criteria for the inner iteration is $\tilde{k}_{max} = 10$ or $\gamma_{min} = 2.0e$-1. The stabilization parameters are $\epsilon = 1.0e$-2 and $\lambda = 1.0e$-2.

Figures 7a and 7b are the inversion results obtained by SD and $l$-BFGS methods. Figures 7c and 7d show the comparison of well log data at $0.1$ km and $0.6$ km. As we can see, the SD method is limited in recovering the deep parts of the model. The $l$-BFGS method provides better inversion result compared to SD method but the deep parts of the inversion result are still not satisfactory. Figure 8a is the reconstructed model by HF Gauss-Newton method with the non-preconditioned CG algorithm (CG-GN method). Figures 8b, 8c and 8d show the inversion results by $l$-BFGS-GN ($M = 5$ and $\mathcal{H}_0 = \mathbf{I}$), DPH-GN and $l$-BFGS-

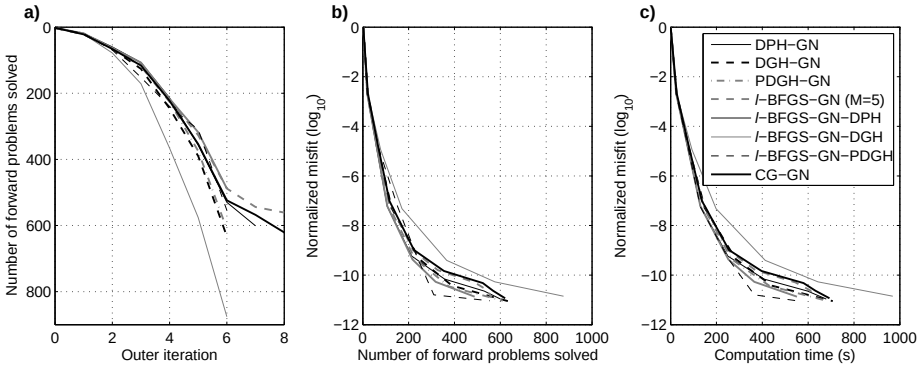


FIG. 5. (a) Number of forward problems solved as the outer iteration proceeds; (b) Normalized misfit $(\log_{10})$ vs. Number of forward problems solved; (c) Normalized misfit $(\log_{10})$ vs. Computation time (s).

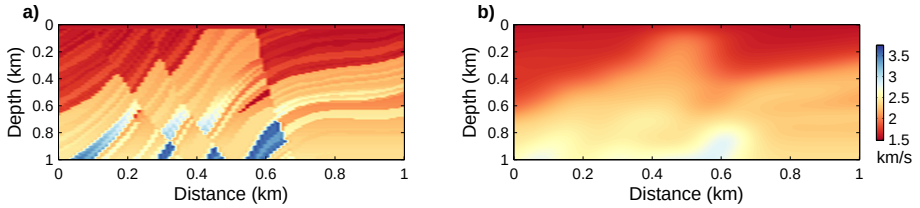


FIG. 6. (a) True P-wave velocity model; (b) Initial P-wave velocity model.

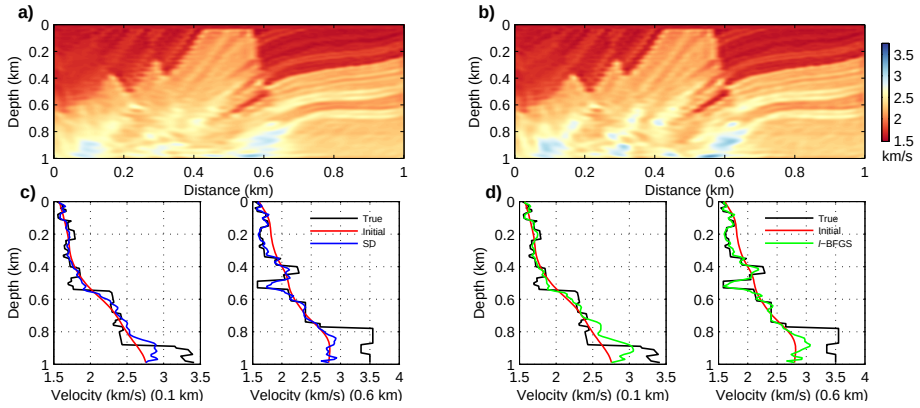


FIG. 7. (a) SD method ($\phi = 0.31$); (b) $l$-BFGS method ($\phi = 0.07$); (c) and (d) are the comparison of the well log data at $0.1$ km and $0.6$ km. The black, red, blue and green lines indicate the true model, initial model, inverted models by SD and $l$-BFGS methods.

GN-DPH methods respectively. Figures 8e and 8f show the well log data comparison at $0.1$ km and $0.6$ km. We first notice that the deep parts of the reconstructed models by HF Gauss-Newton methods have been enhanced obviously compared to SD and $l$-BFGS methods. Furthermore, the inversion results by preconditioned HF Gauss-Newton methods are further improved compared to that by non-preconditioned HF Gauss-Newton method (CG-GN method).

Figures 9a and 9b show the inversion results by DGH-GN and $l$-BFGS-GN-DGH methods with the stabilization parameter $\lambda = 1.0e$-2. The inverted models by DGH-GN and $l$-BFGS-GN-DGH methods are contaminated by artifacts. This is because incorporating the diagonal Gauss-Newton Hessian for preconditioning increases the instability of the in-

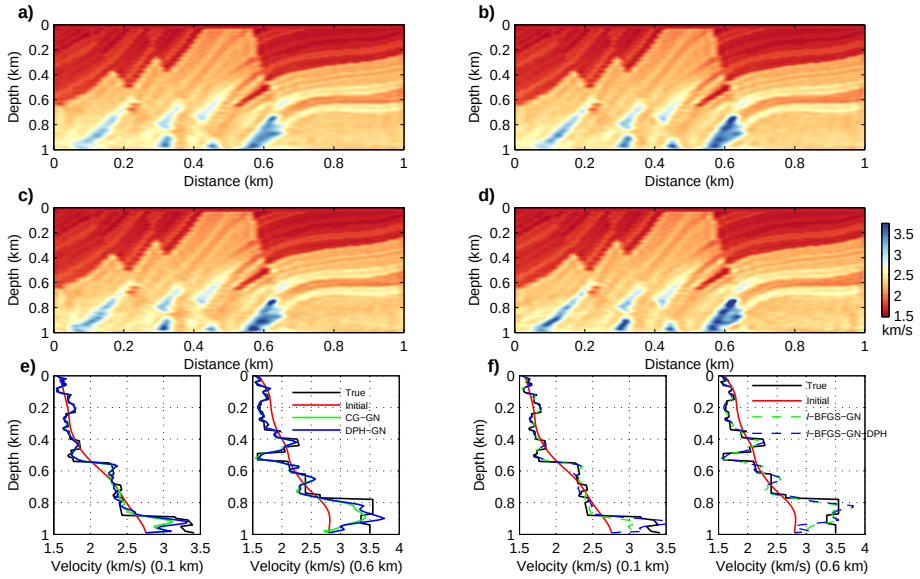


FIG. 8. (a) CG-GN method ($\phi = 4.1e$-3); (b) $l$-BFGS-GN method ($\phi = 1.1e$-3); (c) DPH-GN method ($\phi = 4.4e$-3); (d) $l$-BFGS-GN-DPH method ($\phi = 1.1e$-3); (e) and (f) show the comparison of well log data at $0.1$ km and $0.6$ km. In (e), the green and blue lines indicate the inverted models by CG-GN and DPH-GN methods. In (f), green-dash and blue-dash lines are the inverted models by $l$-BFGS-GN and $l$-BFGS-GN-DPH methods.

version process. Figures 9c and 9d are the inverted models by DGH-GN and $l$-BFGS-GN-DGH methods with the stabilization parameter $\lambda = 5.0e$-2. It can be seen that the DGH-GN and $l$-BFGS-GN-DGH methods with diagonal Gauss-Newton Hessian approximation become more stable and the model can be reconstructed very well. Figures 10a and 10b show the inversion results by PDGH-GN and $l$-BFGS-GN-PDGH methods with the stabilization parameter $\lambda = 5.0e$-2. Figures 10c and 10d show the well log data comparison at $0.1$ km and $0.6$ km. As we can see, the PDGH-GN and $l$-BFGS-GN-PDGH methods based on the proposed pseudo diagonal Gauss-Newton Hessian (equation (23)) can reconstruct the velocity model stably and efficiently. Figure 11 show the model differences between the true model and the inverted models by the above mentioned methods. The $l$-BFGS-GN-DGH and $l$-BFGS-GN-PDGH methods with $\lambda = 5.0e$-2 give the best inversion results.

Figure 12 shows the convergence history of the HF Gauss-Newton methods. We can see that except for DGH-GN and $l$-BFGS-GN-DGH methods ($\lambda = 1.0e$-2), the preconditioned HF Gauss-Newton methods converge faster than the non-preconditioned one (CG-GN method). Generally, the methods with $l$-BFGS preconditioners converge faster than those methods preconditioned by diagonal Hessian approximations. The $l$-BFGS-GN-DGH and $l$-BFGS-GN-PDGH methods with $\lambda = 5.0e$-2 give the fastest convergence rates.

Figure 13a shows the Normalized misfit vs. Number of forward problems solved for the HF Gauss-Newton methods. Figure 13b shows the Normalized misfit vs. Computation time (s). The preconditioned HF Gauss-Newton methods can reconstruct the model more efficiently than the non-preconditioned one (CG-GN method). The $l$-BFGS preconditioning schemes can reduce the computation burden better than the preconditioning methods based on diagonal Hessian approximations. Proper stabilization parameter $\lambda$ should be determined to ensure the effectiveness and stability of the DGH-GN and $l$-BFGS-GN-DGH

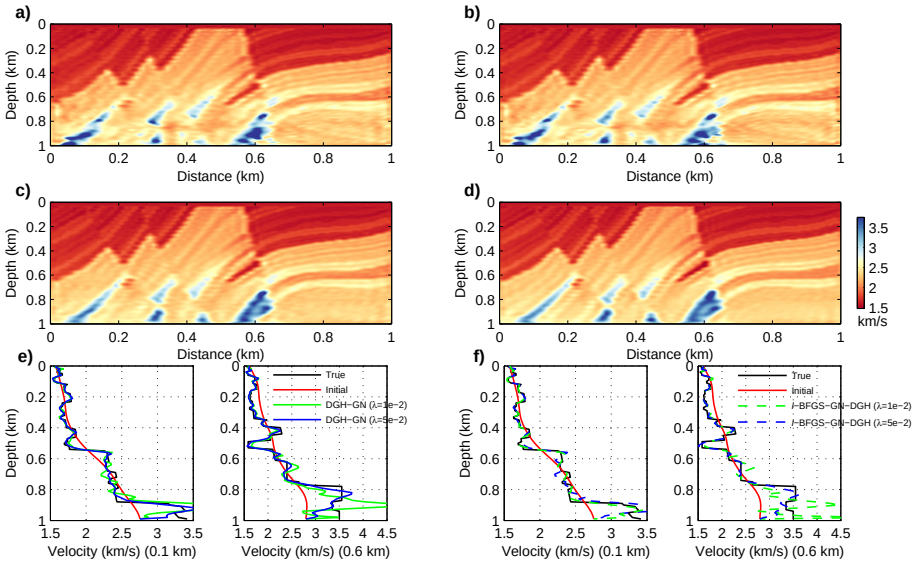


FIG. 9. (a) DGH-GN method ($\lambda = 1.0e$-2 and $\phi = 8.5e$-3); (b) $l$-BFGS-GN-DGH method ($\lambda = 1.0e$-2 and $\phi = 5.2e$-3); (c) DGH-GN method ($\lambda = 5.0e$-2 and $\phi = 2.5e$-3); (d) $l$-BFGS-GN-DGH method ($\lambda = 5.0e$-2 and $\phi = 8.4e$-4); (e) and (f) are the comparison of the well log data at $0.1$ km and $0.6$ km. In (e), the green and blue lines indicate the inverted models by DGH-GN method with $\lambda = 1.0e$-2 and $\lambda = 5.0e$-2. In (f), the green-dash and blue-dash lines indicate the inverted models by $l$-BFGS-GN-PDGH method with $\lambda = 1.0e$-2 and $\lambda = 5.0e$-2.

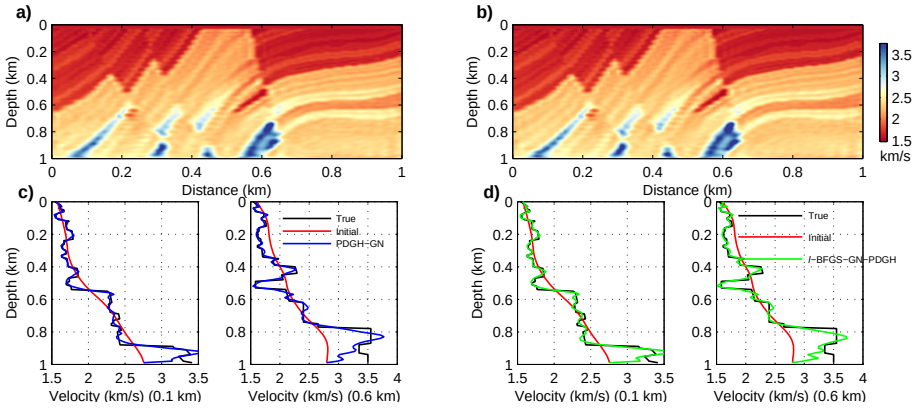


FIG. 10. (a) PDGH-GN ($\lambda = 5.0e$-2 and $\phi = 2.1e$-3); (b) $l$-BFGS-GN-PDGH ($\lambda = 5.0e$-2 and $\phi = 8.3e$-4); (c) and (d) are the comparison of the well log data at $0.1$ km and $0.6$ km. The blue and green lines indicate the inverted models by PDGH-GN and $l$-BFGS-GN-PDGH methods.

methods. Although, $l$-BFGS-GN-DGH method ($\lambda = 5.0e$-2) provides fast convergence rate as shown in Figure 12. More computation cost is required for calculating the receiver-side Green's functions. The $l$-BFGS-GN-PDGH method ($\lambda = 5.0e$-2) based on the proposed pseudo diagonal Gauss-Newton Hessian (equation (23)) does not need additional computational cost compared CG-GN method and it shows the best performance in reducing the computation burden and accelerating the HF Gauss-Newton FWI.

## DISCUSSION

Full-waveform inversion is an ill-posed problem, which means that an infinite number of models matches the data (Virieux and Operto, 2009). Regularization technique

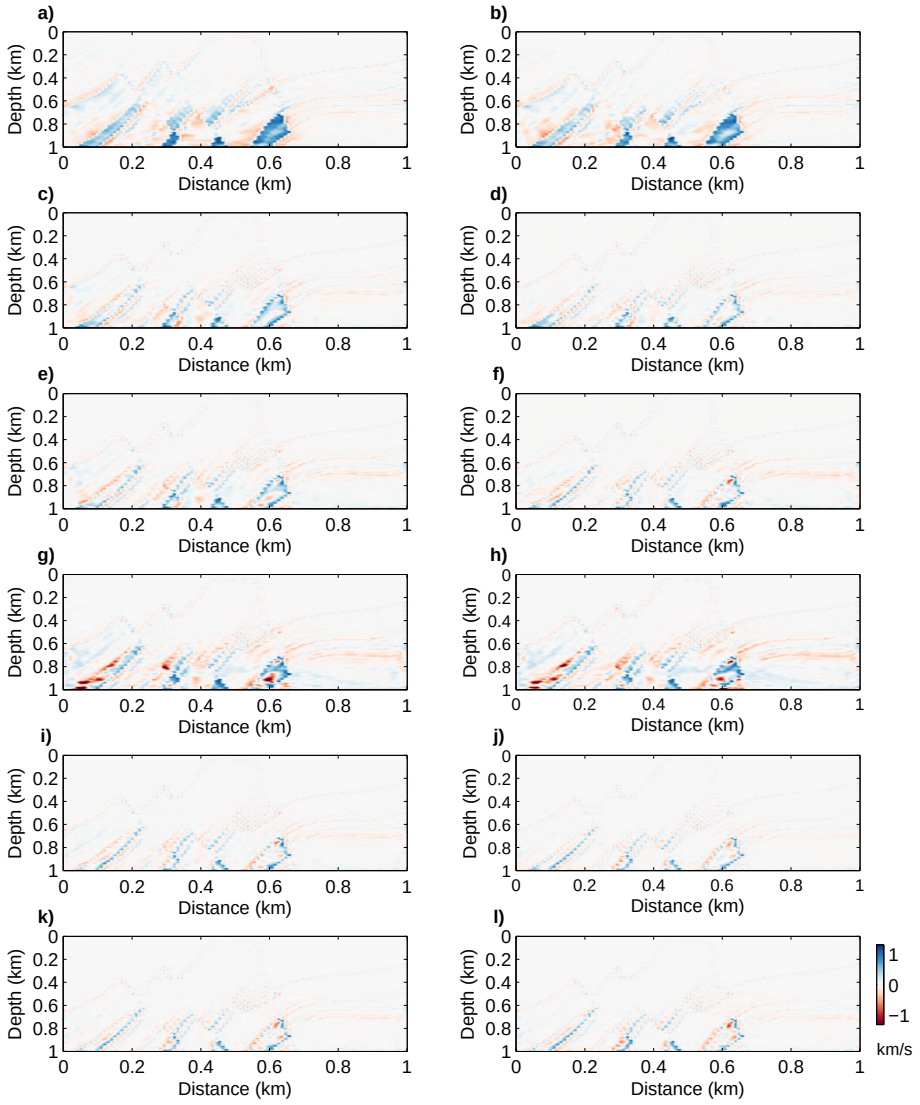


FIG. 11. The model differences between the true model and the inverted models by (a) SD method; (b) $l$-BFGS method; (c) CG-GN method; (d) $l$-BFGS-GN method ($M = 5$ and $\mathcal{H}_0 = \mathbf{I}$); (e) DPH-GN method; (f) $l$-BFGS-GN-DPH method ($M = 5$); (g) DGH-GN method ($\lambda = 1.0e$-2); (h) $l$-BFGS-GN-DGH method ($M = 5$ and $\lambda = 1.0e$-2); (i) DGH-GN method ($M = 5$ and $\lambda = 5.0e$-2); (j) $l$-BFGS-GN-DGH method ($M = 5$ and $\lambda = 5.0e$-2); (k) PDGH-GN method ($\lambda = 5.0e$-2); (l) $l$-BFGS-GN-PDGH method ($M = 5$ and $\lambda = 5.0e$-2).

can alleviate the non-uniqueness of the ill-posed inverse problem and make FWI better posed (Menke, 1984). In this paper, no regularization technique is employed. Hence, for further research, introducing regularization technique, such as Tikhonov regularization (Asnaashari et al., 2013) and Total-variation regularization (Lin, 2015), in the objective function is necessary for improving the performance of the proposed strategies.

In this paper, the numerical examples are carried out with noise-free data set. In practical applications, the seismic data is contaminated by various noises (e.g., random noise and multiples), which will impact the performances of full-waveform inversion significantly. Further studies are needed for testing the sensitivity of the proposed strategies to noisy data set. The pseudo diagonal Gauss-Newton Hessian is proposed by assuming that the sources

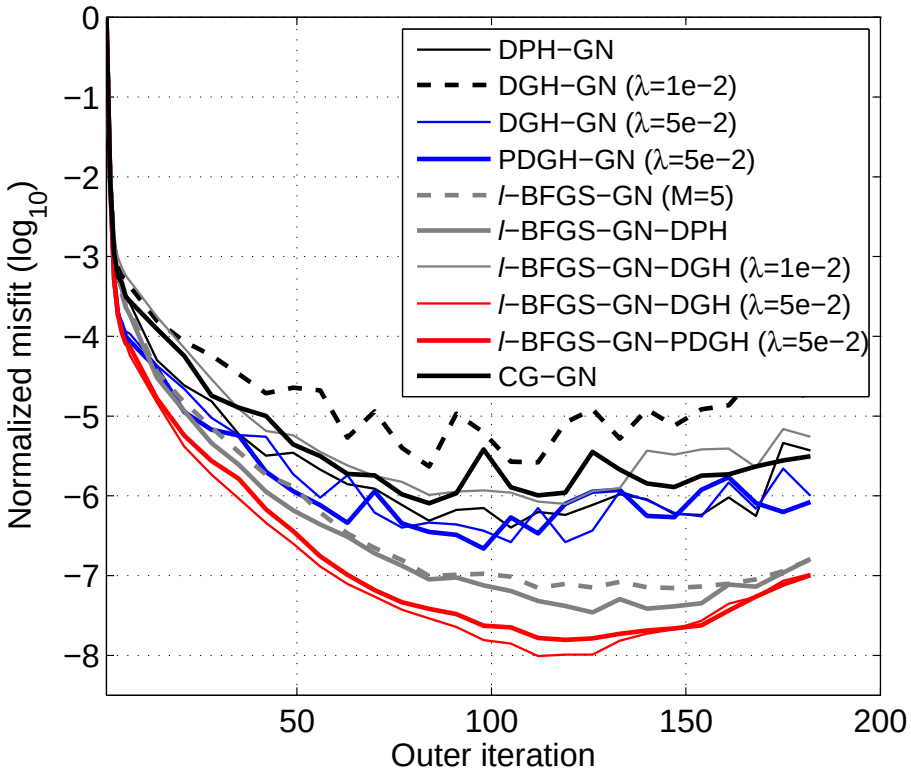


FIG. 12. Comparison of the convergence history for the HF Gauss-Newton methods with different preconditioning schemes. The black-thin-solid, black-bold-dash, blue-thin-solid, blue-bold-solid lines indicate DPH-GN, DGH-GN ($\lambda = 1.0e$-2), DGH-GN ($\lambda = 5.0e$-2) and PDGH-GN ($\lambda = 5.0e$-2). The gray-bold-dash, gray-bold-solid, gray-thin-solid, red-thin-solid, red-bold-solid lines indicate $l$-BFGS-GN, $l$-BFGS-GN-DPH, $l$-BFGS-GN-DGH ($\lambda = 1.0e$-2), $l$-BFGS-GN-DGH ($\lambda = 5.0e$-2) and $l$-BFGS-GN-PDGH ($\lambda = 5.0e$-2) methods. The black-bold-solid line denotes CG-GN method.

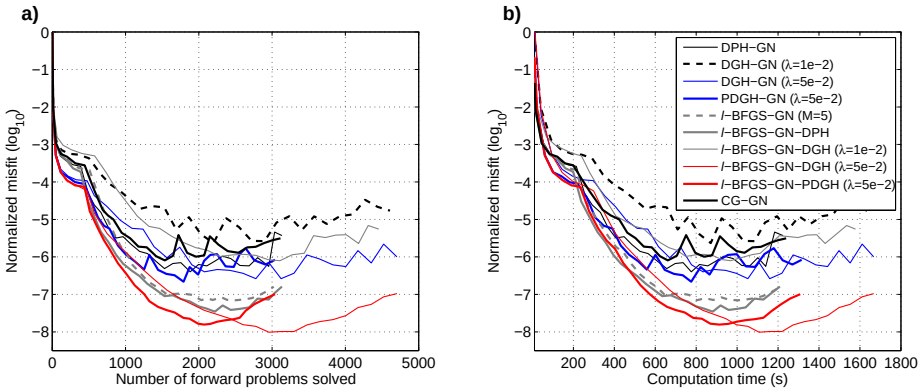


FIG. 13. (a) Normalized misfit ($\log_{10}$) vs. Number of forward problems solved for different preconditioning strategies; (b) Normalized misfit ($\log_{10}$) vs. Computation time (s).

and receivers are co-located, which is only valid for reflection survey. For transmission survey or VSP survey this approximation is not appropriate.

The high computation cost remains an obstacle of FWI for practical applications. One popular approach for mitigating this problem is the phase-encoding (or simultaneous source) method, in which the super-gathers are constructed by summing densely distributed indi-

vidual shots (Vigh and Starr, 2008; Anagaw and Sacchi, 2014; Pan et al., 2015a; Castellanos et al., 2015). In this paper, the traditional shot-profile method is employed for modelling and inversion. Further research should be carried out to test the performances of the proposed methods with phase-encoding technique including random phase-encoding and linear phase-encoding.

## CONCLUSIONS

In this paper, a Hessian-free Gauss-Newton method is carried out for full-waveform inversion in comparison with steepest-descent and $l$-BFGS methods. The HF Gauss-Newton method is implemented with a double-iterative scheme, in which a conjugate-gradient algorithm with a matrix-free fashion is employed for solving the Newton linear system. To accelerate the Hessian-free Gauss-Newton method, we develop different preconditioning schemes for the inner CG algorithm at each outer iteration. A pseudo diagonal Gauss-Newton Hessian is also proposed as preconditioner based on the reciprocal property of the Green's function. We present two numerical examples to show that the preconditioning schemes can improve the convergence rate of Hessian-free Gauss-Newton FWI and reduce the computation cost. It is concluded that the $l$-BFGS preconditioning method with pseudo diagonal Gauss-Newton Hessian as initial guess can speed up the HF Gauss-Newton FWI most efficiently.

## REFERENCES

Anagaw, A. Y., and Sacchi, M. D., 2012, Full waveform inversion with simultaneous sources using the full newton method: SEG Expanded Abstracts, 971–975.

Anagaw, A. Y., and Sacchi, M. D., 2014, Comparison of multifrequency selection strategies for simultaneous-source full-waveform inversion: Geophysics, **79**, R165–R181.

Armijo, L., 1966, Minimization of functions having Lipschitz continuous first partial derivatives: Pacific J. Math, **16**, 1–3.

Asnaashari, A., Brossier, R., Garambois, S., Audebert, F., Thore, P., and Virieux, J., 2013, Regularized seismic full-waveform inversion with prior model information: Geophysics, **78**, R25–R36.

Brossier, R., Operto, S., and Virieux, J., 2009, Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion: Geophysics, **74**, WCC105–WCC118.

Brossier, R., Operto, S., and Virieux, J., 2010, Which data residual norm for robust elastic frequency-domain full-waveform inversion?: Geophysics, **75**, R37–R46.

Broyden, C. G., 1970, The convergence of a class of double-rank minimization algorithms: IMA Journal of Applied Mathematics, **6**, 222–231.

Bunks, C., Saleck, F. M., Zaleski, S., and Chavent, G., 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473.

Byrd, R. H., Lu, P., and Nocedal, J., 1995, A limited memory algorithm for bound constrained optimization: SIAM Journal on Scientific and Statistical Computing, **16**, 1190–1208.

Castellanos, C., Métivier, L., Operto, S., Brossier, R., and Virieux, J., 2015, Fast full waveform inversion with source encoding and second-order optimization methods: Geophysical Journal International, **200**, 720–744.

Davis, T. A., and Duff, I. S., 1997, An unsymmetric pattern multifrontal method for sparse lu factorization: SIAM Journal on Matrix Analysis and Applications, **18**, 140–158.

Demanet, L., Letourneau, P., Boumal, N., Calandra, H., and Snelson, S., 2012, Matrix probing: a randomized preconditioner for the wave equation Hessian: Appl. and Comp. Harmon. Anal., **32**, R25–R36.

Dembo, R. S., and Eisenstat, S. C., 1982, Inexact Newton methods: SIAM J.Number.Anal, **19**, 400–408.

Engquist, B., and Majda, A., 1977, Absorbing boundary conditions for the numerical simulations of waves: Mathematical Computation, **31**, 629–651.

Fichtner, A., and Trampert, J., 2011, Hessian kernels of seismic data functionals based upon adjoint techniques: Geophysical Journal International, **185**, 775–798.

Fletcher, R., 1970, A new approach to variable metric algorithms: The Computer Journal, **13**, 317–322.

Fletcher, R., and Reeves, C. M., 1964, Function minimization by conjugate gradients: Comp. J., **7**, 149–154.

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

Goldfarb, D., 1970, A family of variable-metric methods derived by variational means: Mathematics of Computation, 23–26.

Hu, W., Abubakar, A., and Habashy, T. M., 2009, Simultaneous multifrequency inversion of full-waveform seismic data: Geophysics, **74**, R1–R14.

Hu, W., Abubakar, A., Habashy, T. M., and Liu, J., 2011, Preconditioned non-linear conjugate gradient method for frequency domain full-waveform seismic inversion: Geophysical Prospecting, **59**, 477–491.

Innanen, K. A., 2014, Reconciling seismic AVO and precritical reflection FWI-analysis of the inverse Hessian: SEG Technical Program Expanded Abstracts, 1022–1027.

Jo, C. H., Shin, C., and Suh, J. H., 1996, An optimal 9-point, finite-difference, frequency-space, 2-D scalar wave extrapolator: Geophysics, **61**, 529–537.

Jun, H., Park, E., and Shin, C., 2015, Weighted pseudo-Hessian for frequency-domain elastic full waveform inversion: Weighted pseudo-Hessian for frequency-domain elastic full waveform inversion, **123**, 1–17.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migration: Conference on Inverse Scattering, Theory and Applications, SIAM, Expanded Abstracts, 206–220.

Letourneau, P., Demanet, L., and Calandra, H., 2012, Approximate inversion of the wave-equation Hessian via randomized matrix probing: SEG Technical Program Expanded Abstracts, 1–6.

Lin, Y., 2015, Acoustic- and elastic-waveform inversion using a modified total-variation regularization scheme: Geophysical Journal International, **200**, 489–502.

Ma, Y., and Hale, D., 2012, Quasi-Newton full-waveform inversion with a projected Hessian matrix: Geophysics, **77**, R207–R216.

Marfurt, K., 1984, Accuracy of finite-difference and finite-elements modeling of the scalar and elastic wave equation: Geophysics, **49**, 533–549.

Menke, W., 1984, Geophysical data analysis: Discrete inverse theory: Academic Press.

Métivier, L., Bretaudeau, F., Brossier, R., Virieux, J., and Operto, S., 2014, Full waveform inversion and the truncated Newton method: quantitative imaging of complex subsurface structures: Geophysical Prospecting, **62**, 1–23.

Métivier, L., Brossier, R., Virieux, J., and Operto, S., 2012, The truncated Newton method for full waveform inversion: SEG Technical Program Expanded Abstracts, 1–5.

Nammour, R., and Symes, W., 2009, Approximate constant density acoustic inverse scattering using dip-dependent scaling: SEG Technical Program Expanded Abstracts, 2347–2351.

Nash, S. G., 1985, Preconditioning of truncated-newton methods: SIAM J. Sci. Statist. Comput, **6**, 599–616.

Nash, S. G., 2000, A survey of truncated-newton methods: Journal of computational and applied mathematics, **124**, 45–59.

Nocedal, J., 1980, Updating quasi-Newton matrices with limited storage: Mathematics of Computation, **35**, 773–782.

Nocedal, J., and Wright, S. J., 2006, Numerical Optimization: Springer.

Operto, S., Gholami, Y., Prieux, V., Ribodetti, A., Brossier, R., Metivier, L., and Virieux, J., 2013, A guided tour of multiparameter full waveform inversion with multicomponent data: from theory to practice: The Leading Edge, **32**, 1040–1054.

Operto, S., Virieux, J., Dessa, J. X., and Pascal, G., 2006, Crustal imaging from multifold ocean bottom seismometers data by frequency domain full-waveform tomography: application to the eastern nankai trough: Journal of Geophysical Research, **111**.

Pan, W., Innanen, K. A., and Margrave, G. F., 2014a, A comparison of different scaling methods for least-squares migration/inversion: EAGE Expanded Abstracts, We G103 14.

Pan, W., Innanen, K. A., Margrave, G. F., and Cao, D., 2015a, Efficient pseudo-Gauss-Newton full-waveform inversion in the $\tau$-$p$ domain: Geophysics, **80**, R225–R14.

Pan, W., Innanen, K. A., Margrave, G. F., Fhler, M. C., Fang, X., and Li, J., 2015b, Estimation of elastic constants in HTI media using Gauss-Newton and Full-Newton multi-parameter full waveform inversion: SEG Technical Program Expanded Abstracts, 1177–1182.

Pan, W., Innanen, K. A., Margrave, G. F., Fhler, M. C., Fang, X., and Li, J., 2015c, Estimation of elastic constants in HTI media using Gauss-Newton and Full-Newton multi-parameter full-waveform inversion, submitted.

Pan, W., Margrave, G. F., and Innanen, K. A., 2014b, Iterative modeling migration and inversion (IMMI): Combining full waveform inversion with standard inversion methodology: SEG Technical Program Expanded Abstracts, 938–943.

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., 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, 495–503.

Plessix, R. E., and Mulder, W. A., 2004, Frequency-domain finite-difference amplitude-preserving migration: Geophysical Journal International, **157**, 975–987.

Pratt, R. G., and Chapman, C. H., 1992, Traveltime tomography in anisotropic media-II. application: Geophysical Journal International, **109**, 20–37.

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.

Saad, Y., 2003, Iterative methods for sparse linear systems: SIAM.

Sainath, T. N., Horesh, L., Kingsbury, B., Aravkin, A. Y., and Ramabhadran, B., 2013, Accelerating Hessian-free optimization for Deep Neural Networks by implicit preconditioning and sampling: IEEE Workshop on Automatic Speech Recognition and Understanding, 303–308.

Santosa, F., and Symes, W. W., 1988, Computation of the hessian for least-squares solutions of inverse problems of reflection seismology: Inverse Problems, **4**, 211–233.

Shanno, D. F., 1970, Conditioning of quasi-Newton methods for function minimization: Mathematics of Computation, **24**, 647–656.

Shin, C., Jang, S., and Min, D., 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., Lim, H. Y., Chung, S., and Shin, S., 2001b, Efficient calculation of a partial-derivative wavefield using reciprocity for seismic imaging and inversion: Geophysics, **66**, 1856–1863.

Sirgue, L., and Pratt, R. G., 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: Geophysics, **69**, 231–248.

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.

Tromp, J., Tape, C., and Liu, Q., 2005, Seismic tomography, adjoint methods, time reversal, and banana-doughnut kernels: Geophysical Journal International, **160**, 195–216.

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, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, WCC1–WCC26.

Wu, S., Wang, Y., Zheng, Y., and Chang, X., 2015, Limited-memory BFGS based least-squares pre-stack Kirchhoff depth migration: Geophysical Journal International, **202**, 738–747.