APPLYING A DIAGONAL HESSIAN APPROXIMATION FOR PRECONDITIONING IN 3D ELASTIC FULL WAVEFORM INVERSION

S. Butzer, A. Kurzmann and T. Bohlen

email: *simone.butzer@kit.edu*

keywords: *full waveform inversion, elastic, 3D, adjoint, gradient, preconditioning, diagonal Hessian*

ABSTRACT

Full waveform inversion has the potential to recover high resolution models for multiple parameters. Our approach is based on the gradient method, in which gradients are calculated from forward and adjoint wavefields. Due to the geometrical amplitude decay of these wavefields, we observe high amplitudes in the gradients around source and receiver positions. Thus, a successful inversion requires a good preconditioning. Here, we apply the inverse of a diagonal Hessian approximation for preconditioning, which is a physically founded approach. However, its calculation is computationally expensive, as it requires one additional forward simulation for each receiver, and we therefore calculate it only once for each frequency range. We show the preconditioning for two examples, a transmission geometry example and a surface acquisition example. In transmission geometry, source and receiver artefacts were removed sufficiently and the inversion was successfully performed. Still, the effects are much more profound in the surface geometry example. Here, the gradient in shallow areas is emphasised by the acquisition geometry and the presence of surface waves while its amplitude rapidly decays with depth. The application of our preconditioning approach mitigates these effects and allows a meaningful model update in deeper areas.

INTRODUCTION

Full waveform inversion (FWI) aims to resolve structures of the subsurface in high resolution by minimising the misfit between modeled and observed data. To solve this optimisation problem, we use the conjugate gradient approach (e.g. Tarantola, 1984; Mora, 1987), which uses the gradient of the misfit function to approach its minimum. This approach can be implemented very efficiently with the adjoint method (e.g. Mora, 1987; Plessix, 2006) and is thus realisable for large model and dataset applications. Another class of optimisation methods, the Newton methods, take into account the second derivative of the misfit function, the so-called Hessian matrix. The use of the Hessian can significantly improve the performance of FWI, by speeding up convergence and improving resolution. The Hessian matrix can account for geometrical amplitude effects in the gradient due to source receiver coverage and for limited-bandwidth effects and can thus focus and sharpen the image (Pratt et al., 1998; Brossier et al., 2009). The full-Newton and Gauss-Newton methods explicitely calculate the Hessian or in the latter case the approximate Hessian. However, these methods are computationally very expensive and thus not attractive for realistic problem sizes (Pratt et al., 1998). Quasi-Newton methods, such as the L-BFGS method (Byrd et al., 1995) and the truncated Newton method (Métevier et al., 2012) are computationally more feasible, as they do not calculate the Hessian matrix directly. The L-BFGS method uses changes in gradients and models from recent iterations to approximate the Hessian, whereas the truncated Newton formula is based on second-order adjoint equations.

Another approach, which includes information about the Hessian, is the use of an approximation of the diagonal of the Hessian for preconditioning in the conjugate gradient method (Shin et al., 2001; Brossier et al., 2009). Generally, the gradient shows high amplitudes near sources and receivers due to the geometrical spreading of the forward and adjoint wavefields. Thus, a thorough preconditioning of the gradients is required for a successfull inversion. This can be done by Gaussian tapering around sources and receivers. Still, a more physical and sophisticated approach is the use of the diagonal of the Hessian, which can correctly account for the geometrical amplitude effects in the gradient. In this report we will discuss the theory and implementation of the gradient method with Hessian preconditioning for 3D elastic FWI and show its performance for two simple examples.

THEORY AND IMPLEMENTATION

Newton and conjugate gradient methods

Full waveform inversion (FWI) aims to minimise the misfit beween observed and modeled data. We use the L_2-norm based misfit function E given as

$$E = \frac{1}{2} \sum_{sources} \int dt \sum_{receivers} \delta u_i(\mathbf{x}_s, \mathbf{x}_r, t) \delta u_i(\mathbf{x}_s, \mathbf{x}_r, t) \tag{1}$$

with the i-th component of the displacement residual $\delta u_i = u_i - u_{i,obs}$ at source position $\mathbf{x}_s$ and receiver position $\mathbf{x}_r$. Different optimisation approaches can be used for minimizing the misfit and a good discussion about different Newton methods and gradient methods can be found in Pratt et al. (1998). We will give a short overview here.
A Taylor expansion of the misfit function E around the model parameters $\mathbf{m} = (m_1, ..., m_n)^T$ up to second order leads to:

$$E(\mathbf{m} + \delta\mathbf{m}) = E(\mathbf{m}) + \delta\mathbf{m}^T \nabla_m E(\mathbf{m}) + \frac{1}{2} \delta\mathbf{m}^T \mathbf{H} \delta\mathbf{m} + O(|\delta\mathbf{m}|^3). \tag{2}$$

Hereby $\mathbf{H}$ is defined as the second derivative of the misfit with respect to the model parameters, i.e.,

$$H_{ij} = \frac{\partial^2 E(\mathbf{m})}{\partial m_i \partial m_j} \quad (i = 1, ...n) \;\; (j = 1, ...n). \tag{3}$$

The index n is the total number of model parameters. To find the minimum of the misfit function the deviation of equation 2 with respect to the model perturbation $\delta\mathbf{m}$ is set zero. This gives us the following model update:

$$\delta\mathbf{m} = -\mathbf{H}^{-1} \nabla_m E. \tag{4}$$

The Newton method uses this model update for an iterative solution and updates the model in iteration $k + 1$ with

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \mathbf{H}_k^{-1} \nabla_m E_k. \tag{5}$$

The use of the Hessian leads to a good convergence. However, the calculation and inversion of the large $n \times n$ Hessian matrix is highly expensive. The gradient method thus uses the following simplified update:

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha_k \mathbf{P} \nabla_m E_k. \tag{6}$$

In this case, the model is updated in gradient direction using an appropriate step length α. In this case, the gradient is not scaled and preconditioned by the inverse Hessian. Consequently, a preconditioning operator $\mathbf{P}$ and a the estimation of α are required for the inversion to succeed. The corresponding convergence is lower. A slight improvement in convergence of the gradient method is gained by using the conjugate gradient $\mathbf{c}$ as search direction (Mora, 1987), which is given by

$$\mathbf{c}_k = \mathbf{P} \nabla_m E_k + \beta_k \mathbf{c}_{k-1}, \tag{7}$$

where the scalar β is calculated as given by Mora (1987). This gives as the following model update:

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha_k \mathbf{c}_k. \tag{8}$$

The preconditioning operator used in the gradient method is the replace of the Hessian operator. Thus, it is natural to use some approximation of the Hessian for preconditioning, which can lead to improvements in the gradient method.

Gradient calculation

We use the adjoint-state method to calculate the gradients (e.g. Tarantola, 1984; Mora, 1987; Pratt et al., 1998; Plessix, 2006). Hereby we use a time-frequency approach (Sirgue et al., 2008), where the forward propagation is done in time domain and the gradients are calculated for few discrete frequencies in frequency domain. To transform the wavefields from time to frequency domain, a discrete Fourier transform is performed on the fly. A detailed introduction of our implementation can be found in Butzer et al. (2013). For forward modeling, we use the 3D viscoelastic finite-difference code (SOFI), which is based on a velocity-stress formulation (Bohlen, 2002). The following steps are performed to calculate the gradient direction in each iteration k:

1. forward propagation of source wavefield across themedium
2. calculate residual between modeled and observed seismograms
3. backpropagation of residual wavefield from receivers across the medium with time-reversed residuals acting as source-time function

   in step 1) and 3): discrete Fourier transformation on the fly
   steps 1)-3): these steps are performed for each source
4. gradient ($\nabla_m E_k$) calculated as multiplication of forward wavefield and conjugate backpropagated wavefield in frequency domain and summed up over all frequencies and sources

Afterwards, a preconditioning operator is applied to the gradients and the conjugate gradient direction $\mathbf{c}_k$ (equation 7) is calculated. To estimate an optimal step length α_k we use the misfit value of zero steplength and of two additional test steplengths calculated for a subset of shots. The model can then be updated according to equation 8.
The main computational time in FWI is spent for wavefield modelings. In the gradient method the number of forward modelings is 2$\times$ (number of shots) + 2$\times$ (number shots steplength calculation).

Calculation of the diagonal Hessian approximation

In this section, we will introduce how the diagonal Hessian approximation $\mathbf{H}_D$, that we use for preconditioning, is calculated. Detailed discussions about the calculation of the Hessian matrix can be found in Sheen et al. (2006) and Pratt et al. (1998). The Hessian (equation 3) can be calculated as

$$\mathbf{H} = \mathrm{Re}(\mathbf{J}^t \mathbf{J}^*) + \mathbf{R}. \tag{9}$$

The second term R is generally small (Pratt et al., 1998), and we only use the first term, known as the approximate Hessian. $\mathbf{J}$ is the Jacobian matrix, which is defined as

$$J_{ij} = \frac{\partial u_i}{\partial m_j} \qquad i = 1, 2, ..., w, \quad j = 1, 2, ..., n. \tag{10}$$

The indice i runs over all wavefield parameters w and the indice j runs over all model parameters n. For preconditioning the calculation of $\mathbf{H}$ is restrained to the diagonal elements of the approximate Hessian, with

$$H_{jj} = \sum_i \frac{\partial u_i}{\partial m_j} \frac{\partial u_i^*}{\partial m_j}. \tag{11}$$

The full Hessian is very large with $n \times n$ elements, whereas the diagonal Hessian only consists of n elements. The Jacobian matrix is not explicitely calculated in the gradient method and additional computations are required to calculate it for the diagonal Hessian. The following steps describe, how the Jacobian matrices are constructed and how the diagonal Hessian approximation $\mathbf{H}_D$ is calculated.

1. forward propagation from each source into medium; this is already done for gradient computation
2. backpropagation of delta functions from each receiver into media to find the Green's receiver functions

   in step 1) and 2): discrete Fourier transforms on the fly
3. calculation of Jacobian matrices for each source-receiver combination by multiplication of forward wavefield and conjugate receiver Green's functions in frequency domain
4. calculate diagonal Hessian approximation $\mathbf{H}_D$ as multipliaction of the complex Jacobian matrices with their conjugate, summed up over all frequencies and source-receiver combinations

The calculation of Jacobian matrices for each source-receiver combination is required. Hence, either the Green's receiver functions or the forward propagated wavefields need to be stored. In our implementation, the Hessian calculation is performed in frequency domain, which means, that these wavefields are stored only for few discrete frequencies.
In the gradient calculation, the backpropagated wavefield is generated at all receivers simultanously. For the calculation of $\mathbf{H}_D$ we need the Green's receiver functions. Thus, one forward propagation for each receiver is performed additionally to the gradient calculations. Depending on the number of receivers, this can be very time consuming.
To compute the full diagonal approximate Hessian, the Green's receiver functions are calculated for each spatial direction, which requires $3\times$ (number receivers) modelings in step 2. We use only the component which dominates the forward wavefield.

Application of Hessian preconditioning

After calculation of $\mathbf{H}_D$, as described in the last section, we use this operator to precondition the gradients. The preconditioning operator is then given by:

$$\mathbf{P} = (\mathbf{H}_D + \epsilon\mathbf{I})^{-1}. \tag{12}$$

$\mathbf{I}$ is the identity matrix. Hereby a water level ϵ is added to $\mathbf{H}_D$ to stabilise the inversion. At the moment, we estimate this water level empirically. The inversion of $\mathbf{H}_D$ is straightforward, because we use only the diagonal part of the Hessian.
The Hessian is calculated only once for each frequency stage, because its calculation might become quite expensive. Then the same $\mathbf{P}$ is applied within this frequency stage.

EXAMPLES

Example 1 - Box in transmission geometry

Model and inversion setup Even though the implementation of the Hessian preconditioning is of main interest for its application to FWI of complex models, we choose a relatively simple test to show its effects on the gradients and to prove its performance in the inversion. We consider a transmission geometry test of a box model with the size of $160\times160\times184$ grid points corresponding to 128 m$\times$128 m$\times$147.2 m in x-, y- and z-direction. The true models for the seismic velocities are shown in Figure 1, the density model is chosen constant and known with 2800 kg/m^3. The seismic velocity models contain a box, which is divided into four differently-sized parts with different positiv and negativ velocity variations. The v_p/v_s ratio is not constant. The model does not contain a free surface. As a starting model, the homogeneous background velocities outside the box are used.
Sources and receivers are arranged within x-y-planes, as indicated in Figure 1. We use 12 (3×4) sources

in 88 m depth and 169 (13×13) receivers in 24 m depth. The sources are vertical directed point forces with $\sin^3$-wavelets as source time functions and a dominant frequency of 200 Hz.

In total, we performed 75 iterations, divided into 4 different frequency stages ranging from 160 Hz to

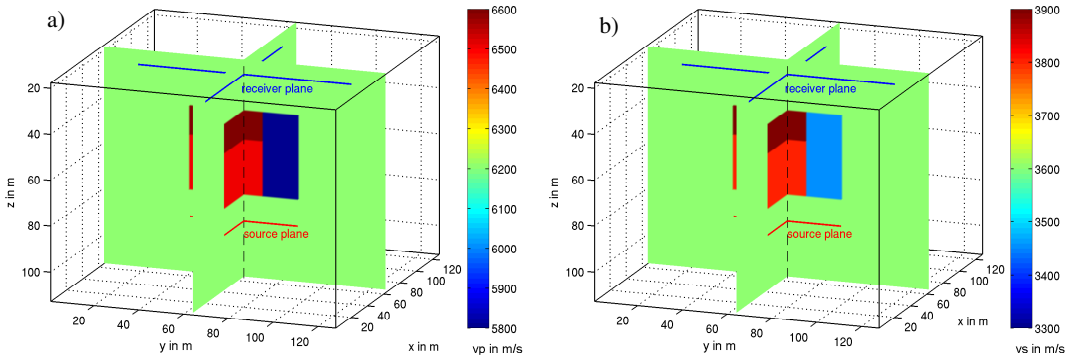

Figure 1: Real box model of v_p (a) and v_s (b) with indication of source and receiver plane.

290 Hz. In each stage five discrete frequencies in 10 Hz intervals were used for inversion and frequencies increased from stage to stage.

For the general gradient method 30 forward modelings are calculated within each iteration. 169 additional forward modelings are required for the calculation of the Green's receiver functions, needed for the diagonal Hessian preconditioning matrix $\mathbf{H}_D$. This matrix was calculated only once for each frequency stage and used for the whole stage. Thus, for the full inversion 667 forward modelings are added to the 2250 forward modelings of the general gradient approach, which is an increase of about 30% of runtime in this example.

Hessian preconditioning Figure 2 shows the effects of the diagonal Hessian preconditioning on the gradients of v_p and v_s for the first iteration, and thus for the first frequency stage. The gradients before preconditioning, normalised to their maximum are shown in Figure 2a) for v_p and b) for v_s. The high amplitudes around sources and receivers are clearly visible. Without preconditioning, the model update is only significant within these areas, and the inversion fails. Figure 2c) and d) shows the logarithm of the normalised diagonal Hessian approximation $\mathbf{H}_D$ for v_p and v_s. The Hessian matrix covers several orders of magnitude and, like the gradient, it shows extremly high values at source and receiver positions. The influence of the geometric amplitude decay of the wavefield is clearly visible. Areas with no or very low wavefield coverage show very low values. This is for example visible in the blue areas of the v_p Hessian. The application of the inverse Hessian in such areas would thus lead to an enormous enhancement of the gradient, even though we have no or very little information in our data. To avoid this, the water level is added to the Hessian, which, at the moment, we determine manually.

$\mathbf{H}_D$ is used as preconditioner according to equation 12 and applied to the gradient. The normalised preconditioned gradient is shown in Figure 2e) and f) for v_p and v_s, respectively. The high amplitudes at the sources are corrected. Most of the high amplitudes around the receivers also vanished, however, some receiver artefacts are still visible. The reason for this is probably the approximation we do by calculating only Green's receiver functions from delta functions applied as vertical forces. Hence, some additional damping at receiver positions might be required. The highest amplitude in the gradient now concentrates on the box area. Some effects of the preconditiong are also visible within the box area, where amplitudes of the preconditioned gradient are higher in the middle part of the box. However, for transmission geometry, these effects are low. Unfortunately, some artefact below the source plane are increased after preconditioning in the gradient of v_s. However, the inversion results show, that these artefacts in the gradients have no significant impact on the final inversion result.

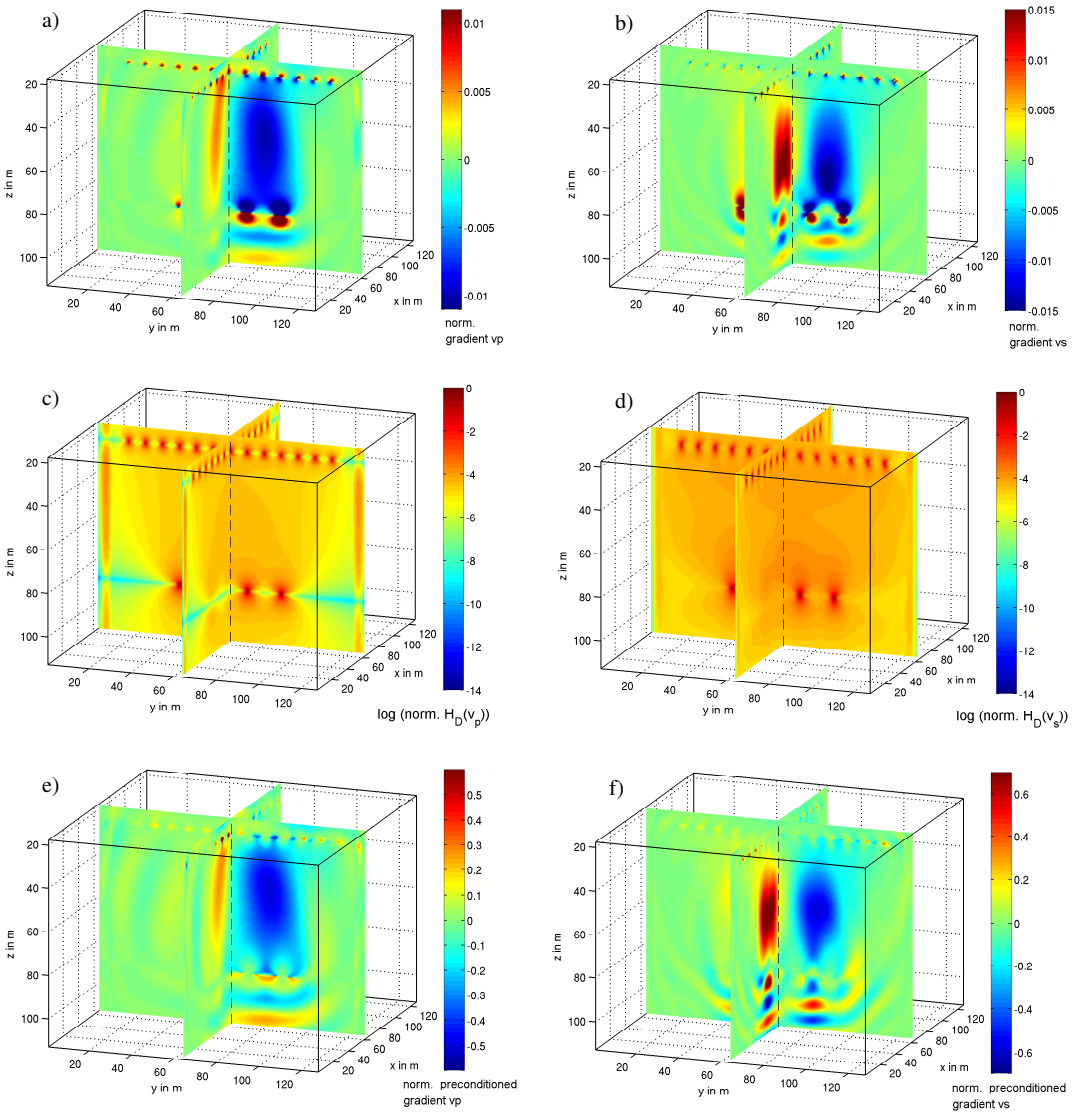

Figure 2: Effects of Hessian preconditioning on the gradients of v_p and v_s in the box model, a) normalised gradient v_p, b) normalised gradient v_s, c) logarithm of normalised $\mathbf{H}_D(v_p)$, d) logarithm of normalised $\mathbf{H}_D(v_s)$, e) normalised preconditioned gradient v_p and f) normalised preconditioned gradient v_s.

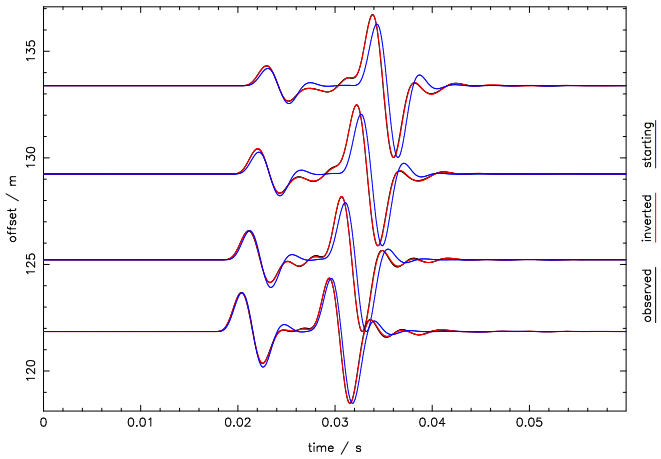

Figure 3: Seismograms of vertical component for observed, starting and final inverted data for few representative traces.

Inversion results Figure 3 shows the data fit of observed, starting and final inverted data for some representative traces for the vertical component. The data is lowpass-filtered with a corner frequency of 290 Hz, which is the maximum frequency used for the inversion. It is visible, that the seismograms of the homogeneous starting model are already relatively close to the data. The final inverted data and the observed data show a nearly perfect fit.

The final inverted models of v_p and v_s are shown for two slices: a horizontal slice in Figure 4 and a vertical slice in Figure 5. For comparison, the real models are plotted. Overall, the box could be successfully reconstructed by the FWI. In the horizontal slice of the final inverted models in Figure 4 b) and d) the three sub-boxes are successfully resolved. Due to the smaller wavelengths of the shear wave, the shape of the sub-boxes is clearer in v_s copared to v_p.

The results plotted in the vertical slice are not as well resolved. The high velocity anomalies on the right side could not be distinguished by the inversion. This is not unexpected when looking at the minimum wavelengths of 21 m for v_s and 35 m for v_p within this high velocity zone, which is only 12 m thick. The resolution of FWI for transmission geometry is about a wavelength and we would therefore require higher frequencies for a better reconstruction.

Example 2 - Surface acquisition geometry

Model and inversion setup The effects of Hessian preconditioning are more profound in surface geometry experiments. We consider a 3D layered onshore surface model as shown in Figure 6 for v_p (a) and v_s (b). The model consists of sedimentary layers over a basin-shaped homogeneous halfspace. The size is of $320\times320\times160$ grid points which corresponds to 256 m$\times$256 m$\times$128 m. At $z = 0$m, the model contains a free surface. The density is homogeneous with 2900 kg/m^3 and remains constant during inversion. For the compressional and shear wave velocities we use a smoothed version of the real model as starting model (see Figure 6 b) and d)). We used 49 (7×7) sources and 81 (9×9) receivers equally distributed along the surface. As source time function, we use a $\sin^3$ function with a dominant frequency of 20 Hz applied as vertical point force.

In this work we will show a preliminary investigation of the effects of Hessian preconditioning on the gradient in the first iteration. The gradient and the diagonal Hessian approximation are calculated for five discrete low frequencies from 8 Hz to 12 Hz.

Figure 7 shows a seismogram section of the vertical component for one exemplary shot lowpass-filtered with 12 Hz. The data is plotted for the real model and the starting model. The wavefield is dominated by surface waves. For these low frequencies, the starting model can already explain the data quite well and

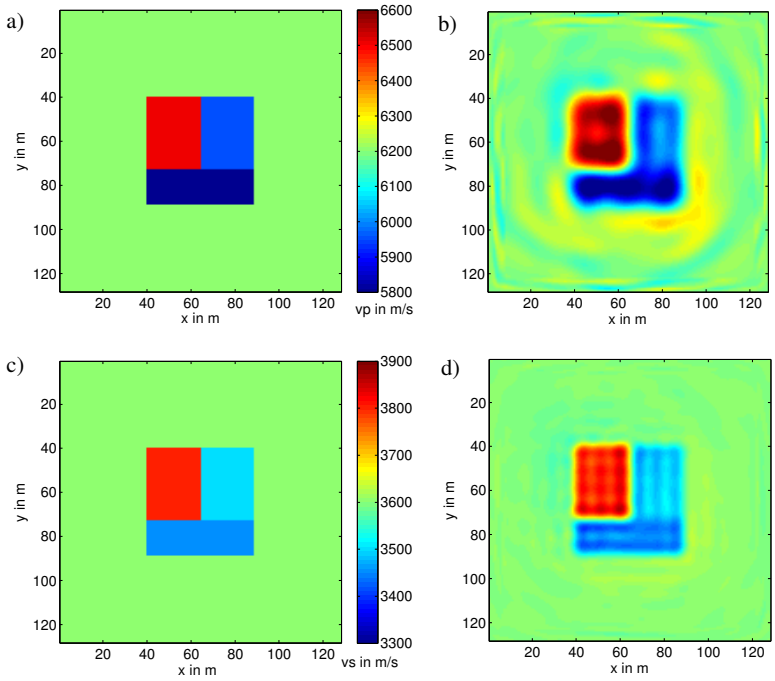

Figure 4: Results box model for horizontal slice at $z = 64\,\text{m}$ with a) real model v_p, b) inverted model v_p, c) real model v_s and d) inverted model v_s.

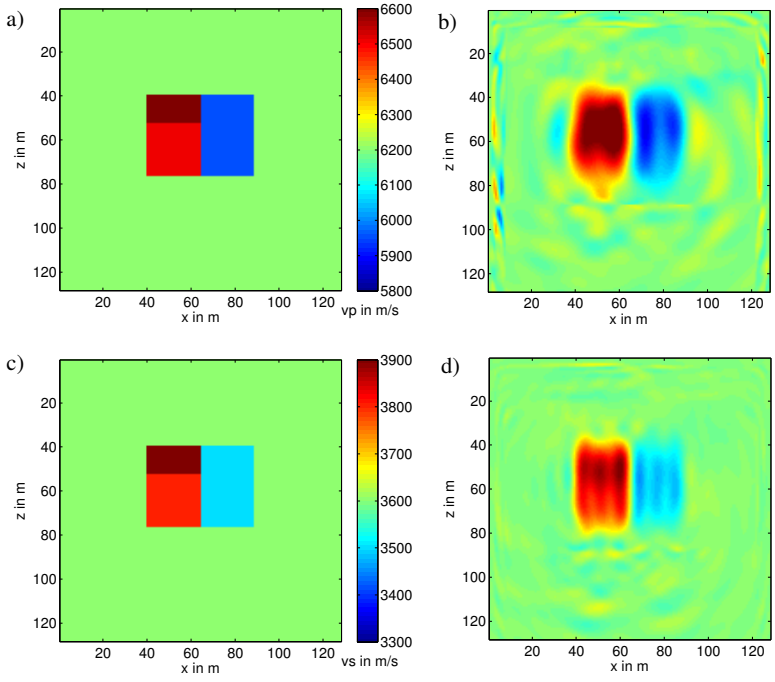

Figure 5: Results box model for vertical slice at $y = 60\,\text{m}$ with a) real model v_p, b) inverted model v_p, c) real model v_s and d) inverted model v_s.

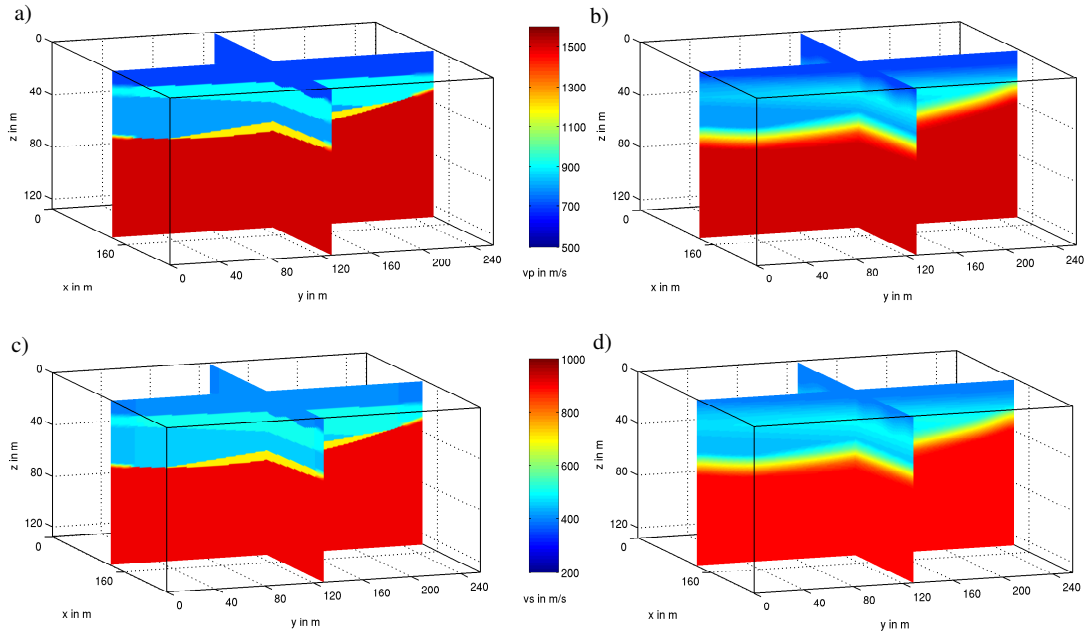

Figure 6: a) Real surface model v_p, b) starting model v_p, c) real surface model v_s and d) starting model v_s.

it is obvious, that no cycle skipping will occur in the inversion. Thus, this frequency range is adequate to start the inversion.

Hessian preconditioning We calculated the gradients and the diagonal Hessian approximation for this frequency range. Hereby, 81 additional forward modelings were required for the calculation of $\mathbf{H}_D$. The results are plotted in Figure 8. The gradients for v_p and v_s are shown in Figure 8 a) and b). The gradients are normalised to their maximum value, and the colorbar is strongly clipped. The highest gradient values are found in shallow areas, while they decrease significantly with depth. A model update without good preconditioning of the gradients would therefore focus on the uppermost part of the model. Figure 8 c) and d) shows the logarithmic diagonal Hessian approximation $\mathbf{H}_D$ for v_p and v_s, respectively. The colorbar is clipped and minimum logarithmic values amount to -19, so that $\mathbf{H}_D$ covers several orders of magnitude. Here, the fast decrease of amplitudes with depth is also visible.
We determined a water level and applied the inverse diagonal Hessian approximation to the gradients. The normalised preconditioned gradients can be seen in Figure 8 e) and f) for v_p and v_s. The highest amplitudes now occur in the middle and deeper structures of the basin. High amplitudes near the surface vanished and the amplitude variation due to geometric wave propagation effects could be corrected. Thus, the inversion will now be able to update the deeper layers of the model.

CONCLUSIONS AND OUTLOOK

A thorough preconditioning is required for gradient based full waveform inversion. We apply a preconditioning operator to the gradients in 3D FWI, which is based on the inverse diagonal Hessian approximation. The calculation of this preconditioning matrix is performed only once for each frequency stage and requires one additional forward simulation for each receiver. We showed its application for a box model in transmission geometry. High amplitudes at sources and receivers could be well corrected. However, the impact of the Hessian preconditioning in the remaining area is low. A successfull inversion was performed. For comparison, the effects of the diagonal Hessian approximation on the gradients of

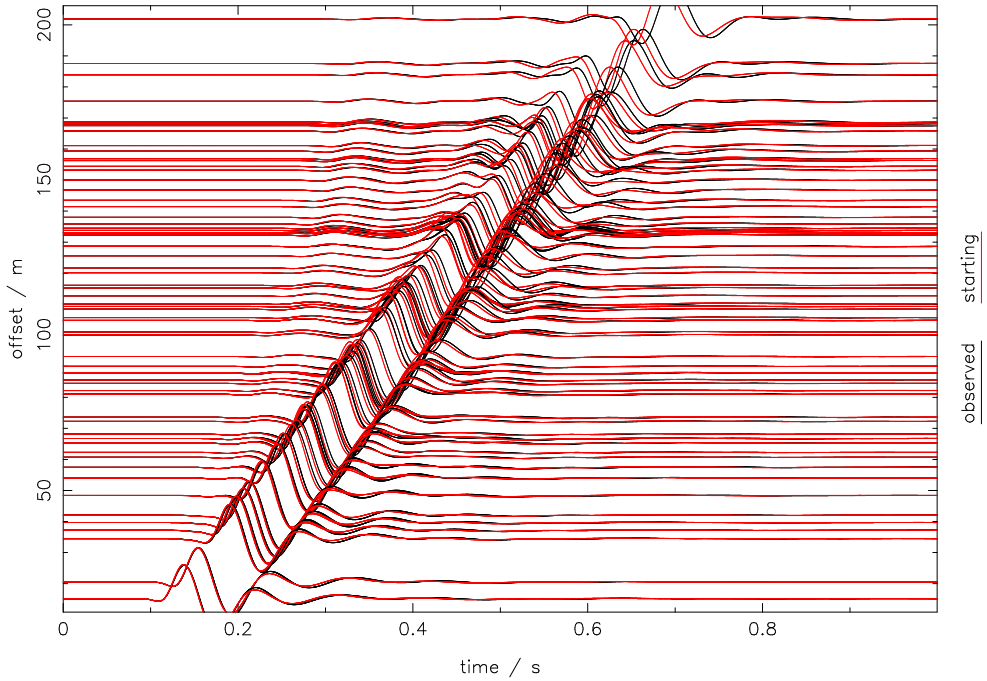

Figure 7: Seismograms of vertical component for observed and starting data for one representative shot lowpass-filtered with 12 Hz.

a surface model were shown. The seismograms of this model are dominated by a surface waves. This leads to gradients and diagonal Hessian matrices which are strongly concentrated near the surface. After preconditioning, however, the gradients concentrate on deeper parts of the model.
In future tests, we will test the performance of Hessian preconditioning in the inversion of the surface model introduced in this work. Additionally we will implement and test the performance of the L-BFGS method which, combined with Hessian preconditioning, should lead to an even better performance.

ACKNOWLEDGMENTS

The work was performed within the project TOAST which is part of the GEOTECHNOLOGIEN program, funded by the German Ministry of Education and Research (BMBF) and the German Research Foundation (DFG), grant 03G0752A
This work was kindly supported by the sponsors of the *Wave Inversion Technology (WIT) Consortium*.
The calculations were performed on the JUROPA cluster at Jülich supercomputing center and on the Cray XE6 at HLRS, Stuttgart.

REFERENCES

Bohlen, T. (2002). Parallel 3-D viscoelastic finite difference seismic modeling. *Computers and Geoscience*, 28:887–889.

Brossier, R., Operto, S., and Virieux, J. (2009). Seismic imaging of complex onshore structures by 2d elastic frequency-domain full-waveform inversion. *Geophysics*, 74(6):WCC105–WCC118.

Butzer, S., Kurzmann, A., and Bohlen, T. (2013). 3d elastic full-waveform inversion of small-scale heterogeneities in transmission geometry. *Geophysical Prospecting*, 61(6):1238–1251.

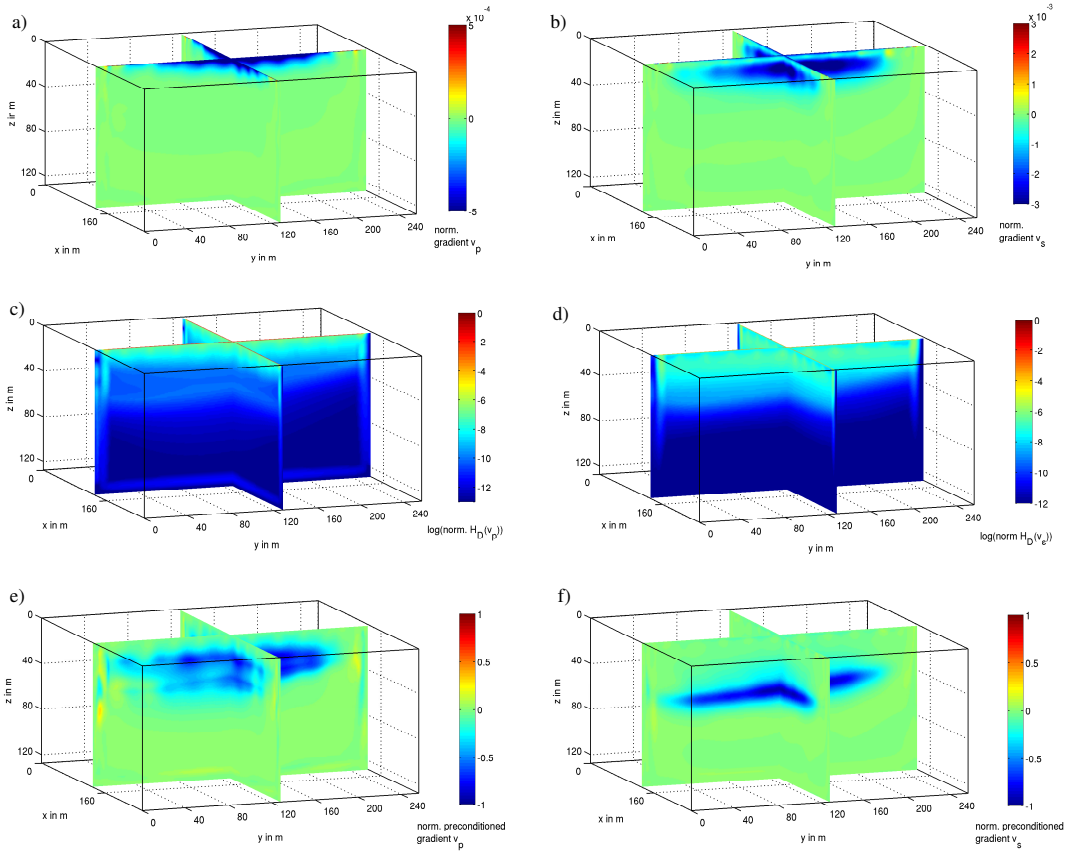

Figure 8: Effects of Hessian preconditioning on the gradients of v_p and v_s in the surface model, a) normalised gradient v_p, b) normalised gradient v_s, c) logarithm of normalised Hessian approximation v_p, d) logarithm of normalised Hessian approximation v_s, e) normalised preconditioned gradient v_p and f) normalised preconditioned gradient v_s.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16(5):1190–1208.

Métevier, L., Brossier, R., Virieux, J., and Operto, S. (2012). The truncated newton method for full waveform inversion. *Journal of Physics: Conference Series*, 386.

Mora, M. (1987). Nonlinear two-dimensional elastic inversion of multioffset data. *Geophysics*, 52(9):1211–1228.

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.

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

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

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

Sirgue, L., Etgen, J., and Albertin, U. (2008). 3d frequency domain waveform inversion using time domain finite difference methods. *70th EAGE Conference and Technical Exhibition*.

Tarantola, A. (1984). Linearized inversion of seismic reflection data. *Geophysical Prospecting*, 32:998–1015.