On cycle-skipping and misfit functions modification for full-wave inversion: comparison of five recent approaches

Arnaud Pladys, Romain Brossier, Yubing Li, Ludovic Métivier

On cycle-skipping and misfit functions modification for full-wave inversion: comparison of five recent approaches

Arnaud Pladys*, Romain Brossier*, Yubing Li‡ and Ludovic Métivier†*

ABSTRACT

Full waveform inversion, a high-resolution seismic imaging method, is known to require sufficiently accurate initial models to converge toward meaningful estimations of the subsurface mechanical properties. This limitation is due to the non-convexity of the least-squares distance with respect to kinematic mismatch. We propose a comparison of five misfit functions promoted recently to mitigate this issue: adaptive waveform inversion, instantaneous envelope, normalized integration, and two methods based on optimal transport. We explain which principles these methods are based on and illustrate how they are designed to better handle kinematic mismatch than a least-squares misfit function. By doing so, we can exhibit specific limitations of these methods in canonical cases. We further assess the interest of these five approaches for application to field data based on a synthetic Marmousi case study. We illustrate how adaptive waveform inversion and the two methods based on optimal transport possess interesting properties, making them appealing strategies applicable to field data. Another outcome is the definition of generic tools to compare misfit functions for full-waveform inversion.

INTRODUCTION

Full waveform inversion (FWI) is a high-resolution seismic imaging method dedicated to reconstructing the mechanical properties of the subsurface (Devaney, 1984; Pratt and Shipp, 1999; Plessix and Perkins, 2010; Raknes et al., 2015; Górszczyk et al., 2017). It is formulated as an iterative process based on minimizing a function measuring the misfit between observed and calculated data over a space of model parameters describing the subsurface. The resolution improvement FWI can procure, compared with standard tomography methods, is used to significantly improve depth-migration images or even produce directly interpretable quantitative estimates of the subsurface mechanical properties (Shen et al., 2018). FWI is applied at multiple scales, from global and regional scales in seismology to exploration scale for the oil & gas industry, and even, more recently, at near-surface scale for geotechnical applications. A thorough review of FWI and its applications can be found in Virieux et al. (2017).

FWI suffers from a significant shortcoming in its classical formulation: the non-convexity of the least-squares (L^2) misfit function on which it is conventionally based.

This non-convexity of the misfit function is an issue because the iterative process on which is based FWI is a local optimization algorithm. Standard size for realistic applications makes global optimization strategies beyond modern high-performance computing platforms current and predictable capabilities. Therefore, if the initial model used is too far away from the global minimum, FWI converges toward a potentially non geologically informative local minimum. This constraint leads to the need for an accurate enough initial model to ensure convergence toward the global minimum of the misfit function.

In a physical sense, the non-convexity of the L^2 misfit function is associated with a phenomenon known as cycle-skipping. It appears when the calculated data are shifted (in time) from more than half a period (corresponding to the signal dominant frequency) compared to the observed data. If the time-shift between observed and calculated data is larger than half a period, the minimization of the L^2 norm between the two signals will "skip" a phase and align the two signals on the closest phase (hence the name, cycle-skipping). This ambiguity translates into an erroneous reconstruction of the velocity model (Virieux and Operto, 2009).

This limitation of FWI has been documented since its origin (Gauthier et al., 1986). To address this limitation in practical cases, the workflow generally relies on data hierarchy (Bunks et al., 1995; Pratt, 1999; Shipp and Singh, 2002; Wang and

* Univ. Grenoble Alpes, ISTerre, F-38000 Grenoble, France
† Univ. Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France
‡ Formerly Univ. Grenoble Alpes, now Aramco Beijing Research Center, Aramco Asia, Beijing, China

Rao, 2009; Brossier et al., 2009). The historical approach consists in interpreting first the lowest frequency available (around 2 to 4 Hz for seismic exploration targets), then progressively introducing higher frequency data, following a multi-scale approach (Sirgue and Pratt, 2004). The lowest frequencies are, by definition, less subject to cycle-skipping. The second level of data hierarchy can then be defined by playing on temporal and/or offset selection of the data. The idea is to reduce the number of propagated wavelengths that are interpreted simultaneously, hence reducing the risk of cycle-skipping. In practice, this second level corresponds to first reconstructing the near-surface and progressively introducing deeper updates referred to as layer stripping approach.

Successful practical applications at the exploration scale often rely on the conjunction of these approaches as well as the design of an accurate initial starting model, obtained, for instance, through reflection tomography or stereotomography (Lambaré, 2008). Nonetheless, the conditions detailed previously to obtain a satisfactory FWI result are not always gathered. For instance, low-frequency data around 2 to 4 Hz are not always available or of sufficient quality. Moreover, obtaining low-frequency can increase the cost of acquisition, or can sometimes not be physically possible, or can even compromise the quality of the high frequency needed to obtain a very high resolution. Accurate initial model building can also be a time-consuming and challenging task requiring strong human expertise as it generally relies on tomography methods based on travel-time or reflected event picking. It also relies on prior information coming from geology or well logs; all of these require human expertise. This makes FWI less robust and reduces its range in terms of applications.

Mitigating the sensitivity to initial model quality has been the motivation for a large number of studies in the past decades. Two main lines of investigations can be identified, both leading to the reformulation of the conventional least-squares FWI problem.

Considering the first line, we regroup methods that can be cast under the frame of "extension strategies". It is not our purpose to give an extensive overview of these methods here, but we try to sketch their main ingredients. The philosophy of extension strategies consists in introducing supplementary degrees of freedom to the FWI problem, which can match the data in the early iterations of the FWI process to avoid cycle-skipping. Relaxing iteratively the use of these artificial degrees of freedom should lead to a correct subsurface model estimation.

Historically, these methods derive from migration velocity analysis (MVA) (Symes, 2008). MVA relies on the scale separation assumption. The subsurface parameters to recover are decomposed as a smooth macro-velocity model and a high wavenumber content reflectivity model. Artificial degrees of freedom are introduced at the reflectivity level by introducing an extra dimension on offset, subsurface offset, or time-lag. The MVA problem is formulated as the iterative update of the macro-velocity model to focus the energy of the "extended" reflectivity model at zero in the artificial dimension. These methods have benefited from in-depth mathematical research work, leading to a clear understanding of their foundations, thanks to the theory of pseudo-differential operators. However, their application to field data is still limited, mainly because of two issues. First, the repeated construction of high-dimensional reflectivity cubes is computationally demanding. Second, the macro-velocity model construction through MVA is complicated as soon as complex data with multi-pathing and multiple reflections are considered.

More recently, another class of extension strategies has emerged. As opposed to model space extension, the artificial degrees of freedom are introduced at the source level, following a source extension strategy (Huang et al., 2018; van Leeuwen and Herrmann, 2013). These methods have shown interesting promises in 2D synthetic case studies. However, their application to 3D field data seems still limited, mainly because of the difficulty of applying these methods in the time-domain. Current solutions either rely on relatively crude approximations (Wang et al., 2016) or on a sophisticated iterative solution, which increases the computational cost of the approach significantly (Aghamiry et al., 2020).

The second investigation line relies on reformulating the FWI problem using an alternative measure of the distance between observed and calculated data, namely a different misfit function. A large variety of approaches have been proposed on this framework. The first proposed along this line is to use cross-correlation measurements (Luo and Schuster, 1991), a strategy later revisited by van Leeuwen and Mulder (2010). The idea behind this is that cross-correlation should give access to the time-shifts between synthetic and observed traces. A misfit function based on the minimization of these time-shifts, resembling a tomography misfit function, should thus be less prone to cycle-skipping. The original approach of Luo and Schuster (1991) was labeled as "wave equation tomography" strategy.

However, when seismic traces contain multiple seismic events, the cross-correlation measurement might fail to give a correct estimation of a potential time-shift. This is why deconvolution based approaches have been later promoted, first by Luo and Sava (2011), then improved by Warner and Guasch (2016). The latter approach has been labeled as "adaptive waveform inversion" (AWI) and is based on a normalized deconvolution of the synthetic and observed seismic traces. It has shown very interesting properties both on synthetic and field data. The deconvolution of the traces yields a Wiener filter, which is then normalized and serves as an input for the misfit function. The misfit function penalizes the energy of the filter away from a bandpass Dirac filter, which would have been obtained in the correct subsurface model. Note that AWI shares some similarities with the extended source approach and can indeed be recast in the frame of these methods (Huang et al., 2018). This indicates that the separation between extended methods and misfit function reformulation methods is not as watertight as one could think. Nevertheless, it is useful to draw a landscape of the investigations around the cycle-skipping issue in FWI.

Another family of misfit function modifications relies on transforming the signal itself prior to comparison through a least-squares distance. Extracting the instantaneous phase and envelope (Fichtner et al., 2008; Bozdağ et al., 2011) has been successfully used in seismology. The goal of the instantaneous phase is to avoid amplitude prediction issues, as earthquake source and receiver calibration are significant challenges in seismology. The use of the envelope to mitigate the cycle-skipping issue has also been developed in the framework of seismic exploration (Wu et al., 2014). An interesting alternative consists of using a normalized integration of the signal, namely the cumulative distribution of the traces. This approach has been promoted by Donno et al. (2013).

Finally, optimal transport distances have also been promoted to derive alternative misfit functions for FWI. The motivation is to benefit from the convexity of the optimal transport distance with respect to translation and dilation, which provides a misfit function convex with respect to time-shifts, this being a good proxy for convexity with respect to seismic velocities (Engquist and Froese, 2014; Métivier et al., 2018). The main difficulty in applying optimal transport in the framework of FWI is that the optimal transport theory is developed to compare probability distributions, therefore positive functions with the same total integral. Seismic data do not fulfill this assumption.

To overcome this difficulty, different options have been promoted. For instance, one can rely on a prior transformation of the signal, such as extraction of positive and negative parts, squaring the data, affine scaling, exponential transform, softmax transform (Engquist and Froese, 2014; Qiu et al., 2017; Yang et al., 2018b; Yang and Engquist, 2018). This has been shown effective in some synthetic cases. However, relevant seismic information might be lost in the process of these transformations.

One solution is to rely on a specific optimal transport distance, which can be extended to comparing non-positive data. This is the Kantorovich-Rubinstein optimal transport (KROT) approach, which has been promoted in Métivier et al. (2016c,a,b), and which has been successfully applied to 3D synthetic elastic data (He et al., 2019b) as well as to field data (Poncet et al., 2018; Messud and Sedova, 2019; Sedova et al., 2019). One interest of this approach is its ability to account for lateral coherency in 2D or 3D shot gathers. One shortcoming is that, even if the valley of attraction is wider, compared with the L^2 approach, the convexity property of the optimal transport distance with respect to time-shifts is lost.

Another option has been promoted more recently. Considering each discrete seismic traces as point clouds and computing the optimal transport distance between synthetic and observed points clouds provide a new distance measurement. This specific optimal transport problem can be cast as a linear assignment problem, for which efficient solvers exist, for point clouds containing a few hundred to thousands of points, a situation we encounter for realistic scale exploration case studies (Métivier et al., 2018, 2019). The benefit of this graph-space optimal transport (GSOT) strategy is its ability to recover the convexity with respect to time-shifts. Compared with the KROT approach, GSOT is a trace-by-trace strategy that does not make it possible to account for lateral coherency. GSOT has been successfully applied to 3D synthetic and field data (He et al., 2019a; Pladys et al., 2019; Li et al., 2019; Górszczyk et al., 2019).

As can be seen, numerous investigations motivated by the inherent ill-posedness of the FWI problem have been lead in parallel. To our knowledge, no cross-comparison has been proposed so far, which is undoubtedly a lack. The first motivation of this study is to start developing tools that could be used to benchmark different FWI strategies. However, beyond a simple comparison of FWI strategies, we would like to highlight specific characteristics that an ideal misfit function should satisfy to render the FWI problem less ill-posed. Cycle-skipping is certainly an issue, but we also show that other criteria than robustness with respect to cycle-skipping should be considered, such as:

- sensitivity to the signal polarity;
- applicability in the framework of complex/multi-arrival data;
- number of tuning parameters and sensitivity to these parameters;
- sensitivity to wrong amplitude prediction and inaccurate wavelet estimation.

To illustrate these properties, we select a series of synthetic case studies of increasing complexity, from time-shifted Ricker traces to a realistic Marmousi II case study (not in inverse crime settings). We restrict our attention to five misfit functions, which have been promoted recently and have shown promising results: adaptive waveform inversion (AWI), instantaneous envelope (IE), normalized integration method (NIM), KROT, and GSOT. We consider extended space strategies out of the scope of this study to keep it reasonably simple, and also because, as stated before, we consider that alternative misfit strategies have shown more promising results than extended space strategies so far in terms of practical applications. The tests that we develop here could, however, be used to benchmark extended space strategies also.

GENERAL FWI FRAMEWORK AND MISFIT FUNCTION FORMULATION

The comparison between misfit functions is made simple by the FWI formalism (reviewed in the following section), more precisely by the adjoint state strategy used to compute the gradient at each iteration of the minimization loop. However, let us recall the main result: a modification of the misfit function results only in modifying the adjoint source. Therefore, implementing different misfit functions in the same FWI code can be done directly by isolating misfit function evaluation and adjoint source computation in different subroutines.

General framework

The FWI problem can be written as

$$\min_m f[m] = F(d_{cal}[m], d_{obs}), \tag{1}$$

where the subsurface parameters are denoted by m, d_{obs} is the observed data, $d_{cal}[m]$ is the synthetic data, and F is a generic function measuring the misfit between d_{obs} and d_{cal}. Under general notation, $d_{cal}[m]$ is obtained through the extraction of the values of wavefield at the receivers location such that

$$d_{cal}[m] = Ru[m]\,, \tag{2}$$

where R is an extraction operator and $u[m]$ is the solution of the wave propagation problem

$$A[m]u = b\,, \tag{3}$$

with $A[m]$ a generic wave propagation operator (from acoustic to visco-elastic).

The solution of the minimization problem 1 is computed through local optimization following the iteration

$$m_{k+1} = m_k + \alpha_k \Delta m_k \tag{4}$$

starting from an initial guess m_0. In eq. 4, α_k is the steplength, which should satisfy the Wolfe criterion (Nocedal and Wright, 2006), and Δm_k is the descent direction, given by

$$\Delta m_k = -P[m_k]\nabla f[m_k]\,, \tag{5}$$

where $\nabla f(m_k)$ is the gradient of the misfit function $f[m]$ and $P[m_k]$ a preconditioner approximating the inverse Hessian operator

$$P[m_k] \simeq H[m_k]^{-1}, \;\; H[m_k] = \nabla^2 f[m_k]. \tag{6}$$

Following the adjoint state strategy (Plessix, 2006), the gradient is given by

$$\nabla f[m] = \left(\frac{\partial A}{\partial m}u, \lambda\right)\,, \tag{7}$$

where $(.,.)$ is the Euclidean scalar product in the wavefield space, and λ is the adjoint field, solution of the adjoint equation

$$A(m)^T\lambda = s\,, \tag{8}$$

where s is the generic adjoint source, given by

$$s = -R^T\left(\frac{\partial F}{\partial d_{cal}}\right)\,. \tag{9}$$

Note that in the case of the L^2 norm, we recover immediately that

$$s = -R^T\left(Ru[m] - d_{obs}\right)\,, \tag{10}$$

i.e. the adjoint source is equal to the residual (difference between observed and calculated data).

Next, we review the formulas for the five misfit functions selected in this study, as well as their corresponding adjoint sources. For convenience, we will introduce the distance measurement function associated with each strategy for a single source/receiver couple, except for the KROT strategy. The calculated and observed data will be denoted by $d_{cal}(t)$ and $d_{obs}(t)$ unless stated otherwise. Except for KROT, the final misfit function is built as a sum over each source/receiver couple of this distance measurement function, and by linearity, the resulting adjoint source is also obtained by summation.

AWI

We give here the AWI formalism. We have

$$F_{AWI}(d_{cal}, d_{obs}) = \frac{\displaystyle\int_0^T |\mathcal{P}(\tau)w(\tau)|^2\,\mathrm{d}\tau}{\displaystyle\int_0^T |w(\tau)|^2\,\mathrm{d}\tau}\,, \tag{11}$$

where $w(t)$ is the Wiener filter which either transforms the calculated $d_{cal}(t)$ into the observed data $d_{obs}(t)$ (forward AWI) or the opposite way around (reverse AWI). Both implementations are discussed in Warner and Guasch (2016). Also, the computation of $w(t)$ can be implemented either in the time-domain or the frequency-domain. In both cases, a water level ε is required to stabilize the deconvolution operation.

The role of the function $\mathcal{P}(\tau)$ is to penalize energy at non-zero time lag. There are several possibilities to define this penalty function. Here we focus only on a Gaussian formulation defined as

$$\mathcal{P}(\tau) = e^{-\tau^2/\sigma^2}\,, \tag{12}$$

where σ is a tuning parameter controlling the width of the Gaussian function away from 0 time-lag. This σ tuning parameter is defined in seconds and corresponds to the maximum expected time-shift between the observed and calculated data.

In the case of a frequency-domain reverse AWI implementation, the adjoint source for a single-trace reads

$$\frac{\partial F_{AWI}}{\partial d_{cal}} = \frac{\int\,(\mathcal{P}(\tau) - 2F(d_{cal}, d_{obs}))\,w(\tau)p(t+\tau)\mathrm{d}\tau}{\int\,w^2(\tau)\mathrm{d}\tau}\,, \tag{13}$$

where

$$p(t) \approx \int \frac{\hat{d}_{obs}(\omega)e^{i\omega t}}{\hat{d}^*_{obs}(\omega)\hat{d}_{obs}(\omega) + \varepsilon}\mathrm{d}\omega\,, \tag{14}$$

with ε defined as

$$\varepsilon = (\max_\omega |d_{obs}(\omega)|)\zeta\,. \tag{15}$$

In eq. 15, ζ is a user-defined damping ratio, ranging from 10^{-2} to 10^{-5} in our experiment. A large ζ will help when trying to tackle large time-shift, with a "smoothing/regularizing" effect. Large ζ is also required if there is noise on the data. A smaller ζ will help preserve small features present in the signal. In terms of computational cost, the overhead associated with the computation of the Wiener filter is negligible, and the AWI strategy can be easily implemented.

IE

The separation of the phase and envelope information of the signal relies on the use of the analytical function defined as follows. For a given time signal $d(t)$, the analytical signal $\widetilde{d}(t)$ is defined as

$$\widetilde{d}(t) = d(t) + i\mathcal{H}[d(t)]\,, \tag{16}$$

where $\mathcal{H}$ is the Hilbert function which can be defined in the time domain as

$$\mathcal{H}[d(t)] = \frac{1}{\pi}P\int_{-\infty}^{+\infty}\frac{d(\tau)}{t-\tau}\mathrm{d}\tau\,, \tag{17}$$

where P stands for the Cauchy principal value. Practically, we do not use the time formulation of the Hilbert function, but rather a frequency domain formulation that gives us the analytical signal in a three-step approach (Marple, 1999):

- Compute the Fourier transform of $d(t)$ using an FFT
- Change the negative frequency to zeros
- Compute the inverse Fourier transform

This directly gives us access to the analytical signal and, by extension, to the Hilbert transform by taking its imaginary part

$$\mathcal{H}[d(t)] = \mathcal{I}[\widetilde{d}(t)] \,. \tag{18}$$

The analytical signal allows to separate the signal as the combinaison of the instantaneous phase $\phi(t)$ and the instantaneous envelope $E(t)$:

$$\widetilde{d}(t) = E(t)e^{i\phi(t)} \,. \tag{19}$$

Thus, the intantaneous enveloppe $E(t)$ can be simply defined as:

$$E(t) = \sqrt{\mathcal{R}[\widetilde{d}(t)]^2 + \mathcal{I}[\widetilde{d}(t)]^2} \,. \tag{20}$$

We can define a new distance-measurement function using instantaneous envelope as

$$F_{IE}(d_{cal}, d_{obs}) = \frac{1}{2}\int_0^T |E_{cal}(t) - E_{obs}(t)|^2 \mathrm{d}t \,, \tag{21}$$

where E_{cal} and E_{obs} are instantaneous envelopes of the calculated and observed data respectively. Following Bozdağ et al. (2011), the adjoint source is defined as:

$$\begin{aligned}\frac{\partial F_{IE}}{\partial d_{cal}} =& \frac{(E_{cal}(t) - E_{obs})d_{cal}(t)}{E_{cal}(t) + \varepsilon} \\ &- \mathcal{H}\left(\frac{(E_{cal}(t) - E_{obs})\mathcal{H}(d_{cal}(t))}{E_{cal}(t) + \varepsilon}\right) \,,\end{aligned} \tag{22}$$

with ε a water level defined as

$$\varepsilon = (\max_t E_{obs}(t))\zeta. \tag{23}$$

Contrary to AWI, in the following experiments, ζ is fixed and taken at $\zeta = 10^{-5}$ for IE. We have verified that the results with IE are not sensitive to this choice.

The instantaneous envelope misfit formulation is straightforward to implement thanks to the algorithm from Marple (1999). No tuning parameter is required, and the computation cost overhead is negligible.

NIM

Donno et al. (2013) consider the least-squares difference between the cumulative distributions Q_{obs} and Q_{cal}. For a given time signal $d(t)$, its normalized cumulative distribution $Q(t)$ is defined by

$$Q(t) = \frac{\displaystyle\int_0^t d(\tau)^2 \mathrm{d}\tau}{\displaystyle\int_0^T d(\tau)^2 \mathrm{d}\tau} \,. \tag{24}$$

The NIM misfit function thus relies on the distance measurement

$$F_{NIM}(d_{cal}, d_{obs}) = \frac{1}{2}\int_0^T |Q_{cal}(\tau) - Q_{obs}(\tau)|^2 \mathrm{d}\tau \,, \tag{25}$$

where $Q_{cal}(t)$ and $Q_{obs}(t)$ are the cumulative distributions associated with $d_{cal}(t)$ and $d_{obs}(t)$ respectively.

The corresponding adjoint source is

$$\begin{aligned}\frac{\partial F_{NIM}}{\partial d_{cal}} =& \frac{2d_{cal}(t)}{\int_0^T Q_{cal}(t)}\left(\int_t^T (Q_{cal}(\tau) - Q_{obs}(\tau))\mathrm{d}\tau \right. \\ &\left. - \int_0^T Q_{cal}(\tau)(Q_{cal}(\tau) - Q_{obs}(\tau))\mathrm{d}\tau\right) .\end{aligned} \tag{26}$$

The NIM implementation is straightforward and does not require any tuning parameters.

KROT

In the frame of the KROT approach, we consider the data as a function of both time and receiver position, such that we denote the calculated and observed data as $d_{cal}(x_r, t)$ and $d_{obs}(x_r, t)$ respectively.

The KROT is based on a particular instance of optimal transport distance, namely the 1-Wasserstein distance. It can be applied to non-positive data, provided mass conservation is satisfied *i.e.*

$$\int_{x_r}\int_0^T d_{cal}(x_r, t) dx_r dt = \int_{x_r}\int_0^T d_{obs}(x_r, t)\mathrm{d}x_r\mathrm{d}t \,. \tag{27}$$

For a given shot in seismic data, this corresponds to the summation over each trace of the mean value in time of the trace. We consider this mean value is equal to 0 (this is the zero-frequency noise, which is usually removed from the data prior to inversion). Therefore the mass conservation assumption is satisfied for seismic data.

On this basis, the KROT distance can be written as

$$F_{KROT}(d_{cal}, d_{obs}) = \max_{\varphi\in \mathrm{Lip}_1}\int_{x_r}\int_0^T \varphi(x_r, t)\big(d_{cal}(x_r, t) - d_{obs}(x_r, t)\big)\mathrm{d}x_r\mathrm{d}t \,, \tag{28}$$

where Lip_1 is the set of 1-Lipschitz functions for the ℓ_1 distance

$$\mathrm{Lip}_1 = \{\varphi(x_r, t), \;\; |\varphi(x_r, t) - \varphi(x'_r, t')| < |x_r - x'_r| + |t - t'|\} \,. \tag{29}$$

The adjoint source is then given by

$$\frac{\partial F_{KROT}}{\partial d_{cal}} = \overline{\varphi}(x_r, t), \tag{30}$$

where

$$\overline{\varphi}(x_r, t) = \arg\max_{\varphi\in \mathrm{Lip}_1}\int_{x_r}\int_0^T \varphi(x_r, t)\left(d_{cal}(x_r, t) - d_{obs}(x_r, t)\right)\mathrm{d}x_r\mathrm{d}t \,. \tag{31}$$

Compared with previous misfit functions, the final misfit is obtained here by summation over shot gather, and not a summation over source/receiver couples (not a trace-by-trace approach).

From the above equations, we see that the computation of the KROT misfit function and its corresponding adjoint source requires solving a constrained maximization problem per shot gather. Details on how to solve this problem are given in Métivier et al. (2016c). The proximal splitting algorithm ADMM is used (Combettes and Pesquet, 2011) and the resulting algorithm has complexity in $O(N \log N)$, where $N = N_r \times N_t$ with N_r the number of receivers and N_t the number of time samples. Compared with the previous misfit functions, the computational cost overhead is non-negligible. Tuning parameters will be associated with a prior scaling of the data to make its maximum amplitude close to 1, and the number of iterations required to solve the constrained maximization problem.

GSOT

Let $(t_i, d(t_i),\ \ i = 1, \dots, N)$ be the discrete graph of the time function $d(t)$. This discrete graph is a point cloud containing N points. The GSOT distance measurement is formulated as

$$F_{GSOT}(d_{cal}, d_{obs}) = \min_{\sigma \in S(N)} \sum_{i=1}^{N} c_{i\sigma(i)}, \tag{32}$$

where c_{ij} is the L^2 distance between the points of the discrete graph of d_{cal} and d_{obs}, namely

$$c_{ij} = |t_i - t_j|^2 + \eta^2 |d_{cal}(t_i) - d_{obs}(t_j)|^2, \tag{33}$$

and $S(N)$ is the ensemble of permutations of $(1 \dots N)$. The function F_{GSOT} corresponds to the 2-Wasserstein distance between the discrete graph of the calculated trace $d_{cal}(t)$ and the observed trace $d_{obs}(t)$.

The scaling parameter η in eq. 33 controls the convexity of the misfit function f_{GSOT} with respect to time-shifts. In practice, we define it as

$$\eta = \frac{\tau}{A}, \tag{34}$$

where τ is a user-defined parameter corresponding to the maximum expected time-shift between observed and calculated data in the initial model, and A is the maximum amplitude discrepancy between observed and calculated data.

The adjoint source of the misfit function $f_{GSOT}[m]$ is computed from $\frac{\partial f_{GSOT}}{\partial_{cal}}$ using the adjoint-state strategy. It is proven in Métivier et al. (2019) the following equality: denoting σ^* the minimizer in eq. 32, we have

$$\frac{\partial F_{GSOT}}{\partial_{cal}} = 2\left(d_{cal} - d_{obs}^{\sigma^*}\right), \tag{35}$$

where

$$d_{obs}^{\sigma^*}(t_i) = d_{obs}(t_{\sigma^*(i)}) \,. \tag{36}$$

In this sense, the GSOT approach can be viewed as a generalization of the L^2 distance: the adjoint source is equal to the difference between calculated and observed data at time samples connected by the optimal assignment σ^*. As the KROT approach, the solution of the problem 32 provides the information to compute both the misfit function and the adjoint source.

The numerical algorithm used to solve the linear assignment problem 32 is the auction algorithm (Bertsekas and Castanon, 1989). For problems involving less than 1000 points, the auction algorithm is very efficient. In seismic exploration, Nyquist sampling yields traces containing a number of points within this order of magnitude. Consequently, Métivier et al. (2019) have designed an efficient numerical strategy, yielding lower computational overhead than the KROT approach. On 3D field data application, we observe 15 to 20% computation time increase for gradient computation on the lowest frequency bands compared with classical L^2 . This computational cost overhead decreases when the frequency band increases as the total complexity of the GSOT problem is $O(\omega^3)$, while the complexity of the wave propagation solver is in $O(\omega^4)$. For more details, the reader can refer to Métivier et al. (2019).

Compared with previous approaches, the computational cost overhead is comparable with AWI, IE, and NIM while being lower than KROT. In terms of implementation, as for KROT, the solution of the assignment problem requires specific solvers, which makes the GSOT implementation less trivial than for AWI, IE, or NIM. In terms of tuning parameters, the more important parameter is the parameter τ, which controls the convexity of GSOT misfit function with respect to time-shifts.

A SIMPLE CONVEXITY ANALYSIS BASED ON TIME-SHIFTED RICKER WAVELETS

We start by investigating the convexity of the proposed misfit functions with respect to time-shifts. We fix a reference signal composed of one Ricker wavelet in the center, seen as the observed data. The calculated data is the same Ricker wavelet, shifted in time with a time-shift going from -1.5 s to 1.5 s. We compute the distance between the reference signal and the calculated signal using the five selected misfit functions, depending on the input time-shift. Results are presented in Figure 1.

The results obtained here with alternative misfit functions might not reflect the performance of the algorithms with total accuracy, both in terms of computational efficiency and inversion results. Algorithms might not have been implemented in the most optimal way or in the way the original authors intended. Subtle choices of tuning parameters might improve the inversion results in some cases. However, the primary purpose of this comparison is to seek to understand how the data is interpreted within each of these strategies and how this affects the inversion results in each case. We intend to provide the reader with sufficient material to infer the main properties and philosophy behind the compared methods.

Let us first analyze the results obtained with L^2 waveform misfit, the reference for FWI. As expected, L^2 misfit displays a narrow basin of attraction, with local minima and a flat part for time-shift superior to 0.4 s. The local minimum appears when the time-shift is larger than 0.12 s, which corresponds to half the Ricker wavelet period. This validates that the L^2 misfit function presents low robustness for shifted-patterns, leading to cycle-skipping when signals are shifted by more than half a period. In such cases, L^2 misfit function does not guarantee convergence toward the global minimum.

We can now compare the selected alternative misfit functions to the L^2 misfit. From the obtained results, we can define two groups. The first one contains GSOT, AWI, and NIM, characterized by a large basin of attraction. The second group contains IE and KROT, characterized by a "slightly" larger basin of attraction than L^2 , but not as wide as the first group members.

Understanding why the first group members exhibit the convexity property is essential. Starting with GSOT, if the input parameters τ is correctly set to the maximum expected time-shift of 1.5 s, the convexity to shifted-patterns is expected as there is a direct link between the τ parameters and the width of the basin of attraction as shown in Métivier et al. (2019).

The same convexity property is observed with AWI. With σ set to 1.5 s, the results are satisfying with a large basin of attraction. Similarly as the τ parameter from GSOT, σ directly controls the convexity to shifted-patterns. Note that we use $\zeta = 10^{-5}$ in this analysis as we predict signal with machine precision.

Finally, to understand the robustness of the NIM approach, we display in Figure 2 the quantities Q_{obs} and Q_{cal} (for three time-shifts, -1.5 s, -0.1 s and in-phase). This makes visible the drastic modification of the signal shape induced by NIM. The NIM cost function boils down to be the area under the curve delimited by $Q_{obs} - Q_{cal}$. We see clearly that this area increases with time-shifts, illustrating the convexity to shifted patterns observed with NIM.

Moving to the second group, to understand why the IE misfit only slightly increases the width of the valley of attraction compared with L^2 , we display in Figure 3 the quantities E_{obs} and E_{cal}. Here we can observe the increase of temporal support of the signal induced by the envelope. This "broader" temporal support of the instantaneous envelope directly translates into the increase of the width of the valley of attraction as IE relies on a L^2 norm between E_{obs} and E_{cal}.

Finally, we present in Figure 4 the function $\overline{\varphi}(t)$ solution of the maximization problem defined in eq. 31, which defines the KROT distance, together with the residuals $d_{obs}(t) - d_{cal}(t)$. We can observe that when Ricker wavelets start to overlap at -0.3 s, we obtain a convexity that classical L^2 cannot achieve. This can be understood by looking at the function $\overline{\varphi}(t)\,[d_{obs}(t) - d_{cal}$ The area below the curve defined by this function corresponds to the KROT misfit function. This area remains constant as long as the two signals do not overlap and monotonically decrease as soon as the two signals overlap, reaching 0 at 0 time-shift.

On a second test, presented in Figure 5, we introduce a second Ricker wavelet that remains in phase. This test aims at validating the robustness to cycle-skipping when multiple arrivals are considered. From the results obtained, we observe that all misfit functions behave similarly as on the previous test except for AWI. In this case, the shape of the misfit function seems affected by oscillations near 0 time-shift, reducing the effective convexity to the one of classical L^2 formulation. This seems to be related to one of the potential issues of deconvolution based misfit function: the sensitivity to cross-talks between multiple events. To analyze this sensitivity of AWI to multi-arrivals, we display the Wiener filters together with the penalty function and the combination of both (Figure 6). In test B (where one wavelet is always in-phase), the Wiener filter presents a strong peak at 0 time-lag due to the in-phase arrivals.

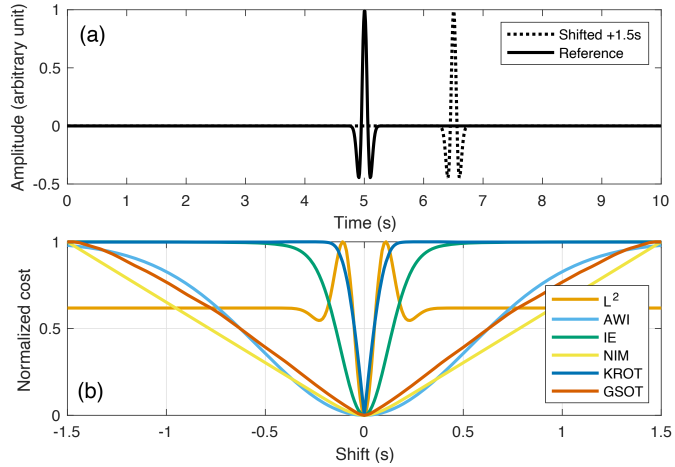

Figure 1: Comparison of several misfit functions in a simple 1D case for one shifted arrival. The arrival is set to be a Ricker wavelet with a central frequency of 4 Hz. (a) represents the signal used for the test (with only one arrival at the center). The fixed reference signal is displayed in continuous black. The shifted signal is displayed in dotted black (here for $+1.5$ s). (b) represents the normalized misfit function values with respect to the time-shift (from -1.5 s to 1.5 s).

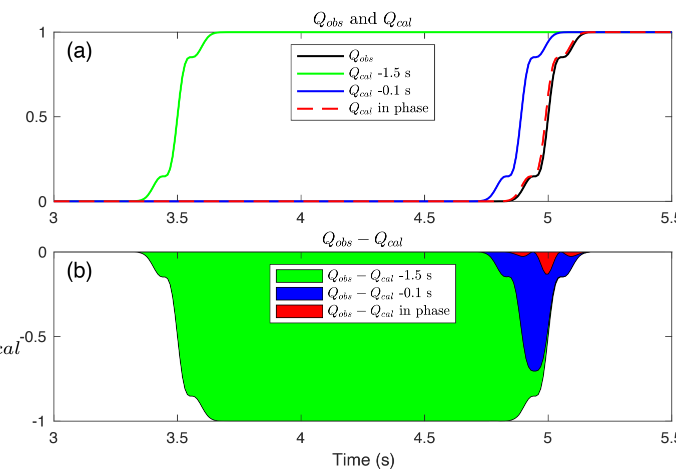

Figure 2: (a) quantities Q_{obs} and Q_{cal} for three time-shifts (-1.5 s in green, -0.1 s in blue and "in phase" in dashed red). (b) the area under the curve for $Q_{obs} - Q_{cal}$ for the three time-shifts.

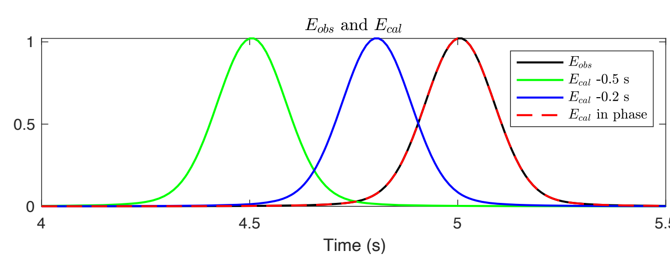

Figure 3: E_{obs} and E_{cal} for three time-shifts (-0.5 s, -0.2 s and in phase).

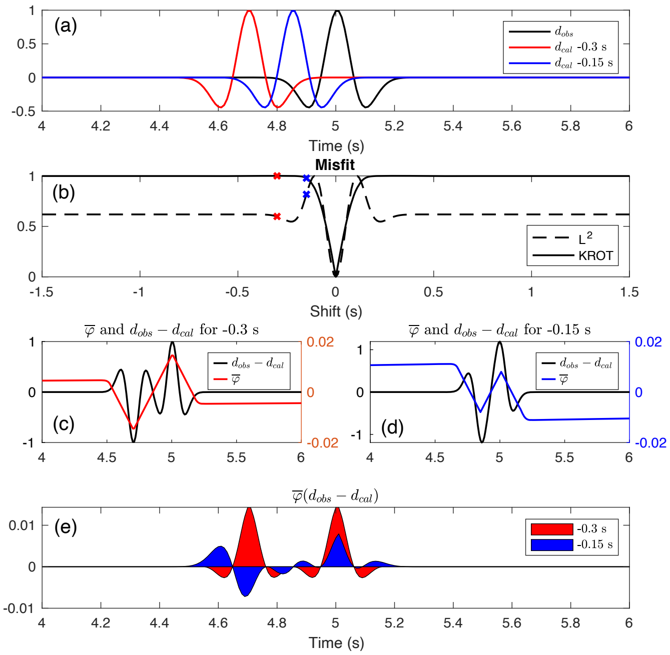

Figure 4: Detail for $\overline{\varphi}(t)$ from KROT. (a) the setup with d_{obs} (in black) and d_{cal} for two time-shifts (-0.3 s in red and -0.15 s in blue). (b) shape of L^2 and KROT misfit function with respect to time-shifts, red and blue cross represent the positions of the two time-shifts selected. (c) and (d) respectively display $\overline{\varphi}(t)$ and $d_{obs} - d_{cal}$ for the two time-shifts of -0.3 s and -0.15 s. (e) the area under the curve for $\overline{\varphi}(t)(d_{obs} - d_{cal})$ quantity for the two time-shifts. This last quantity is used to get the misfit function value after time integration.

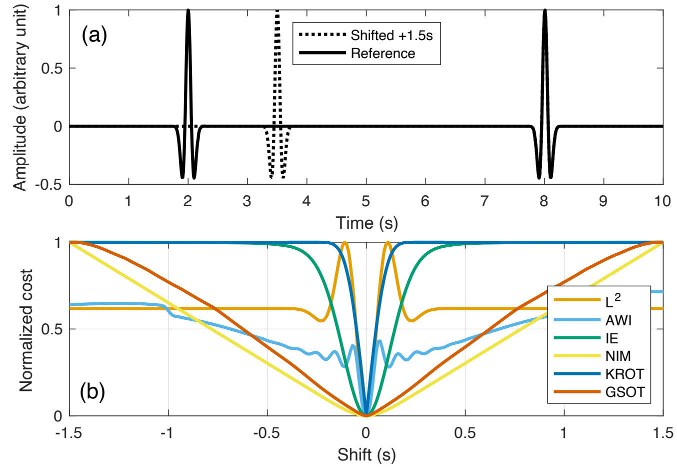

Figure 5: Same as Figure 1 but with two Ricker wavelets with one shifted (left) and one in phase (right).

Because of finite frequency effect, it is not a Dirac delta function but a bandpass Dirac delta function. The oscillations of the bandpass delta function combine in a destructive/constructive manner when the two time-shift peaks (one for each Ricker wavelet) get closer to each other. These interferences are at the origin of the local minima observed.

FWI TESTS ON TWO CANONICAL EXAMPLES

This section attempts to assess the pros and cons of the selected alternative misfit functions on two schematic FWI tests, focusing on a different aspect of the information contained in a dataset. The first test focuses only on transmission with a cross-hole acquisition. The second test focuses mainly on reflection information. These two tests can be seen as a way of assessing if the proposed misfit function can improve the FWI robustness (cycle-skipping in transmission in the first test) while preserving the ability to correctly interpret reflection information (reflector positioning and imaging in the second test)

Both tests are performed in 2D using our 2D/3D time-domain acoustic modeling and inversion code in inverse crime settings (observed and calculated data are computed on the same grid, without noise introduced in the data). Besides, we use a constant density model and invert only for the P-wave velocity model. In both cases, the l-BFGS algorithm is used to minimize the misfit function, with FWI stopping criterion being a line search failure. The source wavelet is a Ricker wavelet with a central frequency $f_{ref} = 3$ Hz. The gradient is smoothed using a Gaussian filter with horizontal and vertical correlation lengths equal to 0.3 times the local wavelength

$$\lambda_{loc}(x, z) = 0.3 \frac{v_P(x, z)}{f_{ref}} . \quad (37)$$

FWI Test 1: transmission configuration

Case study presentation

This first case study focuses on transmitted energy. The exact model is defined as a square of 1000 m sides with homogeneous V_P = 1300 m/s containing a spherical inclusion of

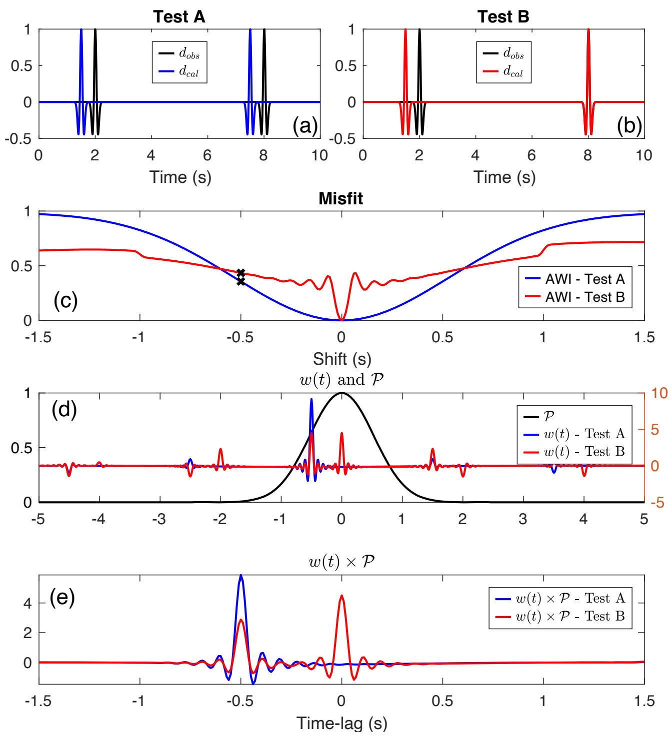

Figure 6: AWI analysis with two setups: test A and test B. (a) test A both wavelets shift, (b) test B only the left wavelet is shifted, the right one being always in phase (similarly to Figure 5). (c) shape of AWI misfit function with respect to time-shift in both cases. The Wiener filters presented under are shown for a time-shift of -0.5 s (black cross on the misfit). (d) Wiener filters ($w(t)$) and the Gaussian penalty function $\mathcal{P}(t)$. (e) the Wiener filters multiplied by the penalty function.

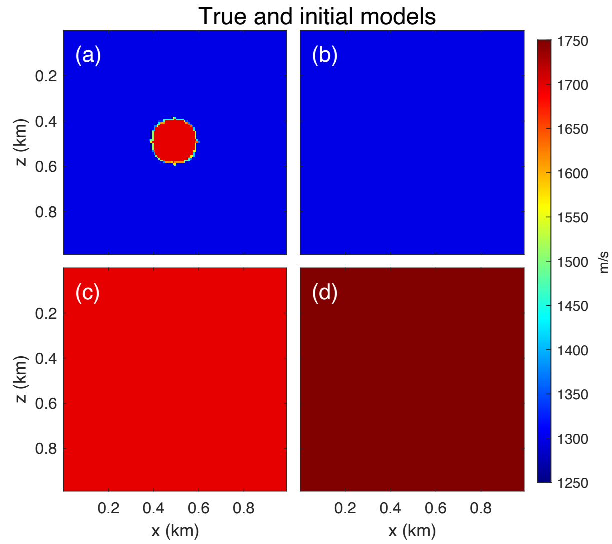

Figure 7: FWI Test 1: (a) true model, (b) initial model 1 with $V_P = 1300$ m/s , (c) initial model 2 with $V_P = 1700$ m/s and (d) initial model 3 with $V_P = 1900$ m/s .

100 m radius in the center with $V_P = 1700$ m/s (Figure 7). The acquisition mimics a crosshole setting, with 96 sources on the left side of the model and 256 receivers on the right side. The spacing is 10 m between sources and 3.8 m between receivers. The boundaries are all set to absorbing layers (Bérenger, 1994) to avoid reflections and only focus on transmitted events. The relatively strong contrast between the background and the anomaly generates an identifiable diffraction pattern in the data. In this experiment, no preconditioning is applied to the gradient. The lower and upper V_P bound constraints are respectively set to 1000 and 2500 m/s .

We introduce three starting homogeneous models (Figure 7). The first is at the true model background velocity (1300 m/s). The second is at $V_P = 1700$ m/s , setting a challenging FWI problem as the starting model is as fast as the inclusion. The third case is even more challenging, with a starting homogeneous V_P model at 1900 m/s .

FWI results are presented in Figure 8 with reconstructed V_P at the final iteration. Figure 9 presents traces for a single source-receiver couple representing the shortest path through the spherical inclusion (straight horizontal path at 500 m depth). Traces are extracted from data generated in the true model, initial model, and final reconstructed model for all misfit functions.

Results from initial model 1

We start the analysis with the "reference" initial model. As shown in Figure 9, this model does not generate cycle-skipping (arrivals in the true model are less than half a period away from the arrivals in the initial model). The objective is to retrieve the high-velocity spherical inclusion in the center of the model. As expected, the L^2 misfit function produces a correct result: the inclusion is retrieved correctly, and the final data are in phase with the true data. The vertical resolution is higher than the horizontal resolution as expected from the cross-hole configuration. This has a lateral smoothing effect on the reconstructed anomaly, which explains why its peak amplitude (around 1500 m/s) is lower than the amplitude of the true anomaly. The five selected misfit functions produce equivalently good results in this configuration. In all cases, the spherical anomaly is reconstructed with a similar resolution, and the data fit is equivalent. In terms of parameter settings, we choose here $\tau = 0.2$ s for GSOT and $\sigma = 0.2$ s for AWI, a choice motivated by the absence of cycle-skipping. For AWI, we use $\zeta = 10^{-2}$ in this transmission test (for all three models) to maximize the kinematic effects of AWI that work better when ζ is relatively high, which acts as a regularization effect.

Results from initial model 2

As can be observed in Figure 9, the second initial model generates clear cycle-skipping in the data. In this case, we expect the L^2 misfit function to fail in reconstructing the anomaly. Indeed, the L^2 fails to converge and reaches the boundary set for the inversion. The final synthetic trace does not match the observed trace. It is interesting to observe that four of the five selected misfit functions succeed in reconstructing the background and the anomaly and produce final synthetic traces in phase with the observed trace in this already quite challenging test. The only alternative misfit function that fails is KROT, which could be expected from the previous section (weak increase of robustness to cycle-skipping). AWI, IE, NIM and GSOT show that the increase in convexity procured by these formulations is enough here to make convergence achievable. The data-fit obtained with these methods is good in this case. In terms of tuning parameters, τ and σ are increased to 0.35 s for GSOT and AWI, according to the time-shift between the reference and the initial traces in the initial model.

Results from initial model 3

Finally, the initial model 3 generates an even more substantial cycle-skipping effect than model 2 (Figure 9). L^2 and KROT still fail to converge to the correct model, as it was already the case starting from model 2.

IE starts to exhibit diagonal cycle-skipping artifacts associated with the longest source/receiver paths in this more challenging setting. This is expected from the time-shift convexity analysis performed before: IE robustness to cycle-skipping is limited. AWI also starts to exhibit artifacts close from the acquisition, while central anomaly is correctly reconstructed (with $\sigma = 0.6$ s). NIM and GSOT (with τ increased to 0.6 s) achieve a relatively satisfactory reconstruction of the background and anomaly, similar to the results obtained from the previous background models.

FWI Test 2: reflection configuration

Case study presentation

This second case study focuses on reflected energy. We consider two different true models, composed of a homogeneous background at 1500 m/s and a velocity layer 100 m thick at 300 m depth (Figure 10). In the first case, the velocity of the layer is set to 1600 m/s , while in the second case, the velocity of the layer is set to 1400 m/s . The starting model is homogeneous at the correct background velocity of 1500 m.s^{-1} (Figure 10). The surface acquisition comprises 96 sources and 512 receivers located close to the surface at 42 m depth. The spacing is 20 m between sources and 3.8 m between receivers. We implement PML absorbing conditions on the bottom and lateral sides of the medium to mimic a medium of infinite extension in these directions and a free surface condition on the top of the model. A simple linear in-depth preconditioner is also applied to compensate for geometrical spreading effects and accelerate the convergence. The lower and upper V_P boundaries for the inversion are respectively set to 1200 m/s and 1800 m/s .

This test analyzes how the reflected data is interpreted by FWI depending on the choice of misfit function. The difference between the two exact models is only the sign of the velocity change at the layer level: in one case, velocity increases; in the second case, it decreases. This induces a change of polarity of the reflected wave, as clearly visible in Figure 11. We want to identify how the different misfit functions are sensitive to this change of polarity.

Results analysis

The reconstructed models are presented in Figure 12. Traces from the observed and synthetic data in the initial and final models for zero offset couple (source and receiver at the same position) located in the middle of the acquisition are presented in Figure 13. For visualization purposes, we cropped over the reflection after the first arrival.

The L^2 results are coherent with the expectation, with a correct reconstruction of the layer in both cases. The L^2 norm is sensitive to amplitude variation and polarity and is expected to interpret reflected events correctly. As the background velocity is known, there is no cycle-skipping in the initial model for the two target models. The data fit in both cases is perfect.

Together with L^2 misfit function, results obtained with IE, KROT, and GSOT are equivalently correct. This is expected from KROT and GSOT, which should behave similarly as L^2 when cycle-skipping does not occur. GSOT relies on $\tau = 0.2$ s in this experiment. This is somehow more surprising from IE, as one could think that the polarity of reflected events might be lost in the envelope extraction process. However, this is not true in this case but might be due to the inverse crime settings we are using. There is indeed a subtle change in the envelope of the observed data between model 1 (positive layer) and model 2 (negative layer) (Figure 14), which is enough to guide the inversion in the right direction. However, this is probably possible only because the first arrival is correctly predicted. Small inaccuracies in predicting the first arrival might be enough to impede a correct reconstruction using IE.

The results obtained with NIM are less satisfactory. In particular, the reconstruction of the negative layer is altered by strong positive artifacts beneath the layer. In the opposite case, negative artifacts also pollute the reconstruction of the positive layer, although the strength of these artifacts seems weaker. Analyzing the data fit shows that NIM has difficulties repro-

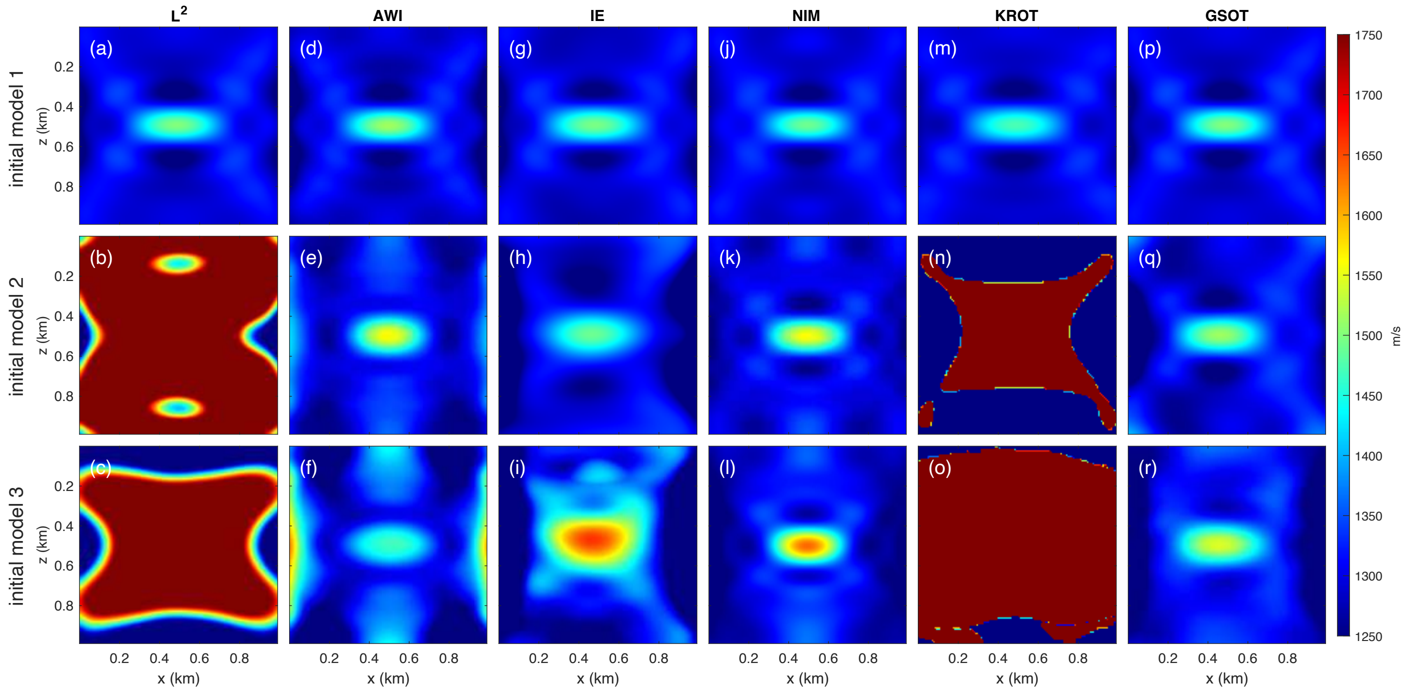

Figure 8: FWI Test 1: V_P results from FWI. First line corresponds to initial model 1, second line to initial model 2 and third line to initial model 3. Each column corresponds to reconstructed V_P model from FWI using respectively L^2 (a-c), AWI (d-f), IE (g-i), NIM (j-l), KROT (m-o) and GSOT (p-r) misfit functions.

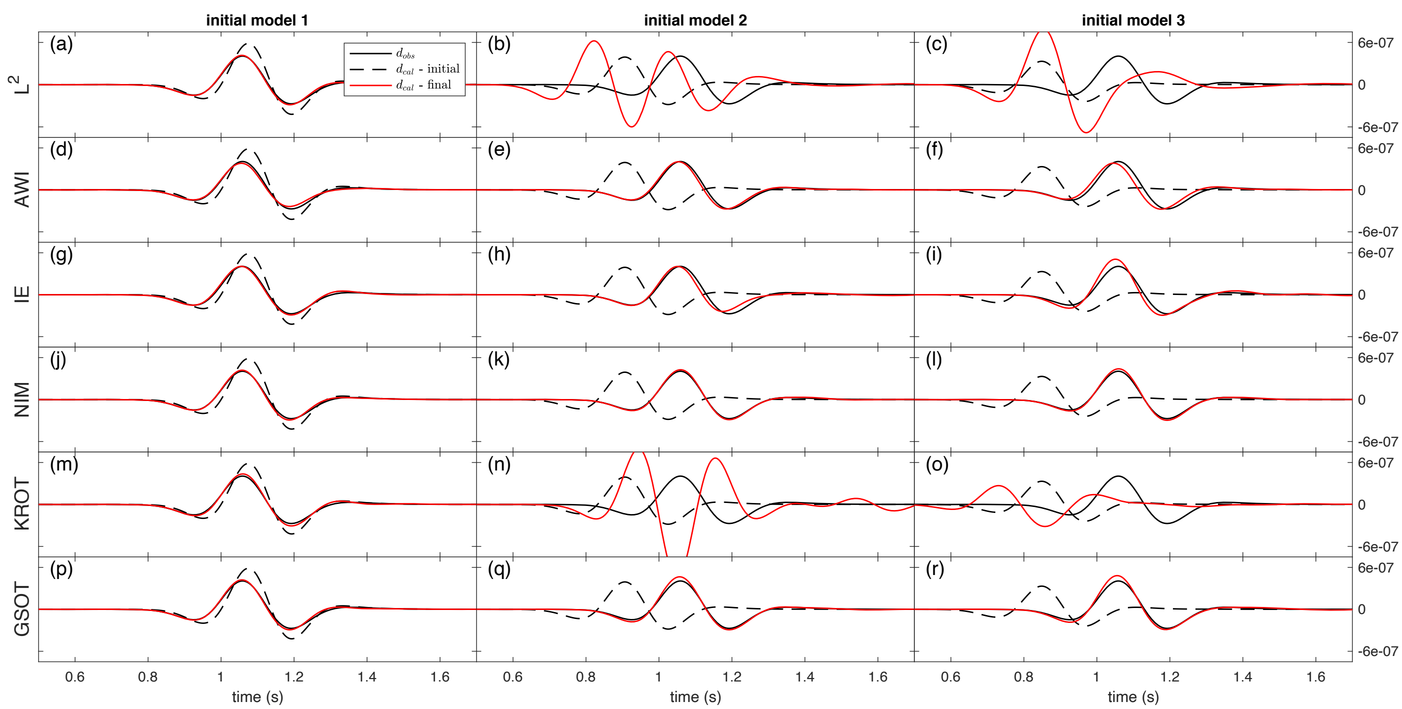

Figure 9: FWI Test 1: Extracted traces along the shortest path (horizontal straight line at 500 m depth passing through the spherical inclusion). Observed data are in solid black, synthetic data in the initial model in dashed black and final reconstructed synthetic data in solid red. First column corresponds to initial model 1, second column to initial model 2 and third column to initial model 3. Each line corresponds to reconstructed V_P model from FWI using respectively L^2 (a-c), AWI (d-f), IE (g-i), NIM (j-l), KROT (m-o) and GSOT (p-r) misfit functions.

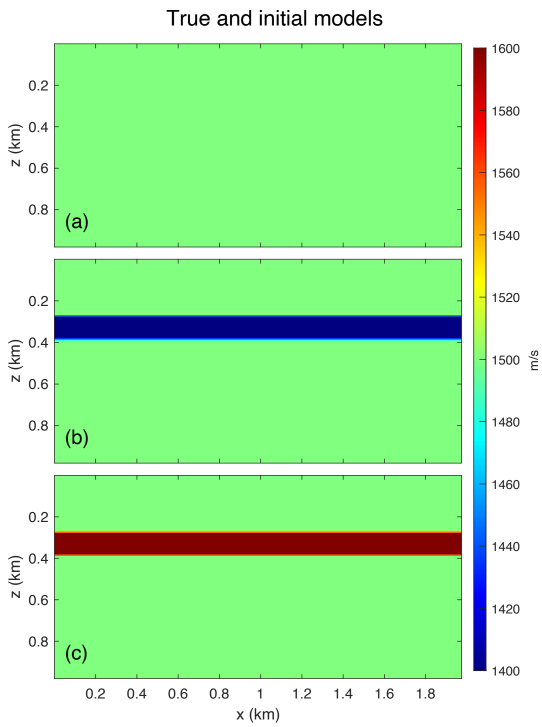

Figure 10: FWI Test 2: (a) Homogeneous initial model ($V_P =$ 1500 m/s). (b) True model with a negative V_P anomaly, (c) true model with a positive V_P anomaly.

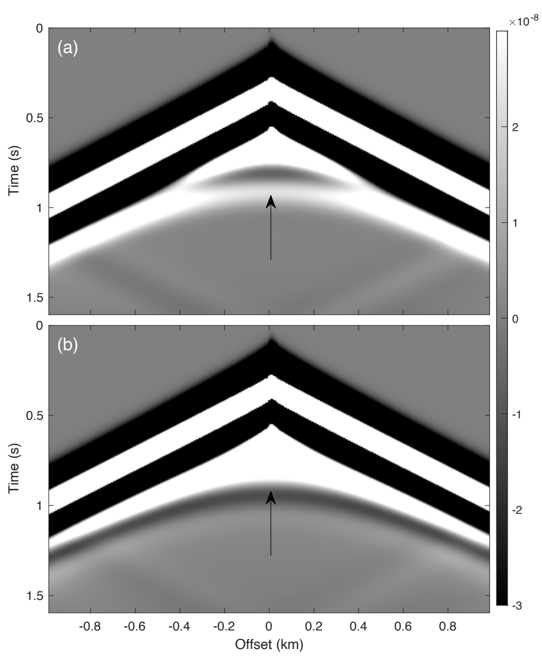

Figure 11: FWI Test 2: 2D common shot gathers for (a) observed data for negative V_P anomaly and (b) positive V_P anomaly. Strong clipping is applied to enhanced the reflected waves which are approximately 100 times smaller in amplitude than the transmitted waves. Polarity reversal of the reflected waves is pointed with black arrows.

ducing the reflection pattern in both cases (Figure 13). Spurious oscillations appear, which can be associated with the positive artifacts observed on the model reconstruction. This lack of sensitivity to the polarity is somehow expected. From NIM formulation, Q_{obs} should be more or less the same independently of model 1 or model 2 being used. This is illustrated in Figure 14, where $Q_{obs}(t)$ is presented for both models (positive and negative layers). We can observe that the difference between the two true models leads to a very marginal modification of Q_{obs} compared to Q_{cal}. This is likely the explanation of the difficulties faced by NIM in interpreting the reflected waves correctly.

Finally, the results obtained with AWI are incorrect for both the negative and positive V_P anomaly. From observing the data, we can see that the direct waves exhibit a clear dominance in amplitude over the reflected events. Therefore, we expect the Wiener filter to be dominated by the direct waves and only show a small imprint of the reflected waves that are of small amplitude ($\approx 1\%$ of peak amplitude). This is illustrated by the Wiener filter shown in Figure 15 for both positive and negative layer models. They indeed present a main event around 0 lag, corresponding to the in-phase direct wave. Around 0.3 s, the imprint of reflected events is very weak but still visible in the Wiener filters. This motivates us to use a large $\sigma = 1$ s to maximize the information coming from the small reflected waves, together with a small $\zeta = 10^{-5}$. However, these settings do not make it possible to obtain satisfactory results with AWI. To have a deeper understanding of why AWI fails to reconstruct a proper V_P model in this case, we perform a sensitivity analysis of the misfit function with respect to the value of V_P in the layer. We compute the AWI misfit value (and L^2 misfit value for reference) between d_{obs} and $d_{cal}(V_P)$. Here d_{obs} corresponds to a shot gather in the center of the acquisition generated in the true model. $d_{cal}(V_P)$ corresponds to data generated in different models similar to the true model, with as only varying parameter the layer velocity (ranging from ± 100 m/s around the layer velocity of the true model). The results of this analysis are presented in Figure 16. The L^2 results are coherent with the expectation: the misfit function is convex with respect to the variation of the layer velocity and presents a minimum when the velocity of the layer used to generated $d_{cal}(V_P)$ is similar to the one of the true model used to generate d_{obs}, so respectively 1400 and 1600 m/s . For AWI, we observed that the minimum is not aligned with the correct velocity (1440 m/s in the first case, 1610 m/s in the second). This exhibits the loss of sensitivity of AWI in this case, explaining the failure of convergence of the FWI. Only reducing the σ below 0.04 s would make AWI behaves more like L^2 and converge to a result similar to the one of NIM. We do not think such a parameterization is interesting as it prevents the advantages introduced by AWI, which is improved convexity, and still introduces artifacts in the reconstructed model and computational overhead compared to L^2 .

This second FWI test is a good illustration of the potential limitation that an alternative misfit function mainly focused on resolving time-shift could introduce. Here AWI and NIM have difficulties providing satisfactory results when the main arrival is correctly predicted. The point to point approaches that classical L^2 procures is here the "reference", making possible to fit the small perturbation properly following the main arrivals. AWI and NIM being more kinematic oriented, it is not surprising that this setup is challenging for such formulations.

MARMOUSI CASE STUDY: TOWARD A MORE REALISTIC CASE STUDY

Common framework

We design a synthetic case study using the Marmousi II P-wave velocity model (Figure 17) (Martin et al., 2006) to continue our analysis on a more realistic FWI configuration. We use a fixed spread surface acquisition model with 128 sources and 169 receivers. The source spacing is 132 m, and the receiver spacing is 100 m. The data is generated with a 4 Hz centered Ricker wavelet high-pass filtered to remove energy below 2 Hz (wavelet is visible in Figure 24). The recording time is set to 7 s. PML absorbing layers are used on the bottom and lateral sides of the model to mimic a medium of infinite extension in these directions, while a free surface condition is applied on top.

In the first case, referred to as "inverse crime inversion", we model the data in the constant density acoustic approximation and use the same grid for modeling and inversion to remain in the inverse crime settings. The mesh spacing is 25 m in this case.

In the second case, referred to as "more realistic inversion", we use the variable density Marmousi II model and a refined 10 m grid to generate the data. White noise bandpassed between 2 Hz and 10 Hz is added to the data to reach a signal to noise ratio of 15%. The inversion is done on a 25 m grid, using a density model derived from the initial V_P model through Gardner's law (Gardner et al., 1974) (Figure 18). A wavelet estimation is done before inversion. Performing the inversion in this more realistic framework, away from the usual inverse crime settings, makes it possible to assess the effects of incorrect amplitude prediction on the different misfit functions and better judge their usability toward field data applications.

The optimization is performed using the l-BFGS algorithm. The regularization of the gradient is defined as 0.3 of the local wavelength. Pseudo-hessian preconditioning is used (Choi and Shin, 2008; Yang et al., 2018a). The lower and upper V_P boundaries for the inversion are respectively set to 1000 m/s and 5200 m/s . Inversion is performed without any frequency continuation approaches or other multi-scale strategies for all the misfit functions considered (including L^2).

Inverse crime inversion

Case study description

We rely here on two starting models. The first one, called $\mathcal{S}500$ (Figure 17 c), is derived from the true V_P model using a Gaussian smoothing with a correlation length of 500 m. This starting model preserves the long wavenumber content of the true model. The second one is a linearly increasing vertical 1D (Figure 17 d) model, ramping from 1500 m/s at seabed to 4500 m/s at depth. This initial model does not contain long-wavelength structures inherited from the true model. The data-fit obtained through this initial model (Figure 19 b) is affected by cycle-skipping. In comparison, the data-fit obtained with $\mathcal{S}500$ is globally better (Figure 19 a), with more in-phase arrivals (especially on 0 to 3 km offset diving waves).

Results starting from $\mathcal{S}500$ initial model

Reconstructed V_P results for all the selected misfit functions are presented in the left column of Figure 20. The associated data-fit obtained after FWI are presented in Figure 21 (for a common shot gather in the middle of the acquisition).

In this model, the L^2 results give, at first order, a good reconstruction of the Marmousi model. However, we can observe on the left part that the horizontal layers are not correctly reconstructed and present an up-shift ($0 < x < 3$ km), associated with a low-velocity anomaly on the shallow left part of the model (around $x = 1$ and $z = 0.8$ km). This corresponds to the part where strong reflections are generated. Because the background velocity is incorrectly predicted in the early iterations, the arrivals corresponding to these reflections are cycle-skipped. As we illustrated earlier, L^2 misfit function being unable to tackle cycle-skipping effects, FWI cannot update the medium correctly to fit these arrivals.

AWI provides a clear improvement over classical L^2 : the horizontal layers are correctly positioned on the left part. The central part at depth ($8 < x < 13$ km, ($z > 2$ km) is improved compared to L^2 , with better contrast and more lateral coherency in the layers structure. Moreover, the low-velocity anomaly on the shallow left part of the model is removed. These results are obtained with $\sigma = 0.25$ s and $\zeta = 10^{-5}$.

IE also improves the reconstructed model. The relatively small improvement in cycle-skipping robustness introduced by the envelope is enough to mitigate the artifacts on the left part of the model ($0 < x < 3$ km) and flattens the layers compared to the L^2 result. One drawback is the slight degradation in the reconstruction of the central part at depth ($9 < x < 13$ km, $z > 2$ km). Still, such a simple formulation is enough to improve the FWI workflow over the classical L^2 in this case.

The case of NIM misfit is interesting. Here, we can see that it fails to converge, producing an erroneous reconstructed model. This illustrates the limitation of NIM when applied to more realistic cases where the data contains multiple arrivals, multiple phases, and potentially mixed phases. Integrating all these pieces of information into a single observable (the cumulative distribution) does not make it possible to reconstruct the subsurface velocity. As we can see in the data-fit, NIM can also not fit the vast majority of the signal.

The KROT misfit function, as AWI and IE, can prevent the appearance of the left side artifacts observed with the L^2 reconstruction ($0 < x < 3$ km). As for IE, the relatively small improvement in terms of attraction valley width provided by KROT is sufficient to improve the results significantly. Besides, KROT can account for the lateral coherency of the data, which might also help stabilize the inversion.

Finally, GSOT also produces a significant improvement over

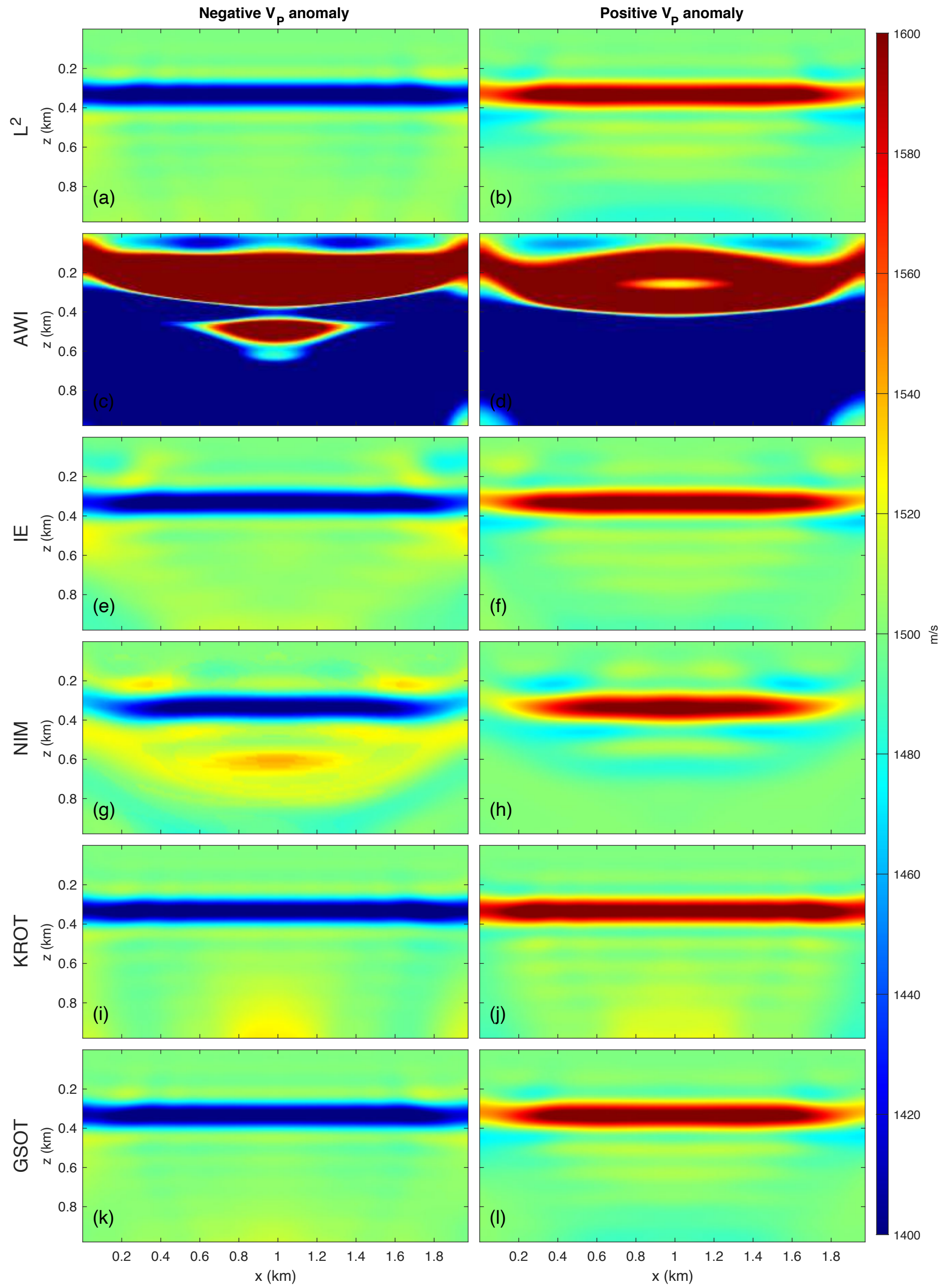

Figure 12: FWI Test 2: FWI layer benchmark results. Left column corresponds to a negative V_P anomaly while the right column to positive V_P . The subfigures under respectively correspond to final reconstructed V_P model obtained with FWI using L^2 (a,b), AWI (c,d), IE (e,f), NIM (g,h), KROT (i,j) and GSOT (k,l).

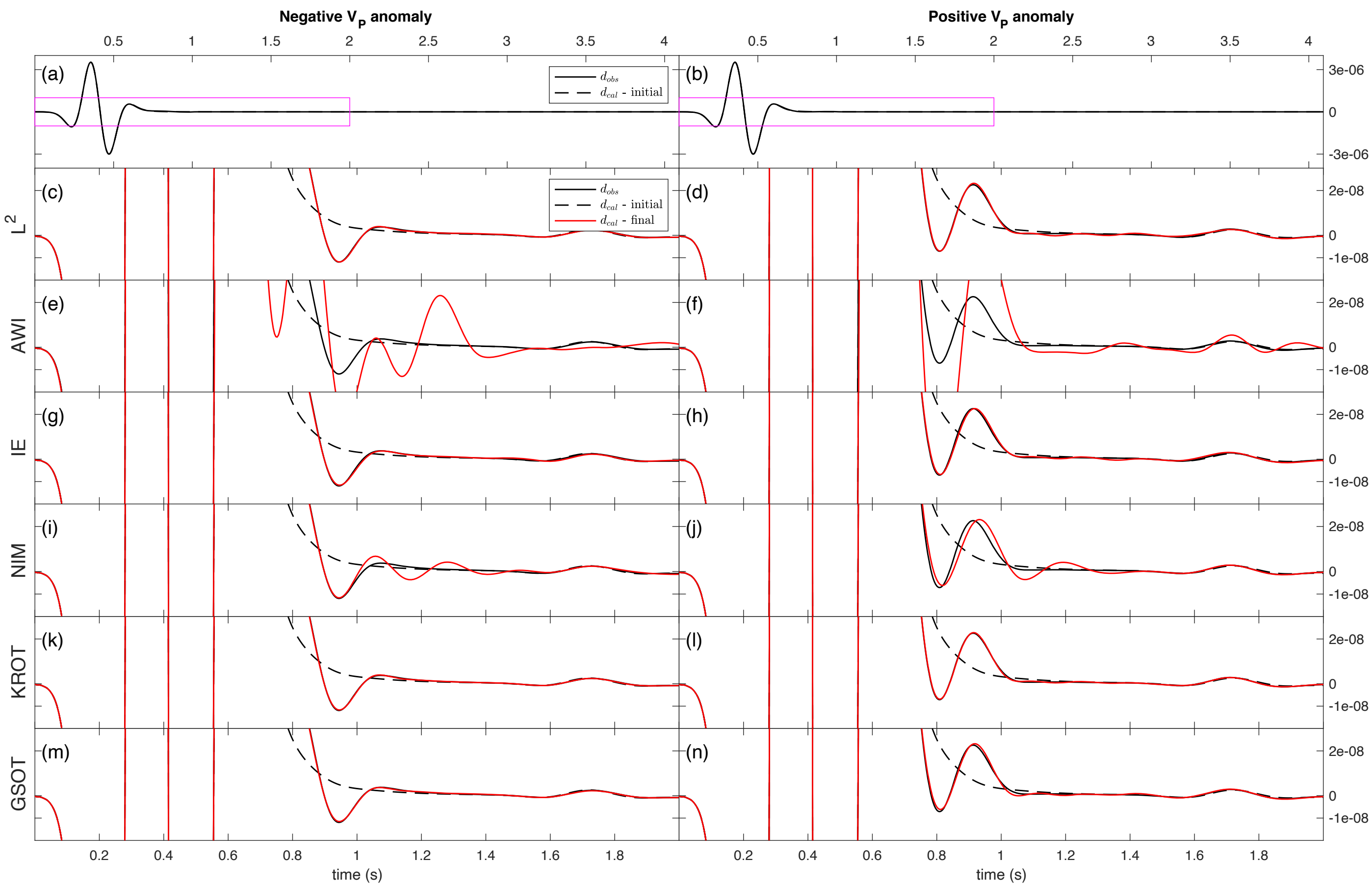

Figure 13: FWI Test 2: Extracted traces for a center shot at zero offset. Observed data are in solid black, synthetic data in the initial model in dashed black, and final reconstructed synthetic data in solid red. The left column corresponds to a negative V_P anomaly while the right column to positive V_P . (a,b) shows the complete traces. Each subfigures under are cropped on the magenta box to emphasize the polarity reversal introduced by the layer. They correspond to traces calculated in the reconstructed V_P model using respectively L^2 (c,d), AWI (e,f), IE (g,h), NIM (i,j), KROT (k,l) and GSOT (m,n) misfits functions.

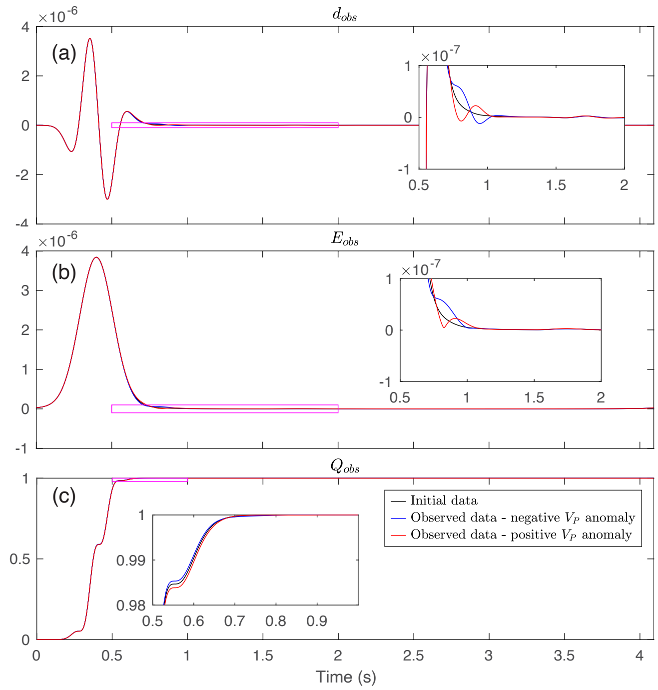

Figure 14: FWI Test 2: Extracted traces for (a) d_{obs} and d_{cal}, (b) E_{obs} and E_{cal}, and (c) Q_{obs} and Q_{cal} for center shot at zero offset. Data associated to negative V_P anomaly are in solid blue, and positive V_P anomaly in solid red. Calculated data is in solid black. Magenta box corresponds to the enlarged area.

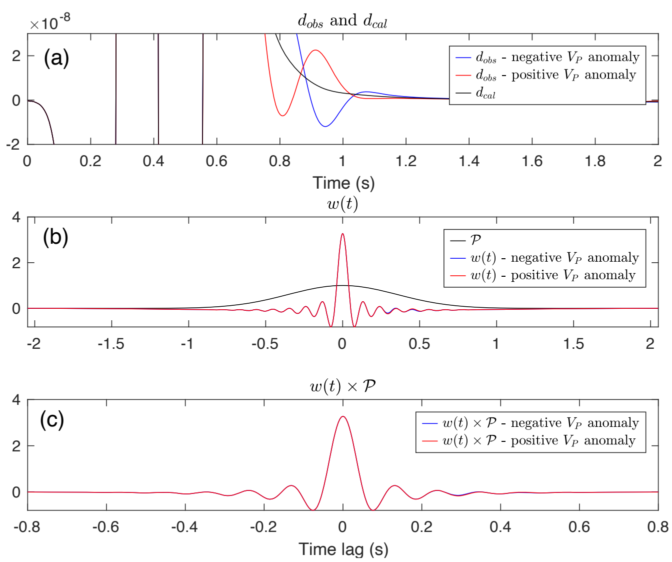

Figure 15: FWI Test 2: (a) Extracted traces d_{obs}, (b) Wiener filters $w(t)$ and $\mathcal{P}$, and (c) $w(t) \times \mathcal{P}$ for center shot at zero offset at first iteration. Data associated to negative V_P anomaly are in solid blue, and positive V_P anomaly in solid red.

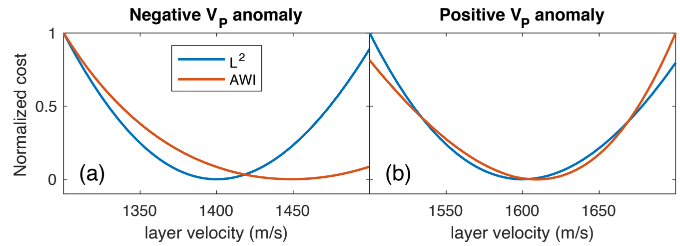

Figure 16: FWI Test 2: Misfit value for L^2 and AWI with respect to the layer velocity for the two inversion cases: (a) negative and (b) positive V_P anomaly. Misfit is calculated between d_{obs} (data in the true medium) and a $d_{cal}(V_P)$ generated in a medium with a correct background $V_P = 1500$ m/s and a as only varying parameter the layer V_P , ranging from ± 100 m/s around the original velocity of layer in d_{obs}.

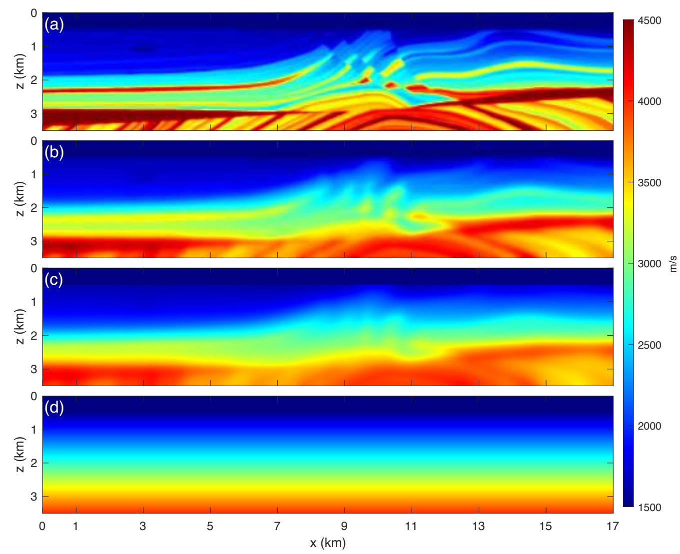

Figure 17: (a) True Marmousi II model. (b) $\mathcal{S}250$ initial model, (c) $\mathcal{S}500$ initial model and (d) 1D initial model.

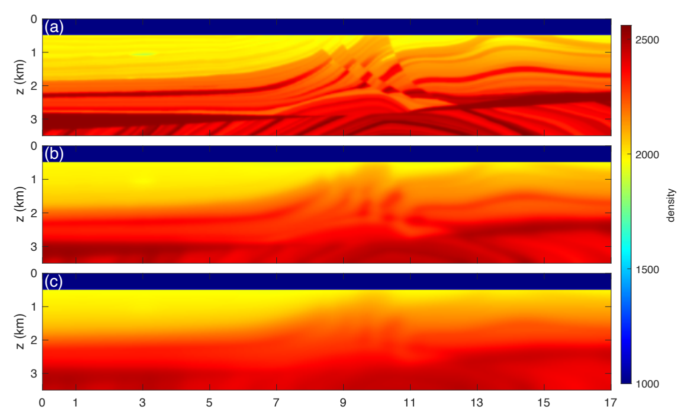

Figure 18: Density model obtained from V_P using Gardner's law for: (a) True Marmousi II model, (b) $\mathcal{S}250$ initial model and (c) $\mathcal{S}500$ initial model.

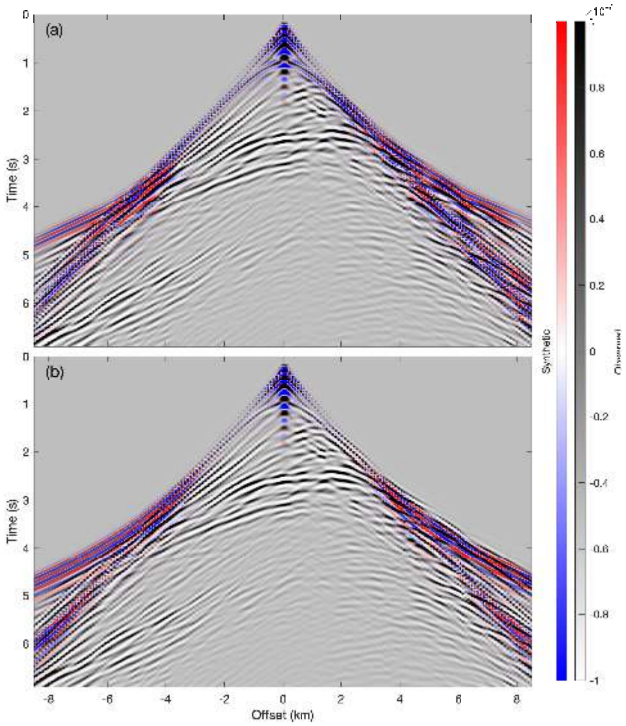

Figure 19: Inverse crime inversion: CSG for field data overlapped by synthetic data in (a) $\mathcal{S}500$ initial model, and (b) 1D initial model. Field data in black and white, overlapped by red to blue synthetic data with transparency. Red and white visible mean out of phase, black and blue mean in phase.

the L^2 result, with almost no artifacts on the reconstructed model. We use $\tau = 0.25$ s, similarly to AWI parameterization. This illustrates that GSOT, as AWI, while being able to significantly enlarge the valley of attraction of the misfit function on simple convexity cases, can also be used in a more realistic framework that mixes transmitted and reflected energy with relatively complex multi-arrival data.

Regarding the data-fit, excepted for NIM, all the misfit functions can provide a good data-fit in this case.

Results starting from 1D initial model

Reconstructed models from the 1D initial model are presented in the right column of Figure 20. The associated data-fit obtained after FWI are presented in Figure 22.

Starting from this initial model, strong artifacts appear on the L^2 results. We observe long-wavelength low-velocity anomalies on both left and right parts of the models typical of cycle-skipping induced artifacts. The data-fit analysis confirms this observation: only early arrivals in the near offset are correctly fitted. Diving waves arriving at larger offsets on the left (5 s and -7.5 km offset) and right parts (4 s and 6 km offset) of the gather are cycle-skipped.

Without any surprise, NIM cannot provide a meaningful estimate of the V_P model, as it is already the case starting from the $\mathcal{S}500$ initial model.

As the 1D initial model generates large time-shifts, and since both IE and KROT are only marginally improving cycle-skipping robustness, it is not surprising to observe artifacts on the associated reconstructed V_P models. IE results present strong artifacts, mainly on the left part of the model ($0 < x < 7$ km), while the shallow right part ($11 < x < 16$ km, $x <= 2$ km) presents an improved reconstruction compared to L^2 . This is confirmed by the data-fit, where all the arrivals on the right part (offset between 1 and 8 km) are correctly predicted, whereas data-fit on the left part (offset between -8 and 0 km) is degradated compared to L^2 data-fit. Conversely, KROT provides a more accurate reconstruction in the left part of the model. The strong low-velocity anomalies observed in the left part ($0 < x < 7$ km) of the L^2 and IE reconstructions are reduced and appear only in the deep part of the model ($z > 2$ km). This is consistent with the data-fit, where we see that using KROT, the long offset diving waves for negative offsets of the shot gather are correctly fitted.

AWI manages to provide a clear improvement over classical L^2 , mainly on the center part of the model ($5 < x < 14$ km), while some artifacts on both sides of the reconstructed model are still present. These parts are more difficult to reconstruct as they are illuminated by waves traveling along the longest paths of the medium, increasing cycle-skipping risk. We set $\sigma = 0.6$ s to try to capture as large as possible time-shifts. To get the best results possible, we used $\zeta = 10^{-2}$ to obtain the helping smoothing effect required to tackle large time-shifts introduced by this initial model. The data-fit obtained with AWI is good, with only some out-of-phase arrivals for late diving waves (around -8 to -6 km offset).

Using GSOT, the reconstructed V_P model also presents a clear improvement over classical L^2 . We use $\tau = 0.6$ s in this case. At the first order, most of the artifacts are removed. Still, some artifacts are present close to the edges ($0 < x < 1$ and $15 < x < 17$ km), which is expected from the lack of illumination in these parts. Some other artifacts are visible in the center part of the structure at depth ($10 < x < 12$ km and $2 < z < 3.5$ km). The data-fit appears to be good, with no out of phase arrivals for all offsets.

Error reduction analysis

Besides this qualitative analysis of the results, we can provide quantitative comparisons by analyzing the data error and model error evolution along with iterations. We use here the following relative L^1 model error definition

$$Err(V_P) = \frac{100}{M} \sum_{i=1}^{M} \frac{|V_{P,i} - V_{P,i}^{true}|}{V_{P,i}^{true}} \quad (38)$$

where V_P^{true} is the true model, M the number of points in the model and i denotes one pixel of the grid used to describe the models at the discrete level.

For the second experiment only (1D initial model), we present the evolution of

- the convergence rate (misfit error with respect to the iterations);
- the L^2 convergence rate (L^2 error with respect to the iterations);

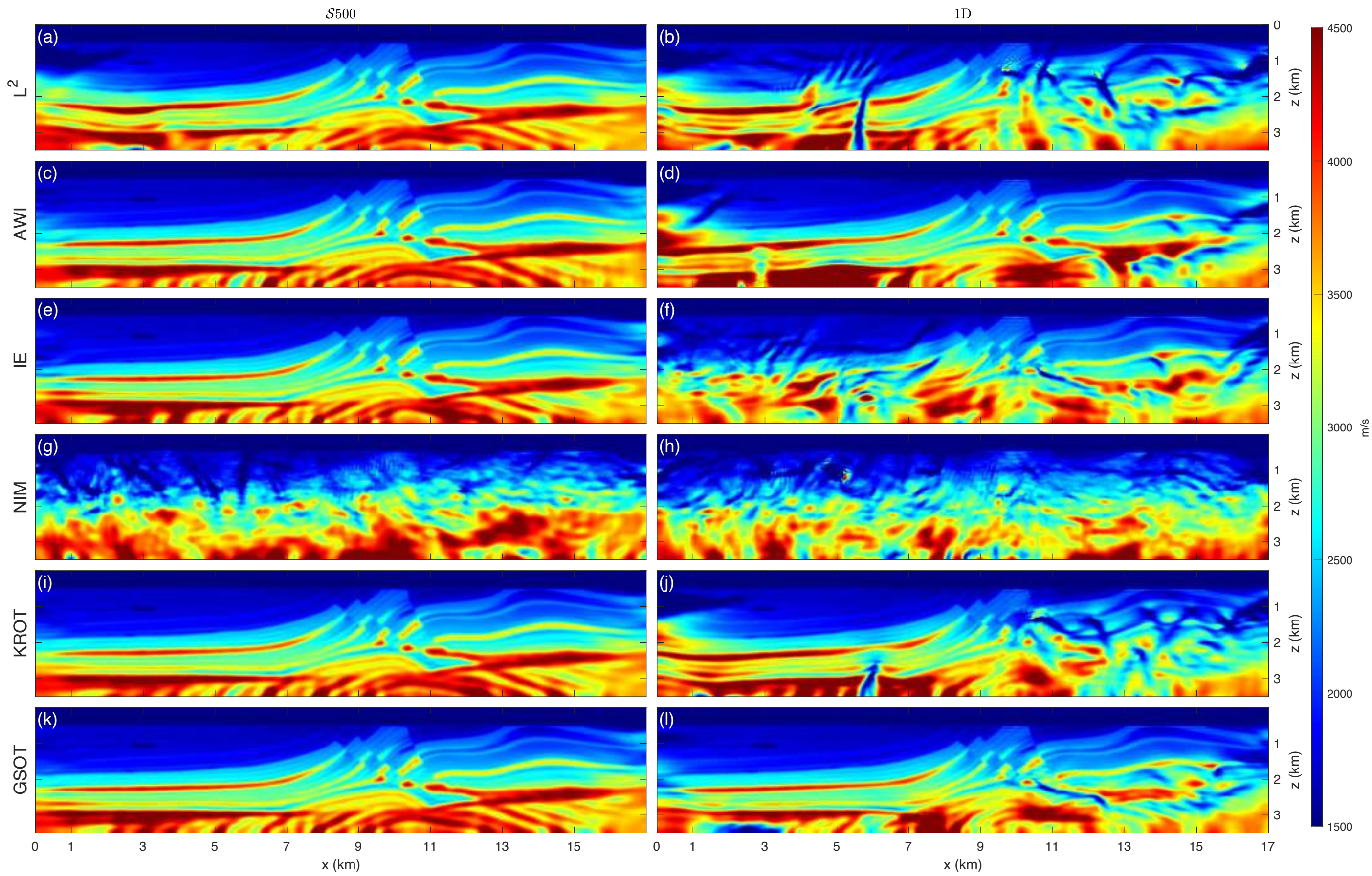

Figure 20: Inverse crime inversion: inverse crime FWI final reconstructed V_P model for Marmousi. Left column corresponds to $\mathcal{S}500$ initial model, right column to 1D initial model. The lines respectively correspond to the final reconstructed V_P model using L^2 (a,b), AWI (c,d), IE (e,f), NIM (g,h), KROT (i,j) and GSOT (k,l).

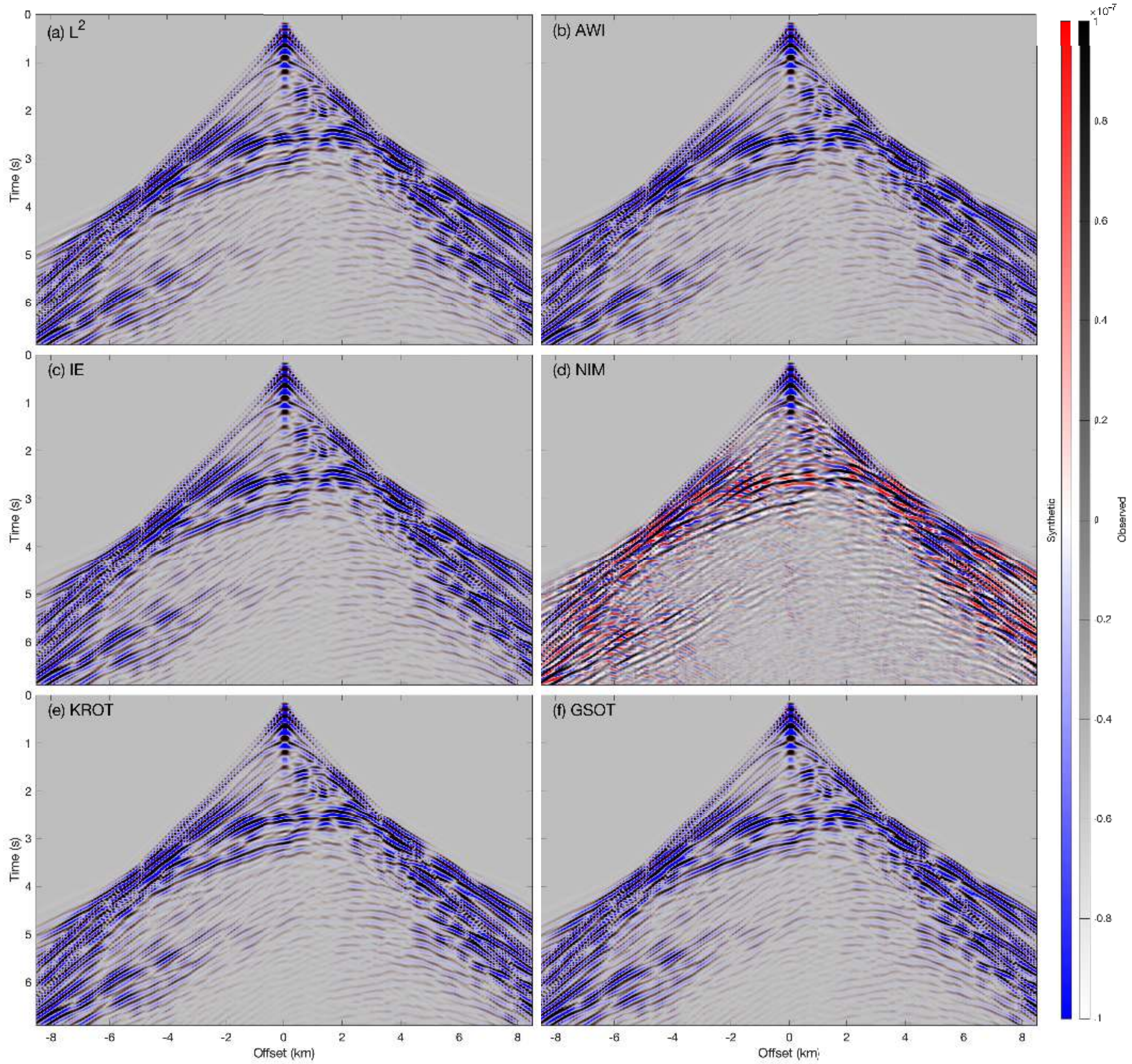

Figure 21: Inverse crime inversion: Overlapped common shot gathers for synthetic data in the final reconstructed V_P model starting from $\mathcal{S}500$ initial model vs field data. Each subfigure corresponds to misfit function, with L^2 (a), AWI (b), IE (c), NIM (d), KROT (e), and GSOT (f).

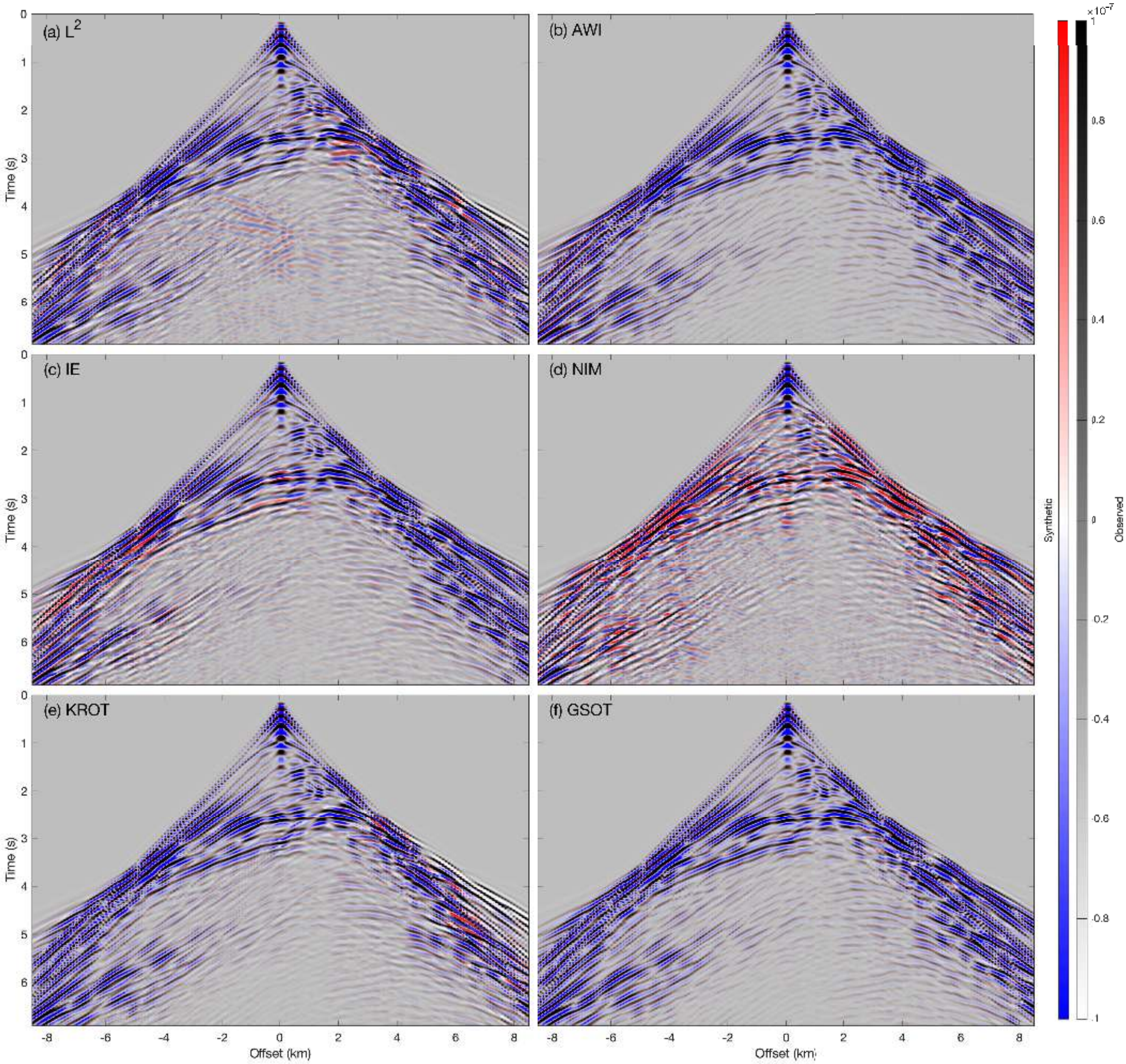

Figure 22: Inverse crime inversion: the same as Figure 21 but starting from 1D initial model.

- the model convergence rate (model error with respect to the iterations);
- the model vs. data convergence rate (model error with respect to the misfit error).

For model error, we truncate the model by 1 km on the left and right sides and 625 m at depth to remove the model areas that are not well illuminated.

For alternative misfit function definition, the L^2-based convergence rate is interesting as moving away from L^2 local minima should be made visible by an increase of the L^2 error with respect to the iterations. Also, the fourth item is interesting, as, ideally, we look for a monotonic decrease of the model error with respect to the misfit error. Besides, to improve the readability, we have excluded from these figures the results corresponding to NIM. The method does not produce reliable results in both cases.

The error reduction analysis is shown in Figure 23. First, we observe that KROT and AWI present a relatively slow convergence rate on the cost evolution, while IE and GSOT have a faster convergence rate. L^2 convergence is in between. KROT follows more or less the same as the L^2 misfit function. This is somehow expected, as the valley of attraction of KROT is expected to be similar to the one of the L^2 misfit function. Note, however, that in the early iterations, KROT displays a small increase of the L^2 error, which clearly states that the two misfit functions follow a different minimization path. IE, AWI, and GSOT display another trend: the L^2 error is increased in the first iterations before being strongly decreased in a second stage. The substantial decrease of the L^2 error appears the latest for AWI (after 100 iterations) and the earliest for GSOT (after 30 iterations). GSOT achieves the smallest L^2 misfit, followed by AWI and KROT. The model convergence rate classifies the misfit functions into two groups: one that does not reduce model error compared to the starting point, with L^2 , IE, and KROT; and a second group that decreases the model error with AWI and GSOT. In the second group, only GSOT provides a constant decrease with respect to the iterations, while AWI start to increases the model error until 100 iterations, followed by a decrease. The final reduction of model error obtained with KROT and IE are smaller than the one attained by the L^2 , still, this does not explicitly compared to better interpretable results overall. AWI and GSOT obtain the best reduction of model error. Finally, looking at the model vs. data convergence, only GSOT provides a quasi-monotonic decrease. For all the others, the model error starts by increasing with the reduction of the misfit.

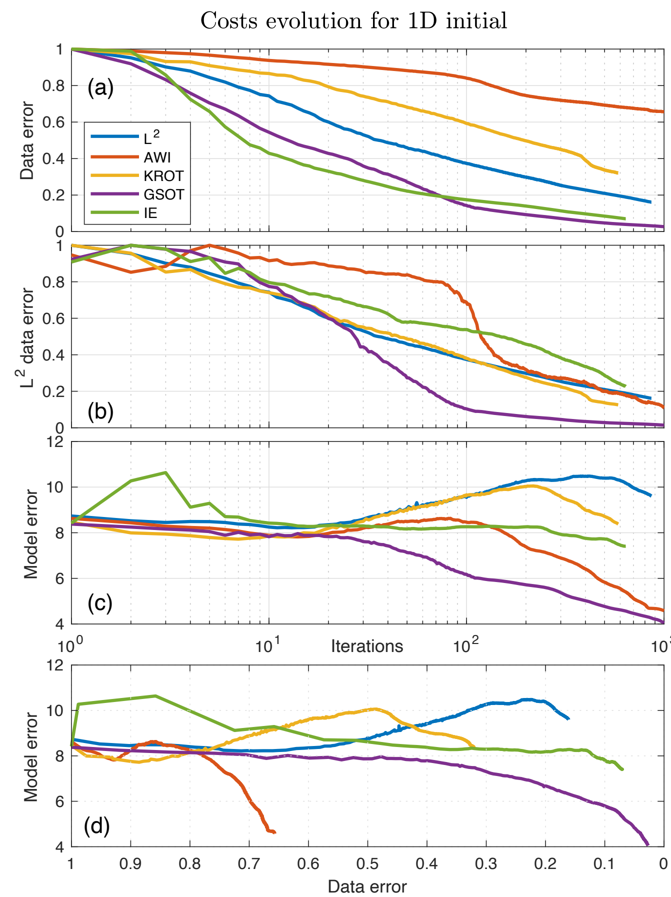

Figure 23: Inverse crime inversion: Costs evolution in the inverse crime Marmousi for 1D initial model. (a) evolution of cost functions over iterations, (b) true L^2 cost evolution over iterations, (c) model error reduction over iterations and finaly (d) model error vs. the data error reduction.

A more realistic inversion

Case study description

Similar to the previous inverse crime inversion, we perform FWI starting from two different initial models. The first one is derived from the true Marmousi model using a lighter Gaussian smoothing, referred to as the $\mathcal{S}250$ model with a correlation length of 250 m (Figure 17 b). The second one is the $\mathcal{S}500$ model already used in the inverse crime settings (Figure 17 c).

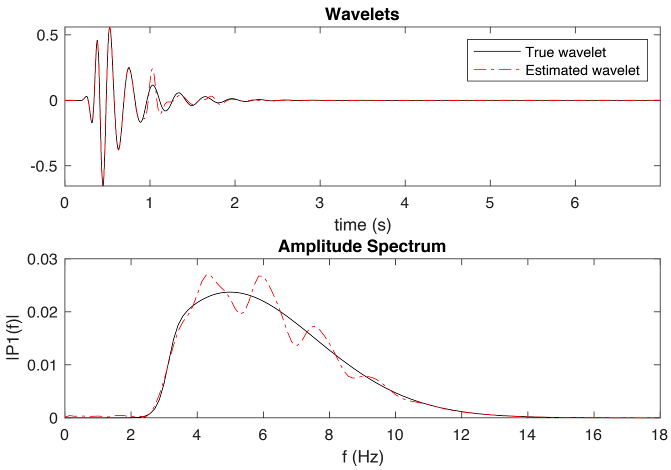

Figure 24: Non inverse crime inversion: (a) comparison of the true wavelet used to generate the observed data versus the inverted wavelet used for inversion in the more realistic inversion. (b) Associated amplitude spectrum of the true wavelet and the inverted wavelet.

We did not re-use the 1D initial model as it proves to be too difficult for any alternative misfit functions to provide convergence in this more realistic case. Not predicting the data to machine precision generates a more challenging benchmark.

The wavelet used for FWI is obtained through a source estimation in the initial model based on short offset (100 m) only to decouple the influence of the initial V_P model as much as possible. The obtained inverted wavelet is presented in Figure 24. We can observe that the inverted wavelet is close to the true wavelet, but some noticeable amplitude and waveform differences are visible. These amplitude effects are induced by the use of a "true" density model for the data generation compared to Gardner's one (Figure 18) used for wavelet estimation and to the white noise added to the data.

The data-fits for these two initial models is presented in (a) of Figures 26 and 27. As expected, the data-fit is better using $\mathcal{S}250$ initial model, while the data-fit generated with $\mathcal{S}500$ initial model displays more out of phase arrivals.

We compare the results obtained using L^2 , AWI, IE, KROT, and GSOT misfit functions. We do not include NIM results here, as we have already shown how the method fails to produce meaningful results in the previous inverse crime settings. A maximum of 500 FWI iterations is performed for both initial models.

Results starting from $\mathcal{S}250$ initial model

Starting from the $\mathcal{S}250$ initial model (Figure 25 left column), the main expected difference with the previous "inverse crime" setup is an inaccurate amplitude prediction (which would be the case if considering field data). Data-fit are presented in Figure 27. Classical L^2 can provide an acceptable result. Good reconstruction in the well-illuminated area is achieved, with no visible artifacts in the center part and only a small low-velocity artifact visible at $x = 2$ km $z = 0.8$ km and a high-velocity artifact at $x = 16$ km $z = 1$ km. The data-fit obtained with L^2 is quite satisfying with most of the arrivals in phase.

This time, IE results are clearly degraded compared to the classical L^2 one. The reconstructed V_P model is tainted with high wavenumber oscillation and strong artifacts. This is an indication that the IE approach is sensitive to a correct amplitude prediction. This validates the interest of a more realistic framework, making us able to detect this kind of limitation. The data-fit presents many out-of-phase arrivals, coherent with the small artifacts present everywhere in the reconstructed V_P .

Again, AWI improves over the L^2 results. We use a relatively small $\sigma = 0.2$ s here as the maximum time-shifts expected are relatively small with this good initial model. We used $\zeta = 10^{-2}$ as noise requires a relatively large amount of damping, moreover as illustrated before, a larger damping value helps when facing challenging FWI setups. The deep center part is improved with a more coherent deep-layer structure. The left ($x = 2$ km and $z = 0.8$ km) and right ($x = 16$ km $z = 1$ km) side artifacts present in L^2 results are also partially mitigated. Surprisingly, the data-fit obtained with AWI is poor for large offset arrivals (from -8 km to -3 km and 3 to 8 km). This degradation of the data-fit is slightly counter-intuitive and does not correlate with the improvement of the reconstructed V_P model observed.

Finally, KROT and GSOT reconstructed models both present similar improvement compared to the L^2 one. We can observe an increase in terms of high wavenumber content. Interestingly, the deep center part ($9 < x < 13$ km, $z > 2$ km), which is the main target of interest of the Marmousi model (an anticlinal structure) is more resolved using KROT and GSOT compared to L^2 . For GSOT, we use $\tau = 0.2$ s in this case. The data-fit obtained with both methods is good, with almost all arrivals in phase. Only some first arrivals between -4 to -2 km offset are still not well explained. The GSOT data-fit appears to be slightly better than the KROT one.

Results starting from $\mathcal{S}500$ initial model

Starting from the $\mathcal{S}500$ model, reconstructed V_P results are presented in Figure 25 right column, while data-fit are presented in Figure 27.

Here, the classical L^2 fails to reconstruct a meaningful V_P model. Many artifacts are present on the model that may come in part from cycle-skipping. This would prevent any interpretation of the reconstructed model. The data-fit present out-of-phase arrivals, even if the majority would appear to be in-phase. This again illustrates potential convergence toward a local minimum that makes possible to fit the data with non-meaningful V_P updates.

With no surprise, IE fails to reconstruct a meaningful V_P estimate. The reconstructed V_P model suffers from many artifacts. The data-fit is clearly degraded compared to L^2 , which is likely explained by the difficulty faced by IE in tackling wrong amplitude predictions compared to classical L^2 .

AWI reconstructed model produces here an improvement over L^2 or IE, with the central part and right part of the Marmousi model more or less retrieved. However, significant artifacts are present in the left part of the model ($1 < x < 6$ km)

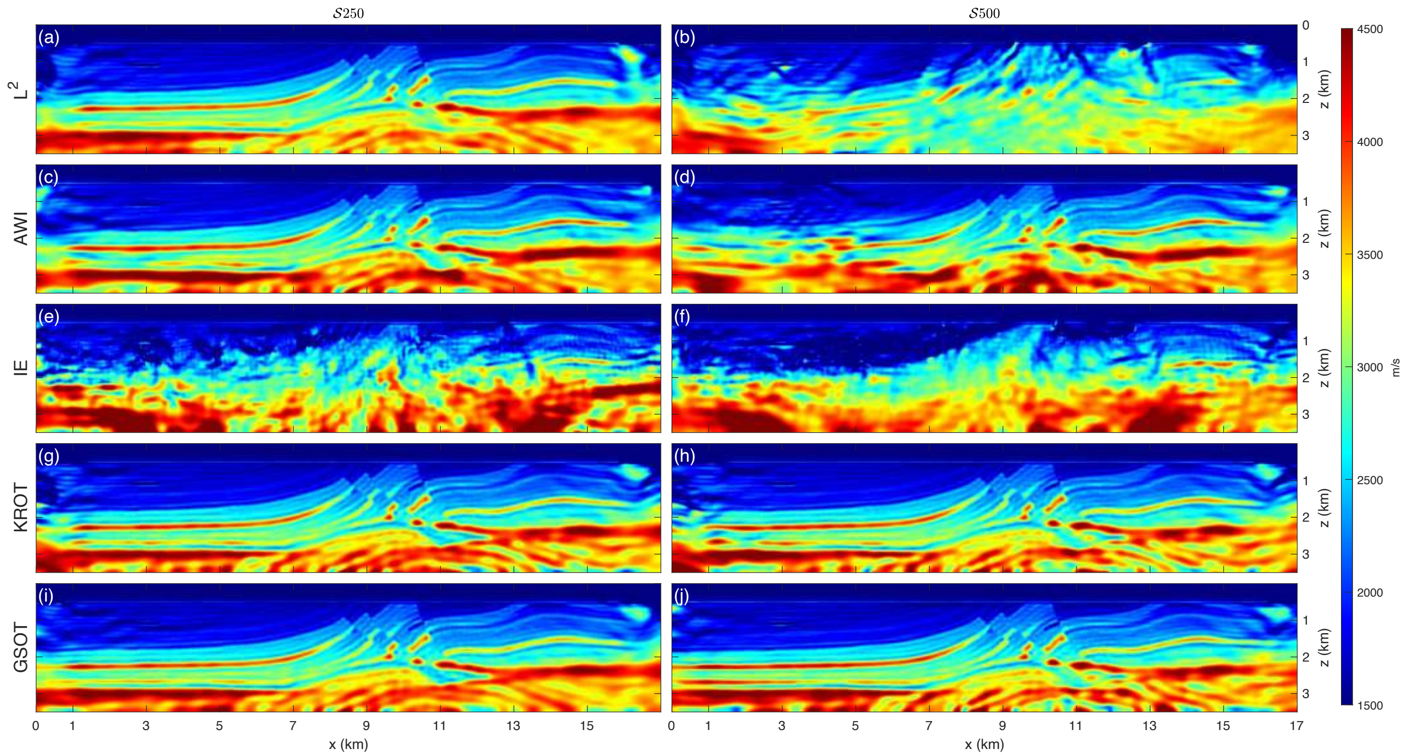

Figure 25: Non inverse crime inversion: More realistic FWI final reconstructed V_P model for Marmousi. Left column corresponds to $\mathcal{S}250$ initial model, right column to $\mathcal{S}500$ initial model. The lines respectively correspond to the final reconstructed V_P model using L^2 (a,b), AWI (c,d), IE (e,f), KROT (g,h) and GSOT (i,j).

associated with an erroneous reconstruction of the central part at depth ($9 < x < 13$ km, $z > 2$ km). Here we increase σ to 0.4 s, and keep $\zeta = 10^{-2}$. The data-fit is degraded with out of phase arrivals for offsets between -8 to -3 km as well as between 2 to 8 km.

KROT produces satisfactory results here. This is interesting as KROT only marginally improves cycle-skipping robustness. Here, it manages to perform well in this complexified case. This is a good indication that the difficulties induced in this more realistic inversion are not only cycle-skipping but also amplitude mismatch (due to density) and noise. As KROT introduces lateral coherency and has a regularizing effect on noise, it is not surprising to observe a better behavior in this case. The data-fit obtained with KROT is good with almost all arrivals in phase, except for some transmitted waves from -3 to -1 km offset and some long offset arrivals around -8 to -7 km.

Finally, GSOT provides a good reconstructed V_P model. The central part ($9 < x < 13$ km, $z > 2$ km) is well reconstructed. The layers show more lateral coherency compared with KROT. Furthermore, left side artifacts are reduced compared to KROT. Again and similarly to AWI, τ is increased to 0.4 s to account for the larger time-shifts introduced by the degraded initial model. The data-fit is also good, with improvement over the KROT for the long offset arrivals around -8 to -7 km.

Error reduction analysis

A similar analysis for the different misfit functions is presented for this inverse crime inversion of Marmousi. The model error is calculated in a similar zone as in the previous experiment.

Starting from the $\mathcal{S}500$ initial model (Figure 28), we observe that KROT and AWI present again a relatively slow convergence rate (AWI being the slower), while L^2 , IE and GSOT have a faster convergence rate. The L^2 data-error is again interesting, with GSOT and KROT performing the most substantial reduction of L^2 data error (with an initial jump to pass a L^2 local minimum for GSOT at the first iteration). While IE increases the cost drastically for the first two iterations, it then fails to reduce the data error. We can note that KROT is not following L^2 misfit function behavior anymore compared to the inverse crime Marmousi case. Regarding AWI, we can observe that it starts to increase the L^2 data error until 20 iterations, then rapidly reduce for 10 iterations, to finish with a constant increase afterward. This time, the model error displays a strong increase for L^2 and IE misfit functions, which is coherent with the artifacts present in the reconstructed V_P models. AWI is also increasing the model error as it is also affected by artifacts, but less drastically than L^2 and IE, which is visible on the reconstructed V_P model. KROT and GSOT manage to decrease the model error continuously. Looking at the model vs. data convergence, only KROT and GSOT present monotonic behavior, while L^2 , IE, and AWI are increasing the model error.

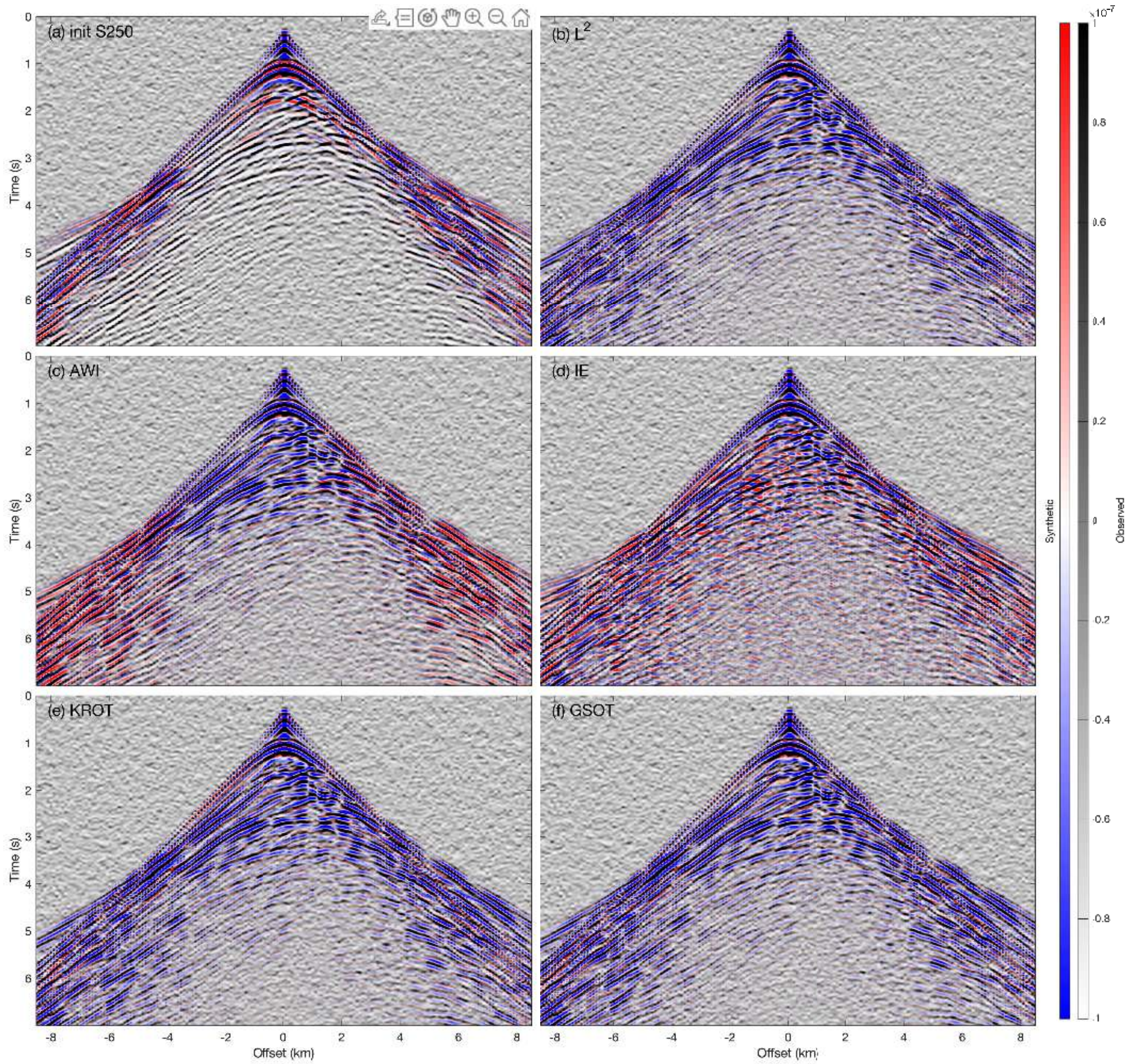

Figure 26: Non inverse crime inversion: Overlapped common shot gathers for synthetic data in the final reconstructed V_P model starting from $\mathcal{S}250$ initial model vs field data. (a) corresponds to the data-fit in the $\mathcal{S}250$ initial model. Then, each subfigure corresponds to misfit function, with L^2 (b), AWI (c), IE (d), KROT (e), and GSOT (f).

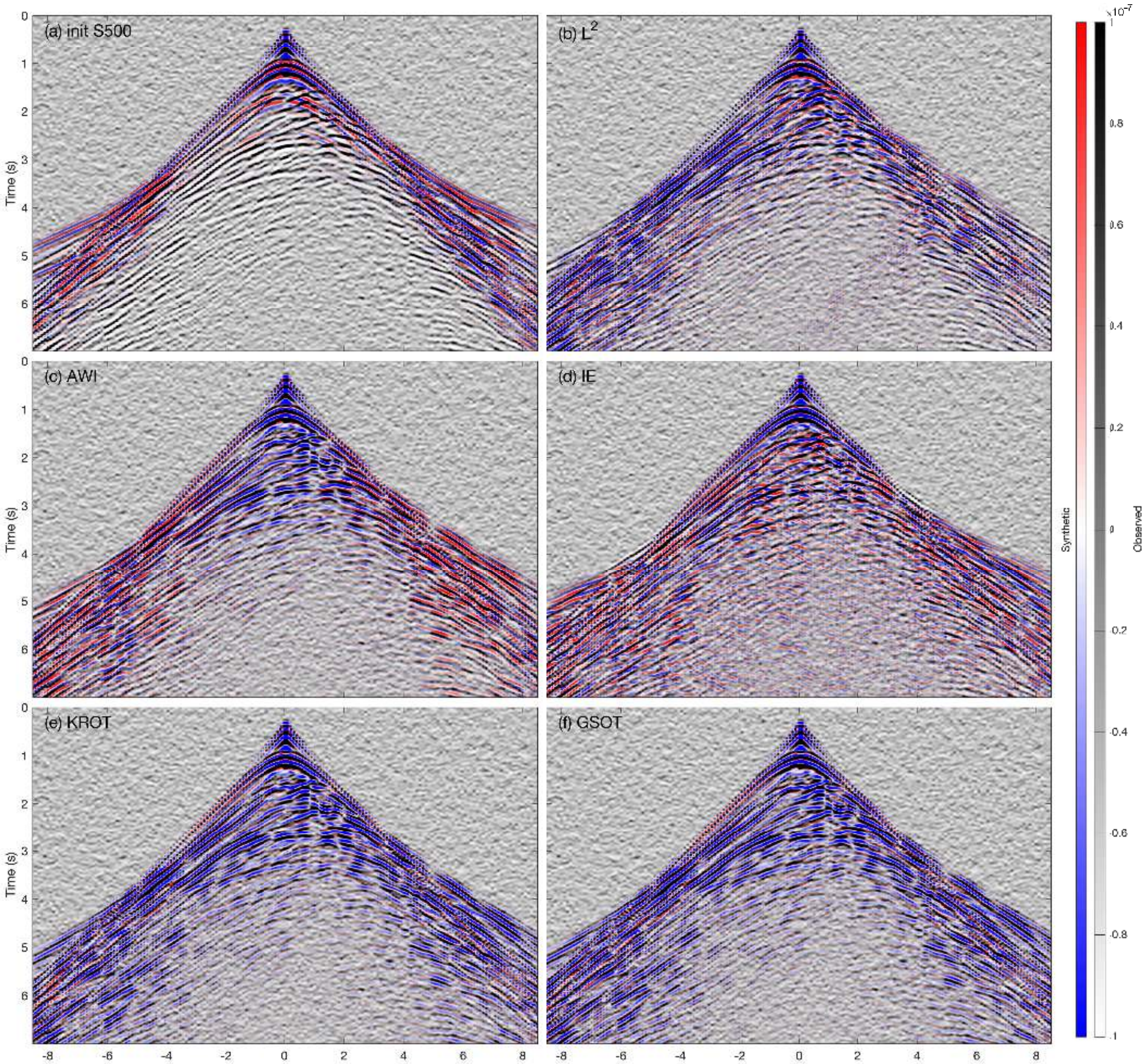

Figure 27: Non inverse crime inversion: Same as Figure 26 but starting from $\mathcal{S}500$ initial model.

Computational cost

The computational overhead induced by the alternative misfit function selected in this review varies from +2 to +30% compared to L^2 misfit. These values are coherent with the values documented in the literature. The key feature here is that even a +30% computational overhead is not a blocking feature and is affordable with modern computing facilities. For us, the key feature is the "physical" performance of the misfit function that translates into an improvement of FWI robustness.

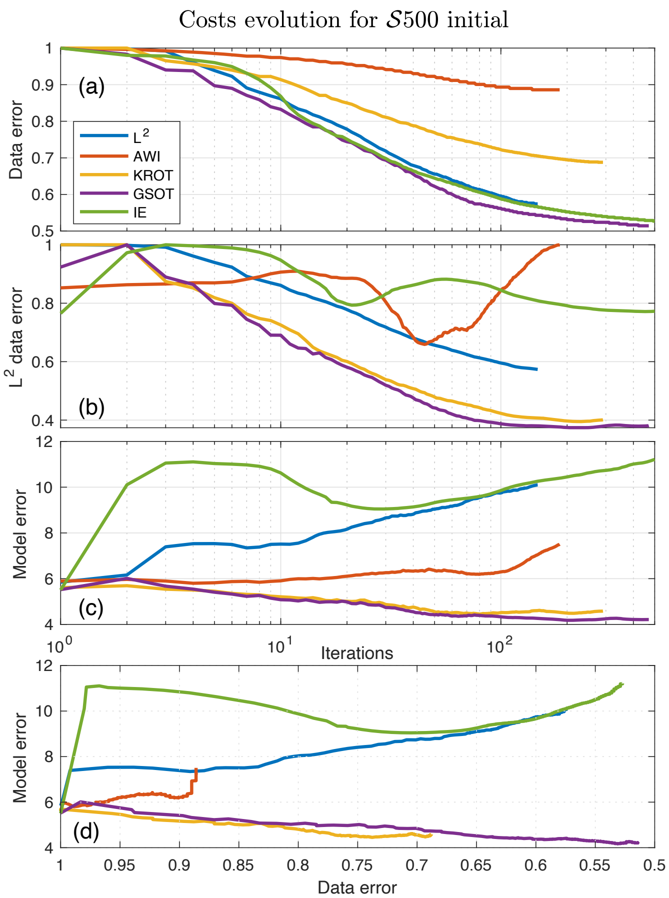

Figure 28: Non inverse crime inversion: Costs evolution in the more realistic Marmousi for $\mathcal{S}500$ initial model. (a) evolution of cost functions over iterations, (b) true L^2 cost evolution over iterations, (c) model error reduction over iterations and finaly (d) model error vs. the data error reduction.

DISCUSSION

Among the five misfit functions compared here, namely NIM, IE, AWI, KROT, GSOT, three of them show a significant improvement in convexity with respect to a time-shift: NIM, AWI, and GSOT. However, when applied to a realistic case (Marmousi), NIM fails to produce a meaningful V_P estimate. Conversely, for AWI, while difficulties are identified on schematic examples, including multiple arrivals (a situation known to be problematic for correlation and deconvolution approaches), satisfactory results are obtained when applied to the Marmousi case, both within and without the inverse crime settings. GSOT also appears as an interesting strategy, providing satisfactory results in all the tests performed here.

Interestingly, while IE and KROT show less robustness to strong cycle-skipping, the small increase in the valley of attraction they provide is sufficient to enhance the velocity reconstruction in the Marmousi test in inverse crime settings. However, IE fails when it comes to non-inverse crime settings, that is when noise corrupts the data, and amplitude prediction cannot be guaranteed anymore. On the contrary, KROT reveals relatively robust to these settings, probably benefiting from its ability to account for the lateral continuity of events in shot-gather representation and for the robustness of optimal transport based distances with respect to the presence of noise (Engquist et al., 2016).

From the experiment performed in this article, KROT, AWI, and GSOT appear as an interesting alternative to the least-squares distance from the perspective of field data application. In cases where no strong cycle-skipping is expected, KROT should perform well, and this is supported by several field data applications already performed on exploration data (Messud and Sedova, 2019; Sedova et al., 2019; Carotti et al., 2020). The computational cost of KROT is relatively higher than that of AWI and GSOT; however, its ability to account for the lateral coherency of the data in shot-gather panels makes it an appealing strategy. For 3D data cubes, cutting it into 2D slices and summing over the slices is a good compromise. To deal with larger kinematics inaccuracy, AWI and GSOT should be preferred options. AWI has already been successfully applied to field data (Warner and Guasch, 2015; Ravaut et al., 2017; Debens et al., 2017; Roth et al., 2018; Guasch et al., 2019; Warner et al., 2019). We, however, show here that it could suffer from some limitations in the case of complex data containing multiple arrivals. GSOT has been mostly applied to synthetic data by now (He et al., 2019a; Provenzano et al., 2020). Nevertheless, field data applications are ongoing (Pladys et al., 2020; Górszczyk et al., 2020).

CONCLUSION

This article is dedicated to comparing misfit function reformulation for FWI, which aims at mitigating cycle-skipping. The first result drawn is that the link between cycle-skipping and the non-convexity with respect to time-shifts of the least-squares distance is evident from the different tests we provide. However, when no such cycle-skipping occurs (sufficiently accurate initial model), least-squares FWI performs well, even for complex data including multiple phases, mixed phases, noise, and when amplitude prediction cannot be performed accurately (as is the case for field data). Therefore, efficient reformulation of the FWI misfit function should not rely only on a better convexity to time-shifts to replace the least-squares norm advantageously but should also exhibit robustness with respect to these settings, which are always met on field data applications.

ACKNOWLEDGMENTS

This study was partially funded by the SEISCOPE consortium (`http://seiscope2.osug.fr`), sponsored by AKERBP, CGG, CHEVRON, EQUINOR, EXXON-MOBIL, JGI, SHELL, SINOPEC, SISPROBE and TOTAL. This study was granted access to the HPC resources of the Froggy platform of the CIMENT infrastructure (`https://ciment.ujf-grenoble.fr`), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA), the OSUG@2020 labex (reference ANR10 LABX56) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche, and the HPC resources of CINES/IDRIS/TGCC under the allocation 046091 made by GENCI."

REFERENCES

Aghamiry, H., A. Gholami, and S. Operto, 2020, Accurate and efficient wavefield reconstruction in the time domain: Geophysics, **85(2)**, A7–A12.

Bérenger, J.-P., 1994, A perfectly matched layer for absorption of electromagnetic waves: Journal of Computational Physics, **114**, 185–200.

Bertsekas, D. P., and D. Castanon, 1989, The auction algorithm for the transportation problem: Annals of Operations Research, **20**, 67–96.

Bozdağ, E., J. Trampert, and J. Tromp, 2011, Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements: Geophysical Journal International, **185**, 845–870.

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

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

Carotti, D., O. Hermant, S. Masclet, M. Reinier, J. Messud, A. Sedova, and G. Lambaré, 2020, Optimal transport full waveform inversion - applications: Presented at the 82^{th} Annual EAGE Meeting (Amsterdam), European Association of Geoscientists & Engineers.

Choi, Y., and C. Shin, 2008, Frequency-Domain Elastic Full Waveform Inversion Using the New Pseudo-Hessian Matrix: Experience Of Elastic Marmousi 2 Synthetic Data: Bulletin of the Seismological Society of America, **98**, 2402–2415.

Combettes, P. L., and J.-C. Pesquet, 2011, Proximal splitting methods in signal processing, *in* Fixed-Point Algorithms for Inverse Problems in Science and Engineering: Springer New York, volume **49** *of* Springer Optimization and Its Applications, 185–212.

Debens, H. A., F. Mancini, M. Warner, and L. Guasch, 2017, Full-bandwidth adaptive waveform inversion at the reservoir: SEG Technical Program Expanded Abstracts 2017, 1378–1382.

Devaney, A., 1984, Geophysical diffraction tomography: Geoscience and Remote Sensing, IEEE Transactions on, **GE-22**, 3–13.

Donno, D., H. Chauris, and H. Calandra, 2013, Estimating the background velocity model with the normalized integration method: EAGE Technical Program Expanded Abstracts 2013, Tu0704.

Engquist, B., and B. D. Froese, 2014, Application of the Wasserstein metric to seismic signals: Communications in Mathematical Science, **12**, 979–988.

Engquist, B., B. D. Froese, and Y. Yang., 2016, Optimal transport for seismic full waveform inversion: Communications in Mathematical Sciences, **14**, 2309–2330.

Fichtner, A., B. L. N. Kennett, H. Igel, and H. P. Bunge, 2008, Theoretical background for continental- and global-scale full-waveform inversion in the time-frequency domain: Geophysical Journal International, **175**, 665–685.

Gardner, G. F., L. Gardner, and A. Gregory, 1974, Formation velocity and density—the diagnostic basics for stratigraphic traps: Geophysics, **39**, 770–780.

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

Górszczyk, A., L. Métivier, and R. Brossier, 2019, Mitigating the nonlinearity of the crustal scale full waveform inversion through the graph space optimal transport misfit function: AGU Fall Meeting Abstracts, S41A–03.

———, 2020, Relaxing the initial model constraint for crustal-scale full-waveform inversion with graph space optimal transport misfit function: Presented at the Expanded Abstracts, 82^{nd} Annual EAGE Meeting (Amsterdam).

Górszczyk, A., S. Operto, and M. Malinowski, 2017, Toward a robust workflow for deep crustal imaging by FWI of OBS data: The eastern nankai trough revisited: Journal of Geophysical Research: Solid Earth, **122**, 4601–4630.

Guasch, L., M. Warner, and C. Ravaut, 2019, Adaptive waveform inversion: Practice: Geophysics, **84(3)**, R447–R461.

He, W., R. Brossier, and L. Métivier, 2019a, 3D elastic FWI for land seismic data: A graph space OT approach: SEG Technical Program Expanded Abstracts 2019, 1320–1324.

He, W., R. Brossier, L. Métivier, and R.-É. Plessix, 2019b, Land seismic multi-parameter full waveform inversion in elastic VTI media by simultaneously interpreting body waves and surface waves with an optimal transport based objective function: Geophysical Journal International, **219**, 1970–1988.

Huang, G., R. Nammour, and W. W. Symes, 2018, Source-independent extended waveform inversion based on space-time source extension: Frequency-domain implementation: Geophysics, **83**, R449–R461.

Lambaré, G., 2008, Stereotomography: Geophysics, **73(5)**, VE25–VE34.

Li, Y., R. Brossier, and L. Métivier, 2019, Joint FWI for imaging deep structures: A graph-space OT approach: SEG Technical Program Expanded Abstracts 2019, 1290–1294.

Luo, S., and P. Sava, 2011, A deconvolution-based objective function for wave-equation inversion: SEG Technical Program Expanded Abstracts, **30**, 2788–2792.

Luo, Y., and G. T. Schuster, 1991, Wave-equation traveltime inversion: Geophysics, **56**, 645–653.

Marple, L., 1999, Computing the discrete-time" analytic" signal via fft: IEEE Transactions on signal processing, **47**, 2600–2603.

Martin, G. S., R. Wiley, and K. J. Marfurt, 2006, Marmousi2: An elastic upgrade for Marmousi: The Leading Edge, **25**, 156–166.

Messud, J., and A. Sedova, 2019, Multidimensional optimal transport for 3d FWI: Demonstration on field data: Presented at the Expanded Abstracts, 81^{th} Annual EAGE Meeting (London).

Métivier, L., A. Allain, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2018, Optimal transport for mitigating cycle skipping in full waveform inversion: a graph space transform approach: Geophysics, **83**, R515–R540.

Métivier, L., R. Brossier, Q. Mérigot, and E. Oudet, 2019, A graph space optimal transport distance as a generalization of L^p distances: application to a seismic imaging inverse problem: Inverse Problems, **35**, 085001.

Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016a, Increasing the robustness and applicability of full waveform inversion: an optimal transport distance strategy: The Leading Edge, **35**, 1060–1067.

———, 2016b, Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion: Geophysical Journal International, **205**, 345–377.

———, 2016c, An optimal transport approach for seismic tomography: Application to 3D full waveform inversion: Inverse Problems, **32**, 115008.

Nocedal, J., and S. J. Wright, 2006, Numerical optimization, 2nd ed.: Springer.

Pladys, A., R. Brossier, M. Irnaka, N. Kamath, and L. Métivier, 2019, Assessment of optimal transport based FWI: 3d OBC valhall case study: SEG Technical Program Expanded Abstracts 2019, 1295–1299.

Pladys, A., R. Brossier, and L. Métivier, 2020, Graph space optimal transport based FWI: 3D OBC valhall case study: Presented at the SEG Technical Program Expanded Abstracts 2020.

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 C. Perkins, 2010, Full waveform inversion of a deep water ocean bottom seismometer dataset: First Break, **28**, 71–78.

Poncet, R., J. Messud, M. Bader, G. Lambaré, G. Viguier, and C. Hidalgo, 2018, Fwi with optimal transport: a 3D implementation and an application on a field dataset: Presented at the Expanded Abstracts, 80^{th} Annual EAGE Meeting (Copenhagen).

Pratt, R. G., 1999, Seismic waveform inversion in the frequency domain, part I: theory and verification in a physical scale model: Geophysics, **64**, 888–901.

Pratt, R. G., and R. M. Shipp, 1999, Seismic waveform inversion in the frequency domain, part II: Fault delineation in sediments using crosshole data: Geophysics, **64**, 902–914.

Provenzano, G., R. Brossier, and L. Métivier, 2020, Joint FWI of diving and reflected waves using a graph space optimal transport distance: synthetic tests on limited-offset surface seismic data: Presented at the SEG Technical Program Expanded Abstracts 2020.

Qiu, L., J. Ramos-Martìnez, A. Valenciano, Y. Yang, and B. Engquist, 2017, Full-waveform inversion with an exponentially encoded optimal-transport norm: SEG Technical Program Expanded Abstracts 2017, 1286–1290.

Raknes, E. B., B. Arntsen, and W. Weibull, 2015, Three-dimensional elastic full waveform inversion using seismic data from the sleipner area: Geophysical Jounal International, **202**, 1877–1894.

Ravaut, C., F. Maao, J. Mispel, A. Osen, M. Warner, L. Guasch, and T. Nangoo, 2017, Imaging beneath a gas cloud in the north sea without conventional tomography: EAGE, 79th Conference and Exhibition, Expanded abstracts, We A3 04.

Roth, T., T. Nangoo, N. Shah, M. Riede, C. Henke, and M. Warner, 2018, Improving seismic image with high resolution velocity model from awi starting with 1d initial model - case study barents sea.

Sedova, A., J. Messud, H. Prigent, G. Royle, and G. Lambaré, 2019, Acoustic land full waveform inversion on a broadband land dataset: the impact of optimal transport: Presented at the Expanded Abstracts, 81^{th} Annual EAGE Meeting (London).

Shen, X., L. Jiang, J. Dellinger, A. Brenders, C. Kumar, M. James, J. Etgen, D. Meaux, R. Walters, and N. Abdullayev, 2018, High-resolution full-waveform inversion for structural imaging in exploration: SEG Technical Program Expanded Abstracts 2018, 1098–1102.

Shipp, R. M., and S. C. Singh, 2002, Two-dimensional full wavefield inversion of wide-aperture marine seismic streamer data: Geophysical Journal International, **151**, 325–344.

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

Symes, W. W., 2008, Migration velocity analysis and waveform inversion: Geophysical Prospecting, **56**, 765–790.

van Leeuwen, T., and F. J. Herrmann, 2013, Mitigating local minima in full-waveform inversion by expanding the search space: Geophysical Journal International, **195(1)**, 661–667.

van Leeuwen, T., and W. A. Mulder, 2010, A correlation-based misfit criterion for wave-equation traveltime tomography: Geophysical Journal International, **182**, 1383–1394.

Virieux, J., A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou, 2017, An introduction to Full Waveform Inversion, *in* Encyclopedia of Exploration Geophysics: Society of Exploration Geophysics, R1–1–R1–40.

Virieux, J., and S. Operto, 2009, An overview of full waveform inversion in exploration geophysics: Geophysics, **74**, WCC1–WCC26.

Wang, C., D. Yingst, P. Farmer, and J. Leveille, 2016, *in* Full-waveform inversion with the reconstructed wavefield method: 1237–1241.

Wang, Y., and Y. Rao, 2009, Reflection seismic waveform tomography: Journal of Geophysical Research, **114**, 1978–2012.

Warner, M., and L. Guasch, 2015, Robust adaptive waveform inversion: SEG Technical Program Expanded Abstracts 2015, 1059–1063.

———, 2016, Adaptive waveform inversion: Theory: Geophysics, **81**, R429–R445.

Warner, M., T. Nangoo, A. Pavlov, and C. Hidalgo, 2019, Extending the velocity resolution of waveform inversion below the diving waves using awi: **2019**, 1–5.

Wu, R.-S., J. Luo, and B. Wu, 2014, Seismic envelope inversion and modulation signal model: Geophysics, **79**, WA13–WA24.

Yang, P., R. Brossier, L. Métivier, J. Virieux, and W. Zhou, 2018a, A Time-Domain Preconditioned Truncated Newton Approach to Multiparameter Visco-acoustic Full Waveform Inversion: SIAM Journal on Scientific Computing, **40**, B1101–B1130.

Yang, Y., and B. Engquist, 2018, Analysis of optimal transport and related misfit functions in full-waveform inversion: GEOPHYSICS, **83**, A7–A12.

Yang, Y., B. Engquist, J. Sun, and B. F. Hamfeldt, 2018b, Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion: Geophysics, **83**, R43–R62.