GEOPHYSICS, VOL. 83, NO. 5 (SEPTEMBER-OCTOBER 2018); P. R515–R540, 18 FIGS.
10.1190/GEO2017-0807.1

Optimal transport for mitigating cycle skipping in full-waveform inversion: A graph-space transform approach

Ludovic Métivier[1], Aude Allain[2], Romain Brossier[3], Quentin Mérigot[4], Edouard Oudet[2], and Jean Virieux[3]

ABSTRACT

Optimal transport distance has been recently promoted as a tool to measure the discrepancy between observed and seismic data within the full-waveform-inversion strategy. This high-resolution seismic imaging method, based on a data-fitting procedure, suffers from the nonconvexity of the standard least-squares discrepancy measure, an issue commonly referred to as cycle skipping. The convexity of the optimal transport distance with respect to time shifts makes it a good candidate to provide a more convex misfit function. However, the optimal transport distance is defined only for the comparison of positive functions, while seismic data are oscillatory. A review of the different attempts proposed in the literature to overcome this difficulty is proposed. Their limitations are illustrated: Basically, the proposed strategies are either not applicable to real data, or they lose the convexity property of optimal transport. On this basis, we introduce a novel strategy based on the interpretation of the seismic data in the graph space. Each individual trace is considered, after discretization, as a set of Dirac points in a 2D space, where the amplitude becomes a geometric attribute of the data. This ensures the positivity of the data, while preserving the geometry of the signal. The differentiability of the misfit function is obtained by approximating the Dirac distributions through 2D Gaussian functions. The interest of this approach is illustrated numerically by computing misfit-function maps in schematic examples before moving to more realistic synthetic full-waveform exercises, including the Marmousi model. The better convexity of the graph-based optimal transport distance is shown. On the Marmousi model, starting from a 1D linearly increasing initial model, with data without low frequencies (no energy less than 3 Hz), a meaningful estimation of the P-wave velocity model is recovered, outperforming previously proposed optimal-transport-based misfit functions.

INTRODUCTION

Full-waveform inversion (FWI) is a high-resolution seismic imaging technique. In its conventional formulation, it is based on the least-squares minimization of the misfit between observed data and data calculated through the solution of wave-propagation equations. Compared with standard seismic imaging techniques, based on the comparison of observables extracted from the data, such as arrival times of seismic events in tomography, FWI aims at interpreting the whole waveform. High-resolution quantitative estimations of subsurface mechanical properties, such as P- and S-wave velocities, density, attenuation, and anisotropy parameters, are expected in the limit of half the shortest propagated wavelength, following the standard diffraction tomography analysis (Devaney, 1982, 1984; Wu and Toksöz, 1987; Sirgue, 2003).

FWI was introduced in the 1980s by Lailly (1983) and Tarantola (1984). The development of high-performance-computing facilities and broadband wide-azimuth seismic-acquisition devices have made possible the successful application of this method to real data, in 2D and 3D geometries, at exploration scale (Operto et al., 2006, 2015; Plessix and Perkins, 2010; Sirgue et al., 2010; Warner et al., 2013; Vigh et al., 2014), and also for global and regional imaging (Fichtner et al., 2010; Tape et al., 2010; Peter et al., 2011; Zhu et al., 2012; Borisov et al., 2015; Bozdağ et al., 2016). Virieux et al.

Manuscript received by the Editor 11 December 2017; revised manuscript received 7 March 2018; published ahead of production 15 June 2018; published online 05 September 2018.
[1]Université Grenoble Alpes, LJK, CNRS, France. E-mail: @ludovic.metivier@ univ-grenoble-alpes.fr.
[2]Université Grenoble Alpes, LJK, France. E-mail: aude.allain@univ-grenoble-alpes.fr; edouard.oudet@imag.fr.
[3]Université Grenoble Alpes, ISTerre, France. E-mail: romain.brossier@univ-grenoble-alpes.fr; jean.virieux@univ-grenoble-alpes.fr.
[4]Université Paris-Sud, LMO, CNRS, France. E-mail: quentin.merigot@math.u-psud.fr.

(2017) give a recent overview of methodological aspects of FWI and its various applications.

The targeted broadband reconstruction of P- and S-wave velocities through FWI requires the ability to recover not only their high-wavenumber content but also their low- to intermediate-wavenumber content. Large- to intermediate-scale perturbations of these parameters mainly influence the kinematics of the wave propagation, affecting essentially the traveltimes of waves, i.e., shifting in time seismic events (Jannane et al., 1989). However, the conventional least-squares function used to measure pointwisely the discrepancy between observed and synthetic data is not convex with respect to these time shifts. Considering, for instance, two Ricker signals, the least-squares misfit function between these two signals, depending on the time shift, exhibits a single global minimum and two local minima. This phenomenon is commonly known as cycle skipping or phase ambiguity (Virieux and Operto, 2009).

Should FWI rely on global optimization or semiglobal optimization schemes, such as Monte Carlo (Jin and Madariaga, 1994; Sambridge and Mosegaard, 2002), genetic algorithms (Sen and Stoffa, 1992; Jin and Madariaga, 1993; Aleardi and Mazzotti, 2016), simulating annealing (Kirkpatrick et al., 1983), or the covariance matrix adaptation evolution strategy (Hansen, 2006), a dense enough sampling of the misfit function would allow for detecting its global minimum. However, real-data applications often involve the reconstruction of hundreds of thousands of discrete parameters in two dimensions, and hundreds of millions of discrete parameters in three dimensions, which makes these strategies beyond current and forthcoming computational capabilities (exascale machines). Thus, FWI has to rely on local optimization schemes: A starting model is updated following the descent directions. In this framework, the nonconvexity of the misfit function is a critical issue. Depending on the quality of the starting model, local optimization strategies might converge toward local minima, which may not be geologically meaningful.

To prevent this issue using a least-squares distance, the initial velocity model should predict the kinematics of main arrivals within half-a-period range. The standard workflow for practical applications of FWI thus consists first in designing an accurate starting velocity model. This is performed mainly through traveltime tomography strategies (for a review, see Nolet, 2008). At the exploration scale, because of the limited offset range, reflection phases mainly drive the velocity model building and provide information at depth (Yilmaz, 1993), although combining different traveltimes improves the sampling (Zhang et al., 1998; Huang and Bellefleur, 2012). Recent high-resolution tomography techniques known as stereotomography (or slope tomography) can be used to increase the resolution of this initial estimation (Billette and Lambaré, 1998; Lambaré, 2008; Prieux et al., 2013; Tavakoli et al., 2017). Starting from these velocity models, hierarchical FWI approaches are designed, decomposing the data from low to high frequencies (Bunks et al., 1995; Pratt, 1999), and possibly from short offset/short time windows to large offset/large time windows, following layer-stripping approaches (Shipp and Singh, 2002; Brossier et al., 2009; Wang and Rao, 2009). Each subset of data is interpreted through FWI, and the resulting velocity model serves as an initial model for the interpretation of the next subset of data. The reason for these hierarchical approaches is that, for low-frequency content and/or short offset/short time windows, the number of propagated wavelength is smaller, reducing the cycle-skipping ambiguity. This strategy works in practice; however, it requires access to low-frequency information rarely available for exploration-scale seismic data. In addition, a careful analysis of the data is required to select the part to be inverted at each level of the hierarchical strategy, which might appear in contradiction with the fundamental purpose of FWI to avoid any prior interpretation, potentially misleading.

Recognizing the inadequacy of the least-squares distance to correctly interpret the time shifts, modifications of the misfit function have been proposed. Crosscorrelation (Tape et al., 2010), as well as instantaneous envelope, coupled with time-windowing techniques (Fichtner et al., 2008; Bozdağ et al., 2011), are popular choices in global- and regional-scale FWI applications. At the exploration scale, the crosscorrelation technique was proposed by Luo and Schuster (1991) to automatically compute the time shift between seismic traces. The FWI problem is then posed either as the minimization of the time shifts, or as the penalization of the time shifts away from zero, within a certain time window (van Leeuwen and Mulder, 2010). In both cases, capturing time shifts through crosscorrelation relies on the assumption of well-separated arrivals, which requires a careful time-window selection; this can be quite challenging for complex multiarrival real data. Another approach consists of computing matching filters for each trace through a linear deconvolution (Luo and Sava, 2011; Warner and Guasch, 2016). The misfit function consists of the penalization of the filter away from a delta function at zero lag. All these approaches promote the direct interpretation of the kinematic attribute of the data, instead of the amplitude. Although these methods enlarge the basin of attraction of the global minimum, relaxing in some way the dependency on the accuracy of the initial model, they can still suffer from cycle skipping, and also from a loss of resolution. The design of adequate penalization functions might also be problematic, due to the high sensitivity of the resulting misfit function to this design (Pladys et al., 2017).

Another class of methods, referred to as extended-domain techniques, are based on a scale separation of the model: The velocity is decomposed as a sum of a smooth part containing the low-wavenumber information (the background velocity) and a reflectivity part containing the high-wavenumber content. Reflectivity hypercubes are constructed through the introduction of one or several redundant parameters in the imaging condition, such as the source position or surface offset (Symes and Kern, 1994), subsurface offset (Shen et al., 2003), time lag (Sava and Fomel, 2006), or common illumination angle (Biondi and Symes, 2004). The misfit function is defined as the minimization of variations of the hypercube along these artificial dimensions: The correct background velocity should yield a unique image with zero contribution along these added dimensions (for a review, see Symes, 2008). Although quite appealing, these strategies suffer from a rather high computational cost associated with the repeated construction of high-resolution images in this extended space.

In this paper, we focus on alternative strategies, based on misfit-function modifications linked to warping techniques. The leading idea is a comparison of data using warping tools, which aims at computing mappings between synthetic and observed data. A first instance is the dynamic warping strategy, introduced by Hale (2013), where local time shifts between synthetic and observed traces are computed based on a spatial lateral coherence. Ma and Hale (2013) propose to minimize these time shifts using a standard least-squares misfit function. Compared with crosscorrelation tech-

niques based only on time information, the estimation of the time shift is more robust, especially when their variations are fast. Similar ideas are presented by Baek et al. (2014) and Zhu (2017), where a warping tool (called a registration tool) is used to modify the synthetic data to get sufficiently close to the observed data to avoid falling within a local minimum. Although these warping techniques are based on the assumption that the observed data can be obtained from locally stretching and squeezing in time the synthetic data, a more systematic way of performing this warping is provided by the optimal transport theory. The mapping between the two functions to be compared is the one that minimizes the sum of the elementary displacements required to map the two functions. When considering time signals, this should yield a convex function with respect to time shifts: The optimal transport effort required to map the data should increase as the time shift increases (Engquist and Froese, 2014). Optimal transport could also consider transport along spatial directions (i.e., receiver locations in a shot-gather representation), although we shall mainly consider in this paper the time axis for better understanding how optimal transport is working.

Although the idea is appealing, the application of optimal transport to FWI is not straightforward, mainly because of the nonpositivity of the seismic data: optimal transport is a mathematical theory developed in the frame of probability distributions; hence, it is applicable to positive functions only. Qiu et al. (2017) and Yang and Engquist (2017) propose several preprocessing of the data to bring the problem back to the comparison of positive quantities. An alternative formulation, based on a dual form of the optimal transport problem, has been proposed by Métivier et al. (2016a, 2016b, 2016c). The case studies presented in these papers show promising results. In particular, the dependence on the accuracy of the initial velocity model is relaxed compared with the conventional least-squares misfit function. However, both strategies (data preprocessing and the dual optimal transport problem) have important drawbacks. Some of the proposed preprocessings are not fully adapted to real-data applications, whereas the others, as well as the dual optimal transport strategy, lose the convexity with respect to time shifts.

From our perspective, this latter feature is a severe drawback because the main interest for using optimal transport is lost. The purpose of this study is to analyze, explain, and illustrate these propositions for the application of optimal transport to oscillatory seismic signals: The outcome is a possible remedy to the convexity loss issue. The related strategy consists of comparing not the data itself through optimal transport, but the graph of the data, following ideas proposed by Thorpe et al. (2016). Usually, a seismic trace is considered as a 1D function of time, with a given sampling rate. After discretization, a seismic trace is thus a 1D vector of values d_n regularly sampled in time: Each value is associated with the amplitude of the seismic signal at a given time step t_n defining a point (t_n, d_n). Consequently, the 1D seismic trace can be considered as a cloud of points in a 2D space, namely, the graph space, where one dimension corresponds to the time axis and the other corresponds to the amplitude axis. Computing the optimal transport distance in the graph space thus amounts to computing the optimal transport distance between observed and synthetic point clouds, in a fully combinatory search. This has the following three advantages:

1) Mathematically, the distance between the synthetic and the observed trace amounts to the comparison of two sets containing the same number of Dirac distributions (assuming that the time-sampling rate is the same for each trace, which is easy to satisfy in practice). The optimal transport problem is thus well-posed because we compare the same number of positive quantities (Dirac values).
2) The geometry of the signal is preserved: No data preprocessing is used to modify it; therefore, the convexity of the misfit function with respect to time shifts is ensured. In other words, the amplitude of the signal becomes a geometric attribute of the transformed data and no information should be lost in this process.
3) No prior assumption on the number of seismic events/arrivals is required: The number of seismic events in the calculated data and the observed data might differ. This is not the case for the more standard warping strategies mentioned above.

Of course, this comes at a price: For each 1D seismic trace, a 2D transport problem has to be solved. If the method is applied to full seismograms, a 3D transport problem would have to be solved (even a 4D problem if the whole data cube is taken into account for a 3D imaging problem). Solving these multidimensional optimal transport problems is computationally expensive and remains a challenge.

In this study, we present promising results in a 2D time-domain acoustic FWI configuration by mapping 1D seismic traces through optimal transport based on the graph space (OT-GS). We aim to illustrate the potentiality of this approach through the investigation of synthetic examples of increasing complexity. In particular, we show that we are able to improve the convexity of the misfit function, with a preserved sensitivity to large time shifts, which makes this approach suitable to mitigate cycle-skipping problems. On the Marmousi 2 experiment, starting from a 1D model, using data missing low-frequency content, we are able to converge toward a meaningful P-wave velocity estimation, whereas the optimal transport strategy based on the Kantorovich-Rubinstein (KR) distance we presented in Métivier et al. (2016a, 2016b, 2016c) does not succeed.

The numerical implementation that is used in this study relies on the solution, for each trace of the data, of a 2D optimal transport problem in the graph space, using the numerical strategy we developed in Métivier et al. (2016a, 2016b, 2016c). It implies an important computational cost increase, which can be mitigated through a decimation of the signal. On the Marmousi experiment, we end up with a computational cost increase of a factor 5.8 for a gradient computation compared with a conventional least-squares approach. However, this factor should be reduced through the use of other optimal transport numerical solvers, such as the ones based on the entropic regularization (Benamou et al., 2015) or the so-called multiscale approach (Mérigot, 2011). This will be the matter of further studies. In this paper, we illustrate the advantageous properties of the graph-transform approach, providing motivation for actively searching how to overcome the cost increase.

The structure of the study is as follows. First, after defining optimal transport, we give a review of the current strategies setup of optimal transport on oscillatory seismic data. In particular, we illustrate why these strategies may not be easily adapted to real-data applications, or end up losing the main interesting property of optimal transport: the sensitivity with respect to time shifts. Next, we introduce the graph-space strategy, and we present why it is fully compatible with the FWI framework. Numerically illustrative simple experiments are then presented to emphasize limits and interests of this approach. These results are further analyzed and discussed for a better understanding of how to account for the amplitude information and how this graph approach is sensitive to the noise

level. Finally, conclusions and perspectives for future investigation, such as reducing the computational complexity, are drawn.

APPLYING OPTIMAL TRANSPORT TO FWI: A REVIEW OF EXISTING STRATEGIES

The optimal transport definition is summarized, and different strategies set up to adapt it to the comparison of oscillatory data are reviewed and compared.

Generalities on optimal transport

Optimal transport, despite its introduction more than two centuries ago by the mathematician G. Monge (Monge, 1781), is a very active field of research in mathematics, as demonstrated by the number of recently published books dedicated to this topic (Ambrosio et al., 2008; Villani, 2008; Santambrogio, 2015). Beyond theoretical results in mathematical analysis that have been obtained using optimal transport tools, the number of numerical applications in image processing, and more recently in seismic imaging, is rapidly increasing. Here, we give a quick overview of the definition of optimal transport based on Kantorovich (1942) relaxation.

Let $f(x)$ and $g(x)$ be two positive functions defined on $X \subset \mathbb{R}^d$, where $d = 1, 2, 3, \ldots$ is the dimension of X, satisfying the mass conservation assumption

$$\int_X f(x)dx = \int_X g(x)dx. \tag{1}$$

The p-Wasserstein distance (i.e., the optimal transport distance) between f and g is defined by

$$W_p(f,g) = \left(\min_{\gamma \in \Pi(f,g)} \int_{X\times X} \gamma(x,x') \|x - x'\|^p dx dx' \right)^{1/p}, \tag{2}$$

where $\Pi(f,g)$ denotes the set of transport plans $\gamma(x,x')$ such that

$$\Pi(f,g) = \left\{ \gamma(x,x'), \int_X \gamma(x,x')dx' = f(x), \int_X \gamma(x,x')dx = g(x') \right\}, \tag{3}$$

and p is a given integer (often chosen as $p = 1$ or 2). Intuitively, f and g can be seen as two distributions of masses, where each entry $f(x)$ (respectively $g(x)$) indicates how much mass is present at position x. The set $\Pi(f,g)$ contains all the transport plans $\gamma(x,x')$ which tell, for each couple (x,x'), how much quantity $f(x)$ at position x should be moved to the position x' to entirely map the distribution of mass f onto the distribution of mass g.

There are infinite such transport plans γ. Solving the linear programming problem in equation 2 amounts to finding the one that minimizes the "effort" required to map f onto g. This effort corresponds to a cost function, which is equal, for each couple (x,x'), to the amount of mass, which has to be transported $\gamma(x,x')$, multiplied by a power of the distance $\|x - x'\|^p$ along which this mass is transported. The distance $\|.\|$, called the ground distance, is usually based on the Euclidean norm $\|.\|$ of $\mathbb{R}^d$; however, any suitable distance on $\mathbb{R}^d$ can be used in practice. Following this interpretation, one can see that if the distribution f is being shifted from the distribution g by a local shift Δx such that

$$f(x) = g(x - \Delta x), \tag{4}$$

one can expect that the optimal transportation cost depends monotonically on the shift Δx. Increasing Δx results in an increase of this optimal transport effort, whereas decreasing Δx decreases it. This monotonicity with respect to the shift Δx is the main reason that the optimal transport distance has attracted interest in the image-processing community because it confers the ability to better recognize similar patterns between images. In spite of the FWI computational cost, it has attracted attention in the FWI community because the convexity to time/space shifts can be seen as a proxy of the convexity with respect to wave velocities (Jannane et al., 1989).

Optimal transport applied to seismic data

We now consider the application of optimal transport to seismic data. In the simplest case, the seismic data are a seismic trace, i.e., an explicit function of time. The space X is thus the time axis in this context. The mass conservation assumption is satisfied by seismic data. Indeed, for a given seismic trace $d(t)$, the total amount of mass translates into an integration over the recording time window T

$$\int_0^T d(t)dt. \tag{5}$$

This corresponds to the zero-frequency content of the signal, which is equal to zero in practice. However, seismic data are oscillatory: The positivity assumption breaks down. This prevents from the direct application of optimal transport to seismic data.

Here, we review and compare the strategies that have been set up to adapt optimal transport to the comparison of seismic data. To this purpose, we focus on the simple Ricker comparison experiment proposed by Engquist and Froese (2014). Two shifted in time Ricker signals, denoted by $d_{\text{obs}}(t)$ (playing the role of observed data) and $d_{\text{cal}}(t;s)$ (playing the role of calculated data) are introduced, such that

$$d_{\text{cal}}(t;s) = d_{\text{obs}}(t - s), \tag{6}$$

where the quantity $s \in \mathbb{R}$ is the time shift.

Each strategy amounts to extract positive quantities from d_{obs} and d_{cal}, or to transform d_{obs} and d_{cal} into positive quantities, which are further compared using an optimal-transport-based misfit function we introduce as $C(s)$. We discuss the advantages and drawbacks of considering these strategies: The corresponding misfit function $C(s)$ is analyzed for better understanding the impact on the convexity with respect to the time shift s. The misfit functions are based on the two-Wasserstein distance W_2. The algorithm used to compute this distance in the 1D case is described in Appendix A: In this simple case, an analytical expression is available (Villani, 2003; Engquist et al., 2016).

The first strategy has been proposed by Engquist and Froese (2014): separately comparing the positive and negative parts of the data using an optimal transport distance (Figure 1a). We introduce the positive and negative part operators $.^+$ and $.^-$ as

$$d^+_{\text{obs}} = \frac{|d_{\text{obs}}| + d_{\text{obs}}}{2}, \qquad d^-_{\text{obs}} = \frac{|d_{\text{obs}}| - d_{\text{obs}}}{2}. \tag{7}$$

The corresponding misfit function is

$$C(s) = W_2(d^+_{\text{obs}}, d^+_{\text{cal}}) + W_2(d^-_{\text{obs}}, d^-_{\text{cal}}). \tag{8}$$

This strategy ensures the convexity of $C(s)$ (Figure 2, dotted black line). However, it faces the following difficulties:

1) The mass conservation is no more satisfied. For the Ricker experiment presented here, this is the case but this cannot be guaranteed for real-data applications: Nothing guarantees that the positive part (respectively the negative part) of the observed data has the same total mass as the positive part (respectively the negative part) of the calculated data, especially if the number of events in confronted traces is not the same or in case of amplitude mismatch.
2) Separating positive and negative parts is not a differentiable operation, which would lead to the definition of a nondifferentiable function, making the use of local optimization techniques to find its minimum impossible.
3) For real-data applications, a phase rotation in the source estimation would lead to misinterpreting positive values of the observed data as negative values and vice versa.

For the next possibilities presented here, the misfit function $C(s)$ can be written as

$$C(s) = W_2(\tilde{d}_{\text{obs}}, \tilde{d}_{\text{cal}}(s)), \tag{9}$$

where $\tilde{d}_{\text{obs}}$ and $\tilde{d}_{\text{cal}}(s)$ are obtained from d_{obs} and d_{cal} from various transformations. Yang and Engquist (2017) propose to apply an affine scaling to the data, such that

$$\begin{aligned} \tilde{d}_{\text{obs}}(t) &= a d_{\text{obs}}(t) + b, \\ \tilde{d}_{\text{cal}}(t;s) &= a d_{\text{cal}}(t;s) + b, \end{aligned} \tag{10}$$

with $b > 0$ chosen sufficiently large so that $\tilde{d}_{\text{obs}}(t)$ and $\tilde{d}_{\text{cal}}(t)$ are positive. This is illustrated in Figure 1b. This strategy is straightforward and ensures the mass conservation. However, it affects the convexity of $C(s)$ with respect to the time shift s, as can be seen in Figure 2 (solid red line). The shape of the misfit function resembles the one of the least-squares misfit function, with a wider valley of attraction and with two flat local minima aside the global minimum. The intuitive interpretation of the optimal transport distance given previously can help us to understand why the convexity is lost. Indeed, adding a constant b to the data amounts to artificially creating mass at each point. Mapping the two signals through an optimal transport strategy thus can be performed through local mass displacement instead of having translational exchanges between the initial and target signals along the time axis. This implies a loss of sensitivity to the time shift and therefore the loss of convexity of the misfit function $C(s)$.

Another possibility is proposed by Qiu et al. (2017), who suggest transforming the data using an exponential function, as illustrated in Figure 1c. In this case, we have

$$\begin{aligned} \tilde{d}_{\text{obs}}(t) &= \alpha \exp \alpha d_{\text{obs}}(t), \\ \tilde{d}_{\text{cal}}(t;s) &= \alpha \exp \alpha d_{\text{cal}}(t;s), \end{aligned} \tag{11}$$

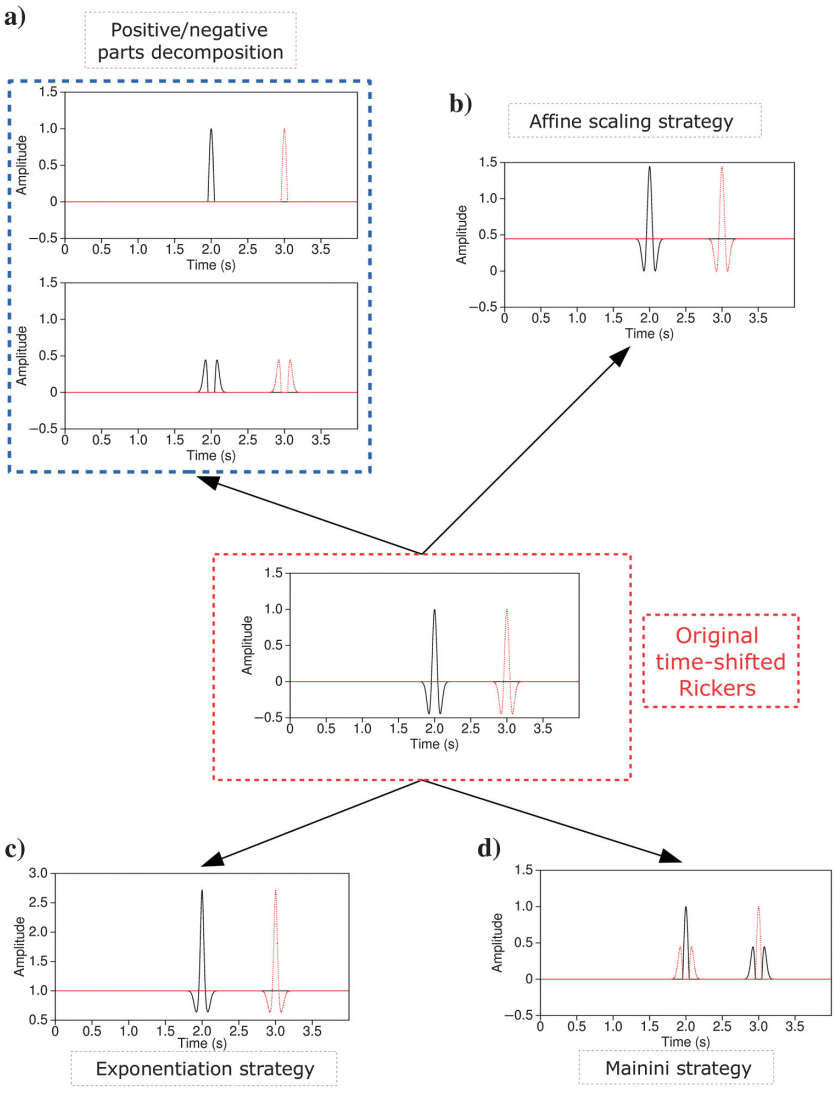

Figure 1. Illustration of conventional strategies to mitigate the nonpositivity issue in optimal transport. The Ricker in the solid black line should be interpreted as the observed signal, whereas the Ricker in the dotted red line should correspond to the calculated signal to which it is compared. (a) The strategy consists in interpreting separately the positive and negative parts of the two Ricker signals. (b) The strategy consists of adding a constant to the two signals rendering them positive. (c) The strategy consists of considering the exponent of the two signals. (d) The strategy uses a decomposition/recomposition of the two signals based on their positive and negative parts following the approach proposed by Mainini (2012).

with a positive constant α. The shape of the corresponding misfit function $C(s)$ is presented in Figure 2, with $\alpha = 1$ (the solid blue line). A single minimum is recovered, and the sensitivity to large time shifts is maintained. However, this strategy faces some difficulties:

1) The exponential transform does not provide the same influence to positive values (high positive values) with respect to negative values (small positive values). This issue could be mitigated through the use of a symmetric transform of the form $\tilde{d}(t) = \alpha \exp \alpha d(t) + \alpha \exp -\alpha d(t)$.
2) The mass conservation assumption is not satisfied. Nothing guarantees that the integral of the exponential of two functions is the same when the integral of the two functions is the same.
3) The exponential transform modifies the amplitude ratio between large amplitude events (direct and diving waves) and smaller amplitude events (reflections for instance).

The latter could result in neglecting reflected events, for instance a severe loss of information, or to bias the phase information.

Another, more systematic approach, originally proposed in the mathematical community by Ambrosio et al. (2011) and Mainini (2012), relies on a particular decomposition/recomposition of the data. This amounts to consider a new observed data and a new synthetic data provided by the following expressions:

$$\tilde{d}_{\text{obs}}(t) = d^{+}_{\text{obs}}(t) + d^{-}_{\text{cal}}(t), \tilde{d}_{\text{cal}}(t) = d^{+}_{\text{cal}}(t) + d^{-}_{\text{obs}}(t), \quad (12)$$

as illustrated in Figure 1d. By construction, $\tilde{d}_{\text{obs}}$ and $\tilde{d}_{\text{cal}}$ are positive and they satisfy the mass conservation assumption. However, as can be seen in Figure 3 (the solid purple line), this modification of the data results in a loss of convexity of the corresponding misfit function $C(s)$. A single minimum at zero time shift $s = 0$ is recovered, with a valley of attraction as wide as the one obtained using the affine scaling strategy. However, the sensitivity to larger time shifts quickly reaches a plateau. This loss of sensitivity can be understood from the data decomposition illustrated in Figure 1d. One can see that this strategy actually amounts to compute the optimal transport

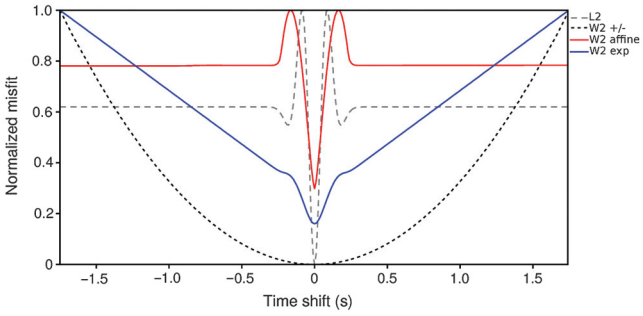

Figure 2. Misfit function between the observed and calculated Ricker signals depending on the time shift, following conventional strategies to mitigate nonpositivity. All of the misfit functions are computed using the two-Wasserstein distance for these examples. The convex curve in the dotted black line corresponds to the positive/negative part separation. The solid red line corresponds to the affine scaling strategy (addition of a constant). The solid blue line corresponds to the exponentiation strategy with the two-Wasserstein distance. As a reference, the curve in the dotted gray line corresponds to the L^2 misfit function.

distance between the positive and the negative parts of $d_{\text{obs}} - d_{\text{cal}}$, the standard L^2 residuals. As soon as these residuals do not overlap in time, the effort to map their positive part to their negative part does not depend on the time shift. Another drawback of this strategy is that it is based on a decomposition of the data in its positive and negative parts: Again, this would lead to the definition of a nondifferentiable misfit function; therefore, it would be impossible to minimize through local optimization techniques.

Finally, in previous studies (Métivier et al., 2016a, 2016b, 2016c), we have developed a method based on the dual of the one-Wasserstein distance, seemingly uncorrelated with the previously mentioned strategies. The motivation for focusing on this distance was twofold:

1) The dual one-Wasserstein distance can be computed for data with negative values. Indeed, the one-Wasserstein distance has a specific dual form, which is a particular instance of the KR norm, defined on the space of signed measures (Bogachev, 2007; Lellmann et al., 2014). In the following, we refer to this distance as the KR distance.
2) We were able to design a fast algorithm for the computation of the KR distance not only for trace-by-trace comparison but also for 2D and 3D full seismograms. Current implementations of the two-Wasserstein metric in the framework of FWI do not seem to provide this possibility.

We can show that this strategy is equivalent to the strategy proposed by Mainini (2012) in the case of the one-Wasserstein distance. A demonstration of this proposition is presented in Appendix B. The main difference in this case is that the Mainini decomposition becomes implicit. Therefore, no explicit decomposition and recomposition of the data using their positive and negative parts is required, yielding a differentiable misfit function, usable in the FWI framework. The shape of the function $C(s)$ using the KR misfit distance is quite similar to the one obtained through the two-Wasserstein distance with the Mainini decomposition, as can be seen in Figure 3 (solid black line).

In light of this comparison, the KR distance appears to us as, somehow, the most suitable for FWI applications, among the different propositions reviewed here. The corresponding misfit function is differentiable; it exhibits a single minimum for the Ricker exam-

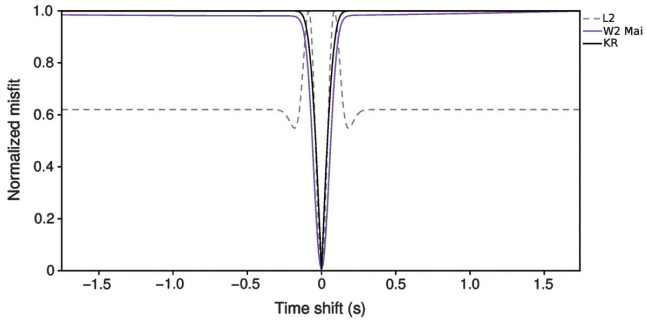

Figure 3. Misfit function between observed and calculated Ricker signals depending on the time shift. The solid purple line corresponds to the Mainini strategy with the two-Wasserstein distance. The solid black line corresponds to the Mainini strategy using the one-Wasserstein distance, i.e., the KR distance between observed and synthetic Ricker. As a reference, the curve in the dotted gray line corresponds to the L^2 misfit function.

ple (this is not the case when using the affine scaling strategy), and the relative amplitudes of the different arrival are not distorted through a nonlinear transformation of the data. The possibility to compare simultaneously entire 2D and 3D shot gathers following the numerical strategy that we introduced in Métivier et al. (2016b, 2016c) is also an important advantage.

Despite these interesting properties, this strategy still loses the convexity with respect to time shifts. In this study, we try to overcome this issue by another transformation of the data, prior to data comparison. This transformation is based on a very simple idea: Instead of applying transport to seismic data as functions of time, the graph of seismic traces could be used.

A GRAPH-TRANSFORMED-BASED OPTIMAL TRANSPORT DISTANCE FOR FWI

The graph formulation is described as well as its adaptation to the FWI approach and its numerical implementation.

Principle: From a discrete signal to a sum of Dirac distributions

We consider a seismic trace $d(t)$, defined as a function on the time interval $[0, T]$. After discretization, the seismic trace becomes a real vector $\mathbf{d} \in \mathbb{R}^N$, where $N \in \mathbb{N}$ is the number of discrete time samples. The discrete graph of $\mathbf{d}$ is defined as the set of N points $\{(t_n, d_n) \in \mathbb{R}^2, n = 1, \ldots, N\}$. Based on this implicit discrete definition of the seismic trace, the following graph transformation $\mathcal{G}$ is introduced, such that

$$\begin{aligned} \mathcal{G}&:\mathbf{d} \to \mathcal{G}(\mathbf{d}) = d^{\mathcal{G}}(x,t), \\ &\mathbb{R}^N \to \mathcal{D}'(\mathbb{R}^2), \end{aligned} \tag{13}$$

with

$$d^{\mathcal{G}}(x,t) = \frac{1}{N}\sum_{n=1}^{N} \delta(t-t_n)\delta(x-d_n), \tag{14}$$

where $\mathcal{D}'(\mathbb{R}^2)$ denotes the space of distributions on $\mathbb{R}^2$ and δ is the Dirac delta distribution. The transformation thus implies that, from a discretized trace, a distribution is built that is a sum of N Dirac distributions in a 2D space, where one dimension (t) is associated with time and the other (x) is associated with amplitude.

Considering two discrete traces $\mathbf{d}_{\text{obs}} \in \mathbb{R}^N$ and $\mathbf{d}_{\text{cal}} \in \mathbb{R}^N$, we are interested in measuring the optimal transport distance $W_p(d^{\mathcal{G}}_{\text{obs}}, d^{\mathcal{G}}_{\text{cal}})$ in the space $\mathcal{D}'(\mathbb{R}^2)$. The positivity assumption is satisfied because the distributions $d^{\mathcal{G}}_{\text{obs}}$ and $d^{\mathcal{G}}_{\text{cal}}$ can take only the values 0 and $1/N$. Provided the time sampling of $\mathbf{d}_{\text{obs}}$ and $\mathbf{d}_{\text{cal}}$ is the same (the same number of discrete values N), their respective total mass is equal to one by definition. For a given $\mathbf{d} \in \mathbb{R}^N$, the integral

$$\int_x \int_t d^{\mathcal{G}}(t,x)dxdt = \frac{1}{N}\sum_{n=1}^{N} \int_x \int_t \delta(t-t_n)\delta(x-d_n)dxdt = 1 \tag{15}$$

tells that the mass conservation assumption is verified.

Practical approach for a differentiable misfit function

The above approach is appealing; however, because it relies on the definition of Dirac distributions, the operator $\mathcal{G}$ is not differentiable. Building a misfit function upon this operator would thus lead again to a nondifferentiability issue.

For this reason, the following smooth version of the graph transform is introduced, based on the approximation of the Dirac distribution through Gaussian functions. Considering again a discretized seismic trace $\mathbf{d} \in \mathbb{R}^N$, the smooth graph transform $\mathcal{G}_\sigma$ is defined as

$$\begin{aligned} \mathcal{G}_\sigma&:\mathbf{d} \to \mathcal{G}_\sigma(\mathbf{d}) = d^{\mathcal{G}_\sigma}(x,t), \\ &\mathbb{R}^N \to \mathcal{C}^\infty(\mathbb{R}^2, \mathbb{R}^+_*), \end{aligned} \tag{16}$$

with

$$d^{\mathcal{G}_\sigma}(x,t) = \frac{1}{2\pi\sigma_x\sigma_t N}\sum_{n=1}^{N} \exp\left(-\frac{(t-t_n)^2}{2\sigma_t^2}\right)\exp\left(-\frac{(x-d_n)^2}{2\sigma_x^2}\right), \tag{17}$$

where $\sigma = (\sigma_t, \sigma_x)$ with quantities (σ_t, σ_x) are user-defined constant scaling parameters, and $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^+_*)$ is the set of strictly positive functions of $(\mathbb{R})$, which are infinitely differentiable. This smooth version of the graph transform $\mathcal{G}$ preserves the positivity, as well as the mass conservation: The total mass of the transformed signal $d^{\mathcal{G}_\sigma}$ is equal to one. Indeed, we have, for a given $\mathbf{d} \in \mathbb{R}^N$

$$\begin{aligned} &\int_t \int_x d^{\mathcal{G}_\sigma}(x,t)dxdt = \frac{1}{2\pi\sigma_x\sigma_t N} \\ &\sum_{n=1}^{N} \int_t \int_x \exp\left(-\frac{(t-t_n)^2}{2\sigma_t^2}\right)\exp\left(-\frac{(x-d_n)^2}{2\sigma_x^2}\right) = 1. \end{aligned} \tag{18}$$

Adaptation to the FWI framework and gradient computation

Considering N_s seismic sources and N_r receivers, a FWI data set is composed of $N_s \times N_r$ seismic traces. We denote the observed and calculated traces $d_{\text{obs},s,r}(t)$ and $d_{\text{cal},s,r}(t)$, respectively. We formulate the following FWI problem:

$$\min_m f(m) = \sum_{s=1}^{N_s}\sum_{r=1}^{N_r} W_p(\mathcal{G}_\sigma(d_{\text{obs},s,r}), \mathcal{G}_\sigma(d_{\text{cal},s,r}[m])), \tag{19}$$

where the synthetic traces $d_{\text{cal},s,r}[m]$ for the source s are extracted from the wavefield $u_s(x,t)$ at the receiver location x_r through the operation

$$d_{\text{cal},s,r}[m](t) = R_r u_s[m] = u_s[m](x_r, t), \tag{20}$$

with the extraction operator R_r, whereas the solution $u_s[m]$ of a wave-propagation problem is denoted in compact form as

$$A(m)u_s(x,t) = b_s(x,t). \tag{21}$$

In equation 21, the operator $A(m)$ stands for any differential operator representing the wave propagation within the subsurface.

As mentioned above, we compute the gradient of the misfit function $f(m)$ using the adjoint-state strategy (Plessix, 2006). Following this method, the standard result states that the gradient $\nabla f(m)$ is given by the sum of the correlations between the incident $u_s[m]$ and the adjoint wavefields $\lambda_s[m]$

$$\sum_{s=1}^{N_s}\left(\frac{\partial A}{\partial m}u_s[m](x,t),\lambda_s[m](x,t)\right), \tag{22}$$

where $(.,.)$ denotes a scalar product in the wavefield space. The adjoint wavefields $\lambda_s[m](x,t)$ are solution of the backprojection equation:

$$A(m)^{\dagger}\lambda_s = \sum_{r=1}^{N_r} R_r^{\dagger}\mu_{s,r}[m], \tag{23}$$

where $.^{\dagger}$ denotes the adjoint operator. The adjoint source terms $\mu_{s,r}[m]$ are given by

$$\mu_{s,r}[m] = \frac{\partial W_p(\mathcal{G}_\sigma(d_{\text{obs},s,r}),\mathcal{G}_\sigma(d_{\text{cal},s,r}[m]))}{\partial d_{\text{cal},s,r}}, \tag{24}$$

where the definition of the misfit function is involved.

In this study, the graph transform of 1D seismic traces induces a 2D optimal transport problem: this requires dedicated and efficient algorithms. For this reason, we base our strategy on the KR distance, for which we have developed an efficient proximal splitting algorithm, making it possible to solve 2D and 3D transport problems for realistic scale seismic data (Métivier et al., 2016b, 2016c). The KR distance is based on the dual formulation of the one-Wasserstein distance stating that, for two functions $f(x)$ and $g(x)$, $x \in X \subset \mathbb{R}^d$, we have

$$W_1(f,g) = \max_{\varphi \in \text{Lip}_1(X)} \int_x \varphi(x)(f(x)-g(x))dx, \tag{25}$$

where $\text{Lip}_1(X)$ is the space of one-Lipschitz functions for the ground distance $\|.\|$ defined on $\mathbb{R}^d$

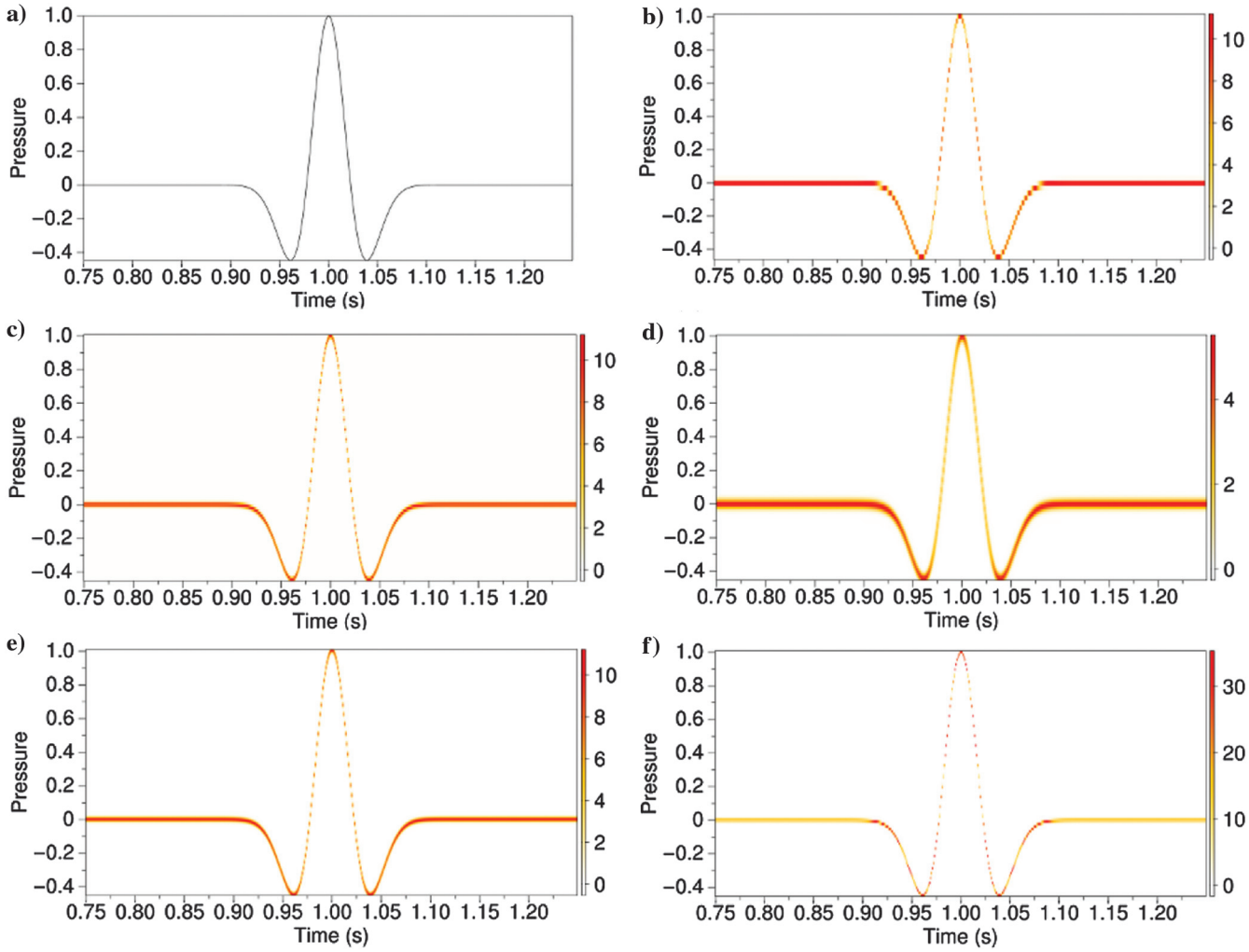

Figure 4. (a) Reference Ricker signal. (b) Representation of this Ricker in the graph space, with $\sigma_t = 0.001$ and $\sigma_x = 0.005$, and 50 discretization points for amplitude axis, (c) 100 discretization points for the amplitude axis, (d) 200 discretization points for the amplitude axis. Representation in the graph space, with 100 discretization points for the amplitude axis, with (e) $\sigma_t = 0.002$ and $\sigma_x = 0.01$, (f) $\sigma_t = 0.0005$ and $\sigma_x = 0.0025$.

$$\mathrm{Lip}_1(X) = \{\varphi(x), \quad \forall\ (x,x') \in X \times X, \\ |\varphi(x) - \varphi(x')| < \|x - x'\|\}. \tag{26}$$

Note that in Métivier et al. (2016b, 2016c), this distance is introduced with additional boundary constraints on the function φ. These boundary constraints lead to a well-posed maximization problem even in the case in which the mass conservation between f and g is not satisfied. As we have seen, the mass conservation is indeed satisfied in the case of seismic data and these boundary constraints can thus be neglected.

The proximal splitting algorithm we use is the ADMM/SDMM method (alternating-direction method of multipliers/simultaneous-direction method of multipliers) (Combettes and Pesquet, 2011). It is dedicated to the solution of convex nonsmooth problems (Métivier et al., 2016c). We recast the linear programming problem arising from the discretization of equation 25 as such a convex nonsmooth problem, with a misfit function composed of two terms. At each iteration, the ADMM/SDMM strategy applies the proximity operators associated with these two terms, for which we have closed-form formulations. In addition, a linear system needs to be solved, which we have demonstrated is equivalent to the second-order finite-difference discretization of Poisson's equation. This linear system is solved with an FFT approach, which has a quasilinear complexity (Swarztrauber, 1974).

Following this KR strategy, the related misfit function is given by

$$\min_m f(m) = \sum_{s=1}^{N_s}\sum_{r=1}^{N_r} \max_{\varphi_{s,r}\in \mathrm{Lip}_1(X)} \int_t \int_x \varphi_{s,r}(x,t)\left(d^{\mathcal{G}_\sigma}_{\mathrm{obs},s,r}(x,t) \right. \\ \left. - d^{\mathcal{G}_\sigma}_{\mathrm{cal},s,r}[m](x,t)\right) dx dt, \tag{27}$$

where $(x,t) \in X \subset \mathbb{R}^2$, with $x \in \mathbb{R}$ related to the amplitude axis, and $t \in \mathbb{R}_+$ related to the time axis. For this particular choice of optimal transport distance, the adjoint source terms $\mu_{s,r}[m]$ are given by

$$(\mu_{s,r}[m])_i = \frac{1}{2\pi\sigma_x\sigma_t N} \int_t \int_x \bar{\varphi}_{s,r}(x,t) \\ \exp\left(-\frac{(t-t_i)^2}{2\sigma_t^2} - \frac{(x-(d_{\mathrm{cal},s,r}[m])_i)^2}{2\sigma_x^2}\right) \frac{x-(d_{\mathrm{cal},s,r}[m])_i}{2\sigma_x^2} dx dt, \tag{28}$$

where the function $\bar{\varphi}_{s,r}(x,t)$ achieves the maximization

$$\bar{\varphi}_{s,r}(x,t) = \underset{\varphi_{s,r}\in \mathrm{Lip}_1(X)}{\mathrm{argmax}} \int_t \int_x \varphi_{s,r}(x,t)\left(d^{\mathcal{G}_\sigma}_{\mathrm{obs},s,r}(x,t) \right. \\ \left. - d^{\mathcal{G}_\sigma}_{\mathrm{cal},s,r}[m](x,t)\right) dx dt. \tag{29}$$

The derivation of this important result is given in Appendix C. It can be interpreted as follows: the optimal transport residual $\bar{\varphi}_{s,r}(x,t)$ coming from the comparison of the data in the graph space is backprojected into the time-domain space through equation 28. This backprojected solution can be considered as the output of the corresponding adjoint operator of the graph-space transformation $\mathcal{G}_\sigma$.

Finally, we can note again that, thanks to the adjoint-state formalism, the change of the misfit function within an existing FWI code only impacts the definition of the adjoint source. This is quite convenient in terms of implementation because only the part of the code dedicated to the misfit-function evaluation and adjoint source computation has to be modified.

Practical implementation details

In practice, both directions of the graph space should be discretized: The time direction has a natural discretization, whereas the amplitude direction requires one. For a given set of traces, the dynamic range could be between the largest positive peak and the lowest negative peak among all the traces. Of course, a trade-off needs to be found between the use of a fine discretization of the amplitude direction in the graph space for an accurate representation of the signal, and the induced computational cost: A fine discretization increases the size of the 2D optimal transport problems to solve and the computational issue.

In addition, an adequate choice of the scaling parameters σ_x and σ_t is required. Too small values can generate numerical accuracy problems as the Gaussian functions representing the Dirac masses become singular. Too large values can be responsible for losing weak amplitude event as the Dirac masses are blurred. In practice, the choice of the number of discretization points along the amplitude axis in the graph space and the settings of the scaling parameters σ_x and σ_t is a calibration step that is done on specifically chosen traces, by comparing them with their graph representation.

Decimating the data in time before transforming it to the graph space will mitigate the computational cost associated with the discretization of the amplitude axis. Time-domain modeling engines on which FWI relies need to satisfy a Courant-Friedrichs-Lewy condition that restricts the time sampling far beyond the Nyquist frequency: This results in a significant oversampling in time of the seismic data. This opens the way to nonnegligible computational gains through decimation of the signal (Yang et al., 2016).

Another practical aspect is related to the design of the ratio between the maximum range along the amplitude axis and the maximum range along the time axis in the graph space. Formally, the optimal transport distance in the graph space is convex with respect to any deformation of the signal along the two dimensions (time and

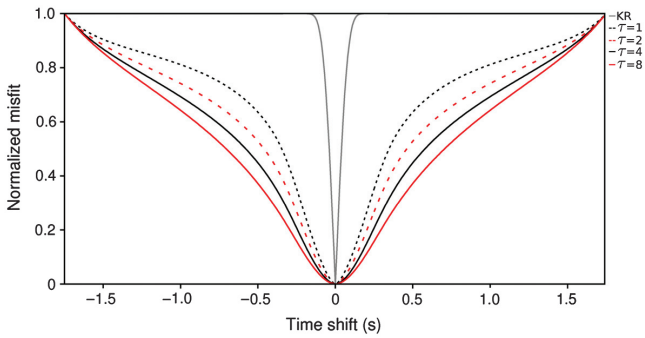

Figure 5. The OT-GS misfit function between observed and calculated Ricker signals, depending on the value of the scaling coefficient τ. $\tau = 1$, dotted black line; $\tau = 2$, dotted red line; $\tau = 4$, solid black line; and $\tau = 8$, solid red line. These misfit functions are compared with the KR misfit function directly applied to the Ricker signals in solid gray line. A single minimum is recovered for all the values of τ. Compared with the KR approach, the sensitivity to large time shifts expected from the use of an optimal transport distance is preserved. As expected, increasing the value of τ increases this sensitivity, whereas decreasing it narrows the valley of attraction.

amplitude). Therefore, the convexity of the distance is ensured not only with respect to time shifts but also with respect to amplitude shifts. For two shifted in time signals and a given ratio between the maximum range along the two axes, for sufficiently large time shifts, it could become less expensive from an optimal transport point of view to displace the Dirac masses along the amplitude axis rather than along the time axis to map the signals. If this is the case, the sensitivity of the optimal transport distance to the time shifts will be drastically reduced. One could adjust the aspect ratio between the amplitude axis and the time axis to favor the displacement along the time axis and to penalize the displacement along the amplitude axis from the optimal transport point of view. In the following, we introduce a control parameter τ, which offers the possibility to play on this aspect ratio to increase the sensitivity to the time shifts rather than on amplitude shifts.

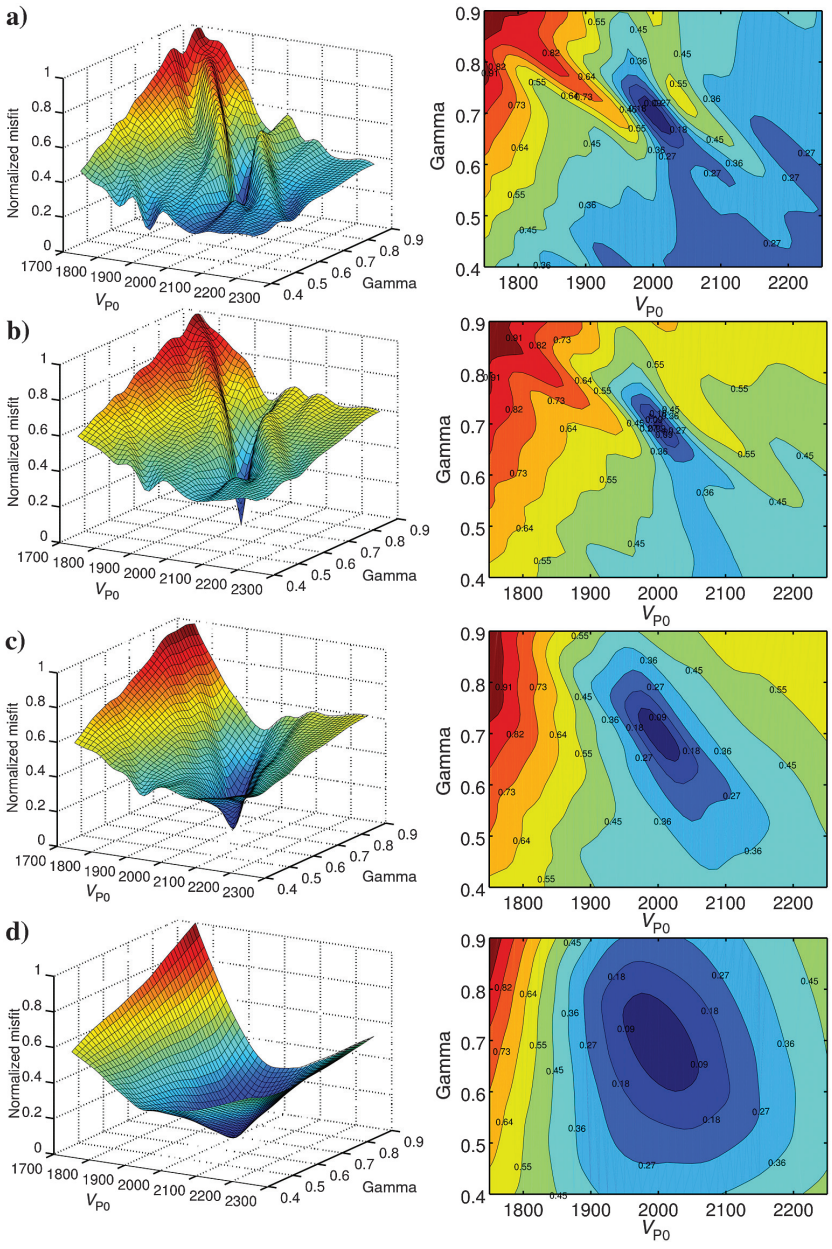

Figure 6. Misfit-function profiles (left column) and associated level sets (right column) for the two-parameters problem: (a) L^2 misfit function, (b) KR misfit function, OT-GS misfit function (c) with $\tau = 1$ and (d) with $\tau = 4$.

NUMERICAL EXPERIMENTS

We now consider four examples of increasing complexity to assess the interest of the proposed OT-GS strategy. First, we go back to the simple shifted-in-time Ricker signals presented previously to investigate the convexity of the OT-GS misfit function with respect to time shifts. We further extend this investigation to the case of a 1D velocity medium linearly increasing in depth: The misfit-function map is computed for the two parameters defining this 1D velocity model. Then, we investigate the behavior of a FWI algorithm based on this misfit function for a crosshole acquisition case study, with a homogeneous medium. We complete these illustrations with an analysis of an experiment based on the Marmousi 2 model (Bourgeois et al., 1991; Martin et al., 2006).

Shifted Ricker example

The shifted Ricker example will be investigated with the misfit function $C(s)$ defined by

$$C(s) = W_1(d^{\mathcal{G}_\sigma}_{\text{obs}}, d^{\mathcal{G}_\sigma}_{\text{cal}}(s)). \qquad (30)$$

In other words, the misfit between time-shifted Ricker signals is expressed by measuring the W_1 distance between their discretized "smoothed" graph representations.

The discrete graph representation of the reference Ricker signal is controlled by the sampling along time and amplitude axes and by the scaling parameters σ_t and σ_x (Figure 4). The reference Ricker is centered on 5 Hz (Figure 4a). The number of time discretization points is set to 2000, and the time discretization step is equal to 0.002 s: It is kept unchanged. Increasing the number of discretization points for the amplitude axis from 50 to 200 (Figure 4b–4d) increases the accuracy of the representation in the graph space of the seismic signal. For instance, staircase effects can be seen when using only 50 discretization points for the amplitude axis, which are removed by using a finer discretization. Choosing higher values for the scaling coefficients σ_t and σ_x produces a blurring effect, which might be detrimental to the representation accuracy in the graph space (Figure 4e). On the other hand, choosing too small values leads to a sparser representation of the signal (Figure 4f), which could yield potential numerical instabilities when computing the gradient of this misfit function.

Based on this comparison, we select a graph representation using 100 discretization points for the amplitude axis and intermediate values for the coefficients σ_t and σ_x, namely, $\sigma_t = 0.001$ and $\sigma_x = 0.005$. We limit the number of itera-

tions of the proximal splitting algorithm used to compute the KR distance in the graph space to 25 to reduce the computational cost. The computation of the misfit function $C(s)$ is performed for several values of the parameter τ, which aims at controlling the sensitivity to time shifts, equal to 1, 2, 4, and 8. The corresponding misfit functions are presented in Figure 5, and they are compared with the one obtained using the KR distance directly on the traces. A single minimum is recovered, for any value of τ, whereas the

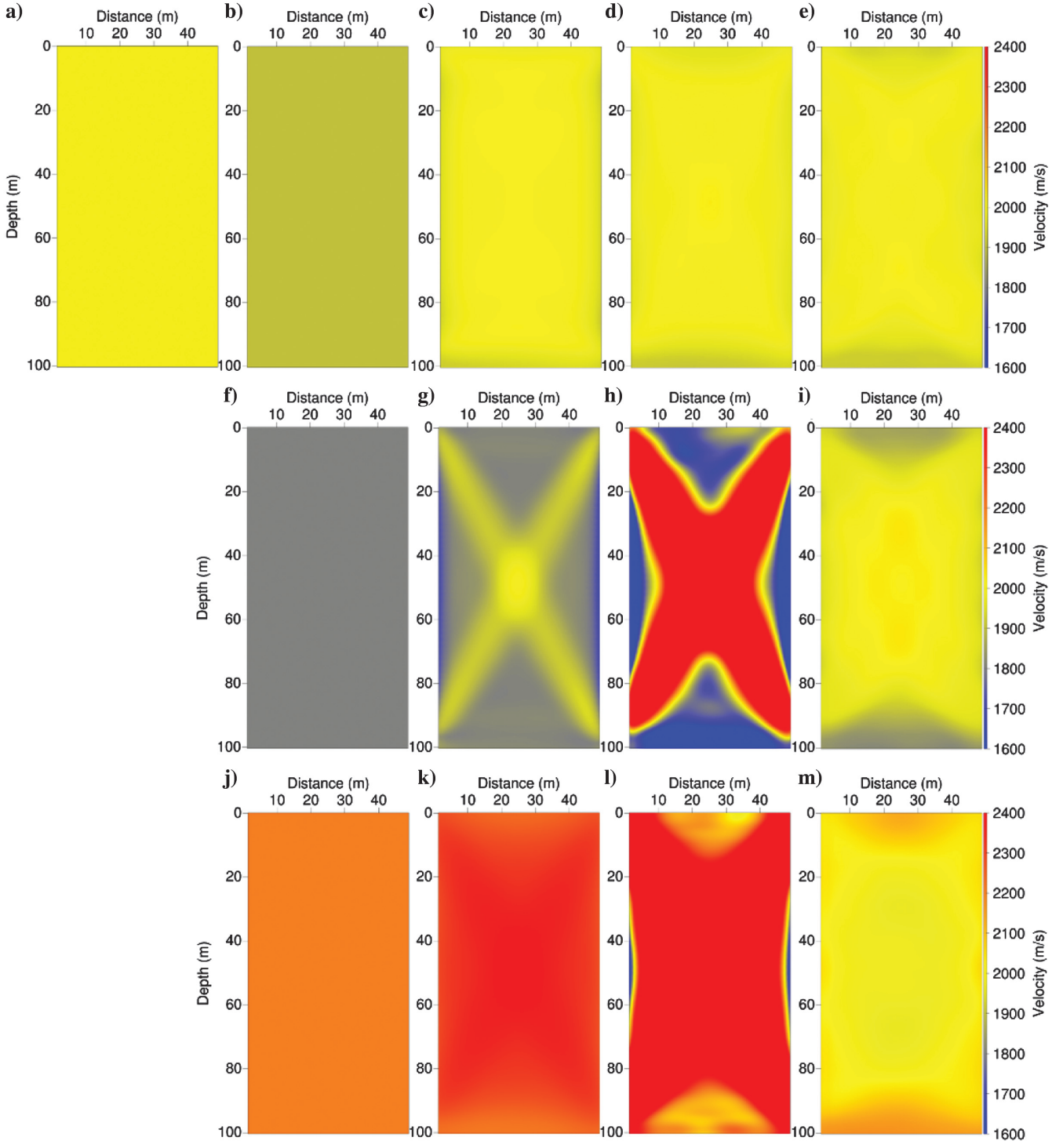

Figure 7. Crosshole experiment: (a) exact V_P model at 2000 m/s. (b) Initial model 1 at 1900 m/s, reconstructed models using (c) the L^2 misfit function, (d) the KR misfit function, and (e) the OT-GS misfit function from initial model 1. (f) Initial model 2 at 1800 m/s, reconstructed models using (g) the L^2 misfit function, (h) the KR misfit function, and (i) the OT-GS misfit function from initial 2. (j) Initial model 3 at 2200 m/s, reconstructed models using (k) the L^2 misfit function, (l) the KR misfit function, and (m) the OT-GS misfit function from initial model 3.

sensitivity to large time shifts, expected from the use of an optimal transport distance, is preserved. This is an encouraging result. In this schematic example, the graph-transport-based approach seems to overcome the difficulty associated with the use of optimal transport distance for nonpositive data. In addition, the control parameter τ appears to play the role that is expected: Increasing its value increases the sensitivity to large time shifts, generating a more convex misfit function, whereas decreasing its value narrows the valley of

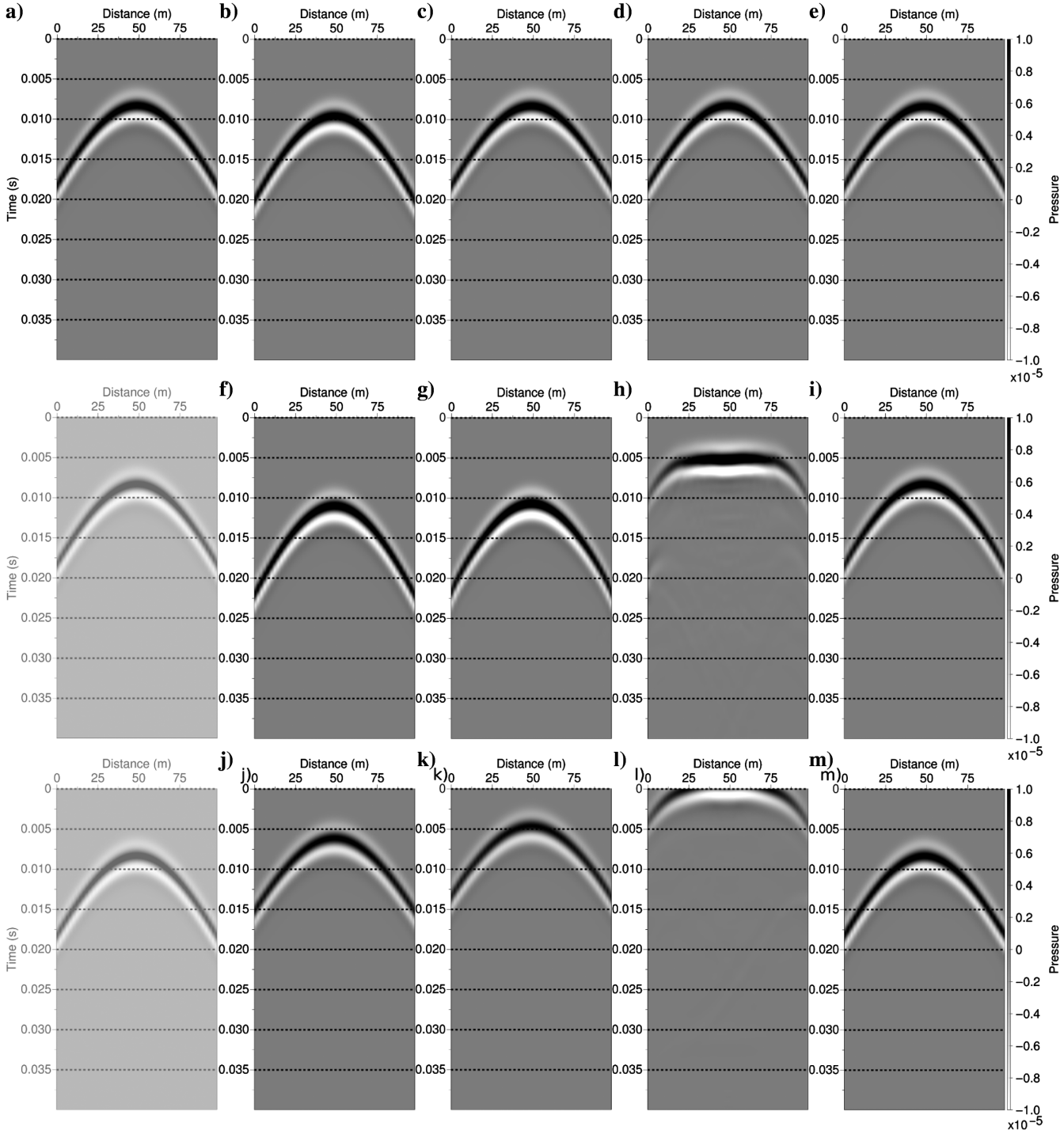

Figure 8. (a) Crosshole experiment: common shot gather (CSG) in the exact model. (b) CSG in the initial model 1, CSG in the model obtained using (c) the L^2 misfit function, (d) using the KR misfit function, (e) using the GT-OS misfit function starting from initial model 1. (f) CSG in the initial model 2, CSG in the model obtained (g) using the L^2 misfit function, (h) using the KR misfit function, (i) using the OT-GS misfit function starting from initial model 2. (j) CSG in the initial model 3, CSG in the model obtained (k) using the L^2 misfit function, (l) using the KR misfit function, (m) using the OT-GS misfit function starting from initial 3. The two shaded shot gather on the left of the second and the third rows are copy of the shot gather in the exact model (a) to make easier the comparison between shots.

attraction of the misfit function for improving the expected resolution. This could, in principle, open the way to a continuous design of hierarchical FWI schemes with large starting values of τ progressively decreased.

Comparison of misfit functions for a two-parameters problem

A two-parameter model allows a simple representation of the misfit function. We consider a linearly increasing velocity model $V_P(z)$ parameterized by a background velocity $V_{P,0}$ and a depth gradient γ such that

$$V_P(z) = V_{P,0} + \gamma z, \tag{31}$$

similar to what was proposed in Mulder and Plessix (2008). The velocity model is defined on a 2D rectangular domain 17 km long and 3.5 km deep. We compute synthetic data in the acoustic approximation using a surface acquisition system with one source located at a depth of 50 m and at $x = 8.45$ km and 168 receivers located at the same depth regularly deployed each 100 m from $x = 0.15$ to 16.85 km. The source wavelet is a Ricker centered on 5 Hz, which has been high-pass filtered to remove its energy of less than 3 Hz. The total recording time is set to 4.6 s.

The data are acquired for a velocity model parameterized with $V_{P,0} = 2000$ m/s and $\gamma = 0.7\text{s}^{-1}$. We compare the L^2 misfit function, the KR misfit function, and the OT-GS misfit function with two values of the scaling coefficient $\tau = 1$ and 4. The range of variations for $V_{P,0}$ and γ is

$$1750 \leq V_{P,0} \leq 2250, \quad 0.4 \leq \gamma \leq 0.9. \tag{32}$$

The sampling of the misfit function for both parameters is $\Delta V_{P,0} = 12.5$ m/s, $\Delta\gamma = 0.025$ s^{-1}.

The settings for the discretization of the graph representation are similar to the one used in the previous experiment. Namely, 2000 points for the time axis and 100 points for the amplitude axis with scaling parameters $\sigma_t = 0.001$ and $\sigma_x = 0.005$. The number of iterations for the proximal splitting algorithm is also limited to 25 as for the previous experiment.

The four misfit functions and their level sets are presented in Figure 6. As can be seen, the L^2 misfit function presents several local minima and the global minimum is surrounded by barriers narrowing the valley of attraction: Converging to the global minimum through local optimization techniques would require a starting model in close vicinity. As was noticed in Métivier et al. (2016c), the use of the KR distance tends to produce a smoother misfit profile with a wider

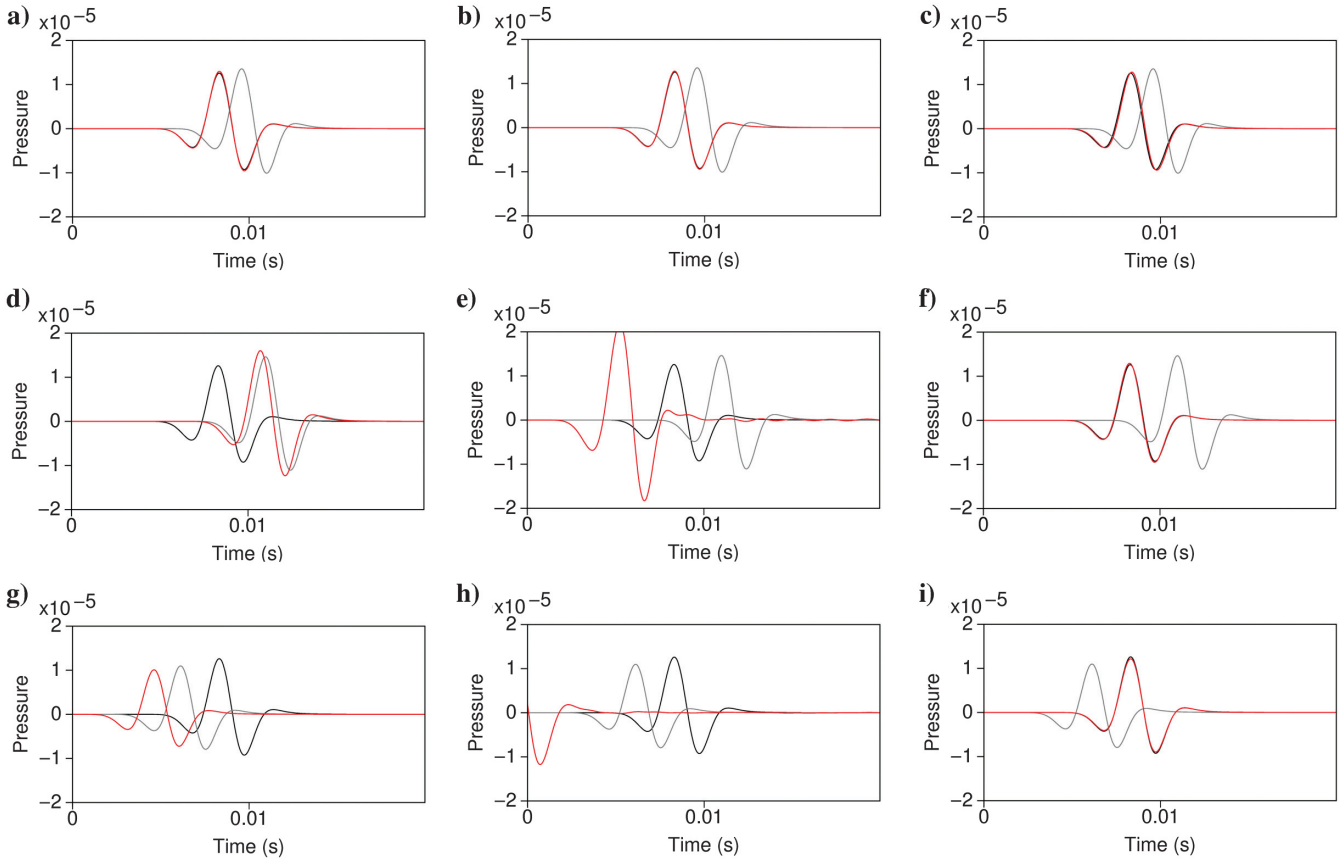

Figure 9. Crosshole experiment: comparison of seismic traces corresponding to the shot and the receiver at $z = 49$ m depth in the left and right boreholes. The seismic traces are computed in the exact model, initial models, and estimated models, depending on the choice of the misfit function. Comparison for the initial model at $V_P = 1900$ m/s, for (a) the L^2 misfit function, (b) the KR misfit function, (c) the OT-GS misfit function. Comparison for the initial model at $V_P = 1800$ m/s, for (d) the L^2 misfit function, (e) the KR misfit function, (f) the OT-GS misfit function. Comparison for the initial model at $V_P = 2200$ m/s, for (g) the L^2 misfit function, (h) the KR misfit function, (i) the OT-GS misfit function.

valley of attraction. However, several local minima are still present. Interestingly, the OT-GS approach provides misfit functions with no local minima: A single global minimum is recovered. Increasing the factor τ from 1 to 4, as in the previous example, enforces the convexity of the misfit function. The attraction valley is enlarged, which favors the convergence to the global minimum, however possibly at the expense of a resolution loss, as has been observed for crosscorrelation functions, for instance (van Leeuwen and Mulder, 2010). This, again, is an encouraging result because the OT-GS approach appears to overcome the limitation of applying the KR misfit function directly to the data, yielding a nonconvex misfit function.

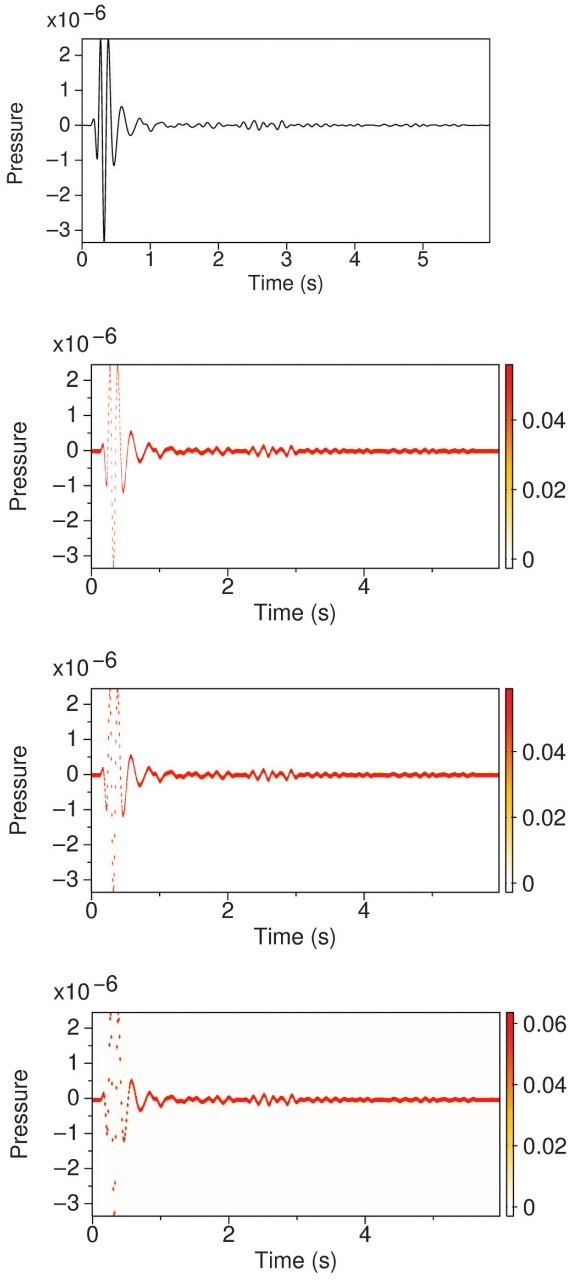

Figure 10. (a) Comparison of a seismic trace computed in the exact Marmousi model and different representations in the graph space depending on the time-decimation rate: (b) no decimation, (c) decimation by a factor two, and (d) decimation by a factor four.

A crosshole tomography experiment

Let us now consider a more realistic example, though still quite simple: A 2D crosshole tomography configuration with 48 sources located in a 100 m deep borehole and 48 receivers located in a second borehole 50 m away from the first borehole should be an interesting illustration of the effects of the various misfit functions. A set of data is acquired in the acoustic approximation in a homogeneous velocity medium at constant velocity $V_P = 2000$ m/s, using a Ricker source centered at 250 Hz.

Three homogeneous initial models are considered: the first one at a velocity $V_P = 1900$ m/s, the second one at a velocity $V_P = 1800$ m/s, and the third one at a velocity $V_P = 2200$ m/s. FWI results obtained using the L^2, KR, and OT-GS misfit functions are compared.

A 2D Gaussian smoothing is applied to the gradient. The correlation length in each direction is equal to a fraction of the local wavelength. This local wavelength is approximated in the current velocity model for a reference frequency set to 250 Hz. The fraction we choose in this example is equal to 0.8.

The corresponding correlation length is between 7.2 and 8.8 m from the slowest to the fastest initial model. Because of this relatively strong smoothing, we rely on a nonlinear conjugate gradient optimizer rather than on the standard l-BFGS solver: The l-BFGS inverse Hessian estimation is impacted by the smoothing. We use the implementation from the SEISCOPE toolbox (Métivier and Brossier, 2016).

For the OT-GS approach, the time-discretization points are 2000 whereas those along the amplitude axis are 100. Values of σ_t and σ_x are equal to 0.001 and 0.005, respectively, as in previous experiments. We also limit the proximal splitting algorithm to compute the KR distance in the graph space to 25 iterations to save computation time. The control parameter τ is set to one in this experiment.

The results we obtain are presented in Figure 7 for the velocity reconstruction and in Figures 8 and 9 for data fitting. Starting from the initial model at 1900 m/s (Figure 7b), the three strategies are able to converge toward a reliable estimation of the P-wave velocity (Figure 7c–7e): no cycle-skipping problems are met. Starting from the initial model at 1800 and at 2200 m/s (Figure 7f and 7j), there are fundamental differences in the velocity reconstruction depending on the misfit we consider (Figure 7g–7m).

The analysis of the data reveals that a cycle-skipping issue indeed occurs in these two initial models (Figure 8a, 8f, and 8j). The initial model at 1800 m/s is too slow: The predicted arrival is delayed by more than a period from the exact one. The initial model at 2200 m/s is too fast: The predicted arrival is advanced by more than a period from the exact one.

For this reason, the L^2 and KR misfit functions are not able to converge toward a correct estimation of the P-wave velocity. Starting from the initial model at 1800 m/s, the L^2 result exhibits a crossed-shape area in which the correct velocity at 2000 m/s is reached (Figure 7g). However, away from this cross-shaped area, the velocity is decreased, while it should be increased: This is a clear sign of cycle skipping. For the KR result (Figure 7h), the same pattern appears amplified: The lower and upper bounds on the P-wave velocity, which have been set to 1000 and 5000 m/s, are reached. Starting from the initial model at 2200 m/s, the L^2 result shows that the whole velocity has been increased whereas it should be decreased (Figure 7k). This is the same for the KR result, except for two low-velocity zones on the edges, near the source and receivers in the middle of the wells (Figure 7l).

The analysis of the data reveals that, starting from the initial models at 1800 and 2200 m/s, the calculated and observed data cannot be correctly put in phase using the L^2 misfit function or the KR misfit function (Figure 8a, 8g, 8h, 8k, and 8l). A closer view is given in Figure 9, where a single trace, corresponding to the source and receiver in the middle of the boreholes, is compared. For each plot, it is computed in the exact model, the initial model, and the result obtained using a given misfit function. As can be seen in Figure 9d, 9e, 9g, and 9h, the L^2 and KR misfit functions are not able to shift the main arrival toward its correct time. The KR results (Figure 9e and 9h) appear more unstable because the main arrival is advanced much more than what is expected to match the observed data. This is consistent with the resulting models, which reach the upper bound set at 5000 m/s in large zones. This instability might be associated with the insensitivity of the KR misfit function with respect to large time shifts, as is illustrated in the 1D example (Figure 3).

Only using the OT-GS misfit function gives a reliable estimation of the P-wave velocity (Figure 7i and 7m). The data computed in the final model are in phase with the exact data (Figure 8i and 8m). The analysis of the central trace in the exact, initial, and final estimations presented in Figure 9c, 9f, and 9i confirms the previous interpretation. Using the OT-GS misfit function, the calculated trace nicely fit the observed data.

This simple experiment confirms the promising potentiality of the OT-GS misfit function to mitigate the cycle-skipping issue.

Marmousi model

The Marmousi model, although still a simple model, could highlight the interest of the OT-GS approach in a more realistic configuration. We consider a modified version of the Marmousi 2 P-wave velocity model for which the bathymetry is flat, presented in Figure 13. The water layer is kept constant at 1500 m/s and assumed to be known. A fixed-spread surface acquisition at a depth of 50 m is used with 128 sources located each 130 m from $x = 0.05$ to 16.7 km. For each source, 168 receivers at the same depth are used, located each 100 m from $x = 0.05$ to 16.8 km. The observed data are acquired using a Ricker-source wavelet centered at 6 Hz and high-pass filtered so as to remove the energy of the signal less than 3 Hz.

A set of data is acquired in the acoustic approximation using the exact Marmousi model. A free-surface condition is implemented at the surface/air interface. The total recording length is set to 6 s. We consider two initial models. The first is obtained by smoothing the exact model using a correlation length of approximately 3 km resulting in a strongly smoothed version of the true model and a significant underestimation of the velocity increase in depth. The second is a 1D model linearly increasing from the ocean bottom at 1500 m/s to the bottom of the model at 3200 m/s, resulting in an even cruder initial guess. For these two models, we want to compare the efficiency of FWI based on the L^2, KR and OT-GS misfit functions.

In Figure 10, we illustrate how a seismic trace is typically represented in the graph space for this more realistic application. We compare one trace extracted from the data corresponding to a source and a receiver located at the center of the model at $x = 8.5$ km, with its representation in the graph space, using three discretizations with 2600, 1300, and 650 points along the time axis, 200 discretization points along the amplitude axis, with the scaling quantities $\sigma_t = 0.001$ and $\sigma_x = 0.0025$. The number of discrete points used to geometrically represent the amplitude variations should be as small as possible to prevent a too drastic increase of the computational time: 200 discrete points seem to be a good

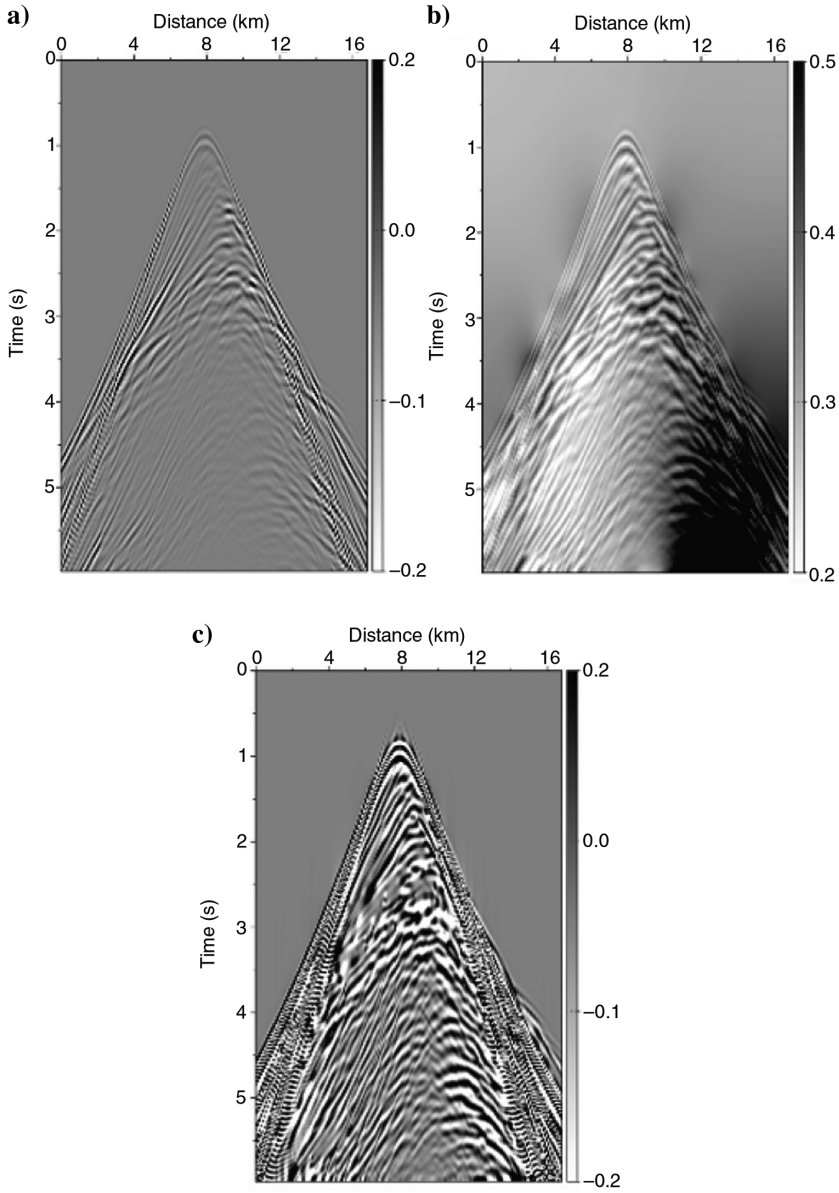

Figure 11. Adjoint source in the 1D initial model corresponding to the shot gather associated with a source located at $x = 8.5$ km for the Marmousi case study: (a) L^2 adjoint source, (b) KR adjoint source, and (c) OT-GS adjoint source.

compromise for the amplitude. The graph transform mainly affects the accuracy of the representation of events with the largest magnitude. The decimation in time increases this effect as the number of points in time becomes too small to properly represent the fast variations of the signal. However, the largest amplitude event corresponds to the direct arrival, which is correctly predicted in this case because the water layer is supposed to be known. For this reason, the accuracy of the representation in the graph space of this large amplitude event might be not crucial in this case. Thus, we select a decimation ratio equal to four, corresponding to 650 points along time. This reduces significantly the computational cost of the OT-GS approach.

Adjoint sources (residuals) associated with L^2 misfit function, the KR misfit function, and the OT-GS misfit function are compared in Figure 11. These adjoint sources are computed in the 1D initial model, for the shot gather associated with the source located at $x = 8.5$ km. For the OT-GS distance, the control parameter τ is set to 1.5 after some trial-and-error calibration. For plotting comparison, the adjoint sources have been normalized to have a maximum absolute value equal to one, whereas the true amplitudes are preserved when backprojected into the medium. The L^2 adjoint source

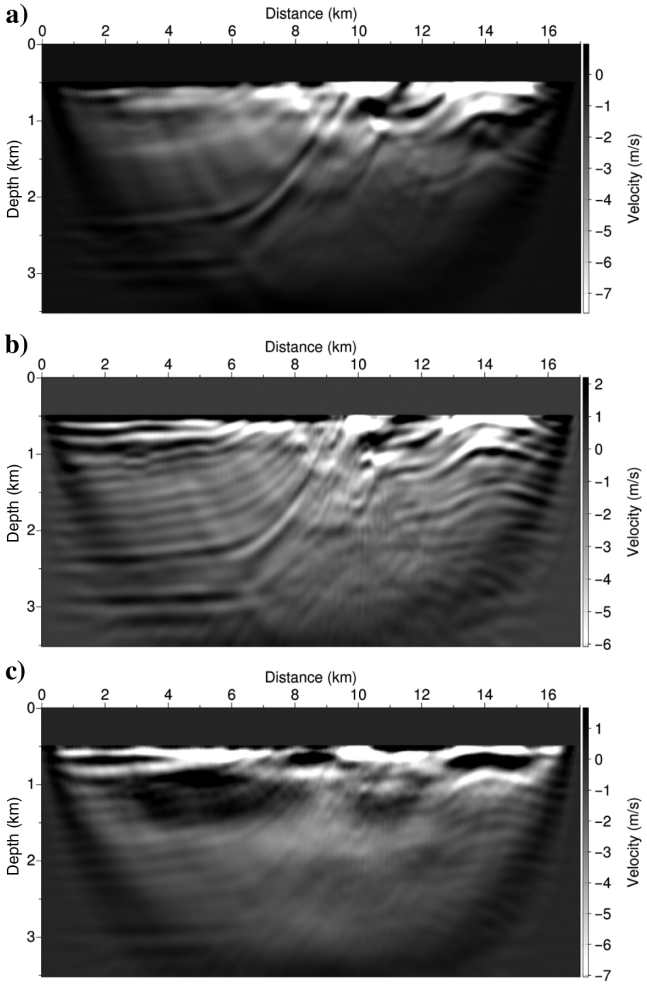

Figure 12. First gradient in the initial 1D model for the Marmousi case study: (a) L^2 gradient, (b) KR gradient, and (c) OT-GS gradient.

is dominated by the mismatch of diving waves at the long offset (Figure 11a). Conversely, the two others give a more important weight to missing reflected events. The KR adjoint source bears a significant imprint of a zero-frequency signal: Actually, its mean is not equal to zero and the color bar has been shifted to enhance the visual comparison with the two other adjoint sources (Figure 11b). Interestingly, the OT-GS adjoint source presents a similar anatomy as the adjoint source associated with the KR adjoint source; however, no zero-frequency signal is introduced: Its mean remains at zero. In addition, the amplitude of residuals associated with the diving wave remains stronger than those associated with reflections, whereas in the adjoint source associated with the KR misfit function, the amplitude of events associated with diving waves and reflections is approximately the same (Figure 11c).

In Figure 12, the first gradients in the 1D initial model are presented, depending on the choice of the misfit function. The L^2 gradient is dominated by shallow updates and the imprint of the strong dipping reflector on the middle left of the model (Figure 12a). The KR gradient mitigates the imprint of this reflector and provides more details on the deeper structure of the model (Figure 12b). The large reflector on the bottom left part of the medium appears at a 3 km depth. In the OT-GS gradient, the imprint of the strong dipping reflector in the medium of the model is removed (Figure 12c). The energy is refocused on smooth shallow updates, while finer scale reflectors are still visible below.

In Figure 13, the P-wave velocity models reconstructed after FWI starting from the two different initial models, using the different misfit functions, are presented. The *l*-BFGS algorithm from the SEISCOPE optimization toolbox is used, together with a Gaussian smoothing of the gradient similar to the one used in the previous crosshole experiment. The correlation length is equal to respectively 0.4 and 0.2 times the local wavelength in the horizontal and vertical directions. A linear norm preserving depth preconditioning is also used to enhance the updates at depth, accounting for the surface acquisition configuration. Following this method, the preconditioned gradient is obtained by multiplying the gradient with the depth. A normalization is then applied to preserve the norm: The preconditioned gradient has the same norm as the original gradient.

No stopping criterion is enforced for the *l*-BFGS algorithm: We let the optimizer free to minimize the misfit function as much as possible. The iterations thus stop on a linesearch failure. In Figure 14, the shot gathers corresponding to the source located at $x = 8.5$ km computed in the exact, initial, and reconstructed P-wave velocity models are presented.

Both initial models generate cycle-skipping effects: Only the direct wave is correctly predicted in the water layer, and the diving waves, which arrive faster at a large offset, are cycle-skipped as shown in Figure 14a, 14b, and 14f. The cycle skipping is more severe for the 1D initial model than for the smooth initial model, especially for the diving waves propagating on the right part of the model. This is consistent with the fact that the smooth initial model preserves a lateral variation with a velocity increase on the bottom right part of the model, which is no longer present in the 1D initial model.

FWI based on the L^2 misfit function fails to recover the P-wave velocity structure starting from the two initial models as expected. In both cases, the initial model is marginally updated. The data computed in the final estimation (Figure 14d and 14e) confirm the presence of cycle-skipping effects: The long-offset diving waves

are incorrectly updated. The diving wave propagating to the left are accelerated, whereas they should be decelerated to fit the observed data.

FWI based on the KR misfit function provides better results when it starts from the first initial model (Figure 13f). The main structures of the model are recovered until 3 km depth despite shallow artifacts on the left and deeper artifacts on the right. These artifacts can be correlated with the data fit presented in Figure 14d, where we can see that the postcritical reflections arriving at later times and larger offsets are incorrectly predicted, on both sides of the model. Starting from the 1D initial model, the reconstruction of the P-wave velocity model is far less satisfactory (Figure 13g). The central structure between $x = 8$ and 11 km starts to be recovered, as well as the shallow structure until 1 km depth on the right part of the model. However, the model is not updated deeper than 2.5 km depth and strong cycle-skipping artifacts can be observed on both sides of the model. This is confirmed by the shot gather presented in Figure 14h. As is observed for the result obtained starting from the first initial model, the postcritical reflections arriving at later times and the large offsets are incorrectly predicted. In addition, a strong reflection appears at 1.5 s, for the receivers located nearby $x = 6$ km. This reflection corresponds to the high-velocity anomaly located at $x = 8$ and 1.5 km depth in Figure 13g. This anomaly might explain a later event in the observed data, arriving at time $t = 3.5$ s on the receivers located nearby $x = 4$ km, as an internal multiple.

FWI using the OT-GS distance yields better results starting from both initial models (Figure 13h and 13i). The velocity structure is relatively well-reconstructed until 3 km depth, even if some artifacts can be identified on the edges of the model, which could be related to a lack of illumination in these zones. In addition, starting from the 1D initial model, the strong reflector located on the right part of the model, at 2.5 km depth, is not reconstructed. This can be well identified in the shot gather presented in Figure 14e and 14i: The postcritical reflections arriving at $t = 5$ s on the right part of the model are correctly predicted only starting from the smooth initial model. A relatively strong amplitude thin-layering artifact also appears at very shallow depth, just below the water layer, on the left part of the model, starting from the two initial models. It is not yet clear why and how they appear: Their presence could be removed through a postprocessing smoothing of the estimated models.

The comparison of OT-GS misfit functions for different values of τ presented for the Ricker examples presented in Figure 5 suggests the possibility of a hierarchical approach based on decreasing values of τ. Finally, we investigate this strategy in the framework of the Marmousi experiment. The model obtained using the OT-GS misfit function with $\tau = 1.5$, starting from the 1D velocity model, is used as a starting model for a sequence of inversions with the OT-GS misfit function with values of τ decreasing to 0.5, 0.25, 0.1, 0.05, and 0.025. For each τ value, the result that is obtained is used as a starting model for the next τ value. We present the models obtained at stages corresponding to $\tau = 0.5$, 0.25, and 0.025 in Figure 15. Interestingly, we see that the velocity structure is significantly improved, especially in depth. The strong reflector in the right part starts being reconstructed. The high-velocity anomaly in the bottom left part at $x = 3$ km and $z = 2.5$ km is also corrected, enhancing the continuity of the deep reflector in the left part of the model. This improvement is confirmed by the shot-gather analysis provided in Figure 16. The exact shot gather is compared with the shot gather in the model obtained from the 1D velocity model using the OT-GS misfit function with $\tau = 1.5$, and the shot gather in the model obtained at the end of the hierarchical approach for decreasing τ values. The reconstruction of the late postcritical reflections is enhanced, in the left part and especially the right part of the gathers. The design of a hierarchical workflow based on decreasing τ values thus seems

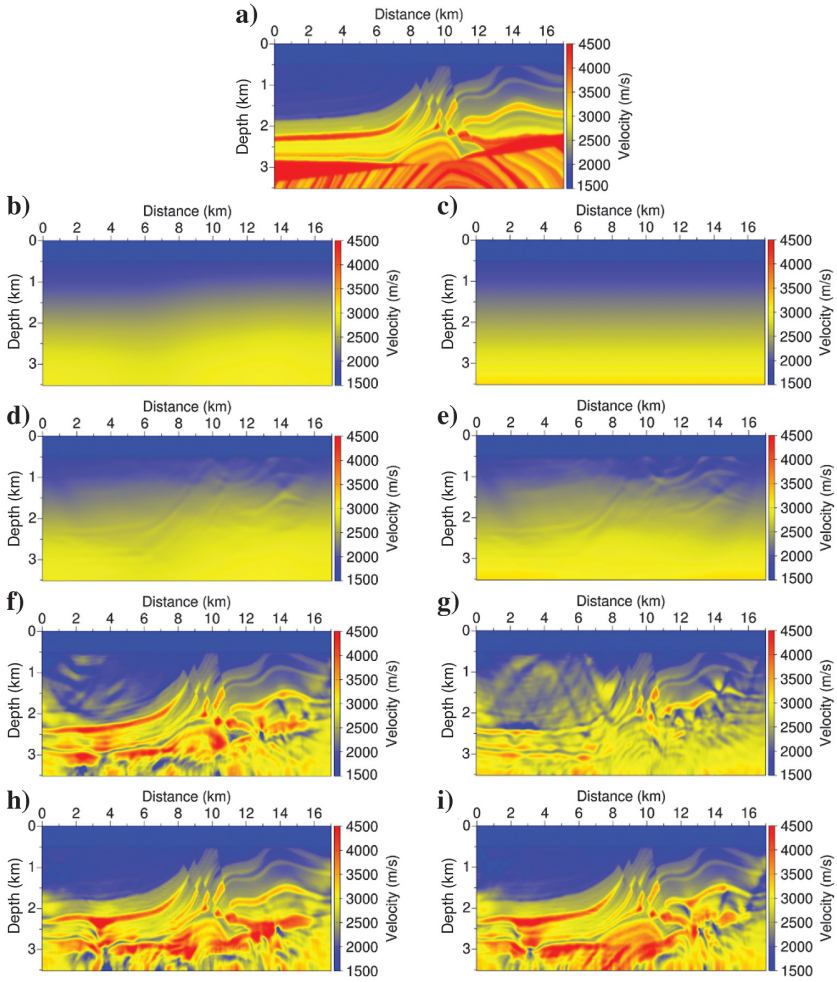

Figure 13. Marmousi experiment, P-wave velocity models. (a) Exact model. (b) Initial model obtained through a Gaussian smoothing of the exact model. (c) Initial 1D linearly increasing model. Results obtained using a L^2 misfit function obtained starting (d) from the smooth initial model, (e) from the 1D initial model. Results obtained using the KR misfit function starting (f) from the smooth initial model, (g) from the 1D initial model. Results obtained using the OT-GS misfit function starting (h) from the smooth initial model and (i) from the 1D initial model.

promising for an accurate broadband reconstruction of the velocity model, starting from a very crude initial model, and data containing no low-frequency information.

DISCUSSION

The numerical experiments illustrate the interest of the OT-GS distance for FWI. In particular, it appears as a promising tool to increase the convexity of the misfit function with respect to the subsurface velocity. However, this preliminary work leaves several questions unanswered. The first is related to the computational cost and the computational complexity of the approach. Basically, at each iteration of the FWI algorithm, in the current implementation, a 2D transport problem has to be solved for each trace of the shot gathers. Based on the numerical solver we have previously developed in Métivier et al. (2016b, 2016c), this implies an algorithmic complexity in $O(N_t \times N_x \times N_{\text{rec}})$, where quantities N_t, N_x, and N_{rec} denote the number of time discrete points, amplitude discrete points, and the number of receivers per source. Because the chosen N_x should be sufficiently large to maintain a sufficient accuracy of

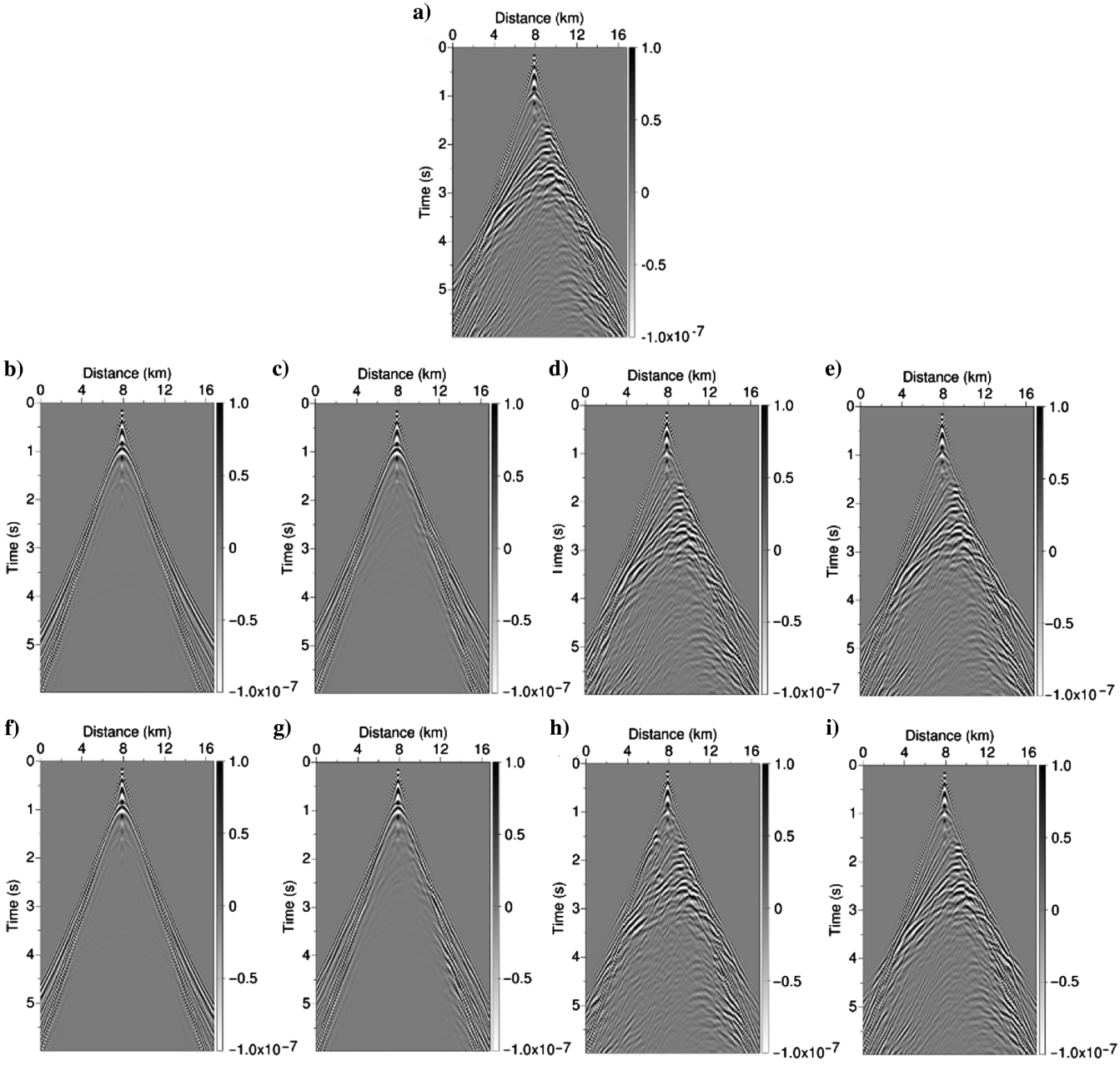

Figure 14. Marmousi experiment, shot gathers computed for a source located at $x = 8.5$ km. (a) Shot gather in the exact model. (b) Shot gather in the smooth initial model. Shot gather in the final model obtained starting from the smooth initial model (c) with the L^2 misfit function, (d) with the KR misfit function, (e) with the graph-transform-based optimal transport misfit function. (f) Shot-gather in the 1D initial model. Shot gather in the final model obtained starting from the 1D initial model (g) with the L^2 misfit function, (h) with the KR misfit function, (i) with OT-GS misfit function.

the data representation in the graph space and the time sampling fine enough for capturing the amplitude evolution, the algorithmic complexity becomes similar as the one requested for the solution of a 3D transport problem using the same algorithm.

This nonnegligible increase can however be mitigated by adjusting different constants parameterizing the problem. First, the data can be decimated in time. On the Marmousi example, the final results are obtained using a decimation in time by a factor four: Only 650 samples are considered. Second, the number of iterations of the proximal splitting algorithm used to solve the optimal transport problem can be reduced. On the Marmousi experiment, only 40 iterations are performed, compared with 100 iterations for the KR misfit function. A cumulative factor of 10 is thus obtained, and the computational cost of the misfit function is thus "only" 20 times larger than the cost associated with the KR misfit function instead of 200. As a more practical comparison, the computational cost of one gradient, on our local cluster, for the Marmousi experiment using the least-squares distance is equal to 32 s, whereas it is equal to 38 s using the KR misfit function and 186 s using the OT-GS distance. The global computational increase is thus by a factor 5.8 compared with the L^2 misfit function and a factor 4.9 compared with the KR misfit function in the settings we have chosen for the Marmousi experiment.

A further question is related to the interpretation of the whole shot gathers in the graph space rather than for each individual trace. There is no theoretical obstacle to proceed in this direction, which would lead to the solution of a 3D optimal transport problem, instead of N_{rec} 2D optimal transport problems, per shot. We have emphasized the interest of the interpretation of the data through optimal transport considering the whole gathers in Métivier et al. (2016b). In particular, this opens the way for accounting for any geometric coherency the data can have in a gather representation, which brings valuable information for constraining the subsurface parameters in the FWI process. This feature is crucial for the KR misfit function: It is far less efficient when applied trace by trace. The reason why we remain in this study in a trace-by-trace formalism is practical. For the Marmousi experiment, solving large 3D optimal transport problem through our proximal splitting technique requires the use of 3D Poisson solvers. Our current implementation, based either on the FISHPACK (Swarztrauber, 1974) solver or on the MUDPACK solver (Adams, 1989), lacks flexibility. Improving our implementation to tackle larger scale 3D Poisson problems will be the matter of further studies.

A hint on the interest for moving toward a full shot-gather interpretation is provided in Figure 17. A comparison of 2D misfit-function profiles obtained using trace-by-trace or full shot-gather strategies is performed in the settings of the second numerical experiment (1D velocity linearly increasing in depth and surface acquisition). The misfit function of the full shot gather is slightly smoother and more convex than the misfit function obtained through a trace-by-trace analysis.

Nonetheless, reducing further the computational cost while still relying on the graph-transform approach will probably require changing the algorithm used to compute the optimal transport distance in the graph space. This work is not straightforward and will require dedicated study. Among the possibilities we are interested in, we will consider using the entropic regularization approach promoted by Benamou et al. (2015) or the semidiscrete approach promoted by Mérigot (2011) and Kitagawa et al. (2017).

Regarding the interpretation of the data, a possibility that is not used in this study, and would require a specific study, consists in reintroducing mass in the graph space to weight the seismic events differently. In the approach proposed in this study, the use of a smooth approximation of the Dirac function for each discrete point leads to the same mass for each point of the graph (equation 17). However, a weighting function $a(x,t)$ in the graph space could be introduced to emphasize the interpretation of particular events. In the Marmousi example, giving more weight to later arrivals might improve the reconstruction at depth. The same weighting function would be required for observed and calculated data to preserve the mass conservation.

Another important question that will need to be assessed in future studies is the robustness of the graph-transform-based optimal transport distance approach with respect to the ability to accurately

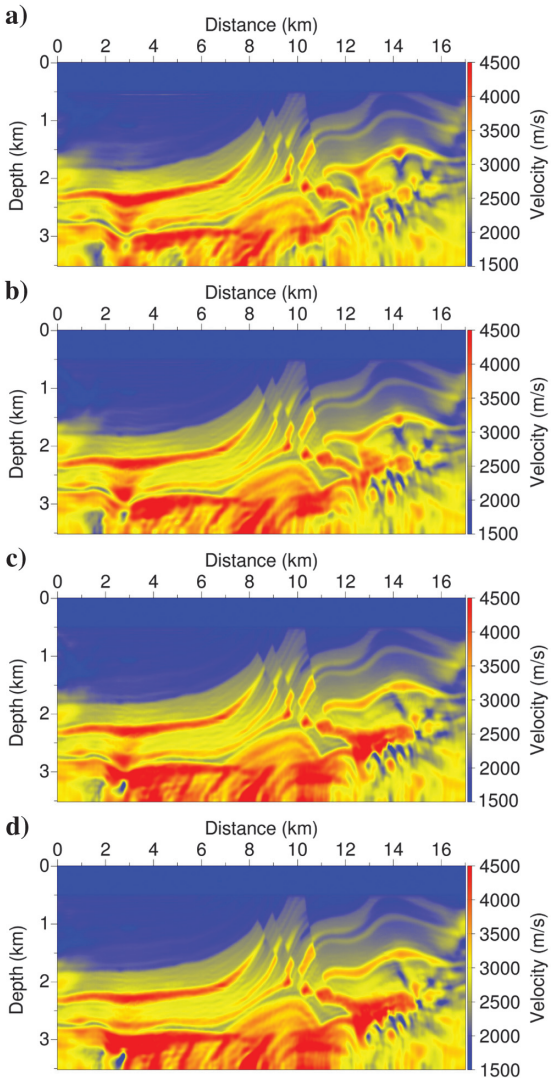

Figure 15. Marmousi experiment, P-wave velocity models, hierarchical approach. (a) Result obtained using the OT-GS misfit function starting from the 1D velocity model with $\tau = 1.5$. (b) Result obtained starting from model (a) using the OT-GS misfit function with $\tau = 0.5$. (c) Result obtained starting from model (b) using the OT-GS misfit function with $\tau = 0.25$. (d) Result obtained starting from model using the OT-GS misfit function with $\tau = 0.025$.

predict the amplitude of the seismic events. The crosshole and Marmousi experiments presented in this study are performed in inverse crime settings. The ability of the OT-GS strategy to handle noise and inaccurate physical modeling of the wave propagation therefore remains to be assessed. In particular, introducing anisotropy in the measure of the ground distance to favor the displacement along the time axis, as controlled by the scaling parameter τ we have introduced, should be in turn responsible for an increased sensitivity of the misfit function to amplitude mismatch. This could be a drawback in the perspective of applications to real data; it is well known that accurate amplitude prediction is difficult.

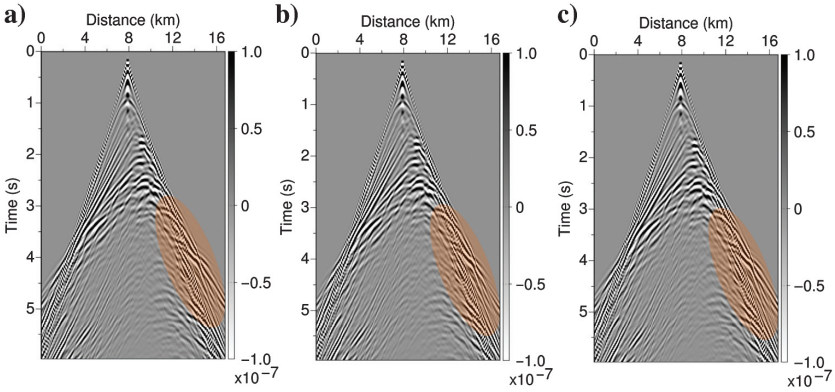

Figure 16. Marmousi experiment, shot gathers computed for a source located at $x = 8.5$ km. (a) Shot gather in the exact model. (b) Shot gather in the model obtained starting from the 1D initial model with the OT-GS misfit function. Shot gather in the model obtained following a hierarchical approach by decreasing the scaling factor τ reaching $\tau = 0.025$ (model from Figure 15d).

This opens several avenues for investigation: Should we rely on a normalization of the signal to focus on the comparison of phases only? Several such normalization strategies have already been proposed to this purpose, one recent instance being the implicit shaping method proposed by Maharramov et al. (2017). This could be used prior to the graph-transform-based optimal transport strategy. Another direction relies on a more accurate prediction of the physics by incorporating attenuation, density effects, as well as elastic and anisotropic effects. This is more ambitious

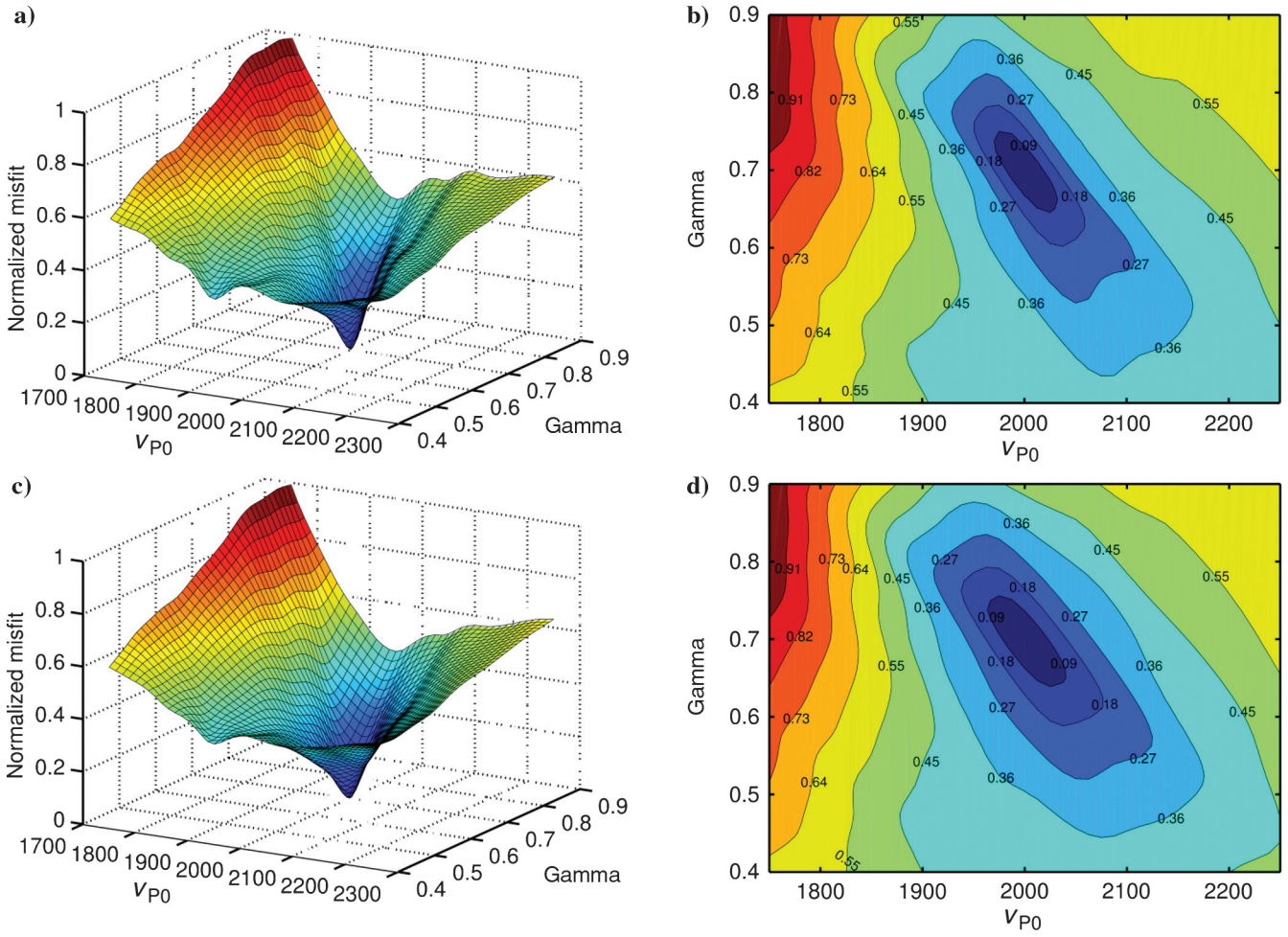

Figure 17. Misfit-function profiles (left column) and associated level sets (right column) for the two-parameter problem: (a) graph-transform-based optimal transport misfit function with $\tau = 1$ in a 1D setting (trace-by-trace comparison), (b) associated level sets, (c) graph-transform-based optimal transport misfit function with $\tau = 1$ in a 2D setting (shot gather comparison), and (d) associated level sets.

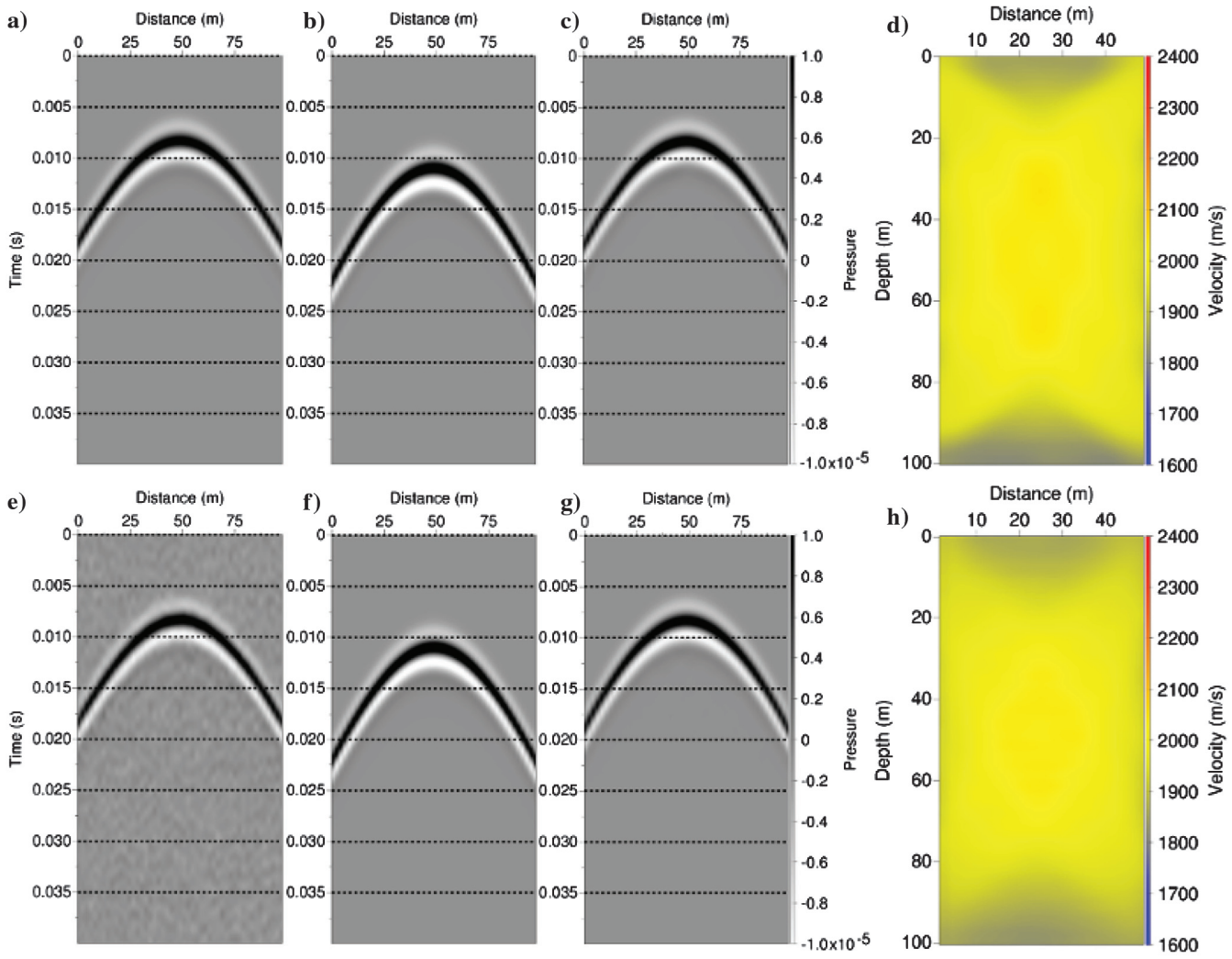

Figure 18. Transmission case study with Gaussian noise on the data. S/N is equal to 10, and the noise has been filtered in the frequency band of the data. (a) Reference data without noise, (b) data in the initial model at $V_P = 1800$ m/s, (c) data in the estimated model with the graph-transform-based optimal transport misfit function, (d) corresponding estimated model. (e) Reference data with noise, (f) data in the initial model at $V_P = 1800$ m/s, (g) data in the estimated model with the graph-transform-based optimal transport misfit function, (h) corresponding estimated model.

and also more challenging. It first implies the design of accurate and flexible modeling tools for viscoelastic wave propagation, including anisotropic effects and correctly accounting for the free-surface condition. We are currently investigating this topic through the design of a spectral element modeling and inversion code (Trinh et al., 2017). Second, it also implies moving toward multiparameter FWI, which requires specific techniques to mitigate crosstalks between parameters (Operto et al., 2013; Métivier et al., 2015; Yang et al., 2018).

As a preliminary investigation on the sensitivity of the misfit function to correct amplitude estimation, we present in Figure 18 results obtained for the crosshole case study, introducing noise in the observed data, with a signal to noise ratio (S/N) equal to 10. The injected noise follows a Gaussian distribution, and it is filtered in the frequency band of the observed data (a cutoff is introduced at 650 Hz). The regularization through the Gaussian smoothing of the gradient is increased in the vertical direction to mitigate the introduction of noise, as would be done for a real application. The results obtained with and without noise using the OT-GS misfit function show that the introduction of noise only marginally degrades the P-wave velocity estimation (Figure 18d and 18h). In particular, the estimation is still better than what is obtained on data without noise, using L^2 or KR misfit functions. This is an encouraging preliminary test because it reveals that the sensitivity to noise might not be a crucial issue for the OT-GS misfit function.

CONCLUSION AND PERSPECTIVES

The interest of applying optimal transport as a measure of distance between seismic data in the framework of FWI relies on its convexity with respect to shifted patterns. This property is expected to provide convex misfit functions with respect to the subsurface velocity: Low-wavenumber perturbations of this parameter mainly influence the kinematics of waves. The main difficulty for applying optimal transport to seismic data comes from its oscillatory nature: Optimal transport is based on the assumption that the compared quantities are positive.

In this study, we first review the different approaches that have been proposed recently to overcome this issue and explain their current limitations. These approaches are mainly based on the prior transformation of the data to make them positive or through an optimal transport extension for signed quantities. They face the following problems:

1) Nondifferentiability: The transformation of the data is a not differentiable operation, yielding a nondifferentiable misfit function.
2) Information loss: The shape of the data is strongly affected, leading to neglect of weaker amplitude events.
3) Convexity loss: The resulting misfit function is no longer convex with respect to time shifts.

Thus, we propose an alternative strategy that consists of comparing the seismic data through the optimal transport in the graph space. Instead of considering the seismic traces as 1D oscillatory functions of time, we consider them as discrete cloud of points in a 2D space. The optimal transport distance can thus be computed between the different cloud of points representing the data. This process amounts to consider the amplitude as a geometric feature of the data. This implies, at least formally, that the shape of the data is not altered by the process.

Several numerical experiments illustrate the properties of this strategy. In particular, in a 2D acoustic setting, the misfit function appears to be more convex with respect to the P-wave velocity than the L^2 misfit function and the KR misfit function promoted in previous studies.

This preliminary work thus indicates that the OT-GS strategy is promising for a better application of the optimal transport distance to FWI. However, further work is required to assess the feasibility of this strategy before its application to real data. A more efficient numerical strategy would be required for large-scale 3D FWI because the computational complexity increases through the introduction of an additional dimension to the data to sample its amplitude. A specific study of the sensitivity of the method with respect to incorrect amplitude prediction is also required. Provided that these two questions receive positive answers, the method might be an interesting tool for real-data applications.

ACKNOWLEDGMENTS

This study was partially funded by the SEISCOPE consortium (http://seiscope2.osug.fr), sponsored by AKERBP, CGG, CHEVRON, EXXON-MOBIL, JGI, SHELL, SINOPEC, STATOIL, TOTAL, and WOODSIDE. 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.

APPENDIX A

FAST COMPUTATION FOR THE TWO-WASSERSTEIN DISTANCE IN THE 1D CASE

For a given positive continuous function $f(x)$, with $x \in \mathbb{R}$ (1D assumption), its cumulative distribution function is defined by

$$C_f(x) = \int_{-\infty}^{x} f(u)du. \tag{A-1}$$

For two positive continuous functions $f(x)$ and $g(x)$, the p-Wasserstein distance between $f(x)$ and $g(x)$ is given analytically by

$$W_p(f,g) = \int_0^1 |C_f^{-1}(x) - C_g^{-1}(x)|^p dx, \tag{A-2}$$

for $p \geq 1$, where C_f^{-1} denotes the inverse cumulative function of C_f, such that

$$C_f^{-1}(C_f(x)) = x.$$

As f is positive and continuous, $C_f(x)$ is monotonically increasing, which makes its inverse $C_f^{-1}(x)$ easy to compute numerically.

APPENDIX B

THE MAININI DECOMPOSITION IN THE PARTICULAR CASE OF THE ONE-WASSERSTEIN DISTANCE

The standard dual formulation of the Kantorovich relaxation 2 is given by

$$W_p(f,g) = \left(\max_{\varphi,\psi} \int_X \varphi(x) f(x) dx + \int_X \psi(x) g(x) dx \right)^{1/p}, \tag{B-1}$$

where φ and ψ satisfy the constraints

$$\varphi(x) + \psi(x') \leq \|x - x'\|^p. \tag{B-2}$$

In the particular case of the one-Wasserstein distance, this dual formulation simplifies to

$$W_1(f,g) = \max_{\varphi \in \mathrm{Lip}_1(X)} \int_X \varphi(x)(f(x) - g(x)dx, \tag{B-3}$$

where $\mathrm{Lip}_1(X)$ is the space of one-Lipschitz functions for the ground distance $\|.\|$ defined in equation 26 (for a proof, see Abrosio et al., 2008; Santambrogio, 2015).

Considering nonpositive functions f and g, we can apply the Mainini decomposition 12 and compute the one-Wasserstein distance between $\tilde{f}$ and $\tilde{g}$ defined by

$$\tilde{f} = f^+ + g^-, \quad \tilde{g} = g^+ + f^-. \tag{B-4}$$

Using the simplified dual formulation B-3, we obtain

$$\begin{aligned}W_1(\tilde{f},\tilde{g}) &= \max_{\varphi\in \mathrm{Lip}_1(X)} \int_X \varphi(x)(\tilde{f}(x)-\tilde{g}(x))dx,\\ &= \max_{\varphi\in \mathrm{Lip}_1(X)} \int_X \varphi(x)(f^+(x)+g^-(x)-g^+(x)-f^-(x))dx,\\ &= \max_{\varphi\in \mathrm{Lip}_1(X)} \int_X \varphi(x)(f^+(x)-f^-(x)-(g^+(x)-g^-(x)))dx,\\ &= \max_{\varphi\in \mathrm{Lip}_1(X)} \int_X \varphi(x)(f(x)-g(x)))dx,\\ &= W_1(f,g). \end{aligned} \tag{B-5}$$

This equality shows that in the case of nonpositive functions satisfying the mass conservation assumption, computing the one-Wasserstein distance through its simplified dual formulation is equivalent to computing the one-Wasserstein distance extended to nonpositive functions through the Mainini decomposition.

APPENDIX C

ADJOINT SOURCE COMPUTATION THROUGH THE LAGRANGIAN APPROACH

For the sake of simplicity, we consider here the case of a unique source $N_s = 1$ because the generalization to an arbitrary number of sources is straightforward by summation. We introduce the Lagrangian operator $L(m, u, d_{\mathrm{cal},r}, \lambda, \mu_r)$ as

$$\begin{aligned}L(m,u,d_{\mathrm{cal},r},\lambda,\mu_r) = &\sum_{r=1}^{N_r} W_p(\mathcal{G}_\sigma(d_{\mathrm{obs},r}),\mathcal{G}_\sigma(d_{\mathrm{cal},r}))+\\ &(A(m)u_s - b_s,\lambda)+\\ &\sum_{r=1}^{N_r}(R_r u - d_{\mathrm{cal},r},\mu_r).\end{aligned} \tag{C-1}$$

For $u[m]$ solution of the wave-propagation problem 21 and $d_{\mathrm{cal},r}[m]$ solution of equation 20, we have

$$L(m,u[m],d_{\mathrm{cal},r}[m],\lambda,\mu_r) = f(m), \tag{C-2}$$

and therefore

$$\frac{\partial L(m,u[m],d_{\mathrm{cal},r}[m],\lambda,\mu_r)}{\partial m} = \nabla f(m). \tag{C-3}$$

Expanding equation C-3 yields

$$\begin{aligned}&\frac{\partial L(m,u[m],d[m]_{\mathrm{cal},r},\lambda,\mu_r)}{\partial u}\frac{\partial u[m]}{\partial m}\\ &+\sum_{r=1}^{N_r}\frac{\partial L(m,u[m],d_{\mathrm{cal},r}[m],\lambda,\mu_r)}{\partial d_{\mathrm{cal},r}}\frac{\partial d_{\mathrm{cal},r}[m]}{\partial m}\\ &+\left(\frac{\partial A}{\partial m}u_s(x,t),\lambda\right) = \nabla f(m).\end{aligned} \tag{C-4}$$

We define $\lambda[m]$ and $\mu_r[m]$, such that

$$\begin{aligned}\frac{\partial L(m,u[m],d[m]_{\mathrm{cal},r},\lambda,\mu_r)}{\partial u} &= 0,\\ \frac{\partial L(m,u[m],d_{\mathrm{cal},r}[m],\lambda,\mu_r)}{\partial d_{\mathrm{cal},r}} &= 0,\end{aligned} \tag{C-5}$$

which yields

$$A(m)^\dagger\lambda = \sum_{r=1}^{N_r} R_r^\dagger \mu_r[m] \tag{C-6}$$

and

$$\mu_r = \frac{\partial W_p(\mathcal{G}_\sigma(d_{\mathrm{obs},r}),\mathcal{G}_\sigma(d_{\mathrm{cal},r}))}{\partial d_{\mathrm{cal},r}}. \tag{C-7}$$

Equations C-6 and C-7 are related to a standard result in FWI: The adjoint sources μ_r are given by the derivative of the misfit function with respect to the calculated data. In this study, we focus on the W_1 distance. Using its dual formulation, it can be computed as

$$\begin{aligned}&W_1(\mathcal{G}_\sigma(d_{\mathrm{obs}}),\mathcal{G}_\sigma(d_{\mathrm{cal}}))\\ &= \max_{\varphi\in \mathrm{Lip}_1(X)}\int_t\int_x \varphi(x,t)(d_{\mathrm{obs}}^{\mathcal{G}_\sigma}(x,t)-d_{\mathrm{cal}}^{\mathcal{G}_\sigma}(x,t))dxdt.\end{aligned} \tag{C-8}$$

We define the function

$$\begin{aligned}&h:(\varphi,\beta)\to h(\varphi,\beta),\\ &\mathrm{Lip}_1(X)\times\mathcal{C}^\infty(\mathbb{R}^2,\mathbb{R}_*^+)\to\mathbb{R},\end{aligned} \tag{C-9}$$

where

$$h(\varphi,\beta) = \int_t\int_x \varphi(x,t)(d_{\mathrm{obs}}^{\mathcal{G}_\sigma}(x,t)-\beta(x,t))dxdt. \tag{C-10}$$

We introduce $\bar{\varphi}(\mathbf{d_{cal}})$, such that for $\mathbf{d_{cal}}\in\mathbb{R}^N$

$$\bar{\varphi}(\mathbf{d_{cal}})(x,t) = \arg\max_{\varphi\in \mathrm{Lip}_1(X)} h(\varphi,\mathcal{G}^\sigma(\mathbf{d_{cal}})), \tag{C-11}$$

and the function $q(\mathbf{d_{cal}})$, such that

$$q:\mathbf{d_{cal}}\to q(\mathbf{d_{cal}}) = h(\bar{\varphi}(\mathbf{d_{cal}}),\mathcal{G}^\sigma(\mathbf{d_{cal}})),\quad \mathbb{R}^N\to\mathbb{R}. \tag{C-12}$$

From the above definitions, we have

$$q(\mathbf{d_{cal}}) = W_1(\mathcal{G}^\sigma(d_{\mathrm{obs}}),\mathcal{G}^\sigma(\mathbf{d_{cal}})). \tag{C-13}$$

We are interested in the expression of the derivatives of q with respect to $\mathbf{d_{cal}}$. For any $\mathbf{v}\in\mathbb{R}^N$, we have

$$q(\mathbf{d_{cal}}+\mathbf{v}) = h(\bar{\varphi}(\mathbf{d_{cal}})+\tfrac{\partial\bar{\varphi}(\mathbf{d_{cal}})}{\partial\mathbf{d_{cal}}}.\mathbf{v},\mathcal{G}^\sigma(\mathbf{d_{cal}})+\tfrac{\partial\mathcal{G}^\sigma(\mathbf{d_{cal}})}{\partial\mathbf{d_{cal}}}.\mathbf{v})+\underset{\mathbf{v}\to 0}{o}(\|\mathbf{v}\|^2). \tag{C-14}$$

From the bilinearity of h, we have

$$q(\mathbf{d_{cal}}+\mathbf{v}) = h(\bar{\varphi}(\mathbf{d_{cal}}),\mathcal{G}^{\sigma}(\mathbf{d_{cal}})) + h\left(\frac{\partial\bar{\varphi}(\mathbf{d_{cal}})}{\partial\mathbf{d_{cal}}}.\mathbf{v},\mathcal{G}^{\sigma}(\mathbf{d_{cal}})\right) + h\left(\bar{\varphi}(\mathbf{d_{cal}}),\frac{\partial\mathcal{G}^{\sigma}(\mathbf{d_{cal}})}{\partial\mathbf{d_{cal}}}.\mathbf{v}\right) + \underset{\mathbf{v}\to 0}{o}(\|\mathbf{v}\|^2). \quad \text{(C-15)}$$

From the definition of $\bar{\varphi}(\mathbf{d_{cal}})$, we have $\partial\bar{\varphi}(\mathbf{d_{cal}})/\partial\mathbf{d_{cal}} = 0$; therefore,

$$q(\mathbf{d_{cal}}+\mathbf{v}) = q(\mathbf{d_{cal}}) + h\left(\bar{\varphi}(\mathbf{d_{cal}}),\frac{\partial\mathcal{G}^{\sigma}(\mathbf{d_{cal}})}{\partial\mathbf{d_{cal}}}.\mathbf{v}\right) + \underset{\mathbf{v}\to 0}{o}(\|\mathbf{v}\|^2). \quad \text{(C-16)}$$

Expanding formula C-16 yields

$$\begin{aligned} q(\mathbf{d_{cal}}+\mathbf{v}) &= q(\mathbf{d_{cal}}) + \int_t\int_x \bar{\varphi}(\mathbf{d_{cal}})(x,t) \left(\sum_{i=1}^{N}\mathbf{v}_i\frac{\partial\mathcal{G}^{\sigma}(\mathbf{d_{cal}})}{\partial d_{\mathrm{cal},i}}\right)dxdt + \underset{\mathbf{v}\to 0}{o}(\|\mathbf{v}\|^2), \\ &= q(\mathbf{d_{cal}}) + \sum_{i=1}^{N}\mathbf{v}_i\int_t\int_x \bar{\varphi}(\mathbf{d_{cal}})(x,t) \left(\frac{\partial\mathcal{G}^{\sigma}(\mathbf{d_{cal}})}{\partial d_{\mathrm{cal},i}}\right)dxdt + \underset{\mathbf{v}\to 0}{o}(\|\mathbf{v}\|^2), \\ &= q(\mathbf{d_{cal}}) + \left\langle \mathbf{v}, \int_t\int_x \bar{\varphi}(\mathbf{d_{cal}})(x,t) \left(\frac{\partial\mathcal{G}^{\sigma}(\mathbf{d_{cal}})}{\partial\mathbf{d_{cal}}}\right)dxdt\right\rangle + \underset{\mathbf{v}\to 0}{o}(\|\mathbf{v}\|^2), \end{aligned} \quad \text{(C-17)}$$

where the scalar product in $\mathbb{R}^N$ is denoted by $\langle .,. \rangle$. From equation C-17, we see that

$$\frac{\partial q}{\partial d_{\mathrm{cal},i}} = \int_t\int_x \bar{\varphi}(\mathbf{d_{cal}})(x,t)\left(\frac{\partial\mathcal{G}^{\sigma}(\mathbf{d_{cal}})}{\partial d_{\mathrm{cal},i}}\right)dxdt. \quad \text{(C-18)}$$

In addition, we have

$$\frac{\partial\mathcal{G}^{\sigma}(\mathbf{d_{cal}})}{\partial d_{\mathrm{cal},i}} = \frac{1}{2\pi\sigma_x\sigma_t N}\exp\left(-\frac{(t-t_i)^2}{2\sigma_t^2}\right) \exp\left(-\frac{(x-d_{\mathrm{cal},i})^2}{2\sigma_x^2}\right)\frac{x-d_{\mathrm{cal},i}}{2\sigma_x^2}, \quad \text{(C-19)}$$

from where comes equation 28.

REFERENCES

Adams, J. C., 1989, MUDPACK: Multigrid portable FORTRAN software for the efficient solution of linear elliptic partial differential equations: Applied Mathematics and Computation, **34**, 113–146, doi: 10.1016/0096-3003(89)90010-6.

Aleardi, M., and A. Mazzotti, 2016, 1D elastic full-waveform inversion and uncertainty estimation by means of a hybrid genetic algorithm: Gibbs sampler approach: Geophysical Prospecting, **65**, 64–85, doi: 10.1111/1365-2478.12397.

Ambrosio, L., N. Gigli, and G. Savaré, 2008, Gradient flows: In metric spaces and in the space of probability measures: Springer Science & Business Media.

Ambrosio, L., E. Mainini, and S. Serfaty, 2011, Gradient flow of the Chapman Rubinstein Schatzman model for signed vortices: Annales de l'Institut Henri Poincaré (C) Non Linear Analysis, **28**, 217–246, doi: 10.1016/j.anihpc.2010.11.006.

Baek, H., H. Calandra, and L. Demanet, 2014, Velocity estimation via registration-guided least-squares inversion: Geophysics, **79**, no. 2, R79–R89, doi: 10.1190/geo2013-0146.1.

Benamou, J.-D., G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, 2015, Iterative Bregman projections for regularized transportation problems: SIAM Journal on Scientific Computing, **37**, A1111–A1138, doi: 10.1137/141000439.

Billette, F., and G. Lambaré, 1998, Velocity macro-model estimation from seismic reflection data by stereotomography: Geophysical Journal International, **135**, 671–690, doi: 10.1046/j.1365-246X.1998.00632.x.

Biondi, B., and W. Symes, 2004, Angle-domain common-image gathers for migration velocity analysis by wavefield-continuation imaging: Geophysics, **69**, 1283–1298, doi: 10.1190/1.1801945.

Bogachev, V. I., 2007, Measure theory: Springer I, II.

Borisov, D., and S. C. Singh, 2015, Three-dimensional elastic full waveform inversion in a marine environment using multicomponent ocean-bottom cables: A synthetic study: Geophysical Journal International, **201**, 1215–1234, doi: 10.1093/gji/ggv048.

Bourgeois, A., M. Bourget, P. Lailly, M. Poulet, P. Ricarte, and R. Versteeg, 1991, Marmousi model and data: 52nd Annual International Conference and Exhibition, EAGE, Extended Abstracts, 5–16.

Bozdağ, E., D. Peter, M. Lefebvre, D. Komatitsch, J. Tromp, J. Hill, N. Podhorszki, and D. Pugmire, 2016, Global adjoint tomography: First-generation model: Geophysical Journal International, **207**, 1739–1766, doi: 10.1093/gji/ggw356.

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, doi: 10.1111/j.1365-246X.2011.04970.x.

Brossier, R., S. Operto, and J. Virieux, 2009, Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion: Geophysics, **74**, no. 6, WCC105–WCC118, doi: 10.1190/1.3215771.

Bunks, C., F. M. Salek, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473, doi: 10.1190/1.1443880.

Combettes, P. L., and J.-C. Pesquet, 2011, Proximal splitting methods in signal processing, *in* H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, eds., Fixed-point algorithms for inverse problems in science and engineering: Springer, Springer optimization and its applications 49, 185–212.

Devaney, A., 1984, Geophysical diffraction tomography: IEEE Transactions on Geoscience and Remote Sensing, **GE-22**, 3–13, doi: 10.1109/TGRS.1984.350573.

Devaney, A. J., 1982, A filtered backprojection algorithm for diffraction tomography: Ultrasonic Imaging, **4**, 336–350, doi: 10.1177/016173468200400404.

Engquist, B., and B. D. Froese, 2014, Application of the Wasserstein metric to seismic signals: Communications in Mathematical Science, **12**, 979–988, doi: 10.4310/CMS.2014.v12.n5.a7.

Engquist, B., B. D. Froese, and Y. Yang., 2016, Optimal transport for seismic full waveform inversion: Communications in Mathematical Sciences, **14**, 2309–2330, doi: 10.4310/CMS.2016.v14.n8.a9.

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, doi: 10.1111/j.1365-246X.2008.03923.x.

Fichtner, A., B. L. N. Kennett, H. Igel, and H. P. Bunge, 2010, Full waveform tomography for radially anisotropic structure: New insights into present and past states of the Australasian upper mantle: Earth and Planetary Science Letters, **290**, 270–280, doi: 10.1016/j.epsl.2009.12.003.

Hale, D., 2013, Dynamic warping of seismic images: Geophysics, **78**, no. 2, S105–S115, doi: 10.1190/geo2012-0327.1.

Hansen, N., 2006, The CMA evolution strategy: A comparing review, *in* J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea, eds., Towards a new evolutionary computation: Springer, Studies in fuzziness and soft computing 192, 75–102.

Huang, J.-W., and G. Bellefleur, 2012, Joint transmission and reflection traveltime tomography using the fast sweeping method and the adjoint-state technique: Geophysical Journal International, **188**, 570–582, doi: 10.1111/j.1365-246X.2011.05273.x.

Jannane, M., W. Beydoun, E. Crase, D. Cao, Z. Koren, E. Landa, M. Mendes, A. Pica, M. Noble, G. Roeth, S. Singh, R. Snieder, A. Tarantola, and D. Trezeguet, 1989, Wavelengths of earth structures that can be resolved from seismic reflection data: Geophysics, **54**, 906–910, doi: 10.1190/1.1442719.

Jin, S., and R. Madariaga, 1993, Background velocity inversion with a genetic algorithm: Geophysical Research Letters, **20**, 93–96, doi: 10.1029/92GL02781.

Jin, S., and R. Madariaga, 1994, Nonlinear velocity inversion by a two-step Monte Carlo: Geophysics, **59**, 577–590, doi: 10.1190/1.1443618.

Kantorovich, L., 1942, On the transfer of masses: Doklady Akademii Nauk USSR, **37**, 7–8.
Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi, 1983, Optimization by simulated annealing: Science, **220**, 671–680, doi: 10.1126/science.220.4598.671.
Kitagawa, J., Q. Mérigot, and B. Thibert, 2017, Convergence of a Newton algorithm for semi-discrete optimal transport: ArXiv e–prints.
Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia, Expanded Abstracts, 206–220.
Lambaré, G., 2008, Stereotomography: Geophysics, **73**, no. 5, VE25–VE34, doi: 10.1190/1.2952039.
Lellmann, J., D. Lorenz, C. Schönlieb, and T. Valkonen, 2014, Imaging with Kantorovich-Rubinstein discrepancy: SIAM Journal on Imaging Sciences, **7**, 2833–2859, doi: 10.1137/140975528.
Luo, S., and P. Sava, 2011, A deconvolution-based objective function for wave-equation inversion: 81st Annual International Meeting, SEG, Expanded Abstracts, 2788–2792.
Luo, Y., and G. T. Schuster, 1991, Wave-equation traveltime inversion: Geophysics, **56**, 645–653, doi: 10.1190/1.1443081.
Ma, Y., and D. Hale, 2013, Wave-equation reflection traveltime inversion with dynamic warping and full waveform inversion: Geophysics, **78**, no. 6, R223–R233, doi: 10.1190/geo2013-0004.1.
Maharramov, M., A. I. Baumstein, Y. Tang, P. S. Routh, S. Lee, and S. K. Lazaratos, 2017, Time-domain broadband phase-only full-waveform inversion with implicit shaping: 87th Annual International Meeting, SEG, Expanded Abstracts, 1297–1301.
Mainini, E., 2012, A description of transport cost for signed measures: Journal of Mathematical Sciences, **181**, 837–855, doi: 10.1007/s10958-012-0718-2.
Martin, G. S., R. Wiley, and K. J. Marfurt, 2006, Marmousi2: An elastic upgrade for Marmousi: The Leading Edge, **25**, 156–166, doi: 10.1190/1.2172306.
Mérigot, Q., 2011, A multiscale approach to optimal transport: Computer Graphics Forum, **30**, 1583–1592, doi: 10.1111/j.1467-8659.2011.02032.x.
Métivier, L., and R. Brossier, 2016, The SEISCOPE optimization toolbox: A large-scale nonlinear optimization library based on reverse communication: Geophysics, **81**, no. 2, F11–F25, doi: 10.1190/geo2015-0031.1.
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, doi: 10.1190/tle35121060.1.
Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016b, Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion: Geophysical Journal International, **205**, 345–377, doi: 10.1093/gji/ggw014.
Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016c, An optimal transport approach for seismic tomography: Application to 3D full waveform inversion: Inverse Problems, **32**, 115008, doi: 10.1088/0266-5611/32/11/115008.
Métivier, L., R. Brossier, S. Operto, and J. Virieux, 2015, Acoustic multiparameter FWI for the reconstruction of P-wave velocity, density and attenuation: preconditioned truncated Newton approach: 85th Annual International Meeting, SEG, Expanded Abstracts, 1198–1203.
Monge, G., 1781, Memoire sur la théorie des déblais et de remblais, *in* Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année: 666–704.
Mulder, W., and R. E. Plessix, 2008, Exploring some issues in acoustic full waveform inversion: Geophysical Prospecting, **56**, 827–841, doi: 10.1111/j.1365-2478.2008.00708.x.
Nolet, G., 2008, A breviary of seismic tomography: Cambridge University Press.
Operto, S., R. Brossier, Y. Gholami, L. Métivier, V. Prieux, A. Ribodetti, and J. Virieux, 2013, A guided tour of multiparameter full waveform inversion for multicomponent data: From theory to practice: The Leading Edge, **32**, 1040–1054, doi: 10.1190/tle32091040.1.
Operto, S., A. Miniussi, R. Brossier, L. Combe, L. Métivier, V. Monteiller, A. Ribodetti, and J. Virieux, 2015, Efficient 3-D frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: Application to Valhall in the visco-acoustic vertical transverse isotropic approximation: Geophysical Journal International, **202**, 1362–1391, doi: 10.1093/gji/ggv226.
Operto, S., J. Virieux, J. X. Dessa, and G. Pascal, 2006, Crustal imaging from multifold ocean bottom seismometers data by frequency-domain full-waveform tomography: Application to the eastern Nankai trough: Journal of Geophysical Research, **111**, B09306, doi: 10.1029/2005JB003835.
Peter, D., D. Komatitsch, Y. Luo, R. Martin, N. Le Goff, E. Casarotti, P. Le Loher, F. Magnoni, Q. Liu, C. Blitz, T. Nissen-Meyer, P. Basini, and J. Tromp, 2011, Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes: Geophysical Journal International, **186**, 721–739, doi: 10.1111/j.1365-246X.2011.05044.x.
Pladys, A., R. Brossier, and L. Métivier, 2017, FWI alternative misfit functions: What properties should they satisfy: 79th Annual International Conference and Exhibition, EAGE, Extended Abstracts, Tu P1 01.
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, doi: http://10.0.4.87/j.1365-246X.2006.02978.x.
Plessix, R. E., and C. Perkins, 2010, Full waveform inversion of a deep water ocean bottom seismometer dataset: First Break, **28**, 71–78, doi: 10.3997/1365-2397.2010013.
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, doi: 10.1190/1.1444597.
Prieux, V., G. Lambaré, S. Operto, and J. Virieux, 2013, Building starting model for full waveform inversion from wide-aperture data by stereotomography: Geophysical Prospecting, **61**, 109–137, doi: 10.1111/j.1365-2478.2012.01099.x.
Qiu, L., J. Ramos-Martìnez, A. Valenciano, Y. Yang, and B. Engquist, 2017, Full-waveform inversion with an exponentially encoded optimal-transport norm: 87th Annual International Meeting, SEG, Expanded Abstracts, 1286–1290.
Sambridge, M., and K. Mosegaard, 2002, Monte Carlo methods in geophysical inverse problems: Reviews of Geophysics, **40**, 1–29, doi: 10.1029/2000RG000089.
Santambrogio, F., 2015, Optimal transport for applied mathematicians: Calculus of variations, PDEs, and modeling, Progress in nonlinear differential equations and their applications: Springer International Publishing.
Sava, P., and S. Fomel, 2006, Time-shift imaging condition in seismic migration: Geophysics, **71**, no. 6, S209–S217, doi: 10.1190/1.2338824.
Sen, M., and P. Stoffa, 1992, Rapid sampling of model space using genetic algorithms: Examples from seismic waveform inversion: Geophysical Journal International, **108**, 281–292, doi: 10.1111/j.1365-246X.1992.tb00857.x.
Shen, P., W. Symes, and C. Stolk, 2003, Differential semblance velocity analysis by wave-equation migration: 73rd Annual International Meeting, SEG, Expanded Abstracts, 2132–2135.
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, doi: 10.1046/j.1365-246X.2002.01645.x.
Sirgue, L., 2003, Inversion de la forme d'onde dans le domaine fréquentiel de données sismiques grand offset: Ph.D. thesis, Université Paris 11-Queen's University.
Sirgue, L., O. I. Barkved, J. Dellinger, J. Etgen, U. Albertin, and J. H. Kommedal, 2010, Full waveform inversion: The next leap forward in imaging at Valhall: First Break, **28**, 65–70, doi: 10.3997/1365-2397.2010012.
Swarztrauber, P. N., 1974, A direct method for the discrete solution of separable elliptic equations: SIAM Journal on Numerical Analysis, **11**, 1136–1150, doi: 10.1137/0711086.
Symes, W. W., 2008, Migration velocity analysis and waveform inversion: Geophysical Prospecting, **56**, 765–790, doi: 10.1111/j.1365-2478.2008.00698.x.
Symes, W. W., and M. Kern, 1994, Inversion of reflection seismograms by differential semblance analysis: Algorithm structure and synthetic examples: Geophysical Prospecting, **42**, 565–614, doi: 10.1111/j.1365-2478.1994.tb00231.x.
Tape, C., Q. Liu, A. Maggi, and J. Tromp, 2010, Seismic tomography of the southern California crust based on spectral-element and adjoint methods: Geophysical Journal International, **180**, 433–462, doi: 10.1111/j.1365-246X.2009.04429.x.
Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266, doi: 10.1190/1.1441754.
Tavakoli, F. B., S. Operto, A. Ribodetti, and J. Virieux, 2017, Slope tomography based on Eikonal solvers and the adjoint-state method: Geophysical Journal International, **209**, 1629–1647, doi: 10.1093/gji/ggx111.
Thorpe, M., S. Park, S. Kolouri, G. Rohde, and D. Slepčev, 2016, A transportation distance for signal analysis: ArXiv e–prints.
Trinh, P. T., R. Brossier, L. Métivier, L. Tavard, and J. Virieux, 2017, Efficient 3D elastic FWI using a spectral-element method: 87th Annual International Meeting, SEG, Expanded Abstracts, 1533–1538.
van Leeuwen, T., and W. A. Mulder, 2010, A correlation-based misfit criterion for wave-equation traveltime tomography: Geophysical Journal International, **182**, 1383–1394, doi: 10.1111/j.1365-246X.2010.04681.x.
Vigh, D., K. Jiao, D. Watts, and D. Sun, 2014, Elastic full-waveform inversion application using multicomponent measurements of seismic data collection: Geophysics, **79**, no. 2, R63–R77, doi: 10.1190/geo2013-0055.1.
Villani, C., 2003, Topics in optimal transportation: Graduate Studies in Mathematics 50, AMS.

Villani, C., 2008, Optimal transport: Old and new: Grundlehren der mathematischen Wissenschaften, Springer.
Virieux, J., A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou, 2017, An introduction to full waveform inversion, *in* V. Grechka and K. Wapenaar, eds., Encyclopedia of exploration geophysics: SEG, R1-1–R1-40.
Virieux, J., and S. Operto, 2009, An overview of full waveform inversion in exploration geophysics: Geophysics, **74**, no. 6, WCC1–WCC26, doi: 10.1190/1.3238367.
Wang, Y., and Y. Rao, 2009, Reflection seismic waveform tomography: Journal of Geophysical Research, **114**, 1978–2012.
Warner, M., and L. Guasch, 2016, Adaptive waveform inversion: Theory: Geophysics, **81**, no. 6, R429–R445, doi: 10.1190/geo2015-0387.1.
Warner, M., A. Ratcliffe, T. Nangoo, J. Morgan, A. Umpleby, N. Shah, V. Vinje, I. Stekl, L. Guasch, C. Win, G. Conroy, and A. Bertrand, 2013, Anisotropic 3D full-waveform inversion: Geophysics, **78**, no. 2, R59–R80, doi: 10.1190/geo2012-0338.1.
Wu, R. S., and M. N. Toksöz, 1987, Diffraction tomography and multisource holography applied to seismic imaging: Geophysics, **52**, 11–25, doi: 10.1190/1.1442237.
Yang, P., R. Brossier, L. Métivier, J. Virieux, and W. Zhou, 2018, A time-domain preconditioned truncated Newton approach to multiparameter visco-acoustic full waveform inversion: SIAM Journal on Scientific Computing, **40**, B1101–B1130, doi: 10.1137/17M1126126.
Yang, P., R. Brossier, and J. Virieux, 2016, Wavefield reconstruction from significantly decimated boundaries: Geophysics, **81**, no. 5, T197–T209, doi: 10.1190/geo2015-0711.1.
Yang, Y., and B. Engquist, 2017, Analysis of optimal transport and related misfit functions in full-waveform inversion: 87th Annual International Meeting, SEG, Expanded Abstracts, 1291–1296.
Yilmaz, Ö., 1993, Seismic data processing: SEG.
Zhang, J., U. S. ten Brink, and M. N. Toksöz, 1998, Nonlinear refraction and reflection travel time tomography: Journal of Geophysical Research, **103**, 29743–29757, doi: 10.1029/98JB01981.
Zhu, H., 2017, Seismogram registration via Markov chain-Monte Carlo optimization and its applications in full-waveform inversion: 87th Annual International Meeting, SEG, Expanded Abstracts, 1336–1341.
Zhu, H., E. Bozdağ, D. Peter, and J. Tromp, 2012, Structure of the European upper mantle revealed by adjoint tomography: Nature Geoscience, **5**, 493–498, doi: 10.1038/ngeo1501.