submitted to *Geophys. J. Int.*

# Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion

L. Métivier [1], R. Brossier[2], Q. Mérigot[3], E. Oudet[4], J. Virieux[5]

[1] *Laboratoire Jean Kuntzmann (LJK), Univ. Grenoble Alpes, CNRS, France. E-mail: ludovic.metivier@ujf-grenoble.fr*

[2] *Institut des sciences de la Terre (ISTerre), Univ. Grenoble Alpes, France. E-mail: romain.brossier@ujf-grenoble.fr*

[3] *Laboratoire CEREMADE, Univ. Paris-Dauphine, CNRS, France. E-mail: quentin@mrgt.fr*

[4] *Laboratoire Jean Kuntzmann (LJK), Univ. Grenoble Alpes, France. E-mail: edouard.oudet@imag.fr*

[5] *Institut des sciences de la Terre (ISTerre), Univ. Grenoble Alpes, France. E-mail: jean.virieux@ujf-grenoble.fr*

**SUMMARY**

Full waveform inversion using the conventional $L^2$ distance to measure the misfit between seismograms is known to suffer from cycle skipping. An alternative strategy is proposed in this study, based on a measure of the misfit computed with an optimal transport distance. This measure allows to account for the lateral coherency of events within the seismograms, instead of considering each seismic trace independently, as is done generally in full waveform inversion. The computation of this optimal transport distance relies on a particular mathematical formulation allowing for the non-conservation of the total energy between seismograms. The numerical solution of the optimal transport problem is performed using proximal splitting techniques. Three synthetic case studies are investigated using this strategy: the Marmousi 2 model, the BP 2004 salt model, and the Chevron 2014 benchmark data. The results emphasize interesting properties of the optimal transport distance. The associated misfit function is less prone to cycle skipping. A workflow is designed to reconstruct accurately the salt structures in the BP 2004 model, starting from an initial model containing no information about these structures. A high resolution P-wave velocity estimation is built from the Chevron 2014 benchmark data, following a frequency continuation strategy. This estimation explains accurately the data. Using the same workflow, full waveform inversion based on the $L^2$ distance converges toward a local minimum. These results yield encouraging perspectives regarding the use of the optimal transport distance for full waveform inversion: the sensitivity to the accuracy of the initial model is reduced, the reconstruction of complex salt structure is made possible, the method is robust to noise, and the interpretation of seismic data dominated by reflections is enhanced.

**Key words:** Optimal transport, controlled source seismology, computational seismology, wave propagation, inverse theory.

## 1 INTRODUCTION

Full waveform inversion (FWI) is a data fitting procedure aiming at computing high resolution estimations of subsurface parameters. The formalism of this method, based on the minimization of the misfit between observed and synthetic data, yields the possibility for estimating any parameter influencing the propagation of seismic waves: P- and S-wave velocities, density, attenuation, anisotropy parameters. In current applications, at the regional or global scale in seismology, and at the exploration scale in seismic imaging, FWI is mainly used as a high resolution velocity model building method (Fichtner et al. 2010; Tape et al. 2010; Peter et al. 2011; Sirgue et al. 2010; Plessix & Perkins 2010; Zhu et al. 2012; Warner et al. 2013; Vigh et al. 2014; Borisov & Singh 2015; Operto et al. 2015). As opposed to conventional tomography methods based on the matching of travel-times only, FWI aims at taking into account the whole recorded signal: all the seismic events (diving waves, pre-and post critical reflections, converted waves) are considered, as well as their amplitude, in the process of estimating the velocity. As a consequence, higher resolution estimates are expected compared to tomography methods, up to the theoretical limit of half the shortest wavelength of the recorded signal (Devaney 1984).

The mismatch between observed and synthetic seismograms is usually computed as the $L^2$ norm of their difference. This is referred to as the $L^2$ distance in the following (the use of the $L^1$ norm and the hybrid $L^1/L^2$ Huber norm has also been promoted for interpreting noisy data in Brossier et al. (2010)). The minimization of this distance is performed through quasi-Newton methods (Nocedal & Wright 2006), involving the computation of the gradient and an approximation of the inverse Hessian operator (Pratt et al. 1998; Métivier et al. 2013, 2014a)

The time-domain formalism of FWI has been introduced by Lailly (1983) and Tarantola (1984). The limitations of FWI as a high resolution velocity model building tool from reflection seismic data have been identified few years after. In Jannane et al. (1989), the sensitivity of the seismic signal with respect to low wavenumber and high wavenumber perturbations of the velocity model is studied. While high wavenumber perturbations have mainly an effect on the amplitude of the signal, low wavenumber variations of the velocity are responsible for shifting in time the seismic traces, mainly influencing the travel-time of the seismic events. Hence, from an inverse problem point of view, reconstructing the large-scale, smooth components of the velocity model, requires to match these travel-time shifts. In addition, this reconstruction should be achieved before injecting high wavenumber in the reconstruction.

Unfortunately, the $L^2$ distance, based on a sample by sample comparison, is not adapted to capture the time shifts between two oscillatory signals. The two signals should have approximately the same shape (prediction of the same events) and the time shift should be no larger than half the period of the

signal. These requirements lead conventional FWI to focus (at least in the first stage of the inversion) on low frequency transmitted waves such as diving waves. These waves sample the subsurface without being reflected, therefore the difference between predicted and observed diving waves should be mainly due to shifts in time of the seismic events. However, if these time shifts are too large, reducing the $L^2$ distance between the signals through a local optimization technique will generate a wrong velocity model which matches the data with one to several phase shifts. This phenomenon is commonly known as cycle skipping. This is the reason why the accuracy of the initial model is of primary importance in conventional FWI: it should be kinematically compatible with the data, i.e. the phase of the main seismic events should be predicted within half a period.

Mitigating this strong dependence on the accuracy of the starting model is a long term issue in FWI. A first strategy, proposed by Pratt (1999) in the frequency-domain, consists in matching the lowest frequency components of the data as a preliminary step. This increases the attraction valley of the misfit function as, in this case, the initial velocity model should only explain the data up to half the period corresponding to the low frequency components that have been extracted. Following a hierarchical approach, the result of this first inversion serves as an initial model for an inversion of data containing higher frequencies. This procedure can be iterated until the whole seismic data has been interpreted. This is the strategy followed for instance in Bunks et al. (1995); Sirgue & Pratt (2004) and Operto et al. (2004).

This hierarchical approach can be complemented with offset and time-windowing strategies. Time-windowing is used to select the diving waves and remove the reflected energy from the observed seismograms. The offset is increased progressively, as large offsets correspond to diving waves traveling across a long distance between the subsurface, therefore containing a large number of oscillations, and more subject to cycle skipping. Time-windowing and offset selection is also known as layer stripping technique: the shallow part of the subsurface is first reconstructed, the depth of investigation being progressively increased by this data selection strategy. Examples of applications can be found for instance in Shipp & Singh (2002); Wang & Rao (2009) in the 2D acoustic approximation, or in Brossier et al. (2009) for the interpretation of onshore data in the 2D elastic approximation.

Despite these successful applications, the hierarchical approach does not really overcome the cycle skipping limitation. Instead, the data interpretation is re-organized in such a way that this limitation does not preclude the estimation of the velocity through FWI. Commonly encountered difficulties for real data application preventing this strategy to produce reliable velocity estimations encompass: the impossibility of building an accurate enough and kinematically compatible initial velocity model, the presence of strong noise corrupting the low frequency part of the data, or offset limitations in the acquisition design.

In the last decades, several attempts have been made to modify the FWI misfit function itself, to avoid comparing the seismic signal using the $L^2$ distance, and to yield a more robust, convex misfit function, less prone to cycle skipping. Two classes of strategies designed to achieve this objective can be identified, referred to as data-domain and image-domain techniques in the following.

The underlying concept of data-domain technique relies so far on a hybridization between tomography methods and FWI. These hybrid methods try to emphasize the matching of travel-times instead of the full signal, to recover the properties of tomography methods, while still benefiting from the expected high resolution power of FWI. One of the first attempt in this direction is the design of the wave-equation tomography (WETT) proposed by Luo & Schuster (1991). This is a tomography method, aiming at matching travel-times. However, while classical tomography methods rely on travel-time picking in the observed data (a possibly heavy pre-processing step) and the computation of travel-times through asymptotic approaches for instance, the travel-times misfit is directly estimated from the cross-correlation of the observed and synthetic traces. This method is interesting as it bridges the gap between tomography and FWI from a formal point of view: a full wave modeling engine is used to compute the synthetic data, and the method can be interpreted as a modification of the FWI misfit function, making possible to use the adjoint formalism to compute the associated gradient, as is commonly done in FWI. Originating from exploration geophysics, this strategy has been adopted by the seismology community as the finite-frequency tomography method (Dahlen et al. 2000; Montelli et al. 2004; Tromp et al. 2005; Nolet 2008).

However, exploiting WETT results as an initial model for FWI is not straightforward. It is well known that the resolution of the tomography method may be too low for producing an accurate enough starting model for FWI (Claerbout 1985). A sufficient accuracy of the initial model is not guaranteed and cycle skipping could still prevent FWI to converge to a reliable estimation. Second, in the presence of non-predicted events (i.e. reflections), the estimation of the time-shifts through cross-correlation collapses. Indeed, evaluating time-shifts between two traces through cross-correlation requires that the signal have approximately the same shape.

While the first difficulty is intrinsic to tomography method, an attempt to enhance the robustness of the automatic travel-time misfit computation through warping has been recently proposed by Ma & Hale (2013). Dynamic image warping is a technology originally designed for pattern recognition in signal processing. In a recent study, Hale (2013) has demonstrated that this method could be applied to determine time shifts between seismograms.

More recently, the design of a misfit function based on deconvolution has been proposed by Luo & Sava (2011). The method has been initially designed to overcome another limitation of cross-correlation based tomography. Luo & Sava (2011) recognize that standard implementations of this

method using a penalization of the nonzero time lags, as proposed for instance by van Leeuwen & Mulder (2010), make the implicit assumption that the seismic data has been acquired with an impulsive source with an infinite spectrum. When applied to real data acquired with band-limited sources, this could result in non negligible artifacts in the gradient. To this purpose, Luo & Sava (2011) propose to compute the travel-time misfit between the synthetic and observed data through a deconvolution of the synthetic data by the observed data, instead of using a cross-correlation of the two signals. This method has started to be applied to realistic scale case-studies in seismic exploration and seems to provide a more robust misfit function, less prone to cycle skipping (Warner & Guasch 2014).

In seismology, other data-domain modifications of the misfit function have been proposed. Fichtner et al. (2008) propose to use a time-frequency analysis of the data through a Gabor transform in order to extract both the travel-times and the amplitude envelope information from the seismic signal. This allows to define a misfit function as a sum of two terms measuring the misfit between travel-times and amplitude envelope separately. Compared to cross-correlation (Luo & Schuster 1991) or dynamic warping (Ma & Hale 2013), the extraction of the travel-times is performed following a more robust technique based on a multi-scale analysis in the time-frequency space. Besides, the information on the amplitude of the signal is not completely discarded as the amplitude envelope is also matched in the inversion process. A similar strategy has been proposed by Bŏzdag et al. (2011) where the amplitude and travel-time information are computed following a Hilbert transform. Compared to the Gabor transform, the Hilbert transform is a purely time-domain related technique, and should thus require less data processing than the Gabor transform. Both strategies can be used in combination with different time-windowing strategies (Maggi et al. 2009). Envelope inversion has also been investigated in the context of exploration seismology (Luo & Wu 2015).

Parallel to the development of these data-domain techniques, the development of image-domain techniques started with the design of Differential Semblance Optimization (DSO) (Symes & Kern 1994) and later on wave equation migration velocity analysis (WEMVA) (Sava & Biondi 2004a,b; Symes 2008). These methods rely on the separability of scales assumption: the velocity model is decomposed as the sum of a smooth background model and a high wavenumber reflectivity model. The reflectivity is related to the smooth background model through an imaging condition: it is the sum for each source of the cross-correlation between the incident wavefield and the back-propagated residuals computed in the smooth background velocity model. This imaging condition can be extended using either an offset selection (Symes & Kern 1994) or an illumination angle selection (Biondi & Symes 2004) in the residuals (the angles are easily accessible when the reflectivity is computed through asymptotic techniques), or a time lag in the cross-correlation (Faye & Jeannot 1986; Sava & Fomel 2006; Biondi & Almomin 2013). Within this framework, an extended image thus consists in a collec-

tion of reflectivity models depending on one of these additional parameters (offset, angle, time lag). This extended image is used to probe the consistency of the smooth background velocity model: the uniqueness of the subsurface implies that for the correct background, the energy should be focused in the image domain, either along the offset/angle dimension, or at zero lag. A new optimization problem is thus defined, either as the penalization of the defocusing of the energy, or as the maximization of the coherency of the energy in the image domain. The corresponding misfit function is minimized iteratively, following standard numerical optimization schemes. The main drawback of these approaches is related to their computational cost. A large number of migration operations has to be performed to build the extended image, and this has to be performed at each iteration of the reconstruction of the smooth background velocity model. This high computational cost seems to have precluded the use of these techniques for 3D waveform inversion up to now. It should also be noted that these methods are based on the assumption that only primary reflections will be used to generate the extended image through migration, which requires non negligible data pre-processing. Locally coherent events in the image-domain associated with, for instance, multiple reflections, would yield inconsistent smooth background velocity models (Lambaré 2002).

Recently, new data-domain modifications of the misfit function based on concepts developed in image processing have emerged. While Baek et al. (2014) promote the use of warping strategies, Engquist & Froese (2014) propose to replace the $L^2$ distance by the Wasserstein distance to compare seismic signals. The Wasserstein distance is a mathematical tool derived from the optimal transport theory, which has already numerous application in computational geometry and image processing (Villani 2003). The underlying idea is to see the comparison of two distributions as an optimal mapping problem. An optimization problem is thus solved to compute the distance between two distributions, also known as the Monge-Kantorovich problem. A cost is associated with all the mappings, accounting for instance for the sum of all the displacements required to map one distribution onto the other. The Wasserstein distance is computed as the minimal cost over the space of all the mappings. These mathematical concepts originate from the work of the French engineer Gaspard Monge at the end of the 18th century, in an attempt to conceive the optimal way of transporting sand to a building site. The Wasserstein distance is then used to define a misfit function measuring the discrepancy between predicted and observed data, which is minimized over the subsurface parameters to be reconstructed. The resulting strategy can thus be seen as a two-level optimization strategy with an outer level for the update of the subsurface parameters and an inner level for the computation of the misfit function using the Wasserstein distance.

In the study proposed by Engquist & Froese (2014), the properties of the Wasserstein distance for the comparison of 1D seismic signals are investigated. In particular, the convexity of the corresponding

misfit function with respect to time-shifts of the signal is emphasized. This can be well understood, as within this context, the measure of the distance is not based on the pure difference of the oscillatory signals, but on all the mappings that can shift and distort the original signal to map the targeted one. Therefore, an information on the travel-time shifts as well as on the amplitude variations of the signal is captured by this distance.

In this study, we are interested in an extension of this method to the comparison of entire seismograms, more precisely common shot-gathers, which are collections of traces corresponding to one seismic experiment. Compared to individual seismic traces, the shot-gathers (which can be seen as 2D images), contain important additional information as lateral coherency corresponds to identifiable seismic events, such as reflections, refraction, or diving waves. Hence, the aim of this study is twofold. The first objective is to present how shot-gathers can be compared using an optimal transport based distance. The second objective consists in demonstrating the interest of using such a distance in the context of FWI through different case studies.

The proposition from Engquist & Froese (2014) is to use the Monge-Ampère formulation of the optimal transport problem for comparing the Wasserstein distance between 1D traces, following earlier studies from Knott & Smith (1984) and Brenier (1991). The computation of the Wasserstein distance is brought back to the solution of the Monge-Ampère problem, a nonlinear system of partial-differential equations, which can be solved efficiently using finite-difference based method (Benamou et al. 2014) or semi-discrete strategies (Mérigot 2011). These are supposed to be amenable strategies for large scale optimal transport problems, however, they may still lack robustness to be extensively used within FWI. A more fundamental difficulty is related to the positivity of the signals and the energy conservation. Two underlying assumptions of the Wasserstein distance is that the compared signals should be positive, and that no energy is lost in the process of mapping one signal to the other. These two assumptions are not verified when comparing seismic signals. First, these are oscillatory signals, and the positivity cannot be satisfied. Second, regarding the energy conservation, aside the difficulty of predicting accurately the signal amplitude which requires an accurate information on the attenuation and the density of the subsurface together with the use of sophisticated forward modeling engines based on the visco-elasto-dynamic equations, there is no fundamental reason that the predicted data contains the same energy as the observed data. Generally, in the simple case of missing reflectors in the models, predicted seismograms will contain less energy than the observed ones. In addition, noise corrupts the data, which is in essence a non-predictable quantity. A strict conservation of the total energy is thus inappropriate for waveform inversion.

A new strategy is introduced in this study to overcome these difficulties. Instead of using the Wasserstein distance, a variant of this distance is used. This variant relies on the dual formulation of

the Monge-Kantorovich problem, and is defined as a maximization problem over the space of bounded functions with variations bounded by the unity (bounded 1-Lipschitz functions). This allows to overcome the restriction associated with the positivity and the strict conservation of the energy between the signals which are compared. An efficient numerical method is designed to compute this distance, making possible the comparison of realistic size seismograms, involving several thousands of time steps and receivers. This method uses an algorithm recently developed in the context of image processing, the Simultaneous Descent Method of Multipliers (SDMM), an instance of the Rockafellar proximal point algorithm (Rockafellar 1976) which is based on proximal splitting techniques (Combettes & Pesquet 2011).

The first synthetic case study on the Marmousi 2 benchmark model (Martin et al. 2006) emphasizes the properties of the misfit function based on the optimal transport distance compared to the misfit function based on the $L^2$ distance. The sensitivity of both strategies to the choice of the starting model is investigated. Better P-wave velocity estimations are systematically recovered when the optimal transport distance is used. The second synthetic case study is based on the BP 2004 model (Billette & Brandsberg-Dahl 2004). The presence of complex salt structures makes this benchmark model challenging for seismic imaging. Most of the energy of the seismic signal is reflected at the interface between the water and these structures, and few percent of the energy travels from the inside of the structures back to the receivers. Starting from a background model containing no information on the presence of the salt structures, a workflow is designed using the optimal transport distance misfit function allowing for a correct reconstruction of the salt bodies. This was not possible using a $L^2$ distance based misfit function. Finally, the third synthetic case study is presented on the benchmark dataset issued by Chevron in 2014. This 2D streamer elastic dataset is challenging for FWI as the maximum offset of $8$ km limits the depth of penetration of the diving waves to the first $3$ kilometers. The quality control on the data, the migrated images and the CIG show that the P-wave velocity estimation obtained with the optimal transport distance is reliable. The Chevron dataset also illustrates that the optimal transport distance is robust to noise, a nice property having its roots in the regularizing properties of the numerical solution of the optimal transport problem which is defined.

In the remainder of the study, the mathematical formalism for the computation of the Wasserstein distance is first introduced. Its definition is given, and a general presentation of the numerical method implemented for its numerical approximation is presented. For the sake of clarity, the technical details regarding the solution of the discrete optimal transport problem are presented in the Appendices B and C. On this basis, a strategy for computing the gradient of the optimal transport distance misfit function using the adjoint-state method is presented, and a numerical illustration on a schematic example using a borehole-to-borehole transmission acquisition is introduced. The three synthetic cases

studies mentioned previously are then presented to outline characteristic properties and performances of FWI based on the optimal transport distance. A discussion and a conclusion are given in the two last sections.

271 ## 2 THEORY

272 ### 2.1 Definition of the Wasserstein distance

Consider two functions $f(x)$ and $g(x)$ defined on a domain $X$ subset of $\mathbb{R}^d$, such that

$$f, g : X \longrightarrow \mathbb{R}, \quad X \subset \mathbb{R}^d, \tag{1}$$

and $M$ a function from $X$ to $X$

$$M : X \longrightarrow X. \tag{2}$$

The $L^p$ Wasserstein distance between $f$ and $g$, denoted by $W^p(f, g)$, is defined by a norm on $\mathbb{R}^d$, denoted by $\|.\|$, an exponent $p \geq 1$, and the constrained minimization problem

$$\begin{cases} W^p(f,g) = \min\limits_{M} \displaystyle\int_{x\in X} \|x - M(x)\|^p f(x)dx, & \text{(3a)} \\ \quad \text{where} \quad \forall A \subset X, \ \displaystyle\int_{x\in A} g(x)dx = \int_{M(x)\in A} f(x)dx. & \text{(3b)} \end{cases}$$

The equation (3b) is a constraint which specifies that $M$ belongs to the ensemble of all the mappings from $f$ to $g$. In this study, we consider the Wasserstein distance defined by the exponent $p = 1$ and the $\ell^1$ distance $\|.\|_1$ on $\mathbb{R}^d$ such that

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad \|x\|_1 = \sum_{i=1}^{d} |x_i|. \tag{4}$$

We denote this distance by $W^1(f, g)$. Instead of using the previous (primal) formulation given by equations (3a) and (3b), which involves a nonlinear constraint associated with energy conservation, the Wasserstein distance $W^1(f, g)$ has the interesting property that it can be computed through the solution of the (dual) linear problem

$$W^1(f,g) = \max_{\varphi \in \mathrm{Lip}_1} \int_{x\in X} \varphi(x)\left(f(x) - g(x)\right) dx, \tag{5}$$

where $\mathrm{Lip}_1$ is the space of 1-Lipschitz functions, such that

$$\forall (x, y) \in X, \quad |\varphi(x) - \varphi(y)| \leq \|x - y\|_1. \tag{6}$$

273 From this definition, one can see that the 1-Lipschitz property (6) ensures bounded variations of the
274 function and precludes fast variations and discontinuities of the function $\varphi$.

The dual definition of the Wasserstein distance $W^1(f, g)$ given in equation (5) can be found in classical optimal transport textbooks such as Evans (1997) or Villani (2008). The maximization problem (5) is well defined if and only if the energy between $f(x)$ and $g(x)$ is conserved in the sense

that

$$E_{f,g} \equiv \int_{x \in X} (f(x) - g(x))\, dx = 0. \tag{7}$$

Indeed, let $\varphi(x) = \alpha \in \mathbb{R}$ be a constant function. This function is 1-Lipschitz, and satisfies

$$\int_{x \in X} \varphi(x)\, (f(x) - g(x))\, dx = \alpha E_{f,g}. \tag{8}$$

Therefore, if (7) is not satisfied, the solution of (5) is the constant function equal to $\infty$ or $-\infty$ depending on the sign of $E_{f,g}$.

As the conservation of the energy can not be guaranteed in seismic imaging ($E_{f,g} \neq 0$ in practice), a generalization of the Wasserstein distance $W^1(f,g)$ for the non-conservative case is considered in this study. This generalization relies on an additional constraint: the function $\varphi(x) \in \mathrm{Lip}_1$ should be also bounded, such that

$$\exists c > 0, \quad \forall x \in X, \quad |\varphi(x)| \leq c. \tag{9}$$

This condition can be seen as a threshold: instead of increasing toward the infinity, the function is limited to reach a fixed, constant value $c$. The space of bounded 1-Lipschitz functions is denoted by $\mathrm{BLip}_1$ in the following. The distance defined between two functions $f$ and $g$ should thus be computed as the solution of the maximization problem

$$\widetilde{W}^1(f,g) = \max_{\varphi \in \mathrm{BLip}_1} \int_{x \in X} \varphi(x)\, (f(x) - g(x))\, dx. \tag{10}$$

Note that some theoretical links exist between the Wasserstein $W^1$ and the distance $\widetilde{W}^1$: see for instance the work of Hanin (1992). A mathematical analysis of this link is, however, beyond the scope of this study.

Common shot-gathers are collections of seismic traces recorded after the explosion of one source, in the time-receiver domain. As such, they can be considered as real functions defined in a two-dimensional space. The observed and calculated shot-gathers are denoted respectively by

$$d^s_{obs}(x_r, t) \quad \text{and} \quad d^s_{cal}[m](x_r, t). \tag{11}$$

The variable $x_r$ is associated with the receiver position and the variable $t$ corresponds to time. The superscript $s$ corresponds to the shot-gather number in a seismic survey containing $S$ shot-gathers. The dependence of the calculated data on to the model parameter $m$ is denoted by $[m]$. The following misfit function is thus introduced

$$f_{\widetilde{W}^1}(m) = \sum_{s=1}^{S} \widetilde{W}^1 \left( d^s_{cal}[m], d^s_{obs} \right), \tag{12}$$

where

$$\widetilde{W}^1\left(d^s_{cal}[m], d^s_{obs}\right) = \max_{\varphi \in \mathrm{BLip}_1} \int_t \int_{x_r} \varphi(x_r, t)\left(d^s_{cal}[m](x_r, t) - d^s_{obs}(x_r, t)\right) dx_r dt. \tag{13}$$

For comparison, the conventional $L^2$ misfit function is

$$f_{L^2}(m) = \sum_{s=1}^{S} \int_t \int_{x_r} \left|d^s_{cal}[m](x_r, t) - d^s_{obs}(x_r, t)\right|^2 dx_r dt. \tag{14}$$

## 2.2 Numerical computation of $\widetilde{W}^1(d_{cal}, d_{obs})$

The numerical computation of the solution to the problem (13) is presented here. The discrete analogous of the distance $\widetilde{W}^1$ distance is defined as

$$\begin{cases} \widetilde{W}^1\left(d_{cal}[m], d_{obs}\right) = \max\limits_{\varphi} \sum\limits_{i=1}^{N_r} \sum\limits_{j=1}^{N_t} \varphi_{ij}\left((d_{cal}[m])_{ij} - (d_{obs})_{ij}\right) \Delta t \Delta x_r & \text{(15a)} \\ \forall (i,j), \quad |\varphi_{ij}| < c, & \text{(15b)} \\ \forall (i,j),\ (k,l) \quad |\varphi_{ij} - \varphi_{kl}| < |(x_r)_i - (x_r)_k| + |t_j - t_l|\,. & \text{(15c)} \end{cases}$$

In (15), $N_r$ and $N_t$ are the number of receivers and discrete time steps respectively, and the standard discrete notations are used

$$(x_r)_i = (i-1) \times \Delta x_r, \quad t_j = (j-1) \times \Delta t, \quad \varphi_{ij} = \varphi((x_r)_i, t_j), \tag{16}$$

where $\Delta x_r$ and $\Delta t$ are the discretization steps in the receiver coordinate and time dimensions respectively.

With these notations, the total number of discrete points for the representation of one shot-gather is $N = N_t \times N_r$. The system (15) defines a linear programming problem involving $2N^2 + 2N$ linear constraints. From a computational point of view, the algorithmic complexity involved for the solution of such a problem would not be affordable for realistic size seismograms, which can involve thousands of receivers positions and discrete time steps, yielding a complexity $N = O(10^6)$. However, an equivalent discrete problem involving only $6N$ linear constraints can be derived by imposing only local constraints on $\varphi$ to enforce the 1-Lipschitz property. This yields the linear programming problem

$$\begin{cases} \widetilde{W}^1\left(d_{cal}[m], d_{obs}\right) = \max\limits_{\varphi} \sum\limits_{i=1}^{N_r} \sum\limits_{j=1}^{N_t} \varphi_{ij}\left((d_{cal}[m])_{ij} - (d_{obs})_{ij}\right) \Delta t \Delta x_r & \text{(17a)} \\ \forall (i,j), \quad |\varphi_{ij}| < c, & \text{(17b)} \\ \forall (i,j), \quad |\varphi_{i+1j} - \varphi_{ij}| < |(x_r)_{i+1} - (x_r)_i| = \Delta x_r & \text{(17c)} \\ \forall (i,j), \quad |\varphi_{ij+1} - \varphi_{ij}| < |t_{j+1} - t_j| = \Delta t. & \text{(17d)} \end{cases}$$

The two linear programming problems (15) and (17) are equivalent. This results from a particular property of the $\ell_1$ distance. The proof of this equivalence is given in appendix A.

The function $h_{d_{cal}[m],d_{obs}}(\varphi)$ is now introduced, such that

$$h_{d_{cal}[m],d_{obs}}(\varphi) = \sum_{i=1}^{N_r} \sum_{j=1}^{N_t} \varphi_{ij} \left( (d_{cal}[m])_{ij} - (d_{obs})_{ij} \right) \Delta t \Delta x_r. \tag{18}$$

Let $K$ be the unit hypercube of $\mathbb{R}^{3N}$

$$K = \left\{ x \in \mathbb{R}^{3N},\ |x_i| \leq 1,\ i = 1, \dots 3N \right\}. \tag{19}$$

The indicator function of $K$, denoted by $i_K$ is defined as

$$i_K(x) = \left| \begin{array}{lll} 0 & \text{if} & x \in K \\ +\infty & \text{if} & x \notin K. \end{array} \right. \tag{20}$$

With these notations, the linear programming problem (17) can be rewritten as

$$W^1 \left( d_{cal}[m], d_{obs} \right) = \max_{\varphi} \ \ h_{d_{cal}[m],d_{obs}}(\varphi) - i_K(A\varphi), \tag{21}$$

where the matrix $A$ is a $3N \times N$, sparse, rectangular matrix representing the constraints on $\varphi$ following the equations (17.b)-(17.d). Assuming an ordering of the discrete vectors $\varphi_{ij}$

$$\varphi = [\varphi_{11}, \varphi_{21}, \dots \varphi_{N_r 1}, \varphi_{12}, \dots, \varphi_{N_r N_t},] \tag{22}$$

the matrix $A$ is such that

$$(A\varphi)_k = \left| \begin{array}{lll} \dfrac{\varphi_k}{c} & \text{for} & k = 1, \dots, N \\ \dfrac{\varphi_{k+1} - \varphi_k}{\Delta x_r} & \text{for} & k = N+1, \dots, 2 \times N \\ \dfrac{\varphi_{k+N_r} - \varphi_k}{\Delta x_t} & \text{for} & k = 2 \times N + 1, \dots, 3 \times N. \end{array} \right. \tag{23}$$

The matrix $A$ is thus a column block matrix with one diagonal block and two bi-diagonal blocks. This pattern is due to the locality of the discrete constraints which are imposed. In the formulation (21), the constraints are encoded in the term $-i_K(A\varphi)$. Indeed, the linear programming problem amounts to the maximization of the function expressed in (21). According to the definition of $i_K$, the corresponding misfit function equals $-\infty$ as soon as one of the linear constraints is not respected, therefore any solution of the maximization problem has to satisfy these constraints.

Rewriting the problem (17) as the problem (21) recasts a linear programming problem into a convex non-smooth optimization problem. The advantage of such a transformation is that there exist efficient techniques to solve such convex non-smooth optimization problems, based on proximal splitting techniques. These techniques use the concept of proximity operator. For the sake of compactness, the definition of proximity operators is given in appendix B, as well as the proximity operator

of the functions $h_{d_{cal}[m],d_{obs}}$ and $i_K$, denoted by $\text{prox}_{h_{d_{cal}[m],d_{obs}}}$ and $\text{prox}_{i_K}$. These operators have a closed-form and can be calculated with a linear complexity, making them inexpensive to compute.

In this study, the problem (21) is solved using the simultaneous direction method of multipliers (SDMM) described in Combettes & Pesquet (2011), which is an instance of the proximal point algorithm (Rockafellar 1976). Following this method, the solution of (21) is obtained through the iterative scheme described in Algorithm 1. At each iteration of this algorithm, the proximity operators $\text{prox}_{h_{d_{cal}[m],d_{obs}}}$ and $\text{prox}_{i_K}$ are invoked, as well as the solution of a linear system involving the square matrix of size $N$

$$Q = I_N + A^T A, \tag{24}$$

where $I_N$ is the identity matrix of size $N$. The solution of these linear systems is the more intensive computational task in the SDMM algorithm, as the application of the proximity operators $\text{prox}_{h_{d_{cal}[m],d_{obs}}}$ and $\text{prox}_{i_K}$ has a linear complexity in number of operations and a negligible cost in terms of memory requirement.

The matrix $Q$ is a sparse square matrix of size $N$, symmetric positive definite by construction, related only to the definition of the equations (17.b)-(17.d). As a consequence, the matrix $Q$ remains constant throughout the whole FWI process. In a first attempt to design an efficient algorithm for the solution of (21), it can be interesting, in a pre-processing step, to factorize this matrix as a product $LL^T$, where $L$ is a lower triangular matrix, using a Cholesky decomposition. Under the assumption $N_t \simeq N_r$, this allows to benefit from a complexity in $O(N^{3/2})$ for the solution of these linear systems through forward and backward substitutions. However, the memory requirement associated with the storage of the factor $L$ is also in $O(N^{3/2})$, which is non negligible for realistic size problems for which the size $N$ can reach $O(10^6)$.

For this reason, an alternative method to solve the linear systems related to $Q$ is designed in this study, which takes advantage of the particular structure of $Q$. This method is adapted from the work of Buzbee et al. (1970). A reduction of the memory requirement from $O(N^{3/2})$ to $O(N)$ is achieved, while maintaining the same computational complexity as forward and backward substitution in $O(N^{3/2})$. In addition, while these operations are intrinsically sequential, the algorithm proposed in this study is based on matrix-vector products which can be easily parallelized. For the sake of compactness, the full description of this strategy is given in Appendix C.

### 2.3 Minimization of the optimal transport distance based misfit function and gradient computation

The minimization of the misfit function (12) is based on conventional quasi-Newton techniques. From an initial estimation $m_0$, these methods construct the sequence

$$m_{k+1} = m_k + \alpha_k \Delta m_k, \tag{25}$$

where $\alpha_k$ is a positive scalar parameter computed through a linesearch strategy (Nocedal & Wright 2006; Bonnans et al. 2006), and $\Delta m_k$ is a model update satisfying

$$\Delta m_k = -H_k \nabla f_{\widetilde{W}_1}(m_k). \tag{26}$$

In equation (26), $\nabla f_{\widetilde{W}_1}(m_k)$ is the gradient of the misfit function (12), and $H_k$ is an estimation of the inverse of its Hessian. In this study, this estimation is computed through the $l$-BFGS approximation (Nocedal 1980). This approximation is based on the values of the gradient at iteration $k$ and the $l$ previous iterations $k-1, \dots k-l+1$.

Therefore, the practical implementation of the proposed strategy in the FWI context only requires the capability of computing the misfit function $f_{\widetilde{W}_1}(m)$ and its gradient $\nabla f_{\widetilde{W}_1}(m)$. To this purpose, the adjoint-state technique is used (Lions 1968; Chavent 1974; Plessix 2006). For the sake of notation simplification, the case of one single shot-gather is considered here ($S = 1$), as the generalization to several shot-gathers is straightforward by summation.

The following Lagrangian function is introduced

$$\mathcal{L}(m, u, d_{cal}, \lambda, \mu) = \widetilde{W}^1 (d_{cal}, d_{obs}) + (F(m,u), \lambda)_{\mathcal{W}} + (Ru - d_{cal}, \mu)_{\mathcal{D}}, \tag{27}$$

where the standard Euclidean scalar product in the wavefield space and the data space is denoted by $(.,.)_{\mathcal{W}}$ and $(.,.)_{\mathcal{D}}$ respectively. The state variables are the incident wavefield, denoted by $u$, and the calculated data, denoted by $d_{cal}$. The adjoint variables are denoted by $\lambda$ and $\mu$. The extraction operator which maps the incident wavefield to the receiver locations is denoted by $R$. The two state equations relating the state variables and the model $m$ are

$$F(m,u) = 0, \quad d_{cal} = Ru. \tag{28}$$

Using the adjoint-state approach, the gradient $\nabla f_{\widetilde{W}_1}(m)$ is given by

$$\nabla f_{\widetilde{W}_1}(m) = \left( \frac{\partial F(m, \overline{u}(m))}{\partial m}, \overline{\lambda} \right), \tag{29}$$

where $\overline{u}(m)$ and $\overline{\lambda}(m)$ are respectively the incident and the adjoint wavefields satisfying the state equation and the adjoint state equation (see Plessix (2006) for the derivation of (29)).

The adjoint-state equations are obtained by canceling the derivatives of the Lagrangian function

with respect to the state variables. This gives

$$\begin{cases} \dfrac{\partial F(m,\overline{u}(m))^T}{\partial u}\lambda = -R^T\mu \\ \mu = \dfrac{\partial \widetilde{W}^1\left(d_{cal},d_{obs}\right)}{\partial d_{cal}}. \end{cases} \tag{30}$$

The first of these two equations involves the adjoint of the wave operator. The wavefield $\lambda$ thus corresponds to the back-propagation of the source term $-R^T\mu$. The second equation relates $\mu$ to the derivatives of the misfit function with respect to the calculated data. Therefore, as already noticed in Brossier et al. (2010) and Luo & Sava (2011) for instance, the modification of the misfit function only impacts the source term of the adjoint wavefield $\lambda$. Remarking that

$$\frac{\partial \widetilde{W}^1\left(d_{cal},d_{obs}\right)}{\partial d_{cal}} = \frac{\partial}{\partial d_{cal}}\left(\max_{\varphi\in \mathrm{BLip}_1}\int \varphi(x_r,t)\left(d_{cal}(x_r,t)-d_{obs}(x_r,t)\right)dx_r dt\right), \tag{31}$$

the secondary adjoint wavefield $\mu$ is simply given by

$$\mu = \arg\max_{\varphi\in \mathrm{BLip}_1}\int \varphi(x_r,t)\left(d_{cal}(x_r,t)-d_{obs}(x_r,t)\right)dx_r dt. \tag{32}$$

The difference between equations (32) and (13) should be emphasized here. The equation (13) defines $\widetilde{W}^1$ as the maximal value of the criterion over the space of bounded 1-Lipschitz functions. The equation (32) defines $\mu$ as the particular bounded 1-Lipschitz function for which this maximal value is reached. This is the meaning to be given to the notations $\max$ and $\arg\max$. Compared to a $L^2$ norm-based misfit function where $\mu$ would be the difference between the observed and calculated seismograms, here $\mu$ is computed as the maximizer of the optimal transport problem designed to compute the $\widetilde{W}^1$ distance between these seismograms.

This has the following consequence regarding the implementation of the proposed strategy. The additional computational cost related to the modification of the misfit function from the standard $L^2$ norm to the $\widetilde{W}^1$ distance is related to the solution of the maximization problem (21). This solution yields not only the misfit function value, which is the value of the criterion to be maximized, but also the adjoint variable $\mu$, which corresponds to the function $\varphi \in \mathrm{BLip}_1$ which achieves this maximization. Hence, one optimal transport problem is solved per source, and its solution allows to compute the misfit function as well as the adjoint variable $\mu$, which is back-propagated following the adjoint state-strategy for getting the adjoint field $\lambda$ . From $\lambda$ and the incident wavefield $u$, the gradient of the misfit function (12) is computed using the equation (29).

### 2.4 Numerical illustration on a simple synthetic study

An illustration of the optimal transport based distance for FWI on a schematic 2D example is now presented. A borehole to borehole transmission acquisition is considered, as presented in Figure 1.

The two boreholes are $2500$ m apart. A single source is used, located at $2500$ m depth in the leftmost borehole. An array of $196$ receivers equally spaced each $25$ m is located in the second borehole, from $50$ m depth to $4900$ m depth. A Ricker source centered on $5$ Hz is used to generate a single shot-gather. The modeling is performed in the acoustic approximation and the pressure wavefield is recorded. The density model is kept constant, equal to $1000$ kg.m$^{-3}$. The velocity of the true medium is homogeneous and set to $v_P^* = 2000$ m.s$^{-1}$. One synthetic shot-gather is computed in a homogeneous medium with a velocity set to $v_P = 1500$ m.s$^{-1}$ and with the correct density.

The convergence of the SDMM algorithm is investigated along $50$ iterations. The bound $c$ corresponding to the constraint (17b) is set to $1$. This pragmatical choice is done in conjunction with a scaling of the residuals prior to the solution of the optimal transport problem. The rationale behind this scaling is that the bound constraint (17b) should be active at the convergence of the SDMM algorithm as the solution of such convex constrained optimization problem lies on the boundary of the convex set. The evolution of $\mu$ throughout the SDMM iterations is presented in Figure 2, and compared to the standard $L^2$ residuals.

The standard residuals (Fig. 2a) present two distinct arrivals: the first one corresponds to the observed data, the second corresponds to the synthetic data. The predicted data arrives later compared to the observed one as the velocity is underestimated. The temporal support of the two arrivals does not overlap, which is a situation typical of cycle skipping: the data is predicted with more than half a phase delay. The SDMM method starts from the initial residual, and converges to an estimation of $\varphi$ where the two distinct arrivals are progressively smoothed. The negative values of the two arrivals are also progressively removed. These negative values correspond to the white part in the initial residuals. In counterpart, the area below the last arrival is set to an almost constant negative value (white zone below the last arrival). To assess the convergence of this maximization problem with linear constraints, the relative evolution of the criterion from the previous to the current iterations is considered. When no progress is observed, the convergence is assumed to be reached. Figure 3 confirms the convergence towards a stationary point after $50$ iterations.

The shape of the optimal transport solution may not be intuitive. To better understand how this can be an approximate solution of the problem (21), consider the situation where the constraints on $\varphi$ would be only to be bounded by $c$, relaxing the 1-Lipschitz constraint. In this case, the solution of the discrete maximization problem would be

$$\varphi_i = \left| \begin{array}{lll} c & \text{if} & d_{cal,i}[m] - d_{obs,i} > 0 \\ -c & \text{if} & d_{cal,i}[m] - d_{obs,i} < 0 \end{array} \right. \tag{33}$$

which would correspond to a discontinuous solution. The effect of the 1-Lipschitz constraint thus consists in smoothing the solution of the maximization problem. This hard constraint forces the SDMM

algorithm to find a trade-off between this imposed regularity and the maximization of the criterion. The selected solution thus starts by giving more weight to the large positive values of the original arrivals (black areas), while the smoothing constraint tends to remove the strong initial oscillations, therefore setting weak positive weights in the position of the negative values of the original arrivals (white areas). Because the zone below the last arrival in the original residuals is slightly negative, the SDMM algorithm sets a negative values in all this area to further maximize the criterion while preserving the smooth property of the solution. Two transverse traces are extracted from the $L^2$ residuals and the solution found by SDMM in Figures 4 and 5. The first is a vertical trace extracted for the receiver located at at $2.5$ km in depth. The second is a horizontal trace extracted at time $t = 2$ s. These traces emphasize the regularity of the optimal transport solution compared to the $L^2$ residuals. The shape of the optimal transport traces resembles the envelope of the $L^2$ traces.

For further analysis of this schematic example, the $L^2$ and $\widetilde{W}^1$ misfit function are evaluated for velocity values going from $v_P = 1500$ m.s$^{-1}$ to $v_P = 2500$ m.s$^{-1}$ with $20$ m.s$^{-1}$ sampling. The results are presented in Figure 6. The $\widetilde{W}^1$ misfit function is evaluated for a number of SDMM iterations going from $5$ to $50$. As expected, the misfit functions all reach the global minimum at $v = 2000$ m.s$^{-1}$. The $L^2$ misfit function presents two secondary minima at $v_P = 1780$ m.s$^{-1}$ and $v_P = 2300$ m.s$^{-1}$. This is an illustration of cycle skipping. For these two values of velocity, the seismogram generated by the Ricker source in $v_P^*$ is matched up to one phase delay. Interestingly, the $\widetilde{W}^1$ misfit function profiles tends to become more and more convex as the value of SDMM iterations increases. The secondary minima still exist, however, they are progressively lifted up, rendering the misfit function closer from a convex function. At the same time, the valley of attraction remains as sharp as for the $L^2$ misfit, which ensures that the "resolution power" of the method is unchanged. This behavior is notably different from the one observed for the cross-correlation based misfit function which ensures more convex misfit function detrimental to the size of the valley of attraction which is significantly broadened, leading to lower resolution methods (van Leeuwen & Mulder 2010).

This schematic example provides a first insight on the behavior of the optimal transport distance for the comparison of seismograms in application to FWI. Using this distance does not prevent from cycle skipping issues, as secondary minima are still present. However, the misfit function tends to be more convex as the numerical approximation of the optimal transport distance converges to a stationary point. In addition, the corresponding back-propagated residuals can be seen as smooth version of the standard $L^2$ residuals, the smoothing operator being related to the computation of the optimal transport distance between the observed and predicted seismograms, and more specifically to the enforcement of the $1$-Lipschitz constraint.

# 3 CASE STUDIES

## 3.1 Settings

The numerical experiments which are presented in this section are based on a 2D acoustic time-domain full waveform inversion code. The wave modeling is performed using a $4$th order (for the Marmousi and BP 2004 case studies) and a $8$th-order (for the Chevron 2014 benchmark dataset) finite-difference stencil for the spatial discretization. A second-order leap-frog scheme is implemented for the time discretization. The three case studies are performed in a marine seismic environment. A free surface condition is implemented at the water/air interface. A windowed sinc interpolation is used to account for receivers not located on grid points in the Chevron case study (Hicks 2002).

The minimization of the misfit function, either the standard $L^2$ misfit function or the $\widetilde{W}^1$ misfit function, is performed using the preconditioned $l$-BFGS method (Nocedal 1980). The SEISCOPE optimization toolbox is used to implement this minimization scheme (Métivier & Brossier 2015). This requires to compute the misfit function and its gradient. The gradient is computed as the cross-correlation in time of the incident wavefield and the adjoint wavefield (equation 29) following the adjoint-state method. A vertical scaling linear in depth is used as a preconditioner for the Marmousi and Chevron case studies. This preconditioning compensates for the loss of amplitude of the gradient in depth associated with geometrical spreading effects when using surface acquisition.

In terms of implementation, the computation of the cross-correlation of the incident and adjoint wavefields requires the capability of accessing to the two wavefields at a given time step. This is a well known difficulty in time-domain FWI or Reverse Time Migration approaches, as the incident wavefield is propagated from an initial condition while the adjoint wavefield is back-propagated from a final condition (Clapp 2009). The strategy implemented in our code consists of first computing the incident wavefield from the initial condition, and storing it at each time steps only at the boundaries of the domain. The incident wavefield is then backpropagated from its final state, reversing in time the boundary conditions which have been stored. The adjoint is backpropagated conjointly with the incident wavefield from its final condition. A more detailed description of this strategy is given in Brossier et al. (2014). The method is based on the assumption that no attenuation is taken into account, otherwise the backpropagation of the incident wavefield is numerically unstable.

Besides, a hybrid MPI/OpenMP is used to execute the code in parallel. The MPI communicator is used to perform the computations associated with each shot-gather in parallel. For each shot-gather, the computation of the incident and adjoint wavefields is further accelerated using OpenMP parallelization of the spatial finite-difference loops. The time cross-correlation loop for the computation of the gradient is also accelerated with OpenMP directives.

In the three following experiments, the computation of the optimal transport distance and the corresponding adjoint source is performed through $50$ iterations of the SDMM method (Algorithm 1). This is a rather pragmatical choice, as it guarantees a manageable additional computational cost (see for instance Table 3 for the Chevron benchmark dataset case studies), while the convergence of the SDMM iterations appears to be reached: although not shown here, the maximization of the criterion and the solution of the optimal transport problem only marginally evolves after $50$ SDMM iterations. As for the previous experiment, the bound $c$ of the constraint (17b) is also set to $1$ and a scaling of the residuals is employed.

### 3.2 Marmousi 2 case study

For the Marmousi 2 case study, a fixed-spread surface acquisition is used, involving $128$ sources located every $125$ m and $168$ receivers located every $100$ m at $50$ m depth. The density model is assumed to be homogeneous, set to the value $\rho_0 = 1000$ kg.m$^{-3}$. The topography of the original Marmousi 2 model is also modified so that the water layer has no horizontal variations (flat bathymetry). This layer is kept fixed to the water P-wave velocity $v_P = 1500$ m.s$^{-1}$ during the inversion.

The observed data is generated using a filtered Ricker wavelet, centered on a $5$ Hz frequency. The low frequency content of this wavelet, below $2.5$ Hz, is removed using a minimum phase Butterworth filter. For real seismic marine data, the noise level below this frequency is too strong for the information to be relevant to constrain the P-wave velocity model. The spectrum and the shape of the resulting wavelet are presented in Figure 7. The spatial discretization step is set to $25$ m to guarantee at least $4$ discretization points by wavelength. The time discretization step is set to $0.0023$ s according to the Courant Friedriech Levy (CFL) condition. The recording is performed over $2000$ time steps, which corresponds to a total recording time of $4.6$ s. In this experiment, a Gaussian filter smoothing with a short correlation length (between $60$ m and $100$ m depending on the local dominant wavelength) is applied to the gradient, to remove fast oscillations which are due to a sparse acquisition design (only one source every $125$ m).

Two initial models are created by smoothing the exact model using a Gaussian filter, with vertical and horizontal correlation lengths equal to $250$ m and $2000$ m respectively. The first model is very close from the exact model, with only smoother interfaces. The second models is more distant from the exact model, as it presents almost only vertical variations, and underestimates the increase of the velocity in depth.

Starting from these two initial models, FWI using the $L_2$ misfit function and the optimal transport distance based misfit function is used to interpret the data. The results are presented in Figure 8. For the first initial model, the results obtained after $100$ iterations are presented (Fig. 8c,d). For the second

initial model, the best results obtained using the two misfit functions are presented (Fig. 8f,g). The exact data as well as the corresponding residuals in the initial and the calculated models are presented in Figure 9.

Starting from the first initial model, both the $L^2$ distance and the optimal transport distance yield estimations very close from the exact model (Fig. 8c,d). However, a difference can be noted regarding the reconstruction of the low velocity zone near $x = 11$ km and $z = 2.5$ km. A high velocity artifact is present in this zone in the estimation obtained with the $L^2$ distance. This is not the case in the estimation obtained with the optimal transport distance.

Starting from the second initial model, FWI based on the $L^2$ distance is unable to provide a satisfactory P-wave velocity estimation (Fig. 8f). This is emphasized by the residuals computed in the corresponding final estimations (Fig. 9f). In comparison, the P-wave velocity estimation obtained using FWI based on the optimal transport distance is significantly closer from the exact model (Fig. 8g). Low velocity artifacts, typical of cycle skipping, can still be seen in depth, below 3 km. Low wavenumber artifacts are also visible on the left part of the model ($x < 1$ km). However, in the central part, the P-wave velocity model is correctly recovered, even starting from this crude approximation. The computed estimation seems to explain correctly the data, as can be seen in Figure 9g . Compared to the results obtained using the first initial model, there are unexplained seismic events, especially for late arrivals around $T = 4$ s. However, most of the data is explained by the computed estimation.

To complete this analysis on the Marmousi case study, the $L^2$ residuals in the two initial models are compared with their optimal transport counterpart (the adjoint variable $\mu$ defined by equation (32)) in Figure 10. The optimal transport residuals are smoother than the $L^2$ residuals, with a lower frequency content. An emphasis of particular seismic events in the optimal transport residuals is also noticeable, compared to the $L^2$ residuals. This is mainly observable for the reflections around 3 s and 8 km offset, and this does not depend on the initial model. The optimal transport thus seems to weight differently the uninterpreted part of the seismograms.

The effect of the modification of the residuals by the optimal transport distance is also emphasized in Figure 11, where two gradients, one associated with the $L^2$ distance, the other with the optimal transport distance, are compared. These gradient are computed in the second initial model, which generates a strong cycle skipping effect with the $L^2$ distance. In order to interpret these gradient as velocity updates, they have been multiplied by $-1$: they represent the first model perturbation used by a steepest descent method. Cycle skipping can be detected in the $L^2$ gradient through the strong shallow low velocity updates, in a zone where the velocity should be increased. The optimal transport distance seems to be able to efficiently mitigate these strong artifacts. The energy in depth is also better balanced. The main interfaces constituting the Marmousi model also appear in this velocity update.

From this first experiment, the optimal transport distance based misfit function appears more robust than the conventional $L^2$ norm-based misfit function. For each initial model, a better P-wave velocity estimation is computed using the optimal transport than using the $L^2$ distance. In particular, correct estimations are obtained in the shallow part located above the depth $z = 3$ km, even starting from a very crude approximation of the exact model. This is a first indication that using the optimal transport distance may be an interesting strategy to mitigate cycle skipping issues in the context of full waveform inversion.

### 3.3 BP 2004 case study

The BP 2004 benchmark model is representative of the geology of the Gulf of Mexico (Billette & Brandsberg-Dahl 2004). This area is characterized by a deep water environment and the presence of complex salt structures. The large P-wave velocity value of the salt structures is responsible for most of the energy of the seismic signal to be reflected back to the receivers from the interface between the water layer and the salt. Only a few percentage of energy of the seismic signal travels within the structure and below before being recorded. This particular configuration makes seismic imaging in the presence of salt structures challenging. The first challenge is to correctly identify and delineate the salt structures. The second challenge consists in correctly imaging zones below the salt structure (sub-salt imaging).

A fixed-spread surface acquisition is used, with $128$ sources and $161$ receivers distant from $125$ m and $100$ m respectively. The depth of the sources and receivers is set to $z = 50$ m. The density model is assumed to be homogeneous such that $\rho_0 = 1000$ kg.m$^{-3}$. The wavelet used to generate the data is based on a Ricker wavelet centered on $5$ Hz. A whitening of the frequency content is performed before a minimum phase Butterworth low-pass and high-pass filters are applied. The spectrum of the resulting wavelet is within an interval from 3 Hz to 9 Hz (Fig. 12). The spatial discretization step is set to $25$ m and the time discretization step is set to $0.0023$ s to respect the CFL condition. The maximum recording time is performed over $4500$ time steps, which corresponds to a recording time of $10.3$ s.

The exact and initial models are presented in Figures 13a and 13b. The left part of the original BP 2004 model has been extracted (Billette & Brandsberg-Dahl 2004). The initial model has been designed such that the imprint of the salt structure has been totally removed: it contains no information on the presence of salt. From this starting model, FWI using a standard $L^2$ distance fails to produce meaningful results, as can be seen in Figure 13c. The time-window is reduced to $4.6$ s to focus the inversion on the shallowest part of the model and reduce cycle skipping issues, however this does not prevent the minimization from converging towards a local minimum far from the exact model. The incorrect P-wave velocity estimation of the starting model prevents the FWI algorithm from locating

the reflectors associated with the top of the salt. Instead, diffracting points are created to match the most energetic events without lateral coherency. In comparison, the same experiment is performed using the optimal transport distance. The results are presented in Figure 13d. As can be seen, the top of the salt structure is correctly delineated. Synthetic shot-gathers corresponding to the source located at $x = 8$ km, computed in the exact model, initial model, $L^2$ estimation, and optimal transport estimation, are presented in Figure 14. This picture shows clearly that the strong reflection coming from the top of salt is inaccurately predicted by the $L^2$ estimation; in particular, the reflected energy which is introduced is discontinuous (Fig.14c). In comparison, the optimal transport estimation yields a correct prediction of this reflection (Fig.14d). The $L^2$ residuals and the optimal transport residuals (the adjoint variable $\mu$ defined by the equation (32)) computed in the initial model are presented in Figure 15. The uninterpreted diving waves appearing in the left bottom corner of the $L^2$ residuals (Fig. 15a) seem to be strongly damped in the corresponding optimal transport residuals. The optimal transport distance seems to rather enhance the reflected events, which is consistent with the previous observations.

Building on this result, a layer stripping workflow is suggested. Five increasing time-windows are defined, with recording time equal to $4.6$ s, $5.75$ s, $6.9$ s, $9.2$ s, and finally $10.3$ s. For each time-window, two to three successive inversions are performed. A Gaussian smoothing with a small correlation length is applied to the model computed after each inversion, which serves as an initial model for the next inversion. This Gaussian smoothing serves only to remove high frequency artifacts appearing in the late iterations of the inversion. Alternative strategies such as Tikhonov regularization or gradient smoothing could have been used instead. A total of $15$ inversions is performed following this process, with in average $221$ iterations of the $l$-BFGS algorithm for each inversion. The stopping criterion is only based on a linesearch failure to give the possibility to the optimizer to minimize as much as possible the misfit function based on the optimal transport distance. The detailed workflow is summarized in Table 1.

The results obtained after the $1^{\text{st}}$, $3^{\text{rd}}$, $6^{\text{th}}$, $9^{\text{th}}$, $12^{\text{th}}$, and $15^{\text{th}}$ inversions are presented in Figure 16. As can be seen, the salt structure is practically entirely recovered at the end of the cycle of inversions (Fig 16f). A continuous progression is achieved from the initial delineation of the top of the salt structure to the full reconstruction of its deeper parts. The subsalt zone, however, whose reconstruction is critical, is not satisfactorily recovered. To this purpose, a possibility would consist in building an initial model from this reconstruction by freezing the salt, which is correctly delineated, and smoothing below the salt. From such an initial model, our previous study show that FWI based on the the $L^2$ distance with a truncated Newton optimization strategy should be able to reconstruct accurately the subsalt region (Métivier et al. 2014a).

A better insight of the reconstruction process is given by the synthetic data computed in intermediate models throughout the different steps of the workflow presented in Figure 17. The shot-gathers are computed for a source located at $x = 8$ km. A particular attention should be accorded to the left part of the seismogram (red rectangles), as this part corresponds to the main salt structure in the exact model. After interpreting correctly the reflections coming from the salt roof (Fig.17a), the transmitted wave traveling within and below the salt is progressively adjusted while deeper reflections are also progressively integrated (Fig.17b to Fig.17f). This behavior is in contrast with standard multi-scale approaches for which the transmitted energy is fitted prior to the reflected energy. However, this may not be inputted to the use of the optimal transport distance. Due to the high velocity contrast, the reflected energy dominates the transmitted energy in the data. This, in conjunction with the layer stripping strategy which focuses the prior steps of the inversion toward short offset data, favors the fit of the reflections prior to the diving waves.

### 3.4 Chevron 2014 case study

In 2014, the Chevron oil company has issued a blind benchmark synthetic dataset for FWI. The aim of such blind benchmark is to provide realistic exploration seismic data to practitioners with which they can experiment various FWI workflow and test methodological developments. As the exact model which has served to build the data is not known, such a case study is closer from an application to field data than synthetic experiments for which the exact model is known.

The Chevron 2014 benchmark dataset is built from a 2D isotropic elastic modeling engine. A frequency-dependent noise has been added to the data to mimic a realistic dataset. Especially, the Signal Over Noise Ratio (SNR) for low frequencies (below 3 Hz) is much less than for higher frequencies. Free surface multiples are incorporated in the data. A streamer acquisition is used, with a maximum of $8$ km offset, with $321$ receivers by sources equally spaced each $25$ m. The depth of the sources and receivers is $z = 15$ m. Among the $1600$ available shots gathers, $256$ have been used in this study, with a distance of $150$ m between each sources. A frequency continuation strategy similar to the one proposed by Bunks et al. (1995) is implemented: Butterworth low-pass and high-pass filters are applied to the selected shot-gathers to generate an ensemble of $15$ datasets with an increasing bandwidth from $2 - 4$ Hz to $2 - 25$ Hz.

The shot-gathers corresponding to the source located at $x = 150$ m are presented for the 1st, 5th, 10th and 15th frequency bands in Figure 18. As mentioned previously, the noise imprint is clearly stronger for the first frequency bands.

The initial model provided by Chevron is presented in Figure 19a. This is a 1D layered model with no horizontal variations except for the water layer on top for which the correct bathymetry has

been incorporated. The P-wave velocity in the water layer is set to $1500$ m.s$^{-1}$. The initial model incorporates an important feature: a low velocity layer is located between the depth $z = 2.3$ km and $z = 3$ km. This velocity inversion and the relatively short available offsets (only $8$ km) prevent diving waves from sampling the deepest part of the model. This makes the benchmark data challenging as only reflection information is available for constraining the deep part of the model.

The workflow which is applied to the Chevron benchmark dataset is the following. Prior to inversion, an estimation of the source wavelet is performed in the initial model, for each frequency band, following the frequency-domain strategy introduced by Pratt (1999). For the first ten frequency bands, $20$ iterations of a preconditioned $l$-BFGS algorithm are performed. For the frequency bands $11$ and $12$, $50$ iterations are performed. For the last three frequency bands, $40$ iterations are performed with a restart of the $l$-BFGS algorithm after the $20$ first iterations. This restart is only due the configuration of the queue of the Blue Gene/Q machine of the IDRIS center, which does not accept jobs running longer than $20$ hours. The restart could be avoided by storing the $l$-BFGS approximation on disk, however this option is not yet implemented in the SEISCOPE optimization toolbox. The spatial and time discretization steps are set to $37.5$ m and $0.004$ s respectively for the $8$ first frequency bands. They are decreased to $25$ m and $0.003$ s respectively for the frequency bands $9$ to $12$. For the last three frequency bands, the discretization step is set to $12.5$m and the time step to $0.001$ s. The misfit function is based on the optimal transport distance. According to the frequency continuation strategy, the P-wave velocity model estimated for one frequency band serves as the initial model for the next frequency band. No regularization is introduced throughout the inversion. However, the model estimated at the end of each inversion is smoothed using a Gaussian filter with a correlation length adapted to the resolution expected after the inversion of each frequency-band. The workflow is summarized in Table 2.

The $256$ shot-gathers are inverted using $1024$ core units of the Blue Gene/Q machine of the IDRIS center. This yields the possibility to assign $16$ threads ($4$ physical thread $\times$ $4$ hyperthreads) for each shot-gather. For such a configuration, the computational times for one gradient depending on the discretization are summarized in Table 3. In particular, we are interested in the additional cost due to the use of the optimal transport distance. The results presented in Table 3 show that the proportion of computational time spent for the solution of the optimal transport problem decreases from $75$ % to $20$ % as the size of the discrete problem increases. This interesting feature is due to the fact the computational complexity of the SDMM algorithm is in $O(N_r^2 \times N_t)$ (see Appendix C), while the computational complexity of the solution of one wave propagation problem is in $O(N_t \times N_x \times N_z)$, $N_x$ and $N_z$ being the number of grid points in the horizontal and vertical dimensions respectively.

The results obtained after inverting the data up to $4$ Hz (frequency band $1$), $10$ Hz (frequency band $8$), $16$ Hz (frequency band $12$) and $25$ Hz (frequency band $15$) are presented in Figure 19b,c,d,e. Three

shallow low velocity anomalies are recovered at approximately $500$ m depth and at the lateral positions $x = 12$ km, $x = 18$ km and $x = 30$ km. An additional small scale low velocity anomaly appears at $x = 14.75$ km and $z = 1$ km in the highest resolution estimation. The original layered structure of the initial model is tilted in the final estimation. The upper (faster) layers bend downward (from left to right), while the low velocity layer at depth $z = 2.5$ km bends upward. Three high velocity anomalies are also recovered on top of the layer above the low velocity layer, at depth $1.8$ km and lateral positions $x = 8$ km, $x = 19$ km, $x = 22$ km. The deeper part of the model, below $3$ km depth, seems less well reconstructed, as it could be expected from the lack of illumination of this zone. However, a curved interface seems to be properly recovered at a depth between $4.5$ and $5$ km. A flat reflector is also clearly visible at the bottom of the model, at depth $z = 5.8$ km.

As the exact model is not known, it is important to perform quality controls of the computed P-wave velocity estimation. A synthetic shot-gather in the model estimated at $25$ Hz is computed and compared to the corresponding benchmark shot-gather in Figure 20. The similarity between the two gathers is important. The kinematic of the diving waves is correctly predicted. Most of the reflected events are in phase. Destructive interference due to free surface effects are also correctly recovered. A slight time-shift can however be observed for the long-offsets diving waves. This time-shift is not in the cycle skipping regime. A similar phenomenon is observed in Operto et al. (2015) where FWI is applied to invert the 3D Valhall data. As mentioned in this study, this time-shift may be due to the accumulation of error with propagating time or an increasing kinematic inconsistency with large scattering angles. The residuals between the two datasets are presented in Figure 21. As both diving and reflected waves are (almost) in phase, the differences are mainly due to amplitude mismatch. This is not surprising as the inversion is based on acoustic modeling. The amplitude mismatch should therefore be the imprint of elastic effects not accounted for in the inversion.

As a second quality control, migrations of the data in the initial model and the estimated models at $10$ Hz and $16$ Hz are performed. The migration results correspond to impedance gradients computed on $30$ Hz low-pass filtered data, with a filter applied on the diving waves to focus on reflection data only. The spatial and time discretization steps are set to $12.5$ m and $0.001$ s respectively. The number of sources is doubled to $512$ (one source each $75$ m) to avoid spatial aliasing. As a post-processing, a polynomial gain is used to balance the energy in depth. The resulting images are presented in Figure 22. The migrated image obtained in the estimated model at $10$ Hz is significantly improved in the shallow part of the model (above $3$ km depth) (Fig. 22b). A significant uplift of this part of the model can be observed. The continuity and the flatness of the reflectors is globally improved. However, the reflectors in the deepest part of the model ($z > 2.5$ km) remain unfocused . The migrated image in the estimated model at $16$ Hz yields a better delineation of these deep reflectors, as indicated by the three

red arrows at the bottom (Fig. 22c). In particular, a continuous tilted reflector appears clearly at $5$ km depth in the left part of the model. This is an indication of a progress in constraining the deep part of the P-wave velocity model, even if this remains challenging as only reflections sample this part of the model.

Another conventional control for assessing the quality of velocity model consists in considering the flatness of CIG. The CIG presented in Figure 23 are obtained by computing migrated images following the previous strategy for different offset ranges. A dip filtering is used in addition, to remove events associated with low-energy $S$-waves. Consistently with what is observed for the migrated images, the curve and the offset extension of the shallowest reflectors is improved by the P-wave velocity model obtained at $10$ Hz (Fig 23b). The P-wave velocity model obtained at $16$ Hz further improves this energy refocusing. Some of the deeper reflectors are also better flatten, as indicated by the bottom arrows in Figure 23c, even if the progress in depth are less significant than the improvement observed in the shallow part.

Finally, a vertical well log of the exact $P$-wave velocity model taken at $x = 39375$ m, at a depth between $1000$ m and $2450$ m is provided in the benchmark data. The corresponding log is extracted from the final estimation obtained at $25$ Hz maximum frequency and compared to this log in Figure 24. This provides another criterion to assess the quality of the estimation. As can be seen in Figure 24, the agreement between the exact and estimated logs is excellent. However, only the shallowest part of the model is constrained here. A deeper exact log would be interesting to have quality control on the deeper part of the model, which is more challenging to recover in this configuration.

To emphasize the benefits provided by using the optimal transport distance, the same frequency continuation workflow is applied to the Chevron 2014 benchmark dataset, with a FWI algorithm based on the conventional $L^2$ distance. The results obtained after the first frequency band and the $8$th frequency band are compared to the results obtained when the optimal transport distance is used in Figure 25. As can be seen, the $L^2$ distance based FWI converges to a local minimum. Already after the first frequency band, the shallow part of the $P$-wave velocity estimation seems incorrect as a strong, flat reflector is introduced at the depth $z = 500$ m. Note that for this simple comparison, no data-windowing strategy is used. As previous experiments in our group indicate, better results using the $L^2$ distance can be obtained for the reconstruction of the shallow part of the model by designing a hierarchical workflow based on the interpretation of transmitted energy first.

To complement this comparison, the residuals associated with the $L^2$ norm and the optimal transport distance in the initial model, for the first frequency band, are presented in Figure 26. This Figure emphasizes the regularization role played by the optimal transport distance. Besides the smoothing effect already detected in the first numerical test, the SDMM algorithm seems to act as a coherency

filter, restoring the continuity of the main seismic events. This feature is particularly important for the interpretation of real data, as the signal over noise ratio of seismic signal below 3 Hz is generally poor.

## 4 DISCUSSION

The method proposed in this study is designed to mitigate issues related to the use of the $L^2$ norm to compare seismic signals in the framework of full waveform inversion. An optimal transport distance is used instead. This change in the measure of the misfit between seismograms appears to bring a more robust strategy, capable of overcoming cycle skipping issues, allowing to better interpret seismic data through FWI. In addition, it seems to facilitate the interpretation of noisy data as it acts as a coherency filter on the residuals which are back-propagated to form the gradient through the adjoint-state method.

Distances based on $L^p$ norms are built as a sum of mismatch over each source and each receiver. As a consequence, these distances consider each seismic traces individually, without accounting for a potential correlation between these traces. However, it is well known from seismic imaging practitioners that shot-gathers, presented in the 2D receiver/time plane, carry much more information than individual traces. Seismic events such as reflection, refraction, conversion, are identifiable on 2D shot-gathers from their lateral coherency in the receiver dimension. In conventional FWI based on $L^p$ distance, this information is used for visualizing the data, but is not accounted for in the inversion. This loss of information is severe and penalizes the inversion. The main advantage of the optimal transport distance presented in this study is its capability of accounting for this lateral coherency in the gather panel. Indeed, the traces of one common shot-gather are now interpreted jointly, through a measure of the distance in the 2D receiver/time plane.

To illustrate this property, a comparison with an alternative strategy based on 1D optimal transport is performed on the Marmousi 2 model. This strategy is closer from the approach promoted by Engquist & Froese (2014): the seismic data is considered as a collection of 1D time signals which are compared independently using a 1D optimal transport distance. The resulting misfit function is a summation over all the traces of this distance between observed and calculated data. The lateral coherency of the seismic event in the receiver dimension is thus not accounted for. This method can be implemented easily using the SDMM method (Algorithm 1). The block tridiagonal system reduces to a tridiagonal system which can be efficiently solved using the Thomas algorithm. The computational complexity of the solution of these 1D optimal transport problem reduces to $O(N_t \times N_r) = O(N)$ (compared to $O(N^{3/2})$ for the 2D optimal transport distance). However this reduction of the complexity comes with a price, as is shown on Figure 27. The reconstruction (Fig. 27d), although more accurate than the reconstruction obtained using the $L^2$ distance (Fig. 27c), is far from being as accurate as the one obtained with the 2D optimal transport distance (Fig. 27e). A strong degradation of the results thus occurs when neglecting the lateral coherency of the events in the receiver dimension.

For further 2D and 3D large size application to real seismic data, the question of the computational cost of the optimal transport distance remains opened. In 3D, as the acquisition comprises inline and

crossline receivers, common shot-gathers should be represented as data cubes, with a coherency of seismic events both in inline and crossline directions. The previous experiment, based on 1D optimal transport, suggests that there is an interest in fully exploiting the lateral coherency of the seismic signal. However, further numerical improvements are required to design a method with a manageable computational time in such a configuration. This could be achieved through a better account of the structure of the matrix $Q$, which is related to a second-order discretization of the Laplacian operator with Neumann boundary conditions. The linear system to be solved at each iteration of the SDMM algorithm could thus be identified as a Poisson equation, for which fast solver exist, either based on Fast Fourier Transform (Swarztrauber 1974), or multigrid methods (Brandt 1977; Adams 1989). If this strategy reveals unfeasible, dimensionality reduction (such as the one presented here from 2D to 1D optimal transport) could still be worthy to investigate, using appropriate regularization techniques. Another option may also consist in changing the formulation of the optimal transport problem to a primal formulation with entropic regularization, as this strategy is indicated to benefit from a reduced computational complexity (Benamou et al. 2015).

Regarding the application of the method, the results obtained on the BP 2004 case study indicate that the measure of the distance between synthetic and observed data through optimal transport distance yields the possibility to better recover salt structures. This may be a first step toward more efficient sub-salt reconstructions. This could be assessed on more realistic datasets than the synthetic BP 2004 model. The Chevron 2012 Gulf Of Mexico dataset could be investigated to this purpose.

An enhancement of the results obtained on the Chevron 2014 benchmark, especially in the deep part of the model, could be possibly obtained by combining the use of optimal transport distance with reflection-based waveform inversion strategies. These methods aim at enhancing the recovery of velocity parameters in zones where the subsurface is mainly sampled by reflected waves rather than transmitted waves. They are based on the scale separability assumption and alternatively reconstruct the smooth velocity and the reflectivity model. This generates transmission kernels between the reflectors and the receivers which provide low wavenumber update of the velocity. The method has been first introduced by Chavent et al. (1994); Plessix et al. (1999), then extended by Xu et al. (2012); Brossier et al. (2015); Zhou et al. (2015). In the Chevron 2014 benchmark dataset, relatively short offsets are used ($8$ km streamer data), and the velocity inversion in the low velocity layer prevents diving waves to penetrate deeply the subsurface. A combination of the optimal transport distance with reflection FWI is thus a potentially interesting investigation.

Another important current issue in FWI is its ability to reconstruct several classes of parameters simultaneously, in a multi-parameter framework. An overview of the challenges associated with this issue is given in Operto et al. (2013). In particular, the importance of an accurate estimation of the inverse

Hessian operator to mitigate as much as possible trade-offs between parameters is emphasized. To this purpose, recent results indicate the interest of using truncated Newton techniques instead of more conventional quasi-Newton optimization strategies (Métivier et al. 2014b, 2015; Castellanos et al. 2015). These techniques rely on an efficient estimation of Hessian-vector products through second-order adjoint state formulas. An extension of this formalism to the case where the optimal transport distance is used instead of the standard $L^2$ should thus be investigated.

## 5 CONCLUSION

A FWI algorithm using a misfit function based on an optimal transport distance is presented in this study. Instead of using the Wasserstein distance, as proposed in (Engquist & Froese 2014), a modified Monge-Kantorovich problem is solved to compute the distance between seismograms, yielding the possibility to account for non-conservation of the energy. The numerical computation of this distance requires the solution of a linear programming problem, which is solved through the SDMM algorithm. This algorithm is based on proximal splitting techniques (Combettes & Pesquet 2011). The main computationally intensive task to be performed within this algorithm is related to the solution of linear systems involving a matrix associated with the constraints of the linear programming problem. An efficient algorithm, based on the work of Buzbee et al. (1970), is set up to solve these linear systems with a complexity in $O(N)$ and $O(N^{3/2})$ in terms of memory requirement and number of operations respectively.

Synthetic experiments emphasize the properties of this distance when applied to FWI. The resulting misfit function is more convex, which helps to mitigate cycle skipping issues related to the use of the more conventional $L^2$ norm. This is illustrated on a simple transmission from borehole to borehole experiment, as well as on the Marmousi 2 case study. From crude initial models, more reliable estimations of the $P$-wave velocity model are obtained using the optimal transport distance.

The property of the optimal transport distance is also tested in the context of salt imaging. The experiment on the BP 2004 case study emphasizes the capability of the method to recover the salt structures from an initial model containing no information about their presence. This yields interesting perspectives in terms of sub-salt imaging.

The experiment on the more realistic Chevron 2014 benchmark dataset emphasizes the satisfactory performances of the method, particularly its robustness to noise. It seems also able to provide a reliable estimation of the $P$-wave velocity in the zone which are sampled by diving waves. In the deepest part where the seismic information is dominated by reflection, the method faces the same difficulties as conventional FWI. This could be overcome by combining the use of the optimal transport distance with reflection FWI strategies.

The proposed method thus seems promising and should be investigated in more realistic configurations, implying 3D waveform inversion. Measuring the misfit between data cubes using the optimal transport distance is a challenging issue, which could yield interesting perspectives for 3D FWI. The introduction of viscous, elastic and anisotropic effects should also be investigated. As the proposed strategy is data-domain oriented, such extension should be straightforward. Finally, specific investigations have to be made to extend the formalism of the method for the computation of second-order

822 derivatives information (Hessian-vector products) through the adjoint-state method. These investiga-
823 tions should be carried on in the perspective of applying this method to multi-parameter FWI.

## APPENDIX A: EQUIVALENCE BETWEEN LINEAR PROGRAMMING PROBLEMS

In this appendix, the proof of equivalence between the linear programming problems (15) and (17) is given. The first of these two problems is the discrete analogous of the problem (10), which uses global constraints to impose the Lipschitz property. The second only uses local constraints to impose the Lipschitz property and is therefore less expensive to solve numerically.

It is straightforward to see that if the global constraints are imposed, the local constraints are satisfied. Interestingly, the reciprocal is also true. To see this, consider a pair of points $v = (x_v, t_v)$ and $w = (x_w, t_w)$ in the 2D grid. A sequence of $N$ points $z_i = (x_i, t_i), \quad i = 1, \ldots, N$, with $z_1 = v$ and $z_N = w$ can be chosen to form a path from $v$ to $w$, such that the points $z_i$ are all adjacent on the grid, with monotonically varying coordinates: this means that each of the sequences $x_i$ and $z_i$ are either increasing or decreasing monotonically. The key is to see that, for such a sequence of points, the $\ell_1$ distance (also known as Manhattan distance) ensures that

$$||w - v||_1 = \sum_i ||z_{i+1} - z_i||_1. \tag{A.1}$$

Now, consider a function $\varphi$ satisfying only the local constraints. The triangle inequality yields

$$||\varphi(w) - \varphi(v)||_1 \leq \sum_i ||\varphi(z_{i+1}) - \varphi(z_i)||_1. \tag{A.2}$$

As the points $z_i$ are adjacent, the local inequality satisfied by $\varphi$ can be used to obtained

$$\sum_i ||\varphi(z_{i+1}) - \varphi(z_i)||_1 \leq \sum_i ||z_{i+1} - z_i||_1. \tag{A.3}$$

Putting together equations (A.2), (A.3) and (A.1) yields

$$||\varphi(w) - \varphi(v)||_1 \leq ||w - v||_1. \tag{A.4}$$

This proves that satisfying the local constraints implies that the global constraints are verified. The linear programming problem (17) is thus the one which is solved to approximate the solution of the continuous problem (10).

## APPENDIX B: PROXIMITY OPERATORS

For a given convex function $f(x)$, its proximity operator $\text{prox}_f$ is defined by

$$\text{prox}_f(x) = \arg\min_y f(y) + \frac{1}{2}\|x - y\|_2^2, \tag{B.1}$$

where the standard Euclidean distance on $\mathbb{R}^d$ is denoted by $\|.\|_2$. Closed-form proximity operators exist for numerous convex functions, which can make them inexpensive to compute. This is the case for the proximity operators of the indicator function $i_K$ and the linear function $h(\varphi)$. The proximity

operator of the indicator function $i_K$ corresponds to the projection on the ensemble $K$ (Combettes & Pesquet 2011).

$$\forall i = 1, \ldots, 3N, \quad \left(\text{prox}_{i_K}(x)\right)_i = \begin{vmatrix} x_i & \text{if} & -1 < x_i < 1 \\ 1 & \text{if} & x_i > 1 \\ -1 & \text{if} & x_i < -1. \end{vmatrix} \tag{B.2}$$

This can be seen as a thresholding operation: any value of $x$ lower than $-1$ (respectively higher than 1) is set to the threshold value $-1$ (respectively 1). The values between $-1$ and 1 remain unchanged. Following the definition (B.1), the proximity operator of the function $h_{d_{cal}[m],d_{obs}}(\varphi)$ is simply

$$\text{prox}_{h_{d_{cal}[m],d_{obs}}}(\varphi) = \varphi - d_{cal}[m] + d_{obs}. \tag{B.3}$$

## APPENDIX C: EFFICIENT SOLUTION OF THE BLOCK TRIDIAGONAL LINEAR SYSTEM WITHIN THE SDMM ALGORITHM

The solution of the problem (21) with the SDMM algorithm involves solving at each iteration a linear system of type

$$Qx = b, \quad (x, b) \in \mathbb{R}^N \times \mathbb{R}^N, \quad Q \in \mathbb{M}_N(\mathbb{R}), \tag{C.1}$$

where $Q$ is defined by the equation (24) and $\mathbb{M}_N(\mathbb{R})$ denotes the ensemble of square matrices of size $N$ with real coefficients. The following ordering is used for the vectors of $\mathbb{R}^N$. Recall that the total size $N$ is the product of the number of time steps $N_t$ and the number of receivers $N_r$. The vectors of $\mathbb{R}^N$ are decomposed in $N_t$ blocks of size $N_r$, such that for all $x \in \mathbb{R}^N$

$$x = [x_1, \ldots x_{N_t}] \in \mathbb{R}^N, \tag{C.2}$$

and

$$\forall i = 1, \ldots, N_t, \quad x_i = [x_{i1}, \ldots x_{iN_r}] \in \mathbb{R}^{N_r}. \tag{C.3}$$

The matrix $Q$ is block tridiagonal such that

$$Q = \begin{pmatrix} F+B & B & & & \\ B & F & B & & \\ & \ddots & \ddots & \ddots & \\ & & B & F & B \\ & & & B & F+B \end{pmatrix}. \tag{C.4}$$

Introducing $\alpha = \frac{1}{\Delta x_r^2}, \beta = \frac{1}{\Delta t^2}$, $B$ is the diagonal matrix

$$B = \text{diag}(-\beta) \in \mathbb{M}_{N_r}(\mathbb{R}), \tag{C.5}$$

and $F$ is the tridiagonal symmetric positive definite matrix

$$F = \begin{pmatrix} 1+\alpha+2\beta & -\alpha & & & & \\ -\alpha & 1+2(\alpha+\beta) & -\alpha & & & \\ & \ddots & \ddots & \ddots & & \\ & & -\alpha & 1+2(\alpha+\beta) & -\alpha & \\ & & & -\alpha & 1+\alpha+2\beta \end{pmatrix} \in \mathbb{M}_{N_r}(\mathbb{R}). \tag{C.6}$$

The matrix $Q$ is thus decomposed in $N_t$ blocks of size $N_r$. The method for block tridiagonal Toeplitz matrices proposed by Buzbee et al. (1970) can be adapted to the solution of this system using the following strategy. First each row of $Q$ is multiplied by $B^{-1}$, which yields the system

$$\begin{cases} (E+I)x_1 + x_2 = b'_1 \\ x_{i-1} + Ex_i + x_{i+1} = b'_i, \quad i = 2, N_t - 1 \\ x_{N_t-1} + (E+I)x_{N_t} = b'_{N_t}, \end{cases} \tag{C.7}$$

where $b'_i = B^{-1}b_i$ and $E = B^{-1}F$. The matrix $E$ is symmetric positive definite by construction, and can be factorized as

$$E = PDP^T, \quad D = \text{diag}(d_j), \quad j = 1, \dots N_r, \quad P^TP = I. \tag{C.8}$$

Using this factorization in (C.7) yields

$$\begin{cases} (D+I)y_1 + y_2 = c_1 \\ y_{i-1} + Dy_i + y_{i+1} = c_i, \quad i = 2, N_t - 1 \\ y_{N_t-1} + (D+I)y_{N_t} = c_{N_t}, \end{cases} \tag{C.9}$$

where

$$y_i = P^Tx_i, \quad c_i = P^Tb'_i, \quad i = 1, \dots N_t. \tag{C.10}$$

The system (C.9) can now be expanded as

$$\begin{cases} (d_j+1)y_{1j} + y_{2j} = c_{1j}, \quad j = 1, N_r \\ y_{i-1j} + d_jy_{ij} + y_{i+1j} = c_{ij}, \quad i = 2, N_t - 1, \quad j = 1, N_r \\ y_{N_t-1j} + (d_j+1)y_{N_tj} = c_{N_tj} \quad j = 1, N_r, \end{cases} \tag{C.11}$$

The vectors $y_{*j}$ and $c_{*j}$ such that

$$\forall j = 1, \dots, N_r, \quad y_{*j} = [y_{1j}, \dots, y_{N_tj}] \in \mathbb{R}^{N_t}, \quad c_{*j} = [c_{1j}, \dots, c_{N_tj}] \in \mathbb{R}^{N_t} \tag{C.12}$$

are introduced. These vectors satisfy the equation

$$K_jy_{*j} = c_{*j}, \tag{C.13}$$

where $K_j$ is the tridiagonal matrix

$$K_j = \begin{pmatrix} d_j+1 & 1 & & & \\ 1 & d_j & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & d_j & 1 \\ & & & 1 & d_j+1 \end{pmatrix}. \tag{C.14}$$

These transformations yield the algorithm 2 to solve the initial system (C.1). As a pre-processing step, the matrix $E$ is factorized as in (C.8), and the eigenvectors are stored in the matrix $P$. The computation cost and the memory requirement of this operation is in $O(N_r^2)$ as $E$ is tridiagonal. The solution of the equation (C.1) is then obtained through the following operations. First, the vectors $b_i$ are multiplied by the diagonal matrix $B^{-1}$ which requires $O(N)$ operations. Second, the vectors $c_i$ are formed following equation (C.10). As the matrix $P$ is full, this requires $O(N_r^2 \times N_t)$ operations. Third, the vectors $y_{*j}$ are computed through the solution of $N_r$ tridiagonal systems of size $N_t$. Tridiagonal systems are efficiently solved through the Thomas algorithm which has a linear complexity (Golub 1996) . Therefore, the computation cost of computing $y_{*j}$ is merely in $O(N_r \times N_t) = O(N)$. The final step consists in computing the vector $x$ from the vectors $y_{*j}$ through the equation (C.10). This requires to multiply each vector $y_i$ by $P$, which costs $O(N_r^2 \times N_t)$ operations. The overall complexity of the algorithm is thus $O(N_r^2 \times N_t)$, and the memory requirement in $O(N)$. In contrast, a Cholesky factorization has the same computational complexity, but requires to store $O(N^{3/2})$ elements. In addition, the forward backward substitution is an intrinsically sequential algorithm, while the most expensive part of the algorithm 2 are the matrix-vector multiplications involving the eigenvectors of the matrix $E$, which can be efficiently parallelized. As a final remark, in the case $N_t < N_r$, the matrices and vectors can be re-organized in $N_r$ blocks of size $N_t$ to yield a complexity in $O(N_t^2 \times N_r)$ instead of $O(N_r^2 \times N_t)$.

## ACKNOWLEDGMENTS

The authors would like to thank gratefully Stéphane Operto for his insightful advice and his availability to comment on these results. The authors also thank Paul Wellington for his help on figure edition, as well as the associated editor and the two reviewers for their helpful comments. This study was partially funded by the SEISCOPE consortium (*http://seiscope2.osug.fr*), sponsored by BP, CGG, CHEVRON, EXXON-MOBIL, JGI, PETROBRAS, SAUDI ARAMCO, SCHLUMBERGER, 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 (ref-

861 erence ANR10 LABX56) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the pro-
862 gramme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche, and the
863 HPC resources of CINES/IDRIS under the allocation 046091 made by GENCI. ”

## 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**(2), 113–146.

Baek, H., Calandra, H., & Demanet, L., 2014. Velocity estimation via registration-guided least-squares inversion, *Geophysics*, **79**(2), R79–R89.

Benamou, J. D., Froese, B. D., & Oberman, A. M., 2014. Numerical solution of the Optimal Transportation problem using the Monge-Ampère equation, *Journal of Computational Physics*, **260**, 107–126.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., & Peyr, G., 2015. Iterative Bregman Projections for Regularized Transportation Problems, *SIAM Journal on Scientific Computing*, **37**(2), A1111–A1138.

Billette, F. J. & Brandsberg-Dahl, S., 2004. The 2004 BP velocity benchmark, in *Extended Abstracts, 67$^{th}$ Annual EAGE Conference & Exhibition, Madrid, Spain*, p. B035.

Biondi, B. & Almomin, A., 2013. Tomographic full waveform inversion (TFWI) by combining FWI and wave-equation migration velocity analysis, *The Leading Edge*, **September, special section: full waveform inversion**, 1074–1080.

Biondi, B. & Symes, W., 2004. Angle-domain common-image gathers for migration velocity analysis by wavefield-continuation imaging, *Geophysics*, **69(5)**, 1283–1298.

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., & Sagastizábal, C. A., 2006. *Numerical Optimization, Theoretical and Practical Aspects*, Springer series, Universitext.

Borisov, D. & Singh, S. C., 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.

Brandt, A., 1977. Multi-level adaptive solutions to boundary-value problems, *Mathematics of Computation*, **31**, 333–390.

Brenier, Y., 1991. Polar factorization and monotone rearrangement of vector-valued functions, *Communications on Pure and Applied Mathematics*, **44**(4), 375–417.

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

Brossier, R., Operto, S., & Virieux, J., 2010. Which data residual norm for robust elastic frequency-domain full waveform inversion?, *Geophysics*, **75**(3), R37–R46.

Brossier, R., Pajot, B., Combe, L., Operto, S., Métivier, L., & Virieux, J., 2014. Time and frequency-domain FWI implementations based on time solver: analysis of computational complexities, in *Expanded Abstracts, 76$^{th}$ Annual EAGE Meeting (Amsterdam)*.

Brossier, R., Operto, S., & Virieux, J., 2015. Velocity model building from seismic reflection data by full waveform inversion, *Geophysical Prospecting*, **63**, 354–367.

Bunks, C., Salek, F. M., Zaleski, S., & Chavent, G., 1995. Multiscale seismic waveform inversion, *Geophysics*, **60**(5), 1457–1473.

Bŏzdag, E., Trampert, J., & Tromp, J., 2011. Misfit functions for full waveform inversion based on instanta-

neous phase and envelope measurements, *Geophysical Journal International*, **185**(2), 845–870.

Buzbee, B. L., Golub, G. H., & Nielson, C. W., 1970. On direct methods for solving poisson's equations, *SIAM Journal on Numerical Analysis*, **7**(4), pp. 627–656.

Castellanos, C., Métivier, L., Operto, S., Brossier, R., & Virieux, J., 2015. Fast full waveform inversion with source encoding and second-order optimization methods, *Geophysical Journal International*, **200(2)**, 720–744.

Chavent, G., 1974. Identification of parameter distributed systems, in *Identification of function parameters in partial differential equations*, pp. 31–48, eds Goodson, R. & Polis, M., American Society of Mechanical Engineers, New York.

Chavent, G., Clément, F., & Gòmez, S., 1994. Automatic determination of velocities via migration-based traveltime waveform inversion: A synthetic data example, *SEG Technical Program Expanded Abstracts 1994*, pp. 1179–1182.

Claerbout, J., 1985. *Imaging the Earth's interior*, Blackwell Scientific Publication.

Clapp, R., 2009. *Reverse time migration with random boundaries*, chap. 564, pp. 2809–2813.

Combettes, P. L. & Pesquet, J.-C., 2011. Proximal splitting methods in signal processing, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, vol. 49 of **Springer Optimization and Its Applications**, pp. 185–212, eds Bauschke, H. H., Burachik, R. S., Combettes, P. L., Elser, V., Luke, D. R., & Wolkowicz, H., Springer New York.

Dahlen, F. A., Hung, S. H., & Nolet, G., 2000. Fréchet kernels for finite-difference traveltimes - I. theory, *Geophysical Journal International*, **141**, 157–174.

Devaney, A., 1984. Geophysical diffraction tomography, *Geoscience and Remote Sensing, IEEE Transactions on*, **GE-22**(1), 3–13.

Engquist, B. & Froese, B. D., 2014. Application of the wasserstein metric to seismic signals, *Communications in Mathematical Science*, **12**(5), 979–988.

Evans, L. C., 1997. Partial differential equations and Monge–Kantorovich mass transfer, *Current developments in mathematics*, pp. 65–126.

Faye, J. P. & Jeannot, J. P., 1986. Prestack migration velocities from focusing depth analysis, in *Expanded Abstracts*, pp. 438–440, Soc. Expl. Geophys.

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

Fichtner, A., Kennett, B. L. N., Igel, H., & Bunge, H. P., 2010. Full waveform tomography for radially anisotropic structure: New insights into present and past states of the Australasian upper mantle, *Earth and Planetary Science Lettters*, **290**(3-4), 270–280.

Golub, G. H., 1996. *Matrix Computation, third edition*, Johns Hopkins Studies in Mathematical Sciences.

Hale, D., 2013. Dynamic warping of seismic images, *Geophysics*, **78**(2), S105–S115.

Hanin, L. G., 1992. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces, *Pro-*

*ceedings of American Mathematical Society*, **115**, 345–352.

Hicks, G. J., 2002. Arbitrary source and receiver positioning in finite-difference schemes using Kaiser windowed sinc functions, *Geophysics*, **67**, 156–166.

Jannane, M., Beydoun, W., Crase, E., Cao, D., Koren, Z., Landa, E., Mendes, M., Pica, A., Noble, M., Roeth, G., Singh, S., Snieder, R., Tarantola, A., & Trezeguet, D., 1989. Wavelengths of Earth structures that can be resolved from seismic reflection data, *Geophysics*, **54**(7), 906–910.

Knott, M. & Smith, C., 1984. On the optimal mapping of distributions, *Journal of Optimization Theory and Applications*, **43**(1), 39–49.

Lailly, P., 1983. The seismic problem as a sequence of before-stack migrations, in *Conference on Inverse Scattering: Theory and Applications*, SIAM, Philadelphia.

Lambaré, G., 2002. The use of locally coherent events in depth processing : a state of the art., in *Extended Abstracts, 72nd annual meeting , (6-10 October 2002, Salt Lake City)*, pp. 2261–2264, Soc. Expl. Geophys.

Lions, J. L., 1968. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris.

Luo, J. & Wu, R.-S., 2015. Seismic envelope inversion: reduction of local minima and noise resistance, *Geophysical Prospecting*, **63**(3), 597–614.

Luo, S. & Sava, P., 2011. A deconvolution-based objective function for wave-equation inversion, *SEG Technical Program Expanded Abstracts*, **30**(1), 2788–2792.

Luo, Y. & Schuster, G. T., 1991. Wave-equation traveltime inversion, *Geophysics*, **56**(5), 645–653.

Ma, Y. & Hale, D., 2013. Wave-equation reflection traveltime inversion with dynamic warping and full waveform inversion, *Geophysics*, **78**(6), R223–R233.

Maggi, A., Tape, C., Chen, M., Chao, D., & Tromp, J., 2009. An automated time-window selection algorithm for seismic tomography, *Geophysical Journal International*, **178**, 257–281.

Martin, G. S., Wiley, R., & Marfurt, K. J., 2006. Marmousi2: An elastic upgrade for Marmousi, *The Leading Edge*, **25**(2), 156–166.

Mérigot, Q., 2011. A multiscale approach to optimal transport, *Computer Graphics Forum*, **30**(5), 1583–1592.

Métivier, L. & Brossier, R., 2015. The seiscope optimization toolbox: A large-scale nonlinear optimization library based on reverse communication, *Geophysics*, pp. in–press.

Métivier, L., Brossier, R., Virieux, J., & Operto, S., 2013. Full Waveform Inversion and the truncated Newton method, *SIAM Journal On Scientific Computing*, **35(2)**, B401–B437.

Métivier, L., Bretaudeau, F., Brossier, R., Operto, S., & Virieux, J., 2014a. Full waveform inversion and the truncated Newton method: quantitative imaging of complex subsurface structures, *Geophysical Prospecting*, **62**, 1353–1375.

Métivier, L., Brossier, R., Labb, S., & Virieux, J., 2014b. Multi-parameter FWI - an illustration of the Hessian operator role for mitigating trade-offs between parameter classes, in *Expanded Abstracts, $6^{th}$ EAGE St-Petersbourg International Conference & Exhibition*.

Métivier, L., Brossier, R., Operto, S., & Virieux, J., 2015. Acoustic multi-parameter FWI for the reconstruc-

tion of P-wave velocity, density and attenuation: preconditioned truncated Newton approach, in *Expanded Abstracts, 85$^{th}$ Annual SEG Meeting (New Orleans)*.

Montelli, R., Nolet, G., Dahlen, F. A., Masters, G., Engdahl, E. R., & Hung, S. H., 2004. Finite-frequency tomography reveals a variety of plumes in the mantle, *Science*, **303**, 338–343.

Nocedal, J., 1980. Updating Quasi-Newton Matrices With Limited Storage, *Mathematics of Computation*, **35**(151), 773–782.

Nocedal, J. & Wright, S. J., 2006. *Numerical Optimization*, Springer, 2nd edn.

Nolet, G., 2008. *A Breviary of Seismic Tomography*, Cambridge University Press, Cambridge, UK.

Operto, S., Ravaut, C., Improta, L., Virieux, J., Herrero, A., & Dell'Aversana, P., 2004. Quantitative imaging of complex structures from dense wide-aperture seismic data by multiscale traveltime and waveform inversions: a case study, *Geophysical Prospecting*, **52**, 625–651.

Operto, S., Brossier, R., Gholami, Y., Métivier, L., Prieux, V., Ribodetti, A., & Virieux, J., 2013. A guided tour of multiparameter full waveform inversion for multicomponent data: from theory to practice, *The Leading Edge*, **Special section Full Waveform Inversion**(September), 1040–1054.

Operto, S., Miniussi, A., Brossier, R., Combe, L., Métivier, L., Monteiller, V., Ribodetti, A., & Virieux, J., 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**(2), 1362–1391.

Peter, D., Komatitsch, D., Luo, Y., Martin, R., Le Goff, N., Casarotti, E., Le Loher, P., Magnoni, F., Liu, Q., Blitz, C., Nissen-Meyer, T., Basini, P., & Tromp, J., 2011. Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes, *Geophysical Journal International*, **186**(2), 721–739.

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**(2), 495–503.

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

Plessix, R. E., Chavent, G., & Roeck, Y.-H. D., 1999. Waveform inversion of reflection seismic data for kinematic parameters by local inversion, *SIAM Journal of Scientific Computing*, **20**, 1033–1052.

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

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

Rockafellar, R. T., 1976. Monotone operators and the proximal point algorithm, *SIAM Journal on Control and Optimization*, **14**(5), 877–898.

Sava, P. & Biondi, B., 2004a. Wave-equation migration velocity analysis. i. theory, *Geophysical Prospecting*, **52**(6), 593–606.

Sava, P. & Biondi, B., 2004b. Wave-equation migration velocity analysis. ii. subsalt imaging examples, *Geophysical Prospecting*, **52**(6), 607–623.

Sava, P. & Fomel, S., 2006. Time-shift imaging condition in seismic migration, *Geophysics*, **71**(6), S209–S217.

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

Sirgue, L. & Pratt, R. G., 2004. Efficient waveform inversion and imaging : a strategy for selecting temporal frequencies, *Geophysics*, **69**(1), 231–248.

Sirgue, L., Barkved, O. I., Dellinger, J., Etgen, J., Albertin, U., & Kommedal, J. H., 2010. Full waveform inversion: the next leap forward in imaging at Valhall, *First Break*, **28**, 65–70.

Swarztrauber, P. N., 1974. A Direct Method for the Discrete Solution of Separable Elliptic Equations, *SIAM Journal on Numerical Analysis*, **11**(6), 1136–1150.

Symes, W. & Kern, M., 1994. Inversion of reflection seismograms by differential semblance analysis: algorithm structure and synthetic examples, *Geophysical Prospecting*, **42**, 565–614.

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

Tape, C., Liu, Q., Maggi, A., & Tromp, J., 2010. Seismic tomography of the southern California crust based on spectral-element and adjoint methods, *Geophysical Journal International*, **180**, 433–462.

Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysics*, **49**(8), 1259–1266.

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

van Leeuwen, T. & Mulder, W. A., 2010. A correlation-based misfit criterion for wave-equation traveltime tomography, *Geophysical Journal International*, **182**(3), 1383–1394.

Vigh, D., Jiao, K., Watts, D., & Sun, D., 2014. Elastic full-waveform inversion application using multicomponent measurements of seismic data collection, *Geophysics*, **79**(2), R63–R77.

Villani, C., 2003. *Topics in optimal transportation*, Graduate Studies In Mathematics, Vol. 50, AMS.

Villani, C., 2008. *Optimal transport : old and new*, Grundlehren der mathematischen Wissenschaften, Springer, Berlin.

Wang, Y. & Rao, Y., 2009. Reflection seismic waveform tomography, *Journal of Geophysical Research*, **114**(B03304), doi:10.1029/2008JB005916.

Warner, M. & Guasch, L., 2014. Adaptative waveform inversion - fwi without cycle skipping - theory, in *76th EAGE Conference and Exhibition 2014*.

Warner, M., Ratcliffe, A., Nangoo, T., Morgan, J., Umpleby, A., Shah, N., Vinje, V., Stekl, I., Guasch, L., Win, C., Conroy, G., & Bertrand, A., 2013. Anisotropic 3D full-waveform inversion, *Geophysics*, **78**(2), R59–R80.

Xu, S., Wang, D., Chen, F., Lambaré, G., & Zhang, Y., 2012. Inversion on reflected seismic wave, *SEG Technical Program Expanded Abstracts 2012*, pp. 1–7.

Zhou, W., Brossier, R., Operto, S., & Virieux, J., 2015. Full waveform inversion of diving & reflected waves for velocity model building with impedance inversion based on scale separation, *Geophysical Journal International*, **202**(3), 1535–1554.

1049 Zhu, H., Bŏzdag, E., Peter, D., & Tromp, J., 2012. Structure of the european upper mantle revealed by adjoint
1050 tomography, *Nature Geoscience*, **5**, 493–498.

## Tables

| Inversion step | Recording time | $l$-BFGS iterations | Smoothing |
|---|---|---|---|
| 1 | 4.6 s | 218 | $r_z = 125$ m, $r_x = 125$ m |
| 2 | 4.6 s | 251 | $r_z = 125$ m, $r_x = 125$ m |
| 3 | 4.6 s | 150 | $r_z = 125$ m, $r_x = 125$ m |
| 4 | 5.75 s | 279 | $r_z = 75$ m, $r_x = 75$ m |
| 5 | 5.75 s | 199 | $r_z = 75$ m, $r_x = 75$ m |
| 6 | 6.9 s | 130 | $r_z = 75$ m, $r_x = 75$ m |
| 7 | 6.9 s | 230 | $r_z = 75$ m, $r_x = 75$ m |
| 8 | 8.05 s | 177 | $r_z = 75$ m, $r_x = 75$ m |
| 9 | 8.05 s | 269 | $r_z = 75$ m, $r_x = 75$ m |
| 10 | 8.05 s | 283 | $r_z = 75$ m, $r_x = 75$ m |
| 11 | 9.2 s | 152 | $r_z = 75$ m, $r_x = 75$ m |
| 12 | 9.2 s | 366 | $r_z = 75$ m, $r_x = 75$ m |
| 13 | 10.35 s | 192 | $r_z = 75$ m, $r_x = 75$ m |
| 14 | 10.35 s | 287 | $r_z = 75$ m, $r_x = 75$ m |
| 15 | 10.35 s | 144 | $r_z = 75$ m, $r_x = 75$ m |

**Table 1.** Workflow followed for the BP 2004 case study.

1052

| Band | Range | Steps | $l$-BFGS iterations | Final smoothing |
|---|---|---|---|---|
| 1 | 2-4 Hz | $\Delta x = 37.5$ m, $\Delta t = 0.004$ s | 20 | $r_z = 112.5$ m, $r_x = 750$ m |
| 2 | 2-4.5 Hz | $\Delta x = 37.5$ m, $\Delta t = 0.004$ s | 20 | $r_z = 112.5$ m, $r_x = 750$ m |
| 3 | 2-5 Hz | $\Delta x = 37.5$ m, $\Delta t = 0.004$ s | 20 | $r_z = 112.5$ m, $r_x = 750$ m |
| 4 | 2-5.5 Hz | $\Delta x = 37.5$ m, $\Delta t = 0.004$ s | 20 | $r_z = 112.5$ m, $r_x = 750$ m |
| 5 | 2-6 Hz | $\Delta x = 37.5$ m, $\Delta t = 0.004$ s | 20 | $r_z = 112.5$ m, $r_x = 750$ m |
| 6 | 2-7 Hz | $\Delta x = 37.5$ m, $\Delta t = 0.004$ s | 20 | $r_z = 112.5$ m, $r_x = 750$ m |
| 7 | 2-8 Hz | $\Delta x = 37.5$ m, $\Delta t = 0.004$ s | 20 | $r_z = 37.5$ m, $r_x = 375$ m |
| 8 | 2-10 Hz | $\Delta x = 37.5$ m, $\Delta t = 0.004$ s | 20 | $r_z = 37.5$ m, $r_x = 375$ m |
| 9 | 2-11 Hz | $\Delta x = 25$ m, $\Delta t = 0.003$ s | 20 | $r_z = 25$ m, $r_x = 250$ m |
| 10 | 2-12 Hz | $\Delta x = 25$ m, $\Delta t = 0.003$ s | 20 | $r_z = 25$ m, $r_x = 250$ m |
| 11 | 2-14 Hz | $\Delta x = 25$ m, $\Delta t = 0.003$ s | 50 | $r_z = 25$ m, $r_x = 250$ m |
| 12 | 2-16 Hz | $\Delta x = 25$ m, $\Delta t = 0.003$ s | 50 | $r_z = 0$ m, $r_x = 250$ m |
| 13 | 2-18 Hz | $\Delta x = 12.5$ m, $\Delta t = 0.001$ s | 40 | $r_z = 0$ m, $r_x = 250$ m |
| 14 | 2-20 Hz | $\Delta x = 12.5$ m, $\Delta t = 0.001$ s | 40 | $r_z = 0$ m, $r_x = 250$ m |
| 15 | 2-25 Hz | $\Delta x = 12.5$ m, $\Delta t = 0.001$ s | 40 | $r_z = 0$ m, $r_x = 125$ m |

**Table 2.** Workflow followed for the Chevron 2014 benchmark case study.

| Frequency bands | $N_x \times N_z$ | $N_t$ | Gradient | Incident | Adjoint + incident | SDMM | % of time for SDMM |
|---|---|---|---|---|---|---|---|
| 1-8 | 20,960 | 2001 | 171 s | 9 s | 33 s | 127s | 74% |
| 9-12 | 47,160 | 2667 | 332 s | 39 s | 121 s | 171 s | 51% |
| 13-15 | 1,886,400 | 8001 | 2455 s | 479 s | 1461 s | 511 s | 20% |

**Table 3.** Computational times for one gradient. This time is decomposed in the following steps: computation of the incident wavefield, backpropagation of the adjoint and the incident wavefields, solution of the optimal transport problem.

**Algorithms**

$y_1^0 = 0$, $y_2^0 = 0$, $z_1^0 = 0$, $z_2^0 = 0$;

**for** $n = 0, 1, \ldots$ **do**

$$\varphi^n = \left(I_N + A^T A\right)^{-1} \left[(y_1^n - z_1^n) + A^T (y_2^n - z_2^n)\right];$$

$$y_1^{n+1} = \text{prox}_{h_{d_{cal}[m], d_{obs}}} (\varphi^n + z_1^n) ;$$

$$z_1^{n+1} = z_1^n + \varphi^n - y_1^{n+1} ;$$

$$y_2^{n+1} = \text{prox}_{i_K} (A\varphi^n + z_2^n) ;$$

$$z_2^{n+1} = z_2^n + A\varphi^n - y_2^{n+1} ;$$

**end**

**Algorithm 1**: SDMM method for the solution of the problem (21).

Pre-processing step: compute the eigenvectors of $E$ and store them in $P$;

**for** $i = 1, \ldots, N_t$ **do**
  $b'_i = B^{-1} b_i$;
**end**

**for** $i = 1, \ldots, N_t$ **do**
  $c_i = P^T b'_i$;
**end**

**for** $j = 1, \ldots, N_r$ **do**
  form $c_{*j}$ from $c = [c_1, \ldots, c_{N_t}]$;
  solve $K_j y_{*j} = c_{*j}$;
**end**

**for** $i = 1, \ldots, N_t$ **do**
  form $c_i$ from $c_{*j}$ from $c = [c_{*1}, \ldots, c_{*N_r}]$;
  $x_i = P c_i$;
**end**

**Algorithm 2**: Efficient solution of the block tridiagonal linear system.

## Figures

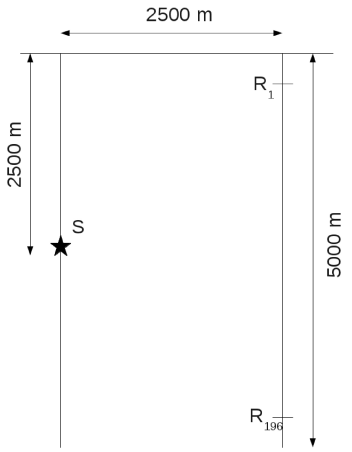


**Figure 1.** Configuration of the borehole to borehole experiment.

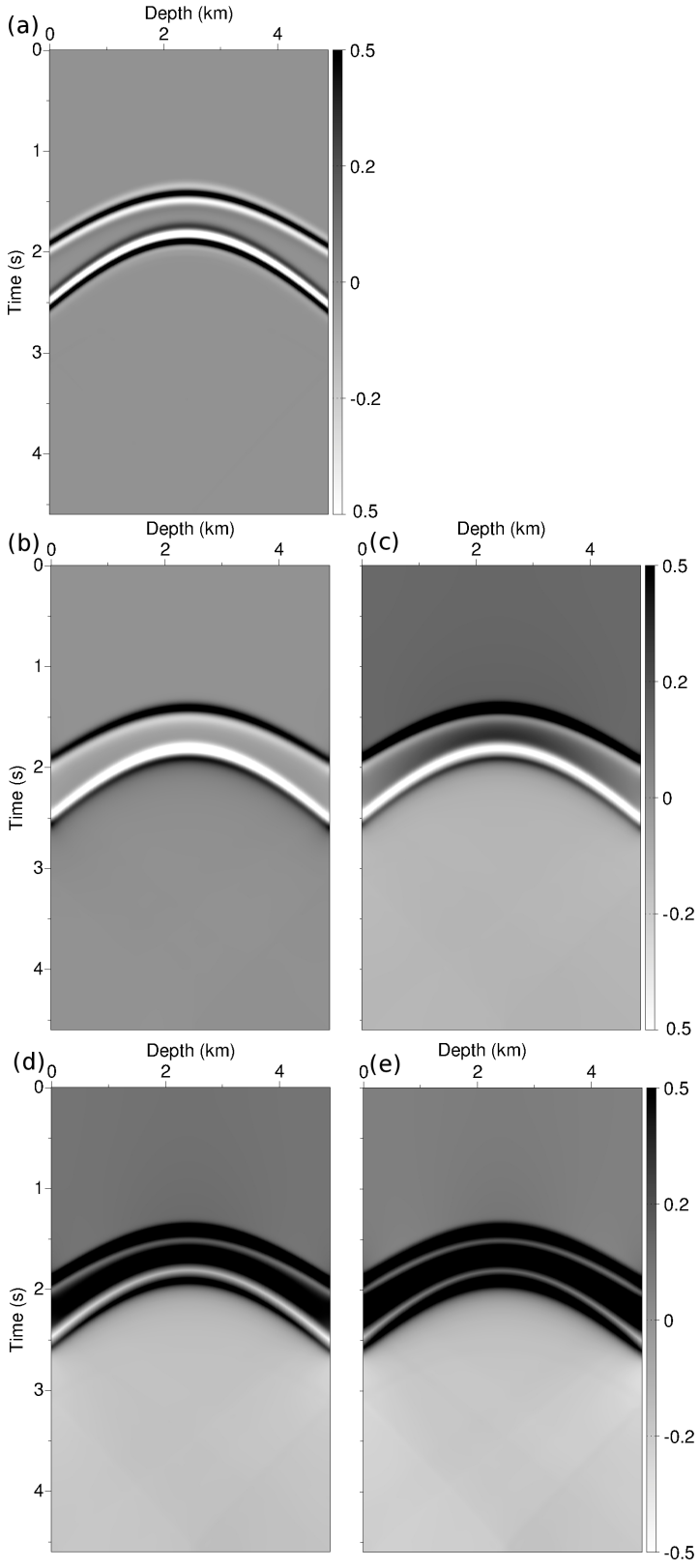


**Figure 2.** L2 residuals (a) and optimal transport based residuals with 5(b), 10 (c), 25 (d), 50 (e) SDMM iterations.

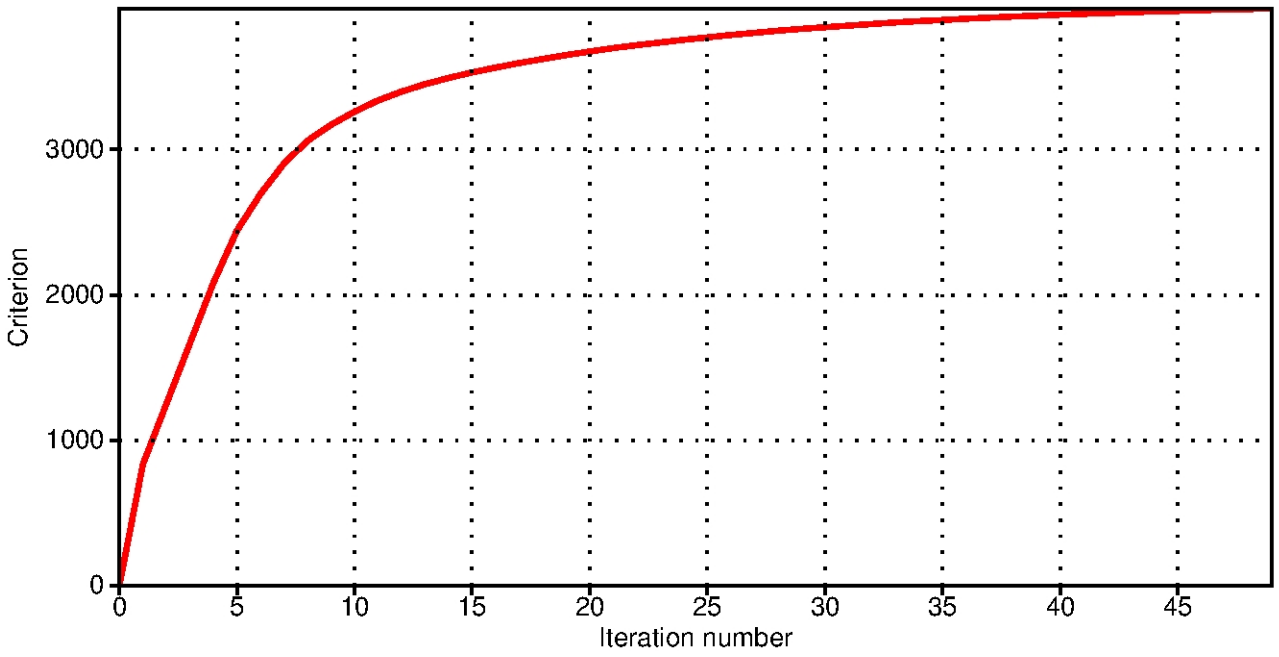


**Figure 3.** Evolution of the criterion maximized by the SDMM method along 50 iterations on the borehole to borehole schematic experiment. The criterion tends asymptotically toward a maximum value of 4000, which suggest that the convergence is reached. This is supported by the evolution of the solution which also seems to have reached a stationary point (Fig. 2).

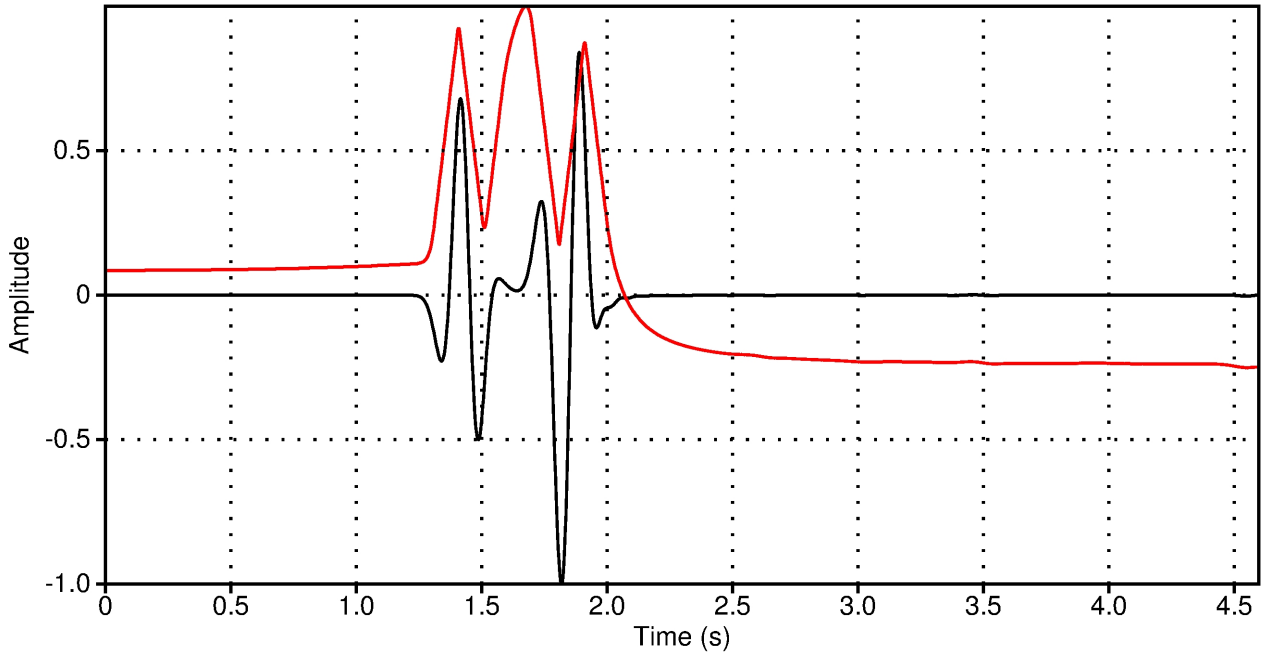


**Figure 4.** Traces extracted at 2.5 km depth from the original residuals (black) and from the solution computed after 50 SDMM iterations for the borehole to borehole schematic experiment.

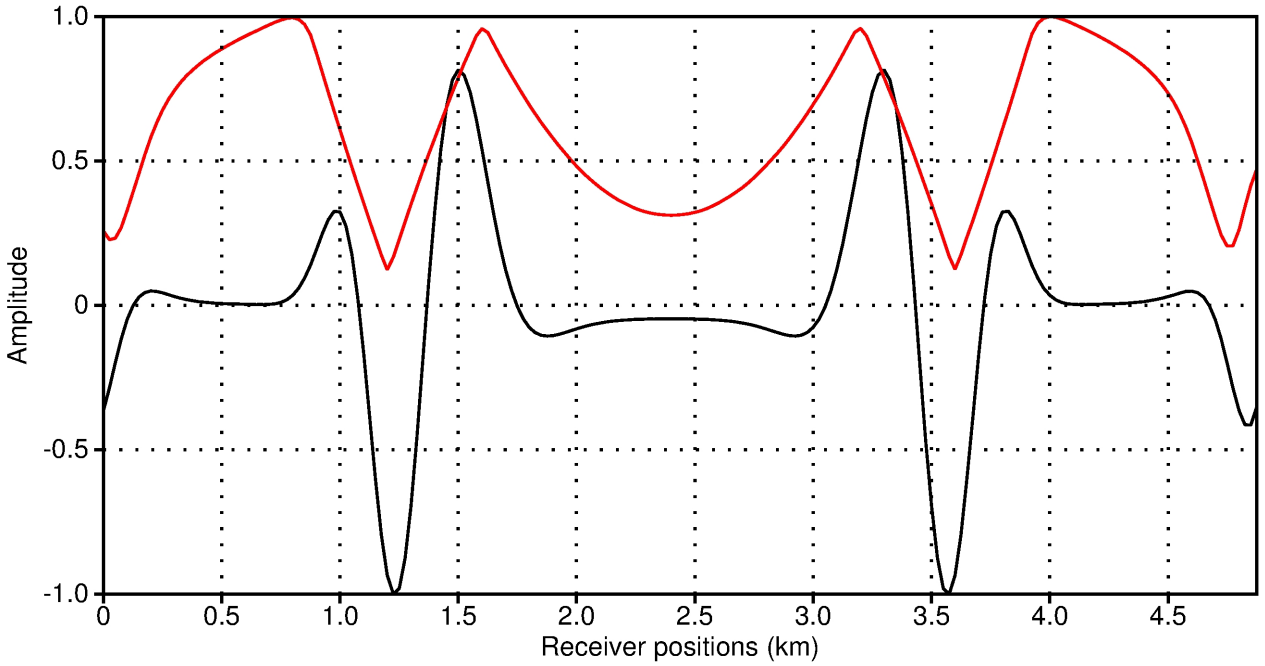


**Figure 5.** Traces extracted at time $t = 2$ s from the original residuals (black) and from the solution computed after 50 SDMM iterations for the borehole to borehole schematic experiment.

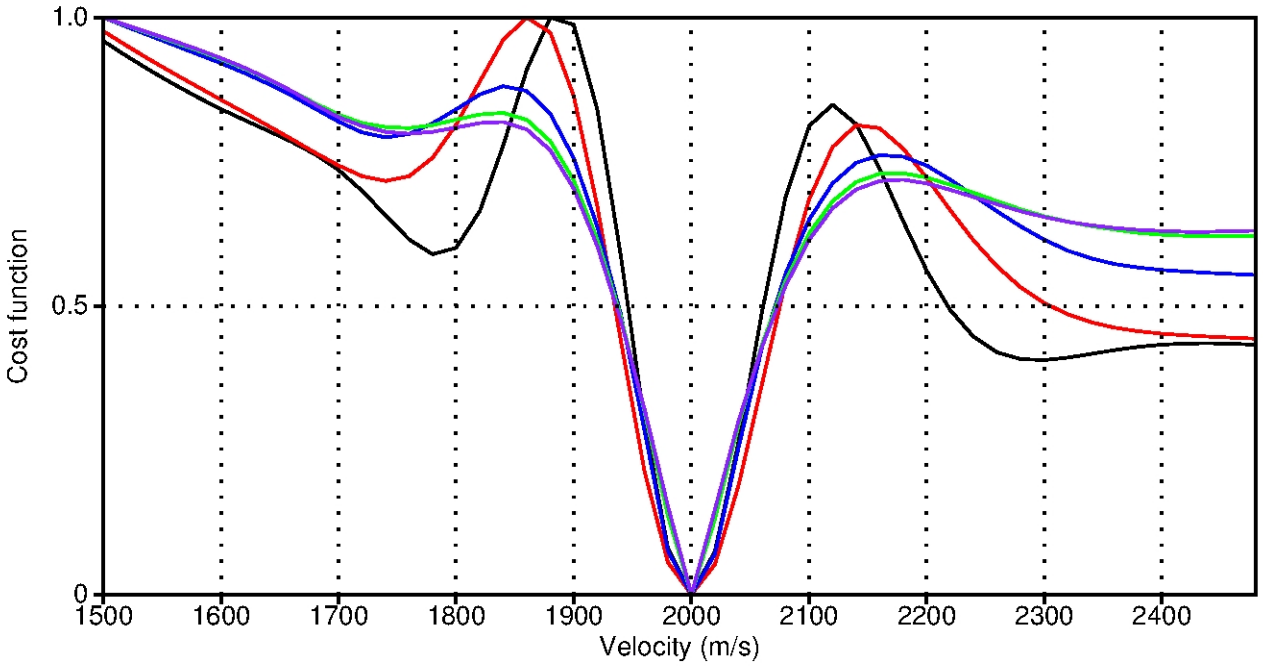


**Figure 6.** $L^2$ misfit function (black) and Wasserstein misfit function obtained with $5$ (red), $10$ (blue), $25$ (green) and $50$ (purple) SDMM iterations. The misfit functions are evaluated for a background velocity value ranging from $1500$ m.s$^{-1}$ to $2500$ m.s$^{-1}$.

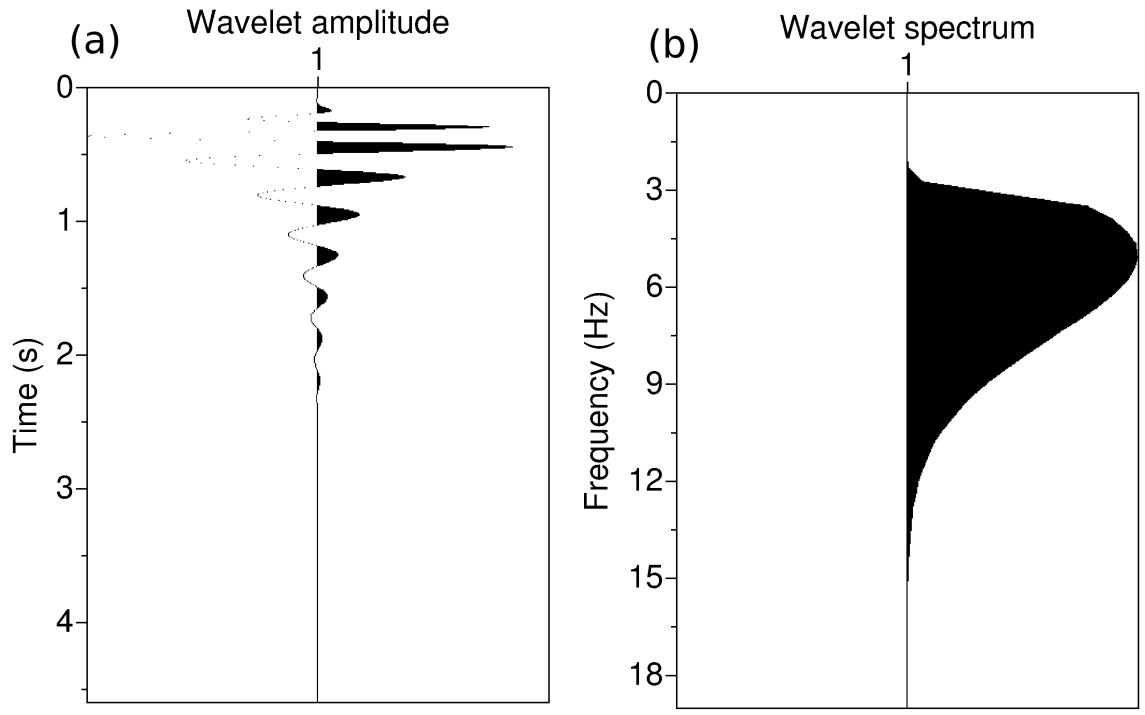


**Figure 7.** Source wavelet used to generate the synthetic dataset on the Marmousi model (a). This source is obtained from a Ricker wavelet centered on 5 Hz after applying a minimum phase Butterworth filter below 2.5 Hz. Corresponding spectrum (b).

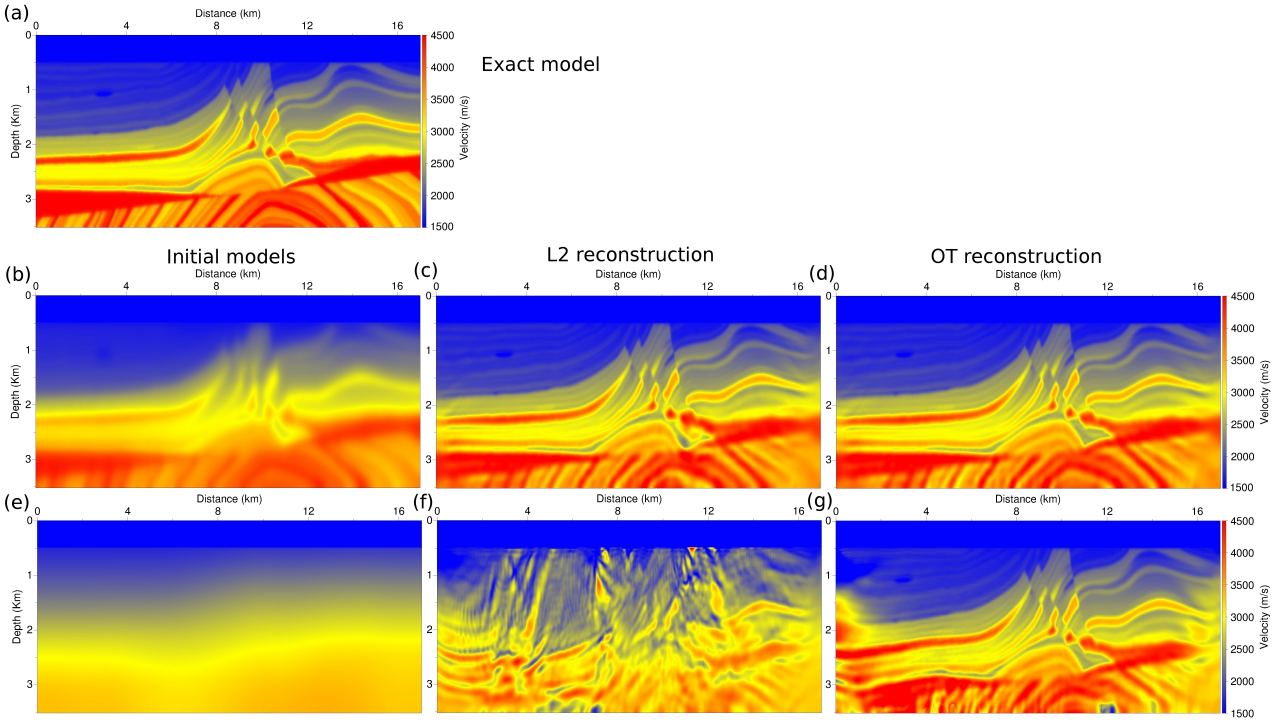


**Figure 8.** Marmousi 2 exact P-wave velocity model (a). Initial P-wave velocity models, computed from the exact model using a Gaussian filter with a correlation length of 250 m (b) and 2000 m (e). Corresponding P-wave velocity estimations with FWI using the $L^2$ misfit function (c),(f). Corresponding P-wave velocity estimations with FWI using the optimal transport distance based misfit function (d),(g).

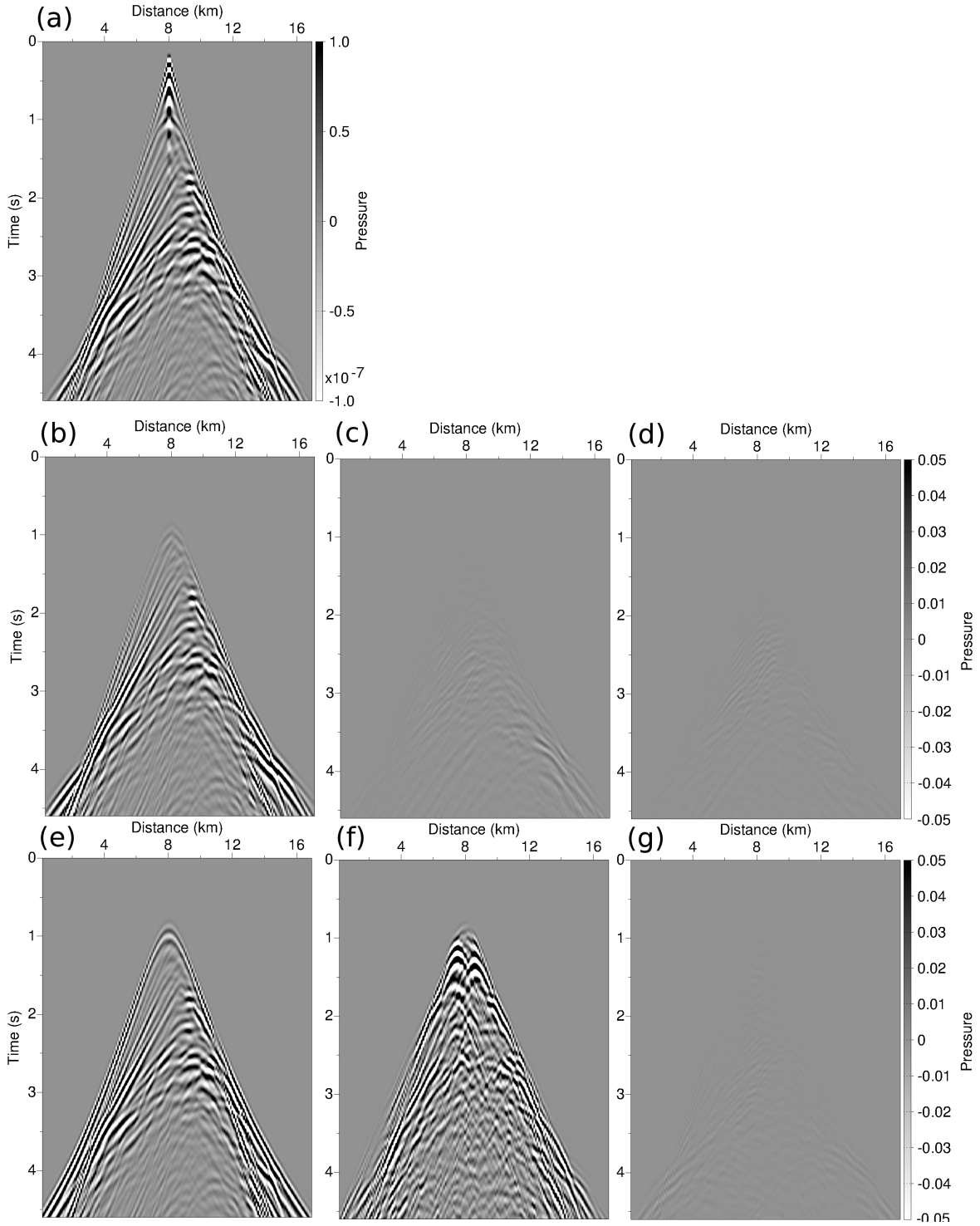


**Figure 9.** Marmousi 2 exact data for the shot-gather corresponding to the source position $x_S = 8$ km (a). Associated residuals in the initial P-wave velocity models (b),(e). Associated residuals in the P-wave velocity models estimated with FWI using the $L^2$ misfit function (c),(f). Associated residuals in the P-wave velocity models estimated with FWI using optimal transport distance based misfit function (d),(g).

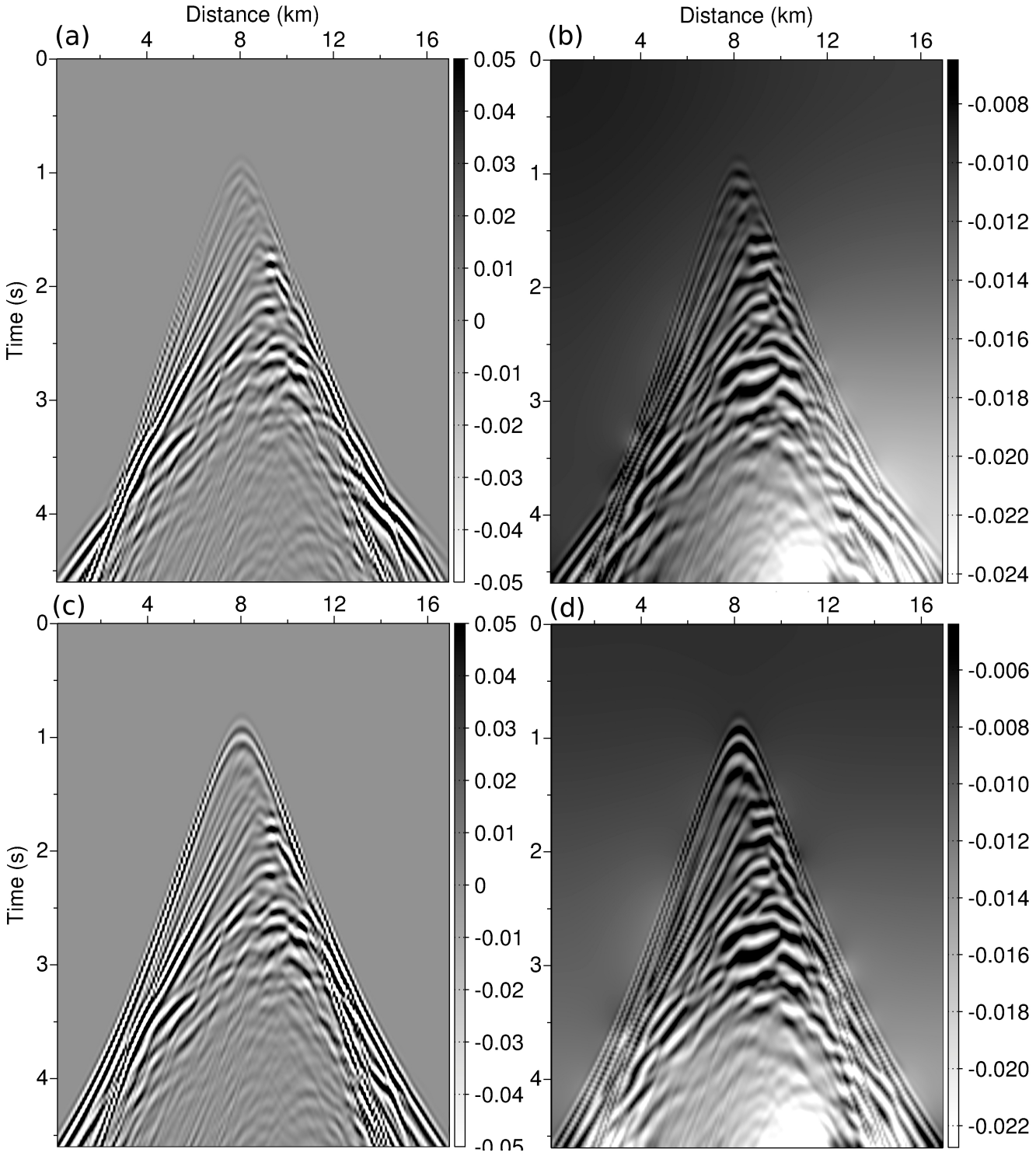


**Figure 10.** $L^2$ residuals in the initial model 1 (a) and 2 (c). Corresponding optimal transport residuals (b),(d).

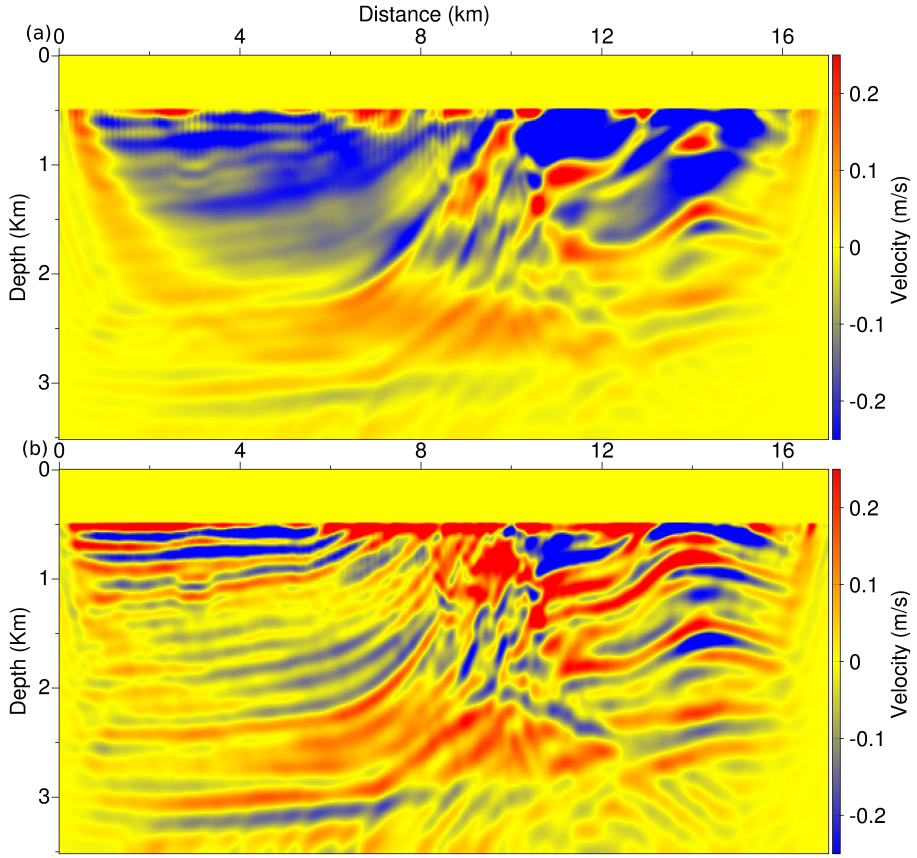


**Figure 11.** Initial descent direction (opposite of the gradient) in the intitial model 3 using the $L^2$ distance (a) and the optimal transport distance (b).

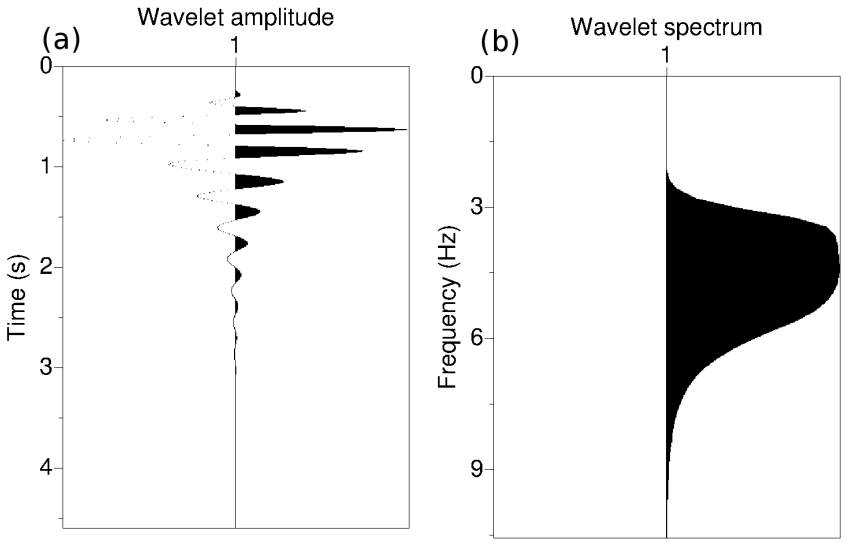


**Figure 12.** Source wavelet used to generate the synthetic dataset on the BP 2004 model (a). This source is obtained from a Ricker wavelet centered on 5 Hz. A whitening of its frequency content is performed before a low-pass and high-pass filter are applied, so that the corresponding spectrum spans an interval from 3 Hz to 9 Hz (b).

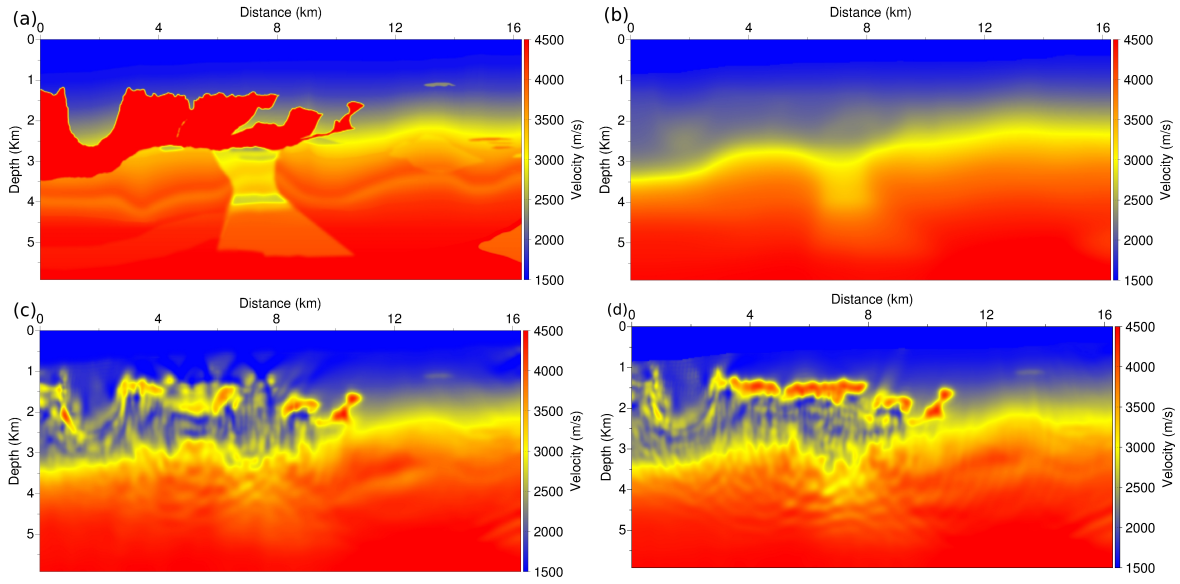


**Figure 13.** BP 2004 exact model (a) and initial model (b). P-wave velocity estimation with a standard $L^2$ norm on short-time window data (4.6 s) (c). The same with the optimal transport distance (d).

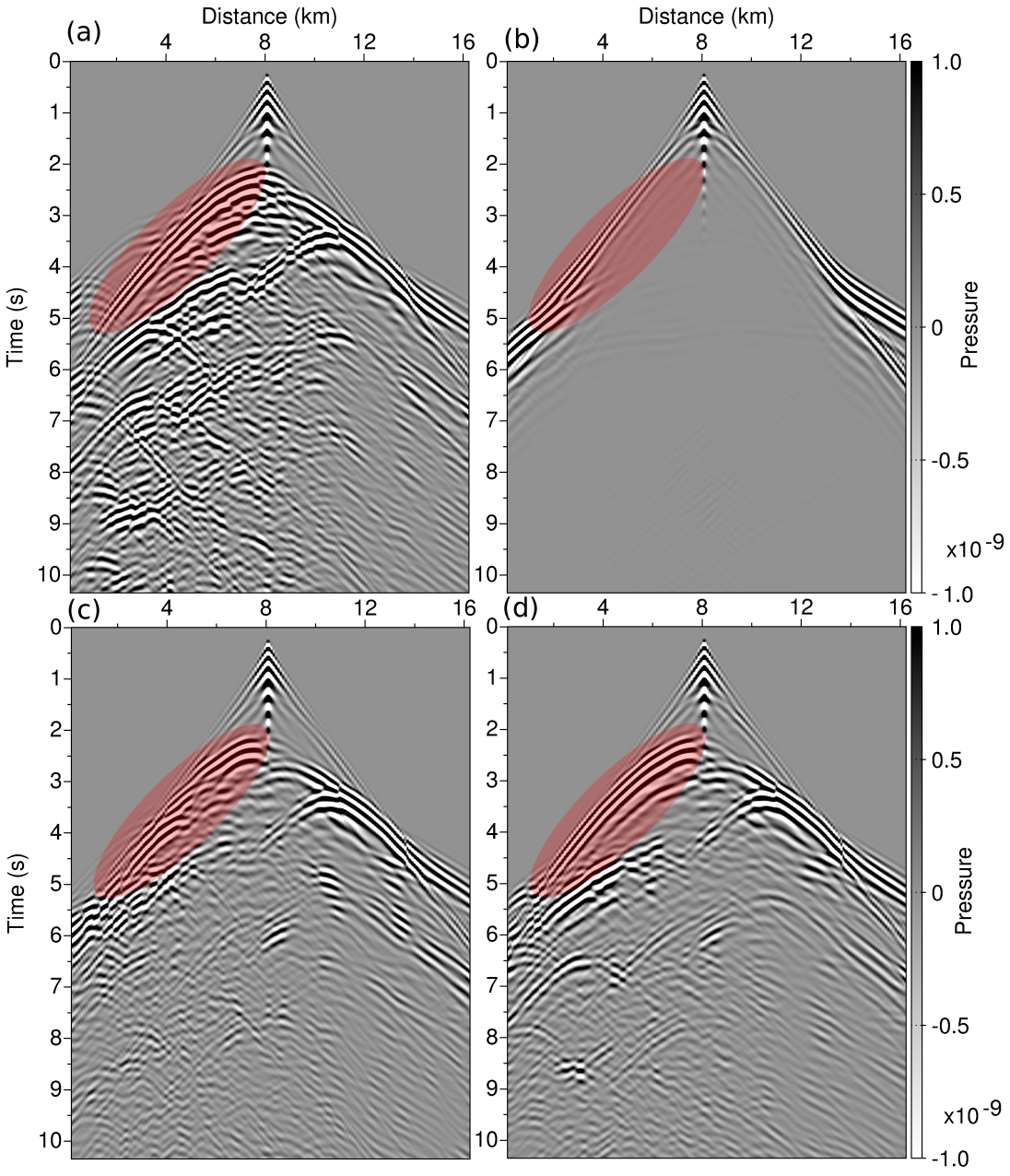


**Figure 14.** BP 2004 exact data (a) and initial data (b). Predicted data in the final model using a standard $L^2$ norm (c). Predicted data in the final model using the optimal transport distance using together with a layer stripping workflow (d). The red ellipses highlight the reflection on the salt roof. This reflection is not present in the initial data (b). Its reconstruction using the $L^2$ distance is discontinuous (c). The use of the optimal transport distance yields a better reconstruction of this event (d).

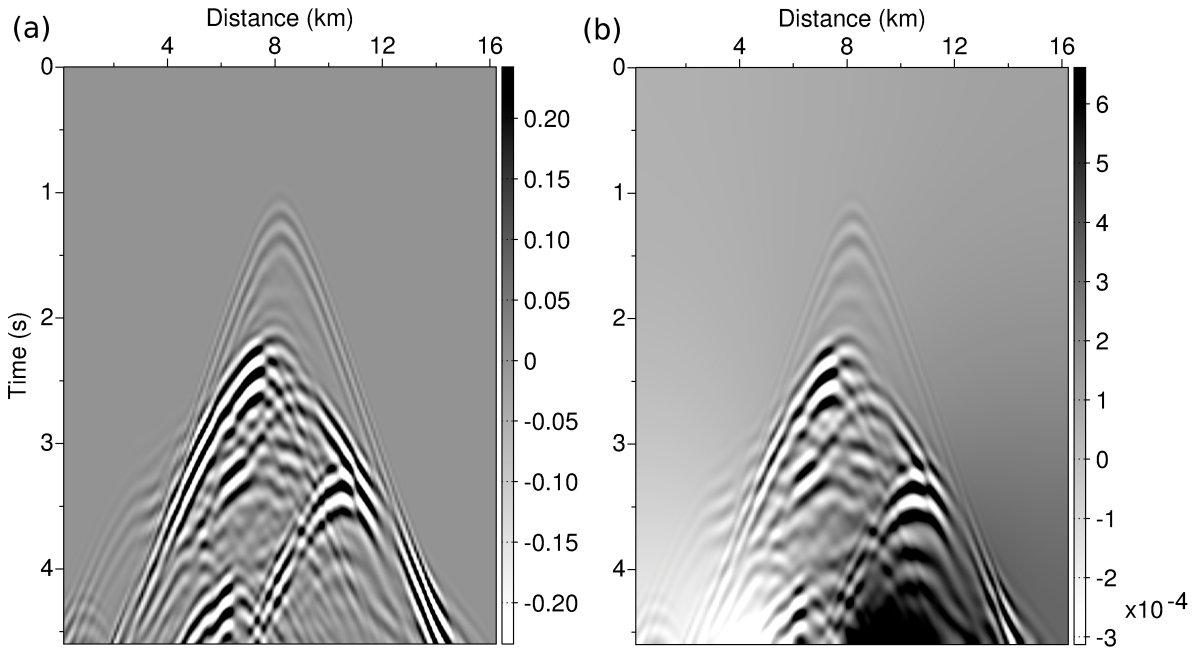


**Figure 15.** BP 2004 case study. $L^2$ residuals in the initial model (a). Optimal transport residuals in the initial model (b).

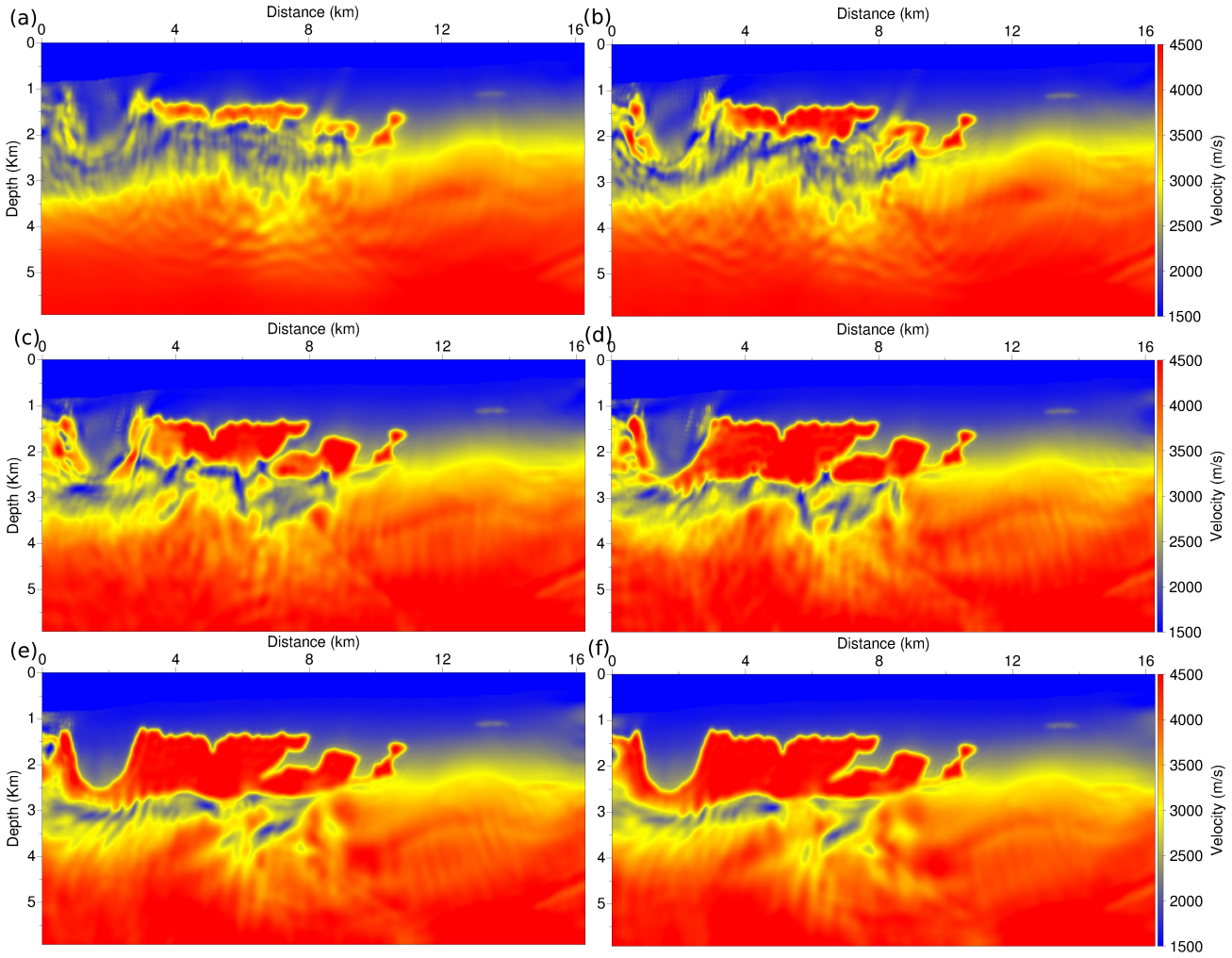


**Figure 16.** BP 2004 P-wave velocity estimation computed after the 1st(a), 3rd (b), 6th (c), 9th (d), 12th (e), and 15th (f) inversion using the optimal transport distance.

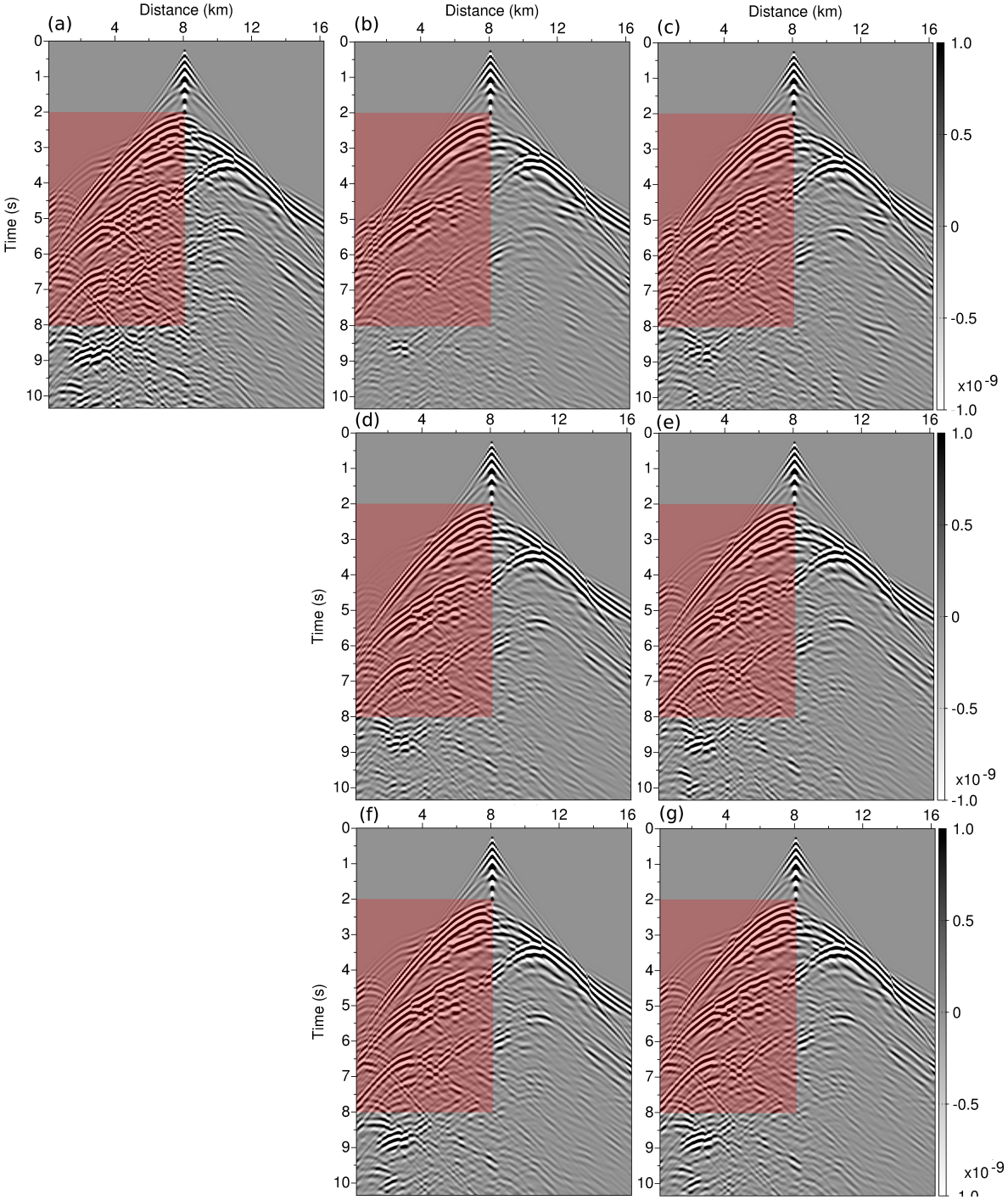


**Figure 17.** Synthetic data in the exact model (a) and in the intermediate models obtained with FWI using an optimal transport distance after the $1^{st}$(b), $3^{rd}$ (c), $6^{th}$ (d), $9^{th}$ (e), $12^{th}$ (f), and $15^{th}$ (g) inversion. The red rectangles highlight the shot-gather zone associated with the diving waves traveling within the salt dome and the reflections generated by deeper interfaces.

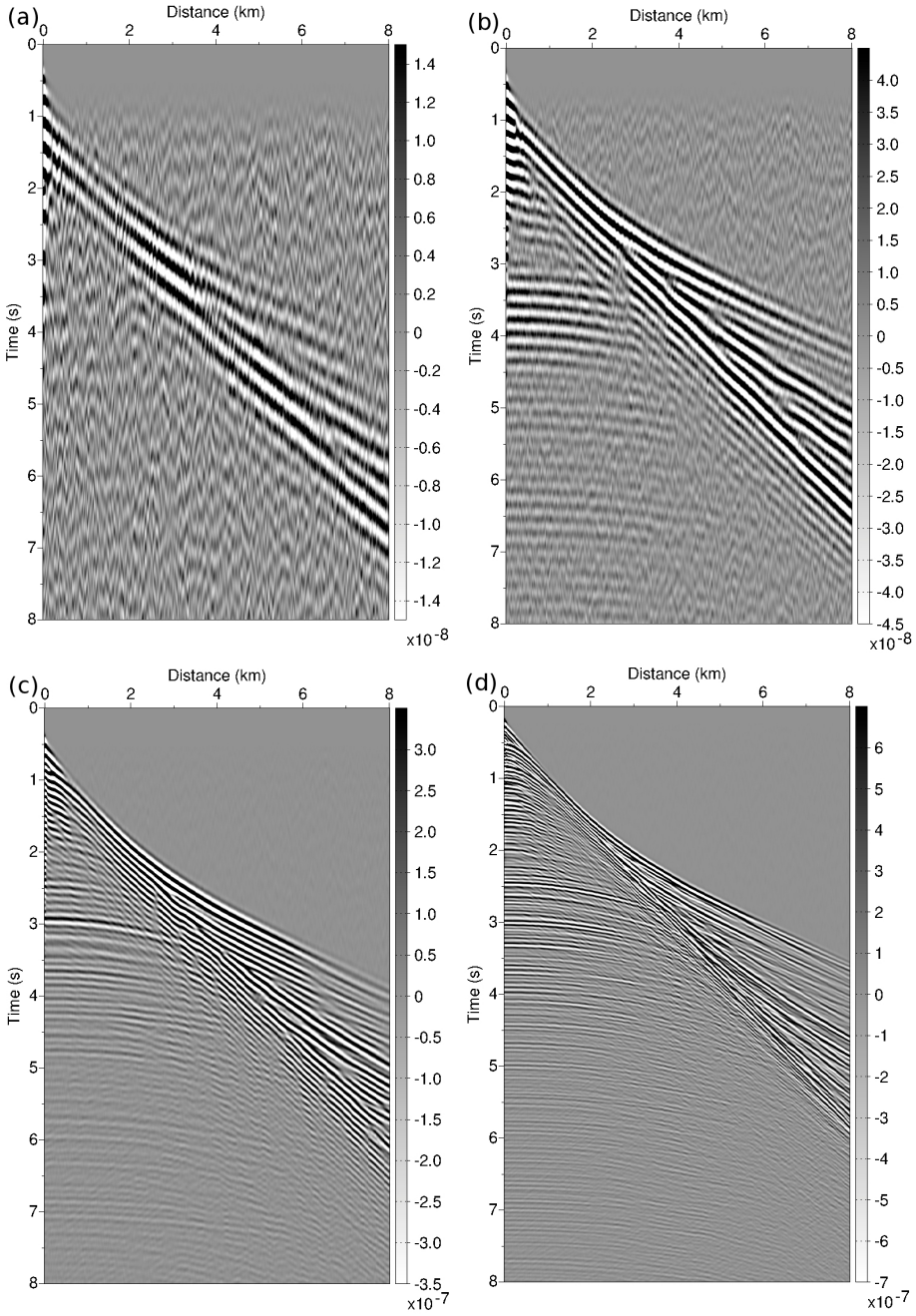


**Figure 18.** Chevron 2014 dataset. Common shot-gather for the source situated at $x = 0$ km for the frequency bands 1 (a), 5 (b), 10 (c), and 15 (d).

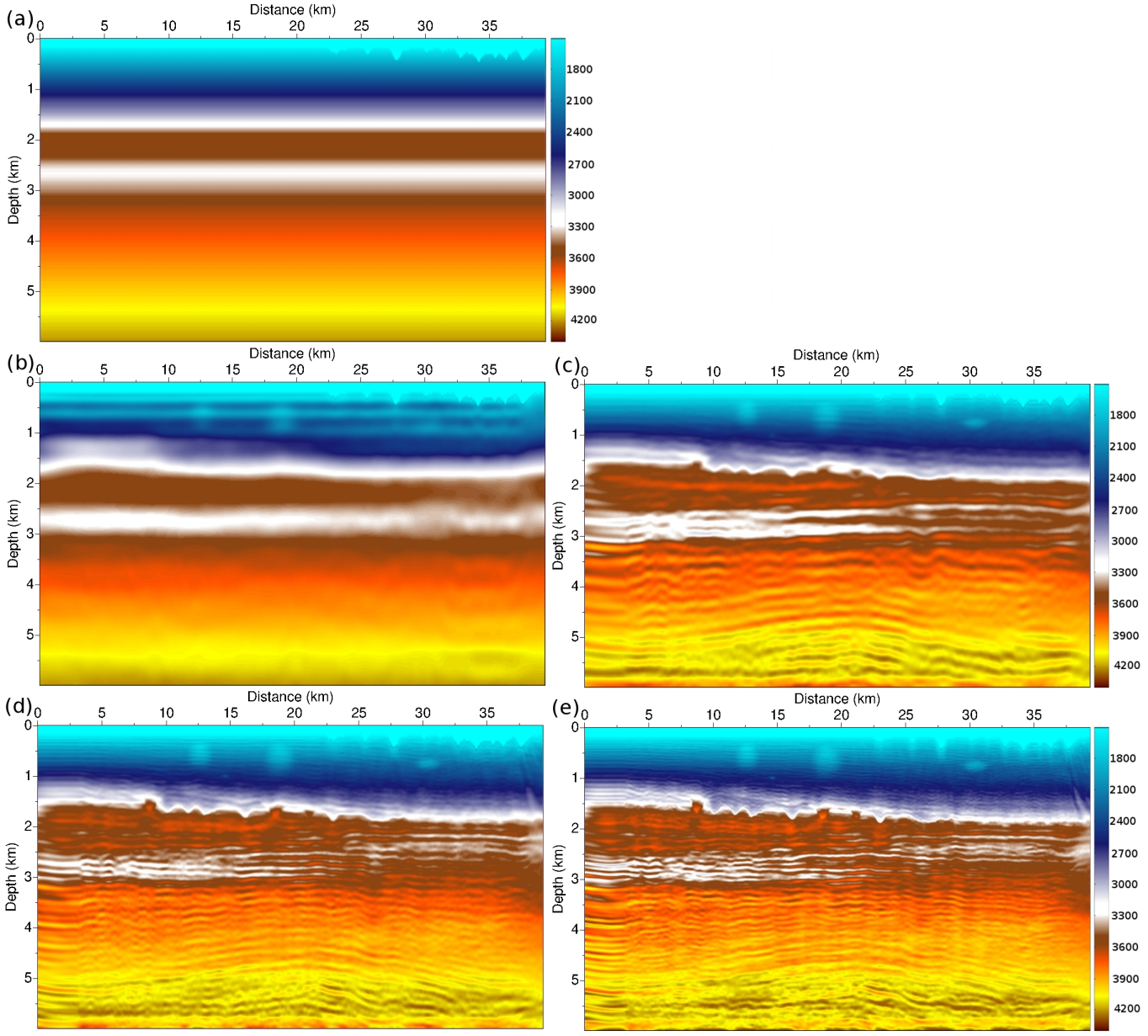


**Figure 19.** Chevron 2014 starting P-wave velocity model (a). Estimated P-wave velocity model at 4 Hz (b), 10 Hz (c), 16 Hz (d), 25 Hz (e).

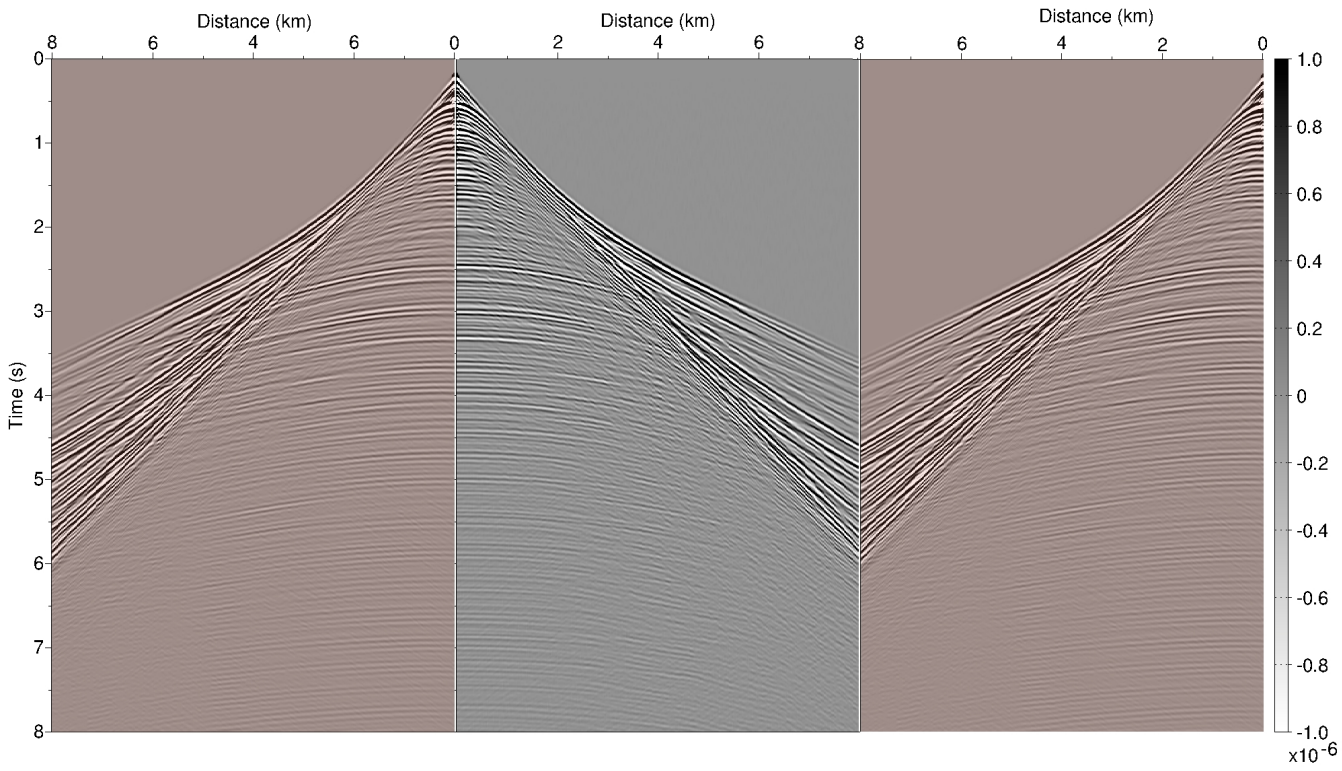


**Figure 20.** Exact common shot-gather for the left most source at 25 Hz, compared to the corresponding synthetic in the final model at 25 Hz (orange panels). The synthetic data is mirrored and placed on both sides of the real data to better compare the match of the different phases.

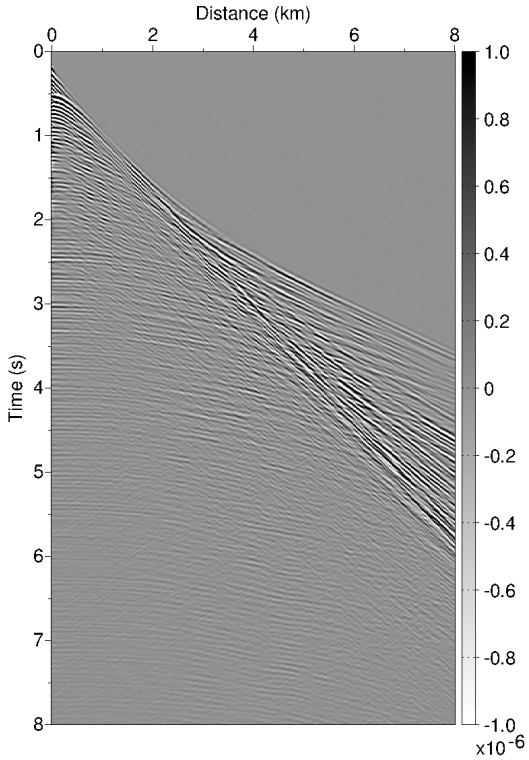


**Figure 21.** Residuals between the exact common shot-gather for the left most source at 25 Hz and the corresponding synthetic common shot-gather.

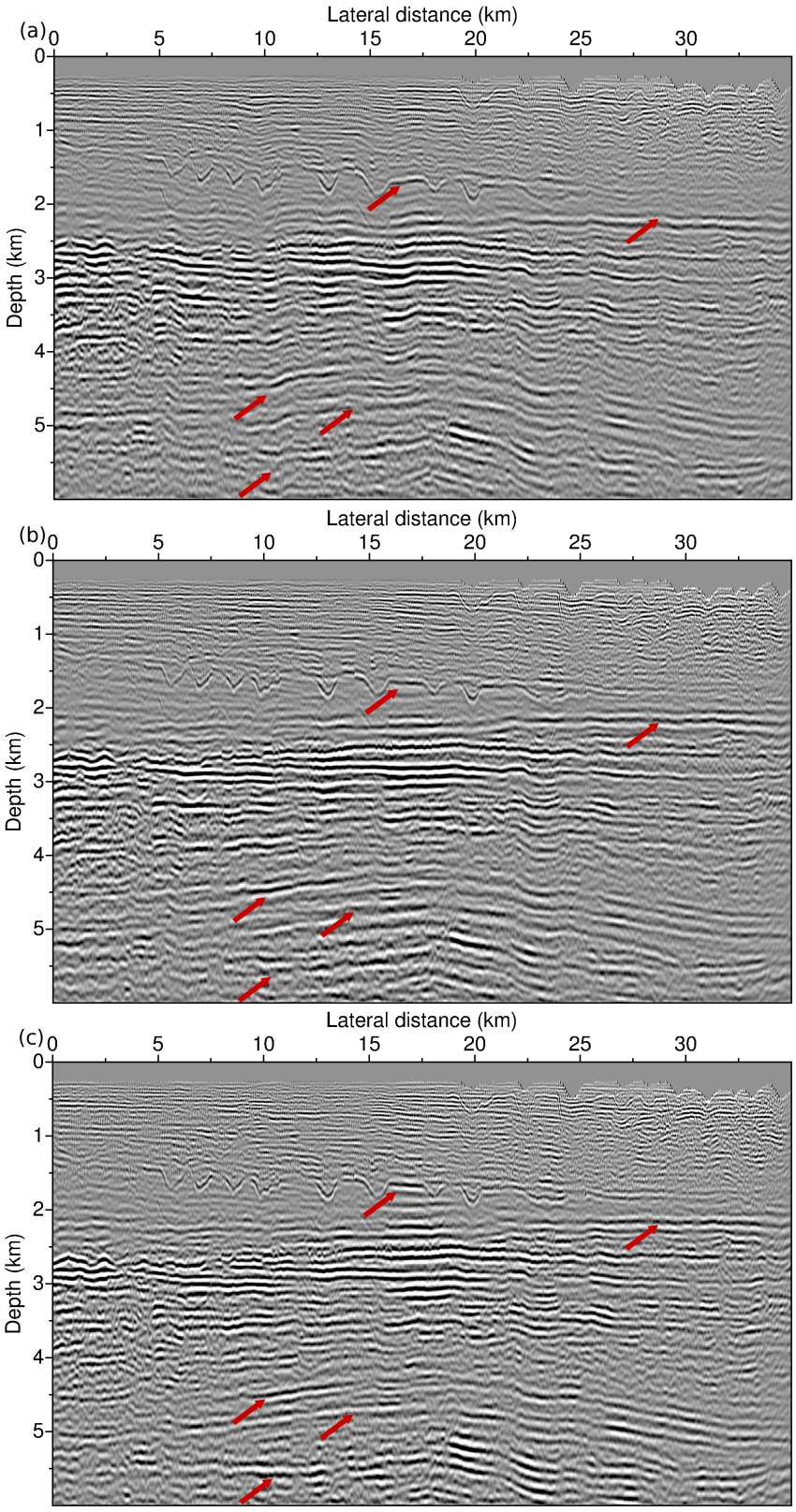


**Figure 22.** Migrated images in the initial model (a), in the model obtained at 10 Hz maximum frequency (b), in the model obtained at 16 Hz maximum frequency (c). Red arrows indicate identifiable improvements of the reconstruction of the reflectors and re-focusing of the energy. Improvements in the shallow part (above 3 km) are already obtained with the 10 Hz P-wave velocity estimation (b). Improvements in the deeper part (below 3 km) are yielded by the P-wave estimation at 16 Hz.

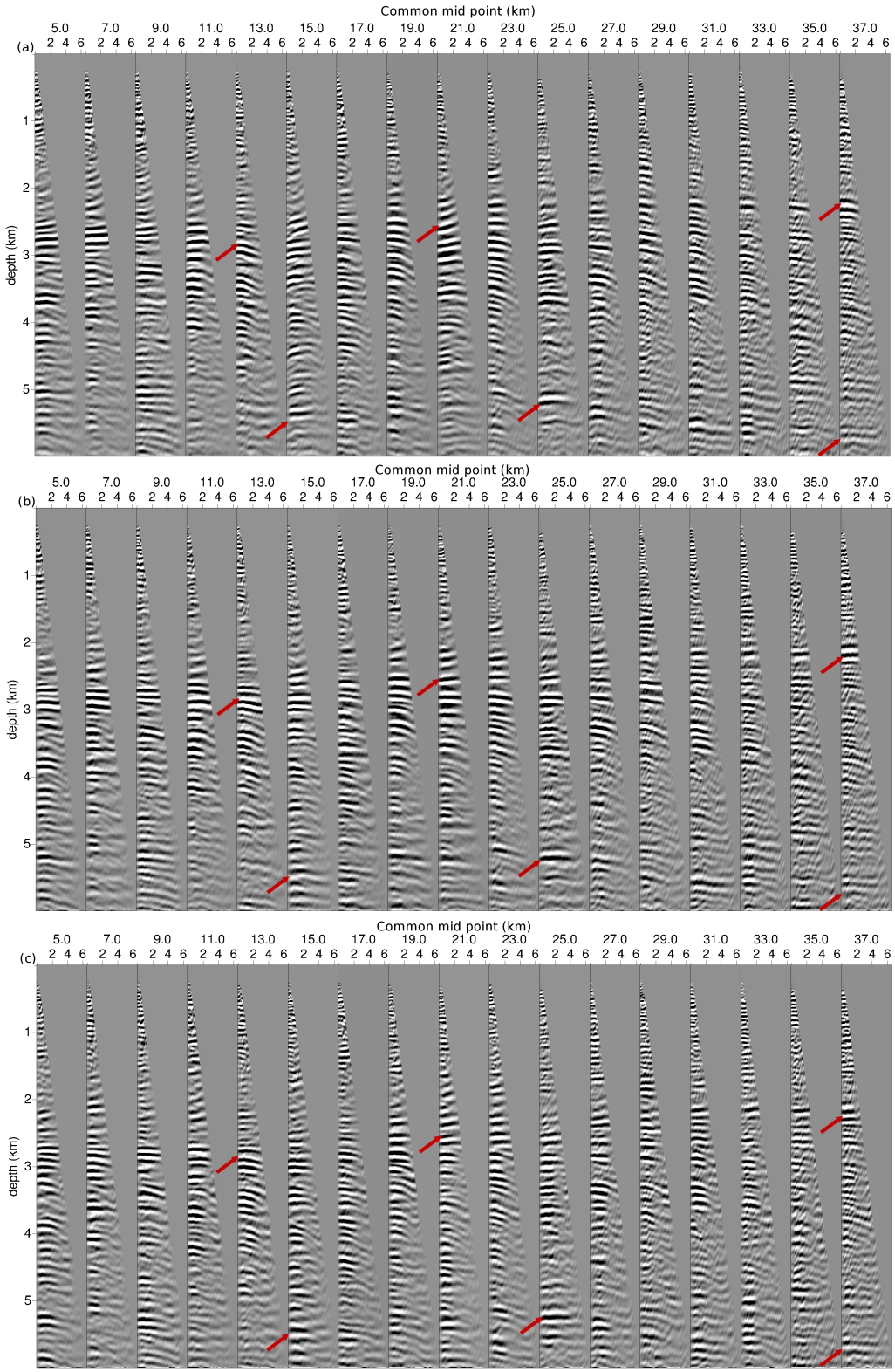


**Figure 23.** CIG in the initial model (a), in the model obtained at 10 Hz maximum frequency (b), in the model obtained at 16 Hz maximum frequency (c). Red arrows indicate identifiable improvement of the CIG continuity in the offset direction. As for the migrated images, improvements in the shallow part (above 3 km) are already obtained with the 10 Hz P-wave velocity estimation (b). Improvements in the deeper part (below 3 km) are yielded by the P-wave estimation at 16 Hz.

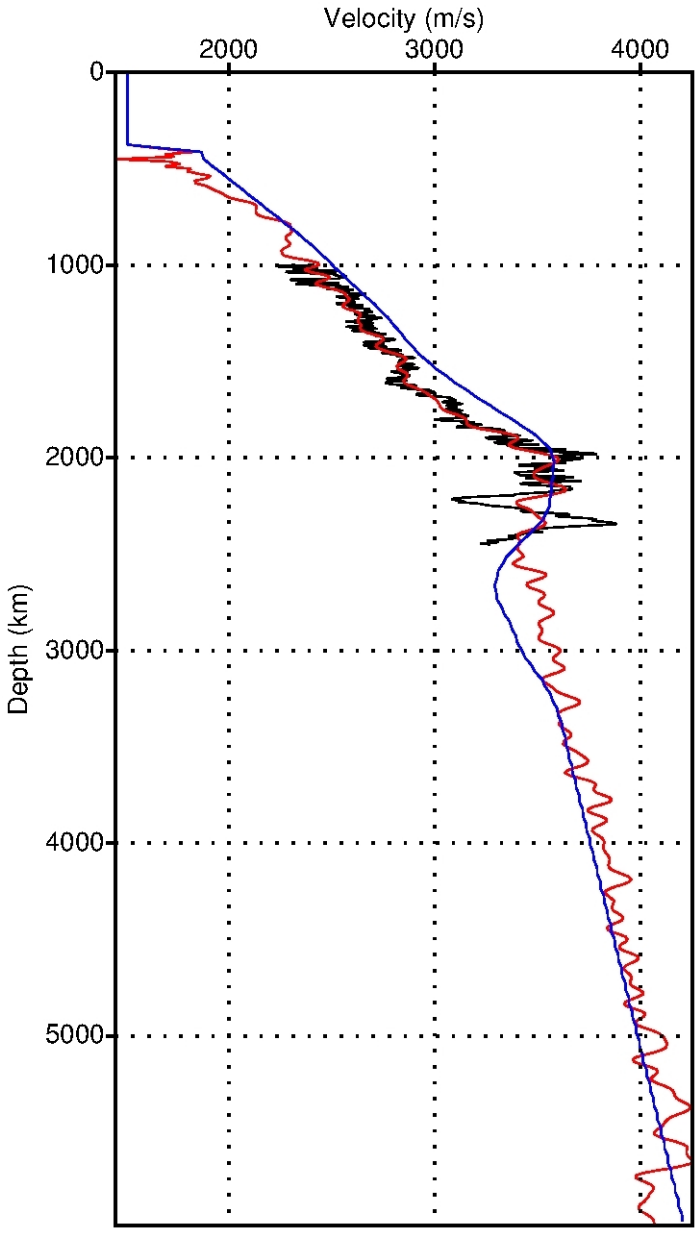


**Figure 24.** Vertical $P$-wave velocity log taken at $x = 39,375$ km. Initial model (blue), exact model (black), estimation at 25 Hz (red).

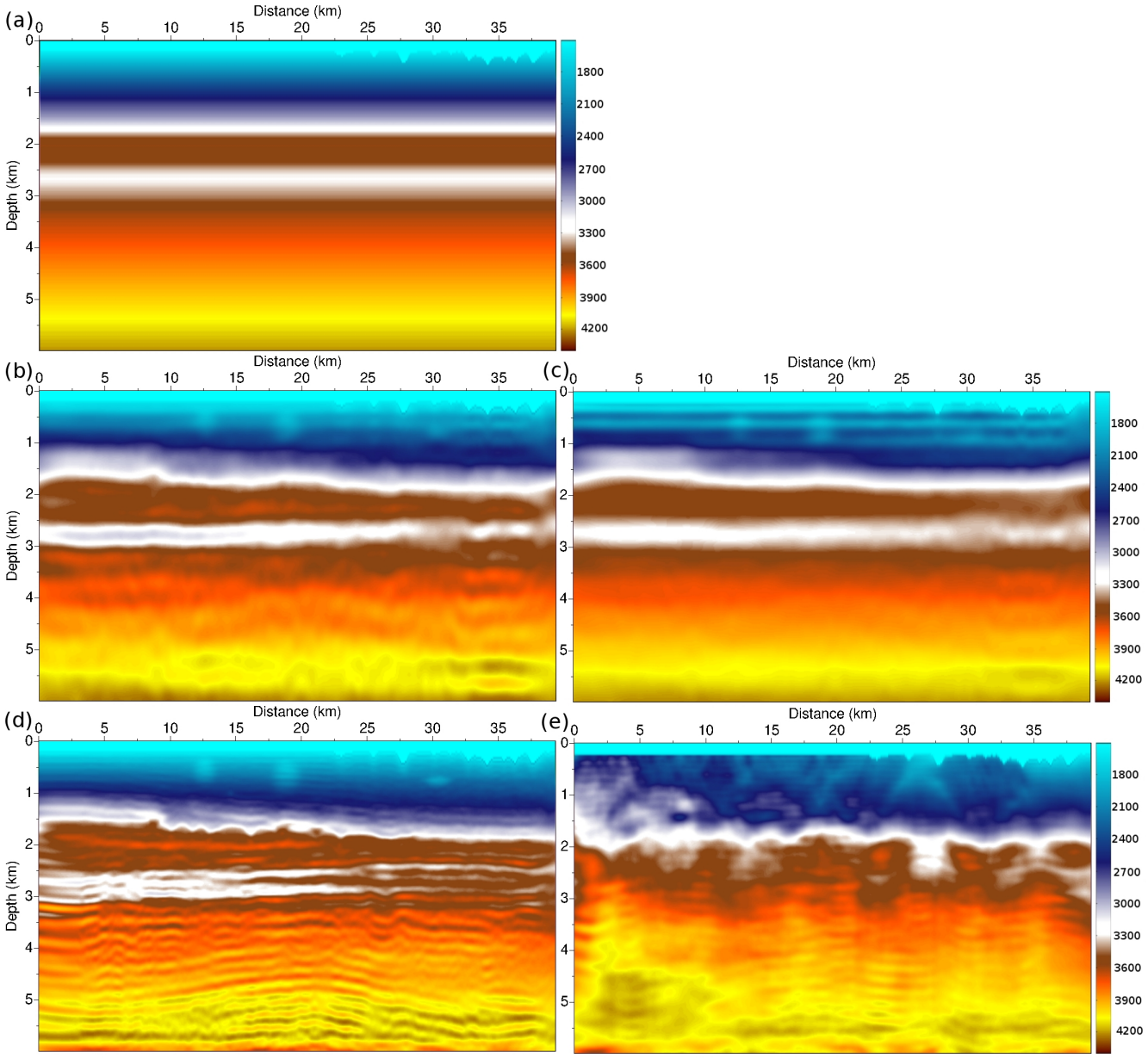


**Figure 25.** Chevron 2014 starting P-wave velocity model (a). Estimated P-wave velocity model at $4$ Hz with the optimal transport distance (b), with the $L^2$ distance (c). Estimated P-wave velocity model at $10$ Hz with the optimal transport distance (d), with the $L^2$ distance (e).

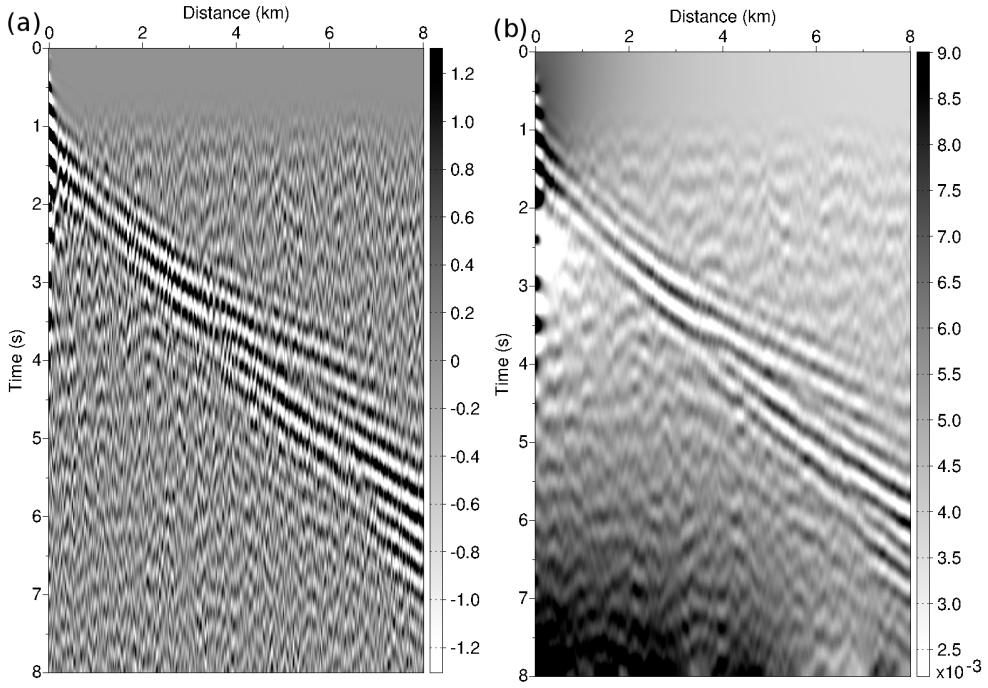


**Figure 26.** Residuals in the initial model for the first frequency band, using the $L^2$ norm misfit function (a), using the optimal transport distance (b).

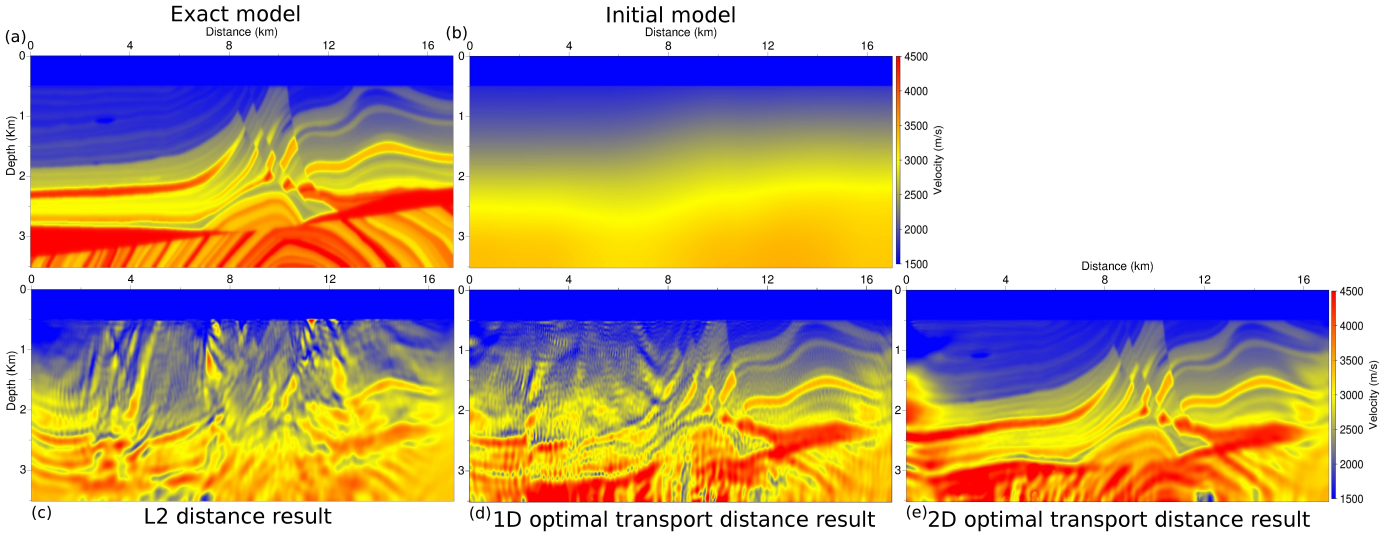


**Figure 27.** Exact Marmousi 2 P-wave velocity model (a). Initial model corresponding to the third initial model of Figure 8 (b). Reconstructed model using the $L^2$ distance (c), using 1D optimal transport distance (d), 2D optimal transport distance (e).