

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

3 L. Métivier¹, R. Brossier², Q. Mérigot³, E. Oudet⁴, J. Virieux⁵

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*

5 **SUMMARY**

6 Full waveform inversion using the conventional L^2 distance to measure the misfit between
7 seismograms is known to suffer from cycle skipping. An alternative strategy is proposed
8 in this study, based on a measure of the misfit computed with an optimal transport dis-
9 tance. This measure allows to account for the lateral coherency of events within the seis-
10 mograms, instead of considering each seismic trace independently, as is done generally
11 in full waveform inversion. The computation of this optimal transport distance relies on a
12 particular mathematical formulation allowing for the non-conservation of the total energy
13 between seismograms. The numerical solution of the optimal transport problem is per-
14 formed using proximal splitting techniques. Three synthetic case studies are investigated
15 using this strategy: the Marmousi 2 model, the BP 2004 salt model, and the Chevron
16 2014 benchmark data. The results emphasize interesting properties of the optimal trans-
17 port distance. The associated misfit function is less prone to cycle skipping. A workflow
18 is designed to reconstruct accurately the salt structures in the BP 2004 model, starting
19 from an initial model containing no information about these structures. A high resolution
20 P-wave velocity estimation is built from the Chevron 2014 benchmark data, following a
21 frequency continuation strategy. This estimation explains accurately the data. Using the
22 same workflow, full waveform inversion based on the L^2 distance converges toward a
23 local minimum. These results yield encouraging perspectives regarding the use of the op-
24 timal transport distance for full waveform inversion: the sensitivity to the accuracy of the
25 initial model is reduced, the reconstruction of complex salt structure is made possible, the
26 method is robust to noise, and the interpretation of seismic data dominated by reflections
27 is enhanced.

28 **Key words:** Optimal transport, controlled source seismology, computational seismology,
29 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

64 signal. These requirements lead conventional FWI to focus (at least in the first stage of the inversion)
65 on low frequency transmitted waves such as diving waves. These waves sample the subsurface with-
66 out being reflected, therefore the difference between predicted and observed diving waves should be
67 mainly due to shifts in time of the seismic events. However, if these time shifts are too large, reducing
68 the L^2 distance between the signals through a local optimization technique will generate a wrong ve-
69 locity model which matches the data with one to several phase shifts. This phenomenon is commonly
70 known as cycle skipping. This is the reason why the accuracy of the initial model is of primary impor-
71 tance in conventional FWI: it should be kinematically compatible with the data, i.e. the phase of the
72 main seismic events should be predicted within half a period.

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

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

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

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

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

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

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

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

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

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

154 Parallel to the development of these data-domain techniques, the development of image-domain
155 techniques started with the design of Differential Semblance Optimization (DSO) (Symes & Kern
156 1994) and later on wave equation migration velocity analysis (WEMVA) (Sava & Biondi 2004a,b;
157 Symes 2008). These methods rely on the separability of scales assumption: the velocity model is de-
158 composed as the sum of a smooth background model and a high wavenumber reflectivity model. The
159 reflectivity is related to the smooth background model through an imaging condition: it is the sum for
160 each source of the cross-correlation between the incident wavefield and the back-propagated residuals
161 computed in the smooth background velocity model. This imaging condition can be extended using
162 either an offset selection (Symes & Kern 1994) or an illumination angle selection (Biondi & Symes
163 2004) in the residuals (the angles are easily accessible when the reflectivity is computed through
164 asymptotic techniques), or a time lag in the cross-correlation (Faye & Jeannot 1986; Sava & Fomel
165 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

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

205 In this study, we are interested in an extension of this method to the comparison of entire seis-
206 mograms, more precisely common shot-gathers, which are collections of traces corresponding to one
207 seismic experiment. Compared to individual seismic traces, the shot-gathers (which can be seen as
208 2D images), contain important additional information as lateral coherency corresponds to identifiable
209 seismic events, such as reflections, refraction, or diving waves. Hence, the aim of this study is twofold.
210 The first objective is to present how shot-gathers can be compared using an optimal transport based
211 distance. The second objective consists in demonstrating the interest of using such a distance in the
212 context of FWI through different case studies.

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

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

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

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

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

268 studies mentioned previously are then presented to outline characteristic properties and performances
269 of FWI based on the optimal transport distance. A discussion and a conclusion are given in the two
270 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

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

and M a function from X to X

$$M : X \longrightarrow X. \quad (2)$$

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 $\|\cdot\|$, an exponent $p \geq 1$, and the constrained minimization problem

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

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 ℓ^1 distance $\|\cdot\|_1$ on \mathbb{R}^d such that

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

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

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

where Lip_1 is the space of 1-Lipschitz functions, such that

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

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 φ .

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

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

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

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

275 Therefore, if (7) is not satisfied, the solution of (5) is the constant function equal to ∞ or $-\infty$ depend-
276 ing 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 $\varphi(x) \in \text{Lip}_1$ should be also bounded, such that

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

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 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

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

277 Note that some theoretical links exist between the Wasserstein W^1 and the distance \widetilde{W}^1 : see for
278 instance the work of Hanin (1992). A mathematical analysis of this link is, however, beyond the scope
279 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

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

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

$$f_{\widetilde{W}^1}(m) = \sum_{s=1}^S \widetilde{W}^1(d_{cal}^s[m], d_{obs}^s), \quad (12)$$

where

$$\widetilde{W}^1(d_{cal}^s[m], d_{obs}^s) = \max_{\varphi \in \text{BLip}_1} \int_t \int_{x_r} \varphi(x_r, t) (d_{cal}^s[m](x_r, t) - d_{obs}^s(x_r, t)) dx_r dt. \quad (13)$$

For comparison, the conventional L^2 misfit function is

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

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

$$\left\{ \begin{array}{l} \widetilde{W}^1(d_{cal}[m], d_{obs}) = \max_{\varphi} \sum_{i=1}^{N_r} \sum_{j=1}^{N_t} \varphi_{ij} ((d_{cal}[m])_{ij} - (d_{obs})_{ij}) \Delta t \Delta x_r \end{array} \right. \quad (15a)$$

$$\left\{ \begin{array}{l} \forall(i, j), \quad |\varphi_{ij}| < c, \end{array} \right. \quad (15b)$$

$$\left\{ \begin{array}{l} \forall(i, j), (k, l) \quad |\varphi_{ij} - \varphi_{kl}| < |(x_r)_i - (x_r)_k| + |t_j - t_l|. \end{array} \right. \quad (15c)$$

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

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

where Δx_r and Δt 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 φ to enforce the 1-Lipschitz property. This yields the linear programming problem

$$\left\{ \begin{array}{l} \widetilde{W}^1(d_{cal}[m], d_{obs}) = \max_{\varphi} \sum_{i=1}^{N_r} \sum_{j=1}^{N_t} \varphi_{ij} ((d_{cal}[m])_{ij} - (d_{obs})_{ij}) \Delta t \Delta x_r \end{array} \right. \quad (17a)$$

$$\left\{ \begin{array}{l} \forall(i, j), \quad |\varphi_{ij}| < c, \end{array} \right. \quad (17b)$$

$$\left\{ \begin{array}{l} \forall(i, j), \quad |\varphi_{i+1j} - \varphi_{ij}| < |(x_r)_{i+1} - (x_r)_i| = \Delta x_r \end{array} \right. \quad (17c)$$

$$\left\{ \begin{array}{l} \forall(i, j), \quad |\varphi_{ij+1} - \varphi_{ij}| < |t_{j+1} - t_j| = \Delta t. \end{array} \right. \quad (17d)$$

283 The two linear programming problems (15) and (17) are equivalent. This results from a particular
 284 property of the ℓ_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

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

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

$$K = \{x \in \mathbb{R}^{3N}, |x_i| \leq 1, i = 1, \dots, 3N\}. \quad (19)$$

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

$$i_K(x) = \begin{cases} 0 & \text{if } x \in K \\ +\infty & \text{if } x \notin K. \end{cases} \quad (20)$$

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

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

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

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

the matrix A is such that

$$(A\varphi)_k = \begin{cases} \varphi_k & \text{for } k = 1, \dots, N \\ \frac{c}{\Delta x_r} - \varphi_k & \text{for } k = N + 1, \dots, 2 \times N \\ \frac{\varphi_{k+N_r} - \varphi_k}{\Delta x_t} & \text{for } k = 2 \times N + 1, \dots, 3 \times N. \end{cases} \quad (23)$$

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

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

296 of the functions $h_{d_{cal}[m],d_{obs}}$ and i_K , denoted by $\text{prox}_{h_{d_{cal}[m],d_{obs}}}$ and prox_{i_K} . These operators have a
 297 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 prox_{i_K} are invoked, as well as the solution of a linear system involving the square matrix of size N

$$Q = I_N + A^T A, \quad (24)$$

298 where I_N is the identity matrix of size N . The solution of these linear systems is the more in-
 299 tensive computational task in the SDMM algorithm, as the application of the proximity operators
 300 $\text{prox}_{h_{d_{cal}[m],d_{obs}}}$ and prox_{i_K} has a linear complexity in number of operations and a negligible cost in
 301 terms of memory requirement.

302 The matrix Q is a sparse square matrix of size N , symmetric positive definite by construction,
 303 related only to the definition of the equations (17.b)-(17.d). As a consequence, the matrix Q remains
 304 constant throughout the whole FWI process. In a first attempt to design an efficient algorithm for the
 305 solution of (21), it can be interesting, in a pre-processing step, to factorize this matrix as a product
 306 LL^T , where L is a lower triangular matrix, using a Cholesky decomposition. Under the assumption
 307 $N_t \simeq N_r$, this allows to benefit from a complexity in $O(N^{3/2})$ for the solution of these linear systems
 308 through forward and backward substitutions. However, the memory requirement associated with the
 309 storage of the factor L is also in $O(N^{3/2})$, which is non negligible for realistic size problems for which
 310 the size N can reach $O(10^6)$.

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

318 **2.3 Minimization of the optimal transport distance based misfit function and gradient** 319 **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

$$m_{k+1} = m_k + \alpha_k \Delta m_k, \quad (25)$$

where α_k is a positive scalar parameter computed through a linesearch strategy (Nocedal & Wright 2006; Bonnans et al. 2006), and Δm_k is a model update satisfying

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

320 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
321 inverse of its Hessian. In this study, this estimation is computed through the l -BFGS approximation
322 (Nocedal 1980). This approximation is based on the values of the gradient at iteration k and the l
323 previous iterations $k-1, \dots, k-l+1$.

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

The following Lagrangian function is introduced

$$\mathcal{L}(m, u, d_{cal}, \lambda, \mu) = \widetilde{W}_1^{-1}(d_{cal}, d_{obs}) + (F(m, u), \lambda)_{\mathcal{W}} + (Ru - d_{cal}, \mu)_{\mathcal{D}}, \quad (27)$$

where the standard Euclidean scalar product in the wavefield space and the data space is denoted by $(\cdot, \cdot)_{\mathcal{W}}$ and $(\cdot, \cdot)_{\mathcal{D}}$ 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 λ and μ . 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

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

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

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

329 where $\bar{u}(m)$ and $\bar{\lambda}(m)$ are respectively the incident and the adjoint wavefields satisfying the state
330 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

$$\begin{cases} \frac{\partial F(m, \bar{u}(m))^T}{\partial u} \lambda = -R^T \mu \\ \mu = \frac{\partial \widetilde{W}^1(d_{cal}, d_{obs})}{\partial d_{cal}}. \end{cases} \quad (30)$$

The first of these two equations involves the adjoint of the wave operator. The wavefield λ thus corresponds to the back-propagation of the source term $-R^T \mu$. The second equation relates μ 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 λ . Remarking that

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

the secondary adjoint wavefield μ is simply given by

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

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

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

347 2.4 Numerical illustration on a simple synthetic study

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

350 The two boreholes are 2500 m apart. A single source is used, located at 2500 m depth in the leftmost
 351 borehole. An array of 196 receivers equally spaced each 25 m is located in the second borehole, from
 352 50 m depth to 4900 m depth. A Ricker source centered on 5 Hz is used to generate a single shot-gather.
 353 The modeling is performed in the acoustic approximation and the pressure wavefield is recorded. The
 354 density model is kept constant, equal to 1000 kg.m^{-3} . The velocity of the true medium is homogeneous
 355 and set to $v_P^* = 2000 \text{ m.s}^{-1}$. One synthetic shot-gather is computed in a homogeneous medium with
 356 a velocity set to $v_P = 1500 \text{ m.s}^{-1}$ and with the correct density.

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

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

$$\varphi_i = \begin{cases} c & \text{if } d_{cal,i}[m] - d_{obs,i} > 0 \\ -c & \text{if } d_{cal,i}[m] - d_{obs,i} < 0 \end{cases} \quad (33)$$

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

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

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

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

411 **3 CASE STUDIES**412 **3.1 Settings**

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

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

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

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

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

452 3.2 Marmousi 2 case study

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

458 The observed data is generated using a filtered Ricker wavelet, centered on a 5 Hz frequency. The
 459 low frequency content of this wavelet, below 2.5 Hz, is removed using a minimum phase Butterworth
 460 filter. For real seismic marine data, the noise level below this frequency is too strong for the information
 461 to be relevant to constrain the P-wave velocity model. The spectrum and the shape of the resulting
 462 wavelet are presented in Figure 7. The spatial discretization step is set to 25 m to guarantee at least
 463 4 discretization points by wavelength. The time discretization step is set to 0.0023 s according to the
 464 Courant Friedrich Levy (CFL) condition. The recording is performed over 2000 time steps, which
 465 corresponds to a total recording time of 4.6 s. In this experiment, a Gaussian filter smoothing with a
 466 short correlation length (between 60 m and 100 m depending on the local dominant wavelength) is
 467 applied to the gradient, to remove fast oscillations which are due to a sparse acquisition design (only
 468 one source every 125 m).

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

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

477 initial model, the best results obtained using the two misfit functions are presented. The exact data as
 478 well as the corresponding residuals in the initial and the calculated models are presented in Figure 9.

479 Starting from the initial model 1, both the L^2 distance and the optimal transport distance yield
 480 estimations very close from the exact model (Fig 8c,d). However, a difference can be noted regarding
 481 the reconstruction of the low velocity zone near $x = 11$ km and $z = 2.5$ km. A high velocity artifact
 482 is present in this zone in the estimation obtained with the L^2 distance. This is not the case in the
 483 estimation obtained with the optimal transport distance.

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

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

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

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

518 3.3 BP 2004 case study

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

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

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

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

558 Building on this result, a layer stripping workflow is suggested. Five increasing time-windows
 559 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
 560 time-window, two to three successive inversions are performed. A Gaussian smoothing with a small
 561 correlation length is applied to the model computed after each inversion, which serves as an initial
 562 model for the next inversion. This Gaussian smoothing serves only to remove high frequency artifacts
 563 appearing in the late iterations of the inversion. Alternative strategies such as Tikhonov regularization
 564 or gradient smoothing could have been used instead. A total of 15 inversions is performed following
 565 this process, with in average 221 iterations of the l -BFGS algorithm for each inversion. The stopping
 566 criterion is only based on a linesearch failure to give the possibility to the optimizer to minimize as
 567 much as possible the misfit function based on the optimal transport distance. The detailed workflow is
 568 summarized in Table 1.

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

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

590 3.4 Chevron 2014 case study

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

596 The Chevron 2014 benchmark dataset is built from a 2D isotropic elastic modeling engine. A
 597 frequency-dependent noise has been added to the data to mimic a realistic dataset. Especially, the Sig-
 598 nal Over Noise Ratio (SNR) for low frequencies (below 3 Hz) is much less than for higher frequencies.
 599 Free surface multiples are incorporated in the data. A streamer acquisition is used, with a maximum
 600 of 8 km offset, with 321 receivers by sources equally spaced each 25 m. The depth of the sources and
 601 receivers is $z = 15$ m. Among the 1600 available shots gathers, 256 have been used in this study, with
 602 a distance of 150 m between each sources. A frequency continuation strategy similar to the one pro-
 603 posed by Bunks et al. (1995) is implemented: Butterworth low-pass and high-pass filters are applied
 604 to the selected shot-gathers to generate an ensemble of 15 datasets with an increasing bandwidth from
 605 2 – 4 Hz to 2 – 25 Hz.

606 The shot-gathers corresponding to the source located at $x = 150$ m are presented for the 1st, 5th,
 607 10th and 15th frequency bands in Figure 18. As mentioned previously, the noise imprint is clearly
 608 stronger for the first frequency bands.

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

611 been incorporated. The P-wave velocity in the water layer is set to 1500 m.s^{-1} . The initial model
 612 incorporates an important feature: a low velocity layer is located between the depth $z = 2.3 \text{ km}$ and
 613 $z = 3 \text{ km}$. This velocity inversion and the relatively short available offsets (only 8 km) prevent diving
 614 waves from sampling the deepest part of the model. This makes the benchmark data challenging as
 615 only reflection information is available for constraining the deep part of the model.

616 The workflow which is applied to the Chevron benchmark dataset is the following. Prior to inver-
 617 sion, an estimation of the source wavelet is performed in the initial model, for each frequency band,
 618 following the frequency-domain strategy introduced by Pratt (1999). For the first ten frequency bands,
 619 20 iterations of a preconditioned l -BFGS algorithm are performed. For the frequency bands 11 and
 620 12, 50 iterations are performed. For the last three frequency bands, 40 iterations are performed with a
 621 restart of the l -BFGS algorithm after the 20 first iterations. This restart is only due the configuration
 622 of the queue of the Blue Gene/Q machine of the IDRIS center, which does not accept jobs running
 623 longer than 20 hours. The restart could be avoided by storing the l -BFGS approximation on disk, how-
 624 ever this option is not yet implemented in the SEISCOPE optimization toolbox. The spatial and time
 625 discretization steps are set to 37.5 m and 0.004 s respectively for the 8 first frequency bands. They are
 626 decreased to 25 m and 0.003 s respectively for the frequency bands 9 to 12. For the last three frequency
 627 bands, the discretization step is set to 12.5m and the time step to 0.001 s. The misfit function is based
 628 on the optimal transport distance. According to the frequency continuation strategy, the P-wave veloc-
 629 ity model estimated for one frequency band serves as the initial model for the next frequency band.
 630 No regularization is introduced throughout the inversion. However, the model estimated at the end of
 631 each inversion is smoothed using a Gaussian filter with a correlation length adapted to the resolution
 632 expected after the inversion of each frequency-band. The workflow is summarized in Table 2.

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

643 The results obtained after inverting the data up to 4 Hz (frequency band 1), 10 Hz (frequency band
 644 8), 16 Hz (frequency band 12) and 25 Hz (frequency band 15) are presented in Figure 19b,c,d,e. Three

645 shallow low velocity anomalies are recovered at approximately 500 m depth and at the lateral positions
 646 $x = 12$ km, $x = 18$ km and $x = 30$ km. An additional small scale low velocity anomaly appears at
 647 $x = 14.75$ km and $z = 1$ km in the highest resolution estimation. The original layered structure of
 648 the initial model is tilted in the final estimation. The upper (faster) layers bend downward (from left to
 649 right), while the low velocity layer at depth $z = 2.5$ km bends upward. Three high velocity anomalies
 650 are also recovered on top of the layer above the low velocity layer, at depth 1.8 km and lateral positions
 651 $x = 8$ km, $x = 19$ km, $x = 22$ km. The deeper part of the model, below 3 km depth, seems less well
 652 reconstructed, as it could be expected from the lack of illumination of this zone. However, a curved
 653 interface seems to be properly recovered at a depth between 4.5 and 5 km. A flat reflector is also
 654 clearly visible at the bottom of the model, at depth $z = 5.8$ km.

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

668 As a second quality control, migrations of the data in the initial model and the estimated models at
 669 10 Hz and 16 Hz are performed. The migration results correspond to impedance gradients computed
 670 on 30 Hz low-pass filtered data, with a filter applied on the diving waves to focus on reflection data
 671 only. The spatial and time discretization steps are set to 12.5 m and 0.001 s respectively. The number
 672 of sources is doubled to 512 (one source each 75 m) to avoid spatial aliasing. As a post-processing, a
 673 polynomial gain is used to balance the energy in depth. The resulting images are presented in Figure
 674 22. The migrated image obtained in the estimated model at 10 Hz is significantly improved in the
 675 shallow part of the model (above 3 km depth) (Fig. 22b). A significant uplift of this part of the model
 676 can be observed. The continuity and the flatness of the reflectors is globally improved. However, the
 677 reflectors in the deepest part of the model ($z > 2.5$ km) remain unfocused. The migrated image in the
 678 estimated model at 16 Hz yields a better delineation of these deep reflectors, as indicated by the three

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

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

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

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

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

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

715 **4 DISCUSSION**

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

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

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

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

749 crossline receivers, common shot-gathers should be represented as data cubes, with a coherency of
750 seismic events both in inline and crossline directions. The previous experiment, based on 1D optimal
751 transport, suggests that there is an interest in fully exploiting the lateral coherency of the seismic
752 signal. However, further numerical improvements are required to design a method with a manageable
753 computational time in such a configuration. This could be achieved through a better account of the
754 structure of the matrix Q , which is related to a second-order discretization of the Laplacian operator
755 with Neumann boundary conditions. The linear system to be solved at each iteration of the SDMM
756 algorithm could thus be identified as a Poisson equation, for which fast solver exist, either based on
757 Fast Fourier Transform (Swarztrauber 1974), or multigrid methods (Brandt 1977; Adams 1989). If
758 this strategy reveals unfeasible, dimensionality reduction (such as the one presented here from 2D to
759 1D optimal transport) could still be worthy to investigate, using appropriate regularization techniques.
760 Another option may also consist in changing the formulation of the optimal transport problem to a
761 primal formulation with entropic regularization, as this strategy is indicated to benefit from a reduced
762 computational complexity (Benamou et al. 2015).

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

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

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

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

789 5 CONCLUSION

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

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

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

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

816 The proposed method thus seems promising and should be investigated in more realistic configu-
 817 rations, implying 3D waveform inversion. Measuring the misfit between data cubes using the optimal
 818 transport distance is a challenging issue, which could yield interesting perspectives for 3D FWI. The
 819 introduction of viscous, elastic and anisotropic effects should also be investigated. As the proposed
 820 strategy is data-domain oriented, such extension should be straightforward. Finally, specific investi-
 821 gations 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.

824 **APPENDIX A: EQUIVALENCE BETWEEN LINEAR PROGRAMMING PROBLEMS**

825 In this appendix, the proof of equivalence between the linear programming problems (15) and (17)
 826 is given. The first of these two problems is the discrete analogous of the problem (10), which uses
 827 global constraints to impose the Lipschitz property. The second only uses local constraints to impose
 828 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 $z_i = (x_i, t_i)$, $i = 1, \dots, N$, 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 ℓ_1 distance (also known as Manhattan distance) ensures that

$$\|w - v\|_1 = \sum_i \|z_{i+1} - z_i\|_1. \quad (\text{A.1})$$

Now, consider a function φ satisfying only the local constraints. The triangle inequality yields

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

As the points z_i are adjacent, the local inequality satisfied by φ can be used to obtain

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

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

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

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

832 **APPENDIX B: PROXIMITY OPERATORS**

For a given convex function $f(x)$, its proximity operator prox_f is defined by

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

where the standard Euclidean distance on \mathbb{R}^d is denoted by $\|\cdot\|_2$. 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).

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

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

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

833 APPENDIX C: EFFICIENT SOLUTION OF THE BLOCK TRIDIAGONAL LINEAR 834 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

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

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 $x \in \mathbb{R}^N$

$$x = [x_1, \dots, x_{N_t}] \in \mathbb{R}^N, \quad (\text{C.2})$$

and

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

The matrix Q is block tridiagonal such that

$$Q = \begin{pmatrix} F + B & B & & & \\ & B & F & B & \\ & & \ddots & \ddots & \ddots \\ & & & B & F & B \\ & & & & B & F + B \end{pmatrix}. \quad (\text{C.4})$$

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

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

and F is the tridiagonal symmetric positive definite matrix

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

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

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

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

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

Using this factorization in (C.7) yields

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

where

$$y_i = P^T x_i, \quad c_i = P^T b'_i, \quad i = 1, \dots, N_t. \quad (\text{C.10})$$

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

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

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

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

are introduced. These vectors satisfy the equation

$$K_j y_{*j} = c_{*j}, \quad (\text{C.13})$$

where K_j is the tridiagonal matrix

$$K_j = \begin{pmatrix} d_j + 1 & 1 & & & \\ & 1 & d_j & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & d_j & 1 \\ & & & & 1 & d_j + 1 \end{pmatrix}. \quad (\text{C.14})$$

835 These transformations yield the algorithm 2 to solve the initial system (C.1). As a pre-processing step,
 836 the matrix E is factorized as in (C.8), and the eigenvectors are stored in the matrix P . The computation
 837 cost and the memory requirement of this operation is in $O(N_r^2)$ as E is tridiagonal. The solution of the
 838 equation (C.1) is then obtained through the following operations. First, the vectors b_i are multiplied by
 839 the diagonal matrix B^{-1} which requires $O(N)$ operations. Second, the vectors c_i are formed following
 840 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
 841 computed through the solution of N_r tridiagonal systems of size N_t . Tridiagonal systems are efficiently
 842 solved through the Thomas algorithm which has a linear complexity (Golub 1996) . Therefore, the
 843 computation cost of computing y_{*j} is merely in $O(N_r \times N_t) = O(N)$. The final step consists in
 844 computing the vector x from the vectors y_{*j} through the equation (C.10). This requires to multiply
 845 each vector y_i by P , which costs $O(N_r^2 \times N_t)$ operations. The overall complexity of the algorithm is
 846 thus $O(N_r^2 \times N_t)$, and the memory requirement in $O(N)$. In contrast, a Cholesky factorization has
 847 the same computational complexity, but requires to store $O(N^{3/2})$ elements. In addition, the forward
 848 backward substitution is an intrinsically sequential algorithm, while the most expensive part of the
 849 algorithm 2 are the matrix-vector multiplications involving the eigenvectors of the matrix E , which
 850 can be efficiently parallelized. As a final remark, in the case $N_t < N_r$, the matrices and vectors can be
 851 re-organized in N_r blocks of size N_t to yield a complexity in $O(N_t^2 \times N_r)$ instead of $O(N_r^2 \times N_t)$.

852 ACKNOWLEDGMENTS

853 The authors would like to thank gratefully Stéphane Operto for his insightful advice and his avail-
 854 ability to comment on these results. This study was partially funded by the SEISCOPE consor-
 855 tium (<http://seiscope2.osug.fr>), sponsored by BP, CGG, CHEVRON, EXXON-MOBIL, JGI, PETRO-
 856 BRAS, SAUDI ARAMCO, SCHLUMBERGER, SHELL, SINOPEC, STATOIL, TOTAL and WOOD-
 857 SIDE. This study was granted access to the HPC resources of the Froggy platform of the CIMENT
 858 infrastructure (<https://ciment.ujf-grenoble.fr>), which is supported by the Rhône-Alpes region (GRANT
 859 CPER07_13 CIRA), the OSUG@2020 labex (reference ANR10 LABX56) and the Equip@Meso
 860 project (reference ANR-10-EQPX-29-01) of the programme Investissements d’Avenir supervised by

⁸⁶¹ the Agence Nationale pour la Recherche, and the HPC resources of CINES/IDRIS under the allocation
⁸⁶² 046091 made by GENCI.”

863 **REFERENCES**

- 864 Adams, J. C., 1989. MUDPACK: Multigrid portable FORTRAN software for the efficient solution of linear
865 elliptic partial differential equations, *Applied Mathematics and Computation*, **34**(2), 113–146.
- 866 Baek, H., Calandra, H., & Demanet, L., 2014. Velocity estimation via registration-guided least-squares inver-
867 sion, *Geophysics*, **79**(2), R79–R89.
- 868 Benamou, J. D., Froese, B. D., & Oberman, A. M., 2014. Numerical solution of the Optimal Transportation
869 problem using the Monge-Ampère equation, *Journal of Computational Physics*, **260**, 107–126.
- 870 Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., & Peyr, G., 2015. Iterative Bregman Projections for
871 Regularized Transportation Problems, *SIAM Journal on Scientific Computing*, **37**(2), A1111–A1138.
- 872 Billette, F. J. & Brandsberg-Dahl, S., 2004. The 2004 BP velocity benchmark, in *Extended Abstracts, 67th*
873 *Annual EAGE Conference & Exhibition, Madrid, Spain*, p. B035.
- 874 Biondi, B. & Almomin, A., 2013. Tomographic full waveform inversion (TFWI) by combining FWI and
875 wave-equation migration velocity analysis, *The Leading Edge*, **September, special section: full waveform**
876 **inversion**, 1074–1080.
- 877 Biondi, B. & Symes, W., 2004. Angle-domain common-image gathers for migration velocity analysis by
878 wavefield-continuation imaging, *Geophysics*, **69**(5), 1283–1298.
- 879 Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., & Sagastizábal, C. A., 2006. *Numerical Optimization, Theoret-*
880 *ical and Practical Aspects*, Springer series, Universitext.
- 881 Borisov, D. & Singh, S. C., 2015. Three-dimensional elastic full waveform inversion in a marine environ-
882 ment using multicomponent ocean-bottom cables: a synthetic study, *Geophysical Journal International*, **201**,
883 1215–1234.
- 884 Brandt, A., 1977. Multi-level adaptive solutions to boundary-value problems, *Mathematics of Computation*,
885 **31**, 333–390.
- 886 Brenier, Y., 1991. Polar factorization and monotone rearrangement of vector-valued functions, *Communica-*
887 *tions on Pure and Applied Mathematics*, **44**(4), 375–417.
- 888 Brossier, R., Operto, S., & Virieux, J., 2009. Seismic imaging of complex onshore structures by 2D elastic
889 frequency-domain full-waveform inversion, *Geophysics*, **74**(6), WCC105–WCC118.
- 890 Brossier, R., Operto, S., & Virieux, J., 2010. Which data residual norm for robust elastic frequency-domain
891 full waveform inversion?, *Geophysics*, **75**(3), R37–R46.
- 892 Brossier, R., Pajot, B., Combe, L., Operto, S., Métivier, L., & Virieux, J., 2014. Time and frequency-domain
893 FWI implementations based on time solver: analysis of computational complexities, in *Expanded Abstracts,*
894 *76th Annual EAGE Meeting (Amsterdam)*.
- 895 Brossier, R., Operto, S., & Virieux, J., 2015. Velocity model building from seismic reflection data by full
896 waveform inversion, *Geophysical Prospecting*, **63**, 354–367.
- 897 Bunks, C., Salek, F. M., Zaleski, S., & Chavent, G., 1995. Multiscale seismic waveform inversion, *Geophysics*,
898 **60**(5), 1457–1473.
- 899 Bözdag, E., Trampert, J., & Tromp, J., 2011. Misfit functions for full waveform inversion based on instantanta-

- 900 neous phase and envelope measurements, *Geophysical Journal International*, **185**(2), 845–870.
- 901 Buzbee, B. L., Golub, G. H., & Nielson, C. W., 1970. On direct methods for solving poisson’s equations,
902 *SIAM Journal on Numerical Analysis*, **7**(4), pp. 627–656.
- 903 Castellanos, C., Métivier, L., Operto, S., Brossier, R., & Virieux, J., 2015. Fast full waveform inversion with
904 source encoding and second-order optimization methods, *Geophysical Journal International*, **200**(2), 720–
905 744.
- 906 Chavent, G., 1974. Identification of parameter distributed systems, in *Identification of function parameters
907 in partial differential equations*, pp. 31–48, eds Goodson, R. & Polis, M., American Society of Mechanical
908 Engineers, New York.
- 909 Chavent, G., Clément, F., & Gómez, S., 1994. Automatic determination of velocities via migration-based
910 traveltine waveform inversion: A synthetic data example, *SEG Technical Program Expanded Abstracts 1994*,
911 pp. 1179–1182.
- 912 Claerbout, J., 1985. *Imaging the Earth’s interior*, Blackwell Scientific Publication.
- 913 Clapp, R., 2009. *Reverse time migration with random boundaries*, chap. 564, pp. 2809–2813.
- 914 Combettes, P. L. & Pesquet, J.-C., 2011. Proximal splitting methods in signal processing, in *Fixed-Point
915 Algorithms for Inverse Problems in Science and Engineering*, vol. 49 of **Springer Optimization and Its
916 Applications**, pp. 185–212, eds Bauschke, H. H., Burachik, R. S., Combettes, P. L., Elser, V., Luke, D. R., &
917 Wolkowicz, H., Springer New York.
- 918 Dahlen, F. A., Hung, S. H., & Nolet, G., 2000. Fréchet kernels for finite-difference traveltimes - I. theory,
919 *Geophysical Journal International*, **141**, 157–174.
- 920 Devaney, A., 1984. Geophysical diffraction tomography, *Geoscience and Remote Sensing, IEEE Transactions
921 on*, **GE-22**(1), 3–13.
- 922 Engquist, B. & Froese, B. D., 2014. Application of the wasserstein metric to seismic signals, *Communications
923 in Mathematical Science*, **12**(5), 979–988.
- 924 Evans, L. C., 1997. Partial differential equations and Monge–Kantorovich mass transfer, *Current developments
925 in mathematics*, pp. 65–126.
- 926 Faye, J. P. & Jeannot, J. P., 1986. Prestack migration velocities from focusing depth analysis, in *Expanded
927 Abstracts*, pp. 438–440, Soc. Expl. Geophys.
- 928 Fichtner, A., Kennett, B. L. N., Igel, H., & Bunge, H. P., 2008. Theoretical background for continental- and
929 global-scale full-waveform inversion in the time-frequency domain, *Geophysical Journal International*, **175**,
930 665–685.
- 931 Fichtner, A., Kennett, B. L. N., Igel, H., & Bunge, H. P., 2010. Full waveform tomography for radially
932 anisotropic structure: New insights into present and past states of the Australasian upper mantle, *Earth and
933 Planetary Science Letters*, **290**(3-4), 270–280.
- 934 Golub, G. H., 1996. *Matrix Computation, third edition*, Johns Hopkins Studies in Mathematical Sciences.
- 935 Hale, D., 2013. Dynamic warping of seismic images, *Geophysics*, **78**(2), S105–S115.
- 936 Hanin, L. G., 1992. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces, *Pro-*

- 937 *ceedings of American Mathematical Society*, **115**, 345–352.
- 938 Hicks, G. J., 2002. Arbitrary source and receiver positioning in finite-difference schemes using Kaiser win-
939 dowed sinc functions, *Geophysics*, **67**, 156–166.
- 940 Jannane, M., Beydoun, W., Crase, E., Cao, D., Koren, Z., Landa, E., Mendes, M., Pica, A., Noble, M., Roeth,
941 G., Singh, S., Snieder, R., Tarantola, A., & Trezeguet, D., 1989. Wavelengths of Earth structures that can be
942 resolved from seismic reflection data, *Geophysics*, **54**(7), 906–910.
- 943 Knott, M. & Smith, C., 1984. On the optimal mapping of distributions, *Journal of Optimization Theory and*
944 *Applications*, **43**(1), 39–49.
- 945 Lailly, P., 1983. The seismic problem as a sequence of before-stack migrations, in *Conference on Inverse*
946 *Scattering: Theory and Applications*, SIAM, Philadelphia.
- 947 Lambaré, G., 2002. The use of locally coherent events in depth processing : a state of the art., in *Extended*
948 *Abstracts, 72nd annual meeting , (6-10 October 2002, Salt Lake City)*, pp. 2261–2264, Soc. Expl. Geophys.
- 949 Lions, J. L., 1968. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod,
950 Paris.
- 951 Luo, J. & Wu, R.-S., 2015. Seismic envelope inversion: reduction of local minima and noise resistance,
952 *Geophysical Prospecting*, **63**(3), 597–614.
- 953 Luo, S. & Sava, P., 2011. A deconvolution-based objective function for wave-equation inversion, *SEG Techni-*
954 *cal Program Expanded Abstracts*, **30**(1), 2788–2792.
- 955 Luo, Y. & Schuster, G. T., 1991. Wave-equation travelttime inversion, *Geophysics*, **56**(5), 645–653.
- 956 Ma, Y. & Hale, D., 2013. Wave-equation reflection travelttime inversion with dynamic warping and full wave-
957 form inversion, *Geophysics*, **78**(6), R223–R233.
- 958 Maggi, A., Tape, C., Chen, M., Chao, D., & Tromp, J., 2009. An automated time-window selection algorithm
959 for seismic tomography, *Geophysical Journal International*, **178**, 257–281.
- 960 Martin, G. S., Wiley, R., & Marfurt, K. J., 2006. Marmousi2: An elastic upgrade for Marmousi, *The Leading*
961 *Edge*, **25**(2), 156–166.
- 962 Mérigot, Q., 2011. A multiscale approach to optimal transport, *Computer Graphics Forum*, **30**(5), 1583–1592.
- 963 Métivier, L. & Brossier, R., 2015. The seiscopes optimization toolbox: A large-scale nonlinear optimization
964 library based on reverse communication, *Geophysics*, pp. in–press.
- 965 Métivier, L., Brossier, R., Virieux, J., & Operto, S., 2013. Full Waveform Inversion and the truncated Newton
966 method, *SIAM Journal On Scientific Computing*, **35**(2), B401–B437.
- 967 Métivier, L., Bretaudeau, F., Brossier, R., Operto, S., & Virieux, J., 2014a. Full waveform inversion and the
968 truncated Newton method: quantitative imaging of complex subsurface structures, *Geophysical Prospecting*,
969 **62**, 1353–1375.
- 970 Métivier, L., Brossier, R., Labb, S., & Virieux, J., 2014b. Multi-parameter FWI - an illustration of the Hessian
971 operator role for mitigating trade-offs between parameter classes, in *Expanded Abstracts, 6th EAGE St-*
972 *Petersbourg International Conference & Exhibition*.
- 973 Métivier, L., Brossier, R., Operto, S., & Virieux, J., 2015. Acoustic multi-parameter FWI for the reconstruc-

- 974 tion of P-wave velocity, density and attenuation: preconditioned truncated Newton approach, in *Expanded*
 975 *Abstracts, 85th Annual SEG Meeting (New Orleans)*.
- 976 Montelli, R., Nolet, G., Dahlen, F. A., Masters, G., Engdahl, E. R., & Hung, S. H., 2004. Finite-frequency
 977 tomography reveals a variety of plumes in the mantle, *Science*, **303**, 338–343.
- 978 Nocedal, J., 1980. Updating Quasi-Newton Matrices With Limited Storage, *Mathematics of Computation*,
 979 **35**(151), 773–782.
- 980 Nocedal, J. & Wright, S. J., 2006. *Numerical Optimization*, Springer, 2nd edn.
- 981 Nolet, G., 2008. *A Breviary of Seismic Tomography*, Cambridge University Press, Cambridge, UK.
- 982 Operto, S., Ravaut, C., Imbrota, L., Virieux, J., Herrero, A., & Dell’Aversana, P., 2004. Quantitative imaging of
 983 complex structures from dense wide-aperture seismic data by multiscale traveltimes and waveform inversions:
 984 a case study, *Geophysical Prospecting*, **52**, 625–651.
- 985 Operto, S., Brossier, R., Gholami, Y., Métivier, L., Prioux, V., Ribodetti, A., & Virieux, J., 2013. A guided tour
 986 of multiparameter full waveform inversion for multicomponent data: from theory to practice, *The Leading*
 987 *Edge, Special section Full Waveform Inversion*(September), 1040–1054.
- 988 Operto, S., Miniussi, A., Brossier, R., Combe, L., Métivier, L., Monteiller, V., Ribodetti, A., & Virieux, J.,
 989 2015. Efficient 3-D frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data:
 990 application to Valhall in the visco-acoustic vertical transverse isotropic approximation, *Geophysical Journal*
 991 *International*, **202**(2), 1362–1391.
- 992 Peter, D., Komatitsch, D., Luo, Y., Martin, R., Le Goff, N., Casarotti, E., Le Loher, P., Magnoni, F., Liu, Q.,
 993 Blitz, C., Nissen-Meyer, T., Basini, P., & Tromp, J., 2011. Forward and adjoint simulations of seismic wave
 994 propagation on fully unstructured hexahedral meshes, *Geophysical Journal International*, **186**(2), 721–739.
- 995 Plessix, R. E., 2006. A review of the adjoint-state method for computing the gradient of a functional with
 996 geophysical applications, *Geophysical Journal International*, **167**(2), 495–503.
- 997 Plessix, R. E. & Perkins, C., 2010. Full waveform inversion of a deep water ocean bottom seismometer dataset,
 998 *First Break*, **28**, 71–78.
- 999 Plessix, R. E., Chavent, G., & Roeck, Y.-H. D., 1999. Waveform inversion of reflection seismic data for
 1000 kinematic parameters by local inversion, *SIAM Journal of Scientific Computing*, **20**, 1033–1052.
- 1001 Pratt, R. G., 1999. Seismic waveform inversion in the frequency domain, part I: theory and verification in a
 1002 physical scale model, *Geophysics*, **64**, 888–901.
- 1003 Pratt, R. G., Shin, C., & Hicks, G. J., 1998. Gauss-Newton and full Newton methods in frequency-space
 1004 seismic waveform inversion, *Geophysical Journal International*, **133**, 341–362.
- 1005 Rockafellar, R. T., 1976. Monotone operators and the proximal point algorithm, *SIAM Journal on Control and*
 1006 *Optimization*, **14**(5), 877–898.
- 1007 Sava, P. & Biondi, B., 2004a. Wave-equation migration velocity analysis. i. theory, *Geophysical Prospecting*,
 1008 **52**(6), 593–606.
- 1009 Sava, P. & Biondi, B., 2004b. Wave-equation migration velocity analysis. ii. subsalt imaging examples, *Geo-*
 1010 *physical Prospecting*, **52**(6), 607–623.

- 1011 Sava, P. & Fomel, S., 2006. Time-shift imaging condition in seismic migration, *Geophysics*, **71**(6), S209–S217.
- 1012 Shipp, R. M. & Singh, S. C., 2002. Two-dimensional full wavefield inversion of wide-aperture marine seismic
1013 streamer data, *Geophysical Journal International*, **151**, 325–344.
- 1014 Sirgue, L. & Pratt, R. G., 2004. Efficient waveform inversion and imaging : a strategy for selecting temporal
1015 frequencies, *Geophysics*, **69**(1), 231–248.
- 1016 Sirgue, L., Barkved, O. I., Dellinger, J., Etgen, J., Albertin, U., & Kommedal, J. H., 2010. Full waveform
1017 inversion: the next leap forward in imaging at Valhall, *First Break*, **28**, 65–70.
- 1018 Swarztrauber, P. N., 1974. A Direct Method for the Discrete Solution of Separable Elliptic Equations, *SIAM*
1019 *Journal on Numerical Analysis*, **11**(6), 1136–1150.
- 1020 Symes, W. & Kern, M., 1994. Inversion of reflection seismograms by differential semblance analysis: algo-
1021 rithm structure and synthetic examples, *Geophysical Prospecting*, **42**, 565–614.
- 1022 Symes, W. W., 2008. Migration velocity analysis and waveform inversion, *Geophysical Prospecting*, **56**, 765–
1023 790.
- 1024 Tape, C., Liu, Q., Maggi, A., & Tromp, J., 2010. Seismic tomography of the southern California crust based
1025 on spectral-element and adjoint methods, *Geophysical Journal International*, **180**, 433–462.
- 1026 Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysics*, **49**(8),
1027 1259–1266.
- 1028 Tromp, J., Tape, C., & Liu, Q., 2005. Seismic tomography, adjoint methods, time reversal and banana-doughnut
1029 kernels, *Geophysical Journal International*, **160**, 195–216.
- 1030 van Leeuwen, T. & Mulder, W. A., 2010. A correlation-based misfit criterion for wave-equation travelttime
1031 tomography, *Geophysical Journal International*, **182**(3), 1383–1394.
- 1032 Vigh, D., Jiao, K., Watts, D., & Sun, D., 2014. Elastic full-waveform inversion application using multicompo-
1033 nent measurements of seismic data collection, *Geophysics*, **79**(2), R63–R77.
- 1034 Villani, C., 2003. *Topics in optimal transportation*, Graduate Studies In Mathematics, Vol. 50, AMS.
- 1035 Villani, C., 2008. *Optimal transport : old and new*, Grundlehren der mathematischen Wissenschaften, Springer,
1036 Berlin.
- 1037 Wang, Y. & Rao, Y., 2009. Reflection seismic waveform tomography, *Journal of Geophysical Research*,
1038 **114**(B03304), doi:10.1029/2008JB005916.
- 1039 Warner, M. & Guasch, L., 2014. Adaptive waveform inversion - fwi without cycle skipping - theory, in *76th*
1040 *EAGE Conference and Exhibition 2014*.
- 1041 Warner, M., Ratcliffe, A., Nangoo, T., Morgan, J., Umpleby, A., Shah, N., Vinje, V., Stekl, I., Guasch, L., Win,
1042 C., Conroy, G., & Bertrand, A., 2013. Anisotropic 3D full-waveform inversion, *Geophysics*, **78**(2), R59–R80.
- 1043 Xu, S., Wang, D., Chen, F., Lambaré, G., & Zhang, Y., 2012. Inversion on reflected seismic wave, *SEG*
1044 *Technical Program Expanded Abstracts 2012*, pp. 1–7.
- 1045 Zhou, W., Brossier, R., Operto, S., & Virieux, J., 2015. Full waveform inversion of diving & reflected waves
1046 for velocity model building with impedance inversion based on scale separation, *Geophysical Journal Inter-
1047 national*, **202**(3), 1535–1554.

1048 Zhu, H., Bözdag, E., Peter, D., & Tromp, J., 2012. Structure of the european upper mantle revealed by adjoint
1049 tomography, *Nature Geoscience*, **5**, 493–498.

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

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

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

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.

1054 **Algorithms**

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

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

$$\left| \begin{array}{l} \varphi^n = (I_N + A^T A)^{-1} [(y_1^n - z_1^n) + A^T (y_2^n - z_2^n)]; \\ y_1^{n+1} = \text{prox}_{h_{d_{cal}[m], d_{obs}}}(\varphi^n + z_1^n); \\ z_1^{n+1} = z_1^n + \varphi^n - y_1^{n+1}; \\ y_2^{n+1} = \text{prox}_{i_K}(A\varphi^n + z_2^n); \\ z_2^{n+1} = z_2^n + A\varphi^n - y_2^{n+1}; \end{array} \right.$$

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  $i = 1, \dots, N_t$  do
  |  $b'_i = B^{-1}b_i$ ;
end
for  $i = 1, \dots, N_t$  do
  |  $c_i = P^T b'_i$ ;
end
for  $j = 1, \dots, N_r$  do
  | form  $c_{*j}$  from  $c = [c_1, \dots, c_{N_t}]$ ;
  | solve  $K_j y_{*j} = c_{*j}$ ;
end
for  $i = 1, \dots, N_t$  do
  | form  $c_i$  from  $c_{*j}$  from  $c = [c_{*1}, \dots, c_{*N_r}]$ ;
  |  $x_i = P c_i$ ;
end

```

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

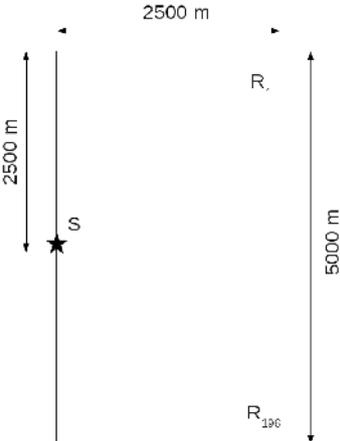


Figure 1. Configuration of the borehole to borehole experiment.

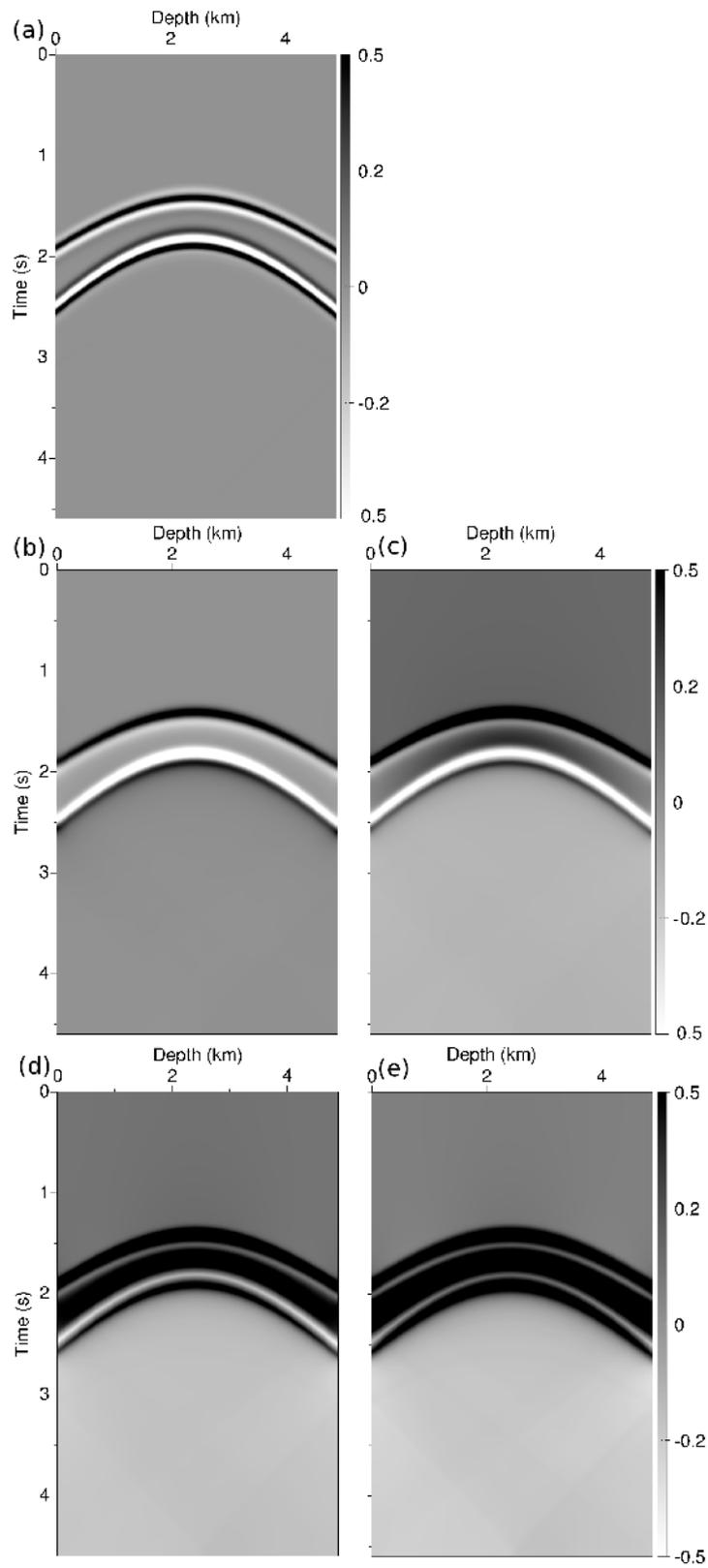


Figure 2. L2 residuals (a) and optimal transport based residuals with 5(b), 10 (c), 25 (d), 50 (e) SDMM iterations.

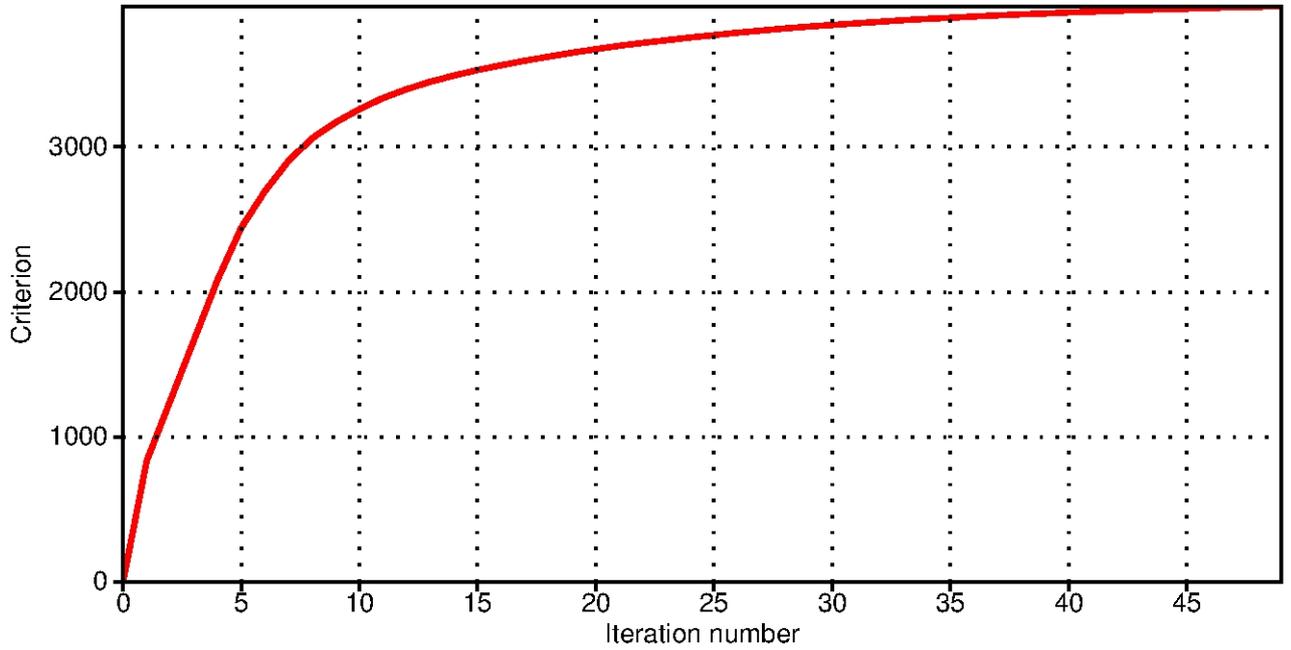


Figure 3. Evolution of the criterion maximized by the SDMM method along 50 iterations on the borehole to borehole schematic experiment. The criterion tends asymptotically toward a maximum value of 4000, which suggest that the convergence is reached. This is supported by the evolution of the solution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

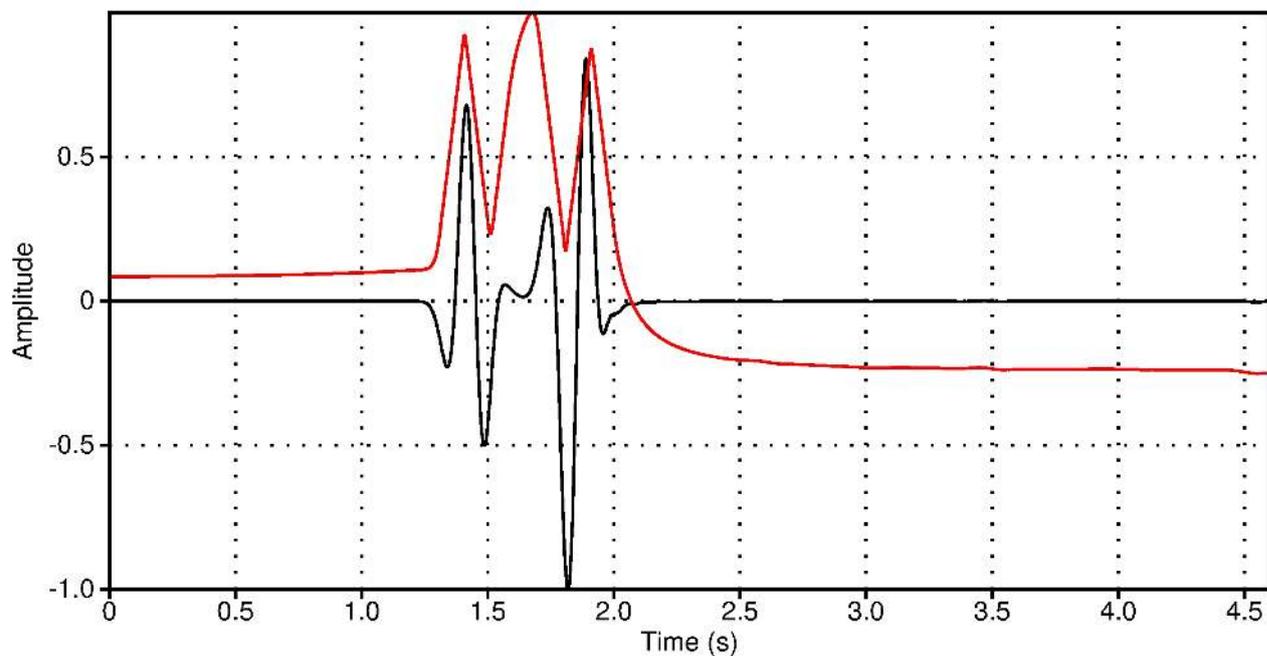


Figure 4. Traces extracted at 2.5 km depth from the original residuals (black) and from the solution computed after 50 SDMM iterations for the borehole to borehole schematic experiment.

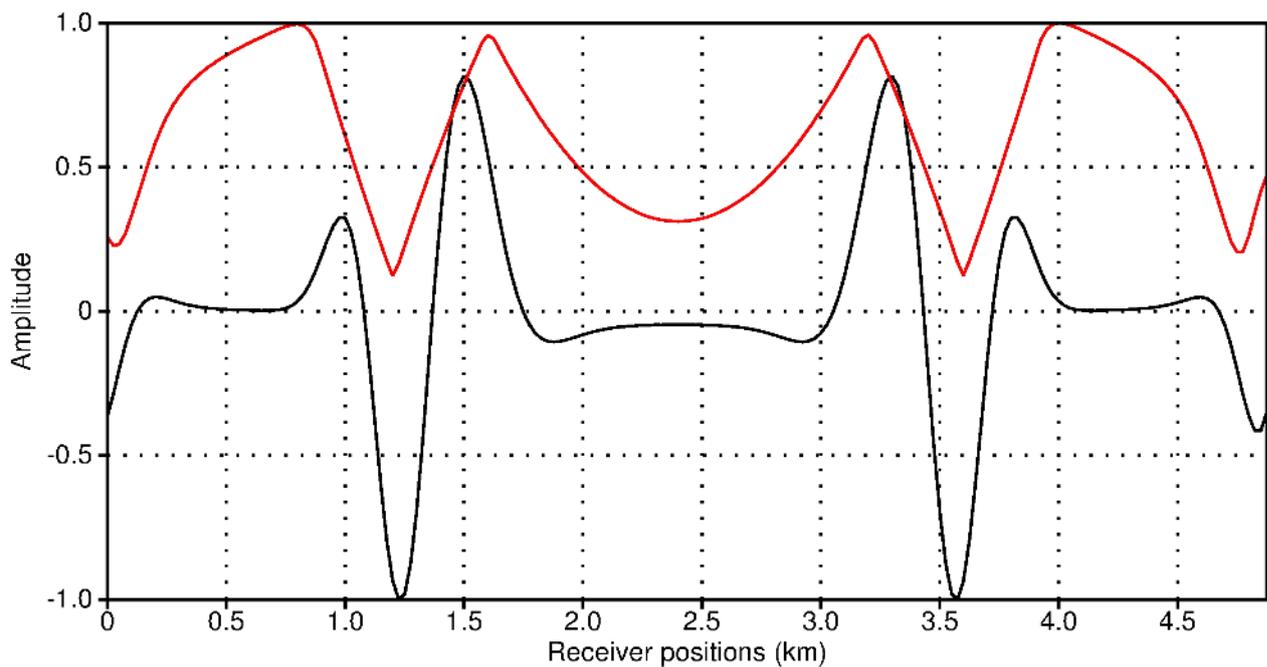


Figure 5. Traces extracted at time $t = 2$ s from the original residuals (black) and from the solution computed after 50 SDMM iterations for the borehole to borehole schematic experiment.

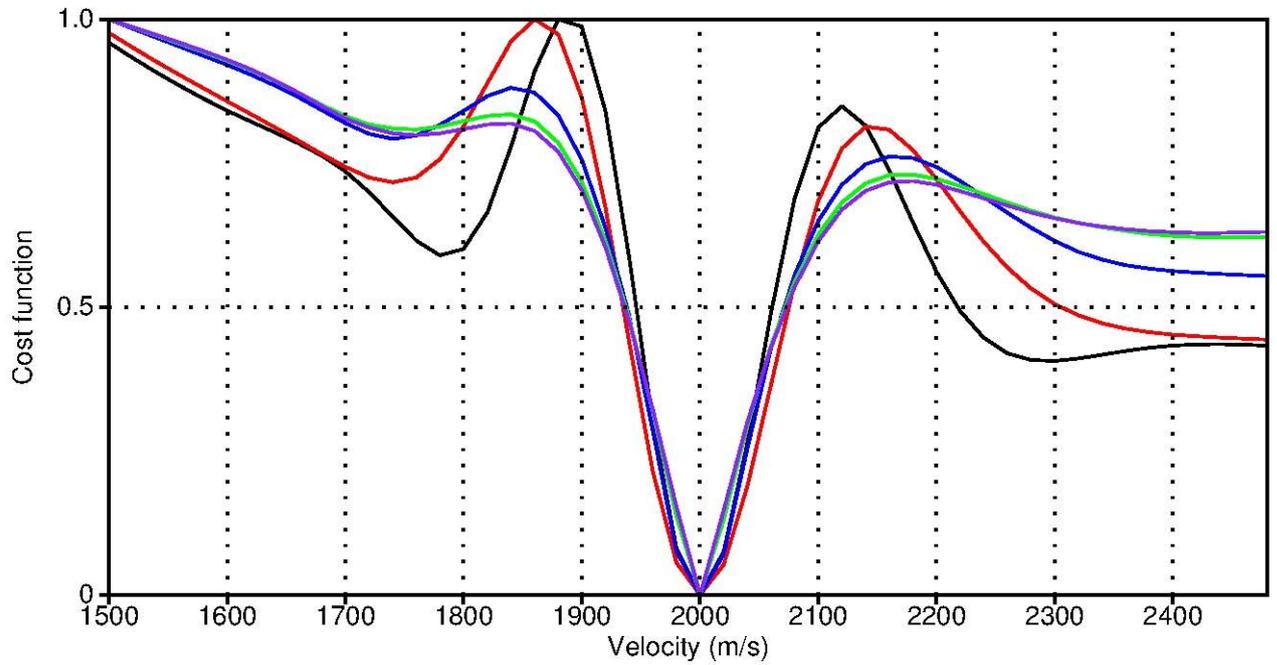


Figure 6. L^2 misfit function (black) and Wasserstein misfit function obtained with 5 (red), 10 (blue), 25 (green) and 50 (purple) SDMM iterations. The misfit functions are evaluated for a background velocity value ranging from 1500 m.s^{-1} to 2500 m.s^{-1} .

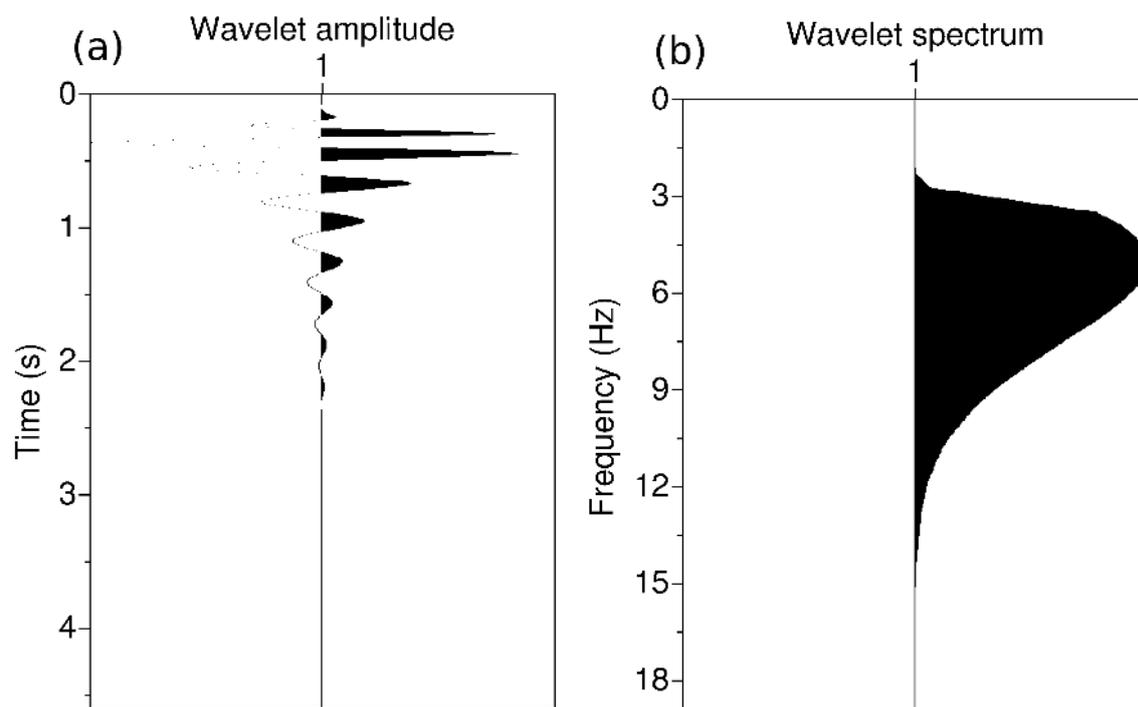


Figure 7. Source wavelet used to generate the synthetic dataset on the Marmousi model (a). This source is obtained from a Ricker wavelet centered on 5 Hz after applying a minimum phase Butterworth filter below 2.5 Hz. Corresponding spectrum (b).

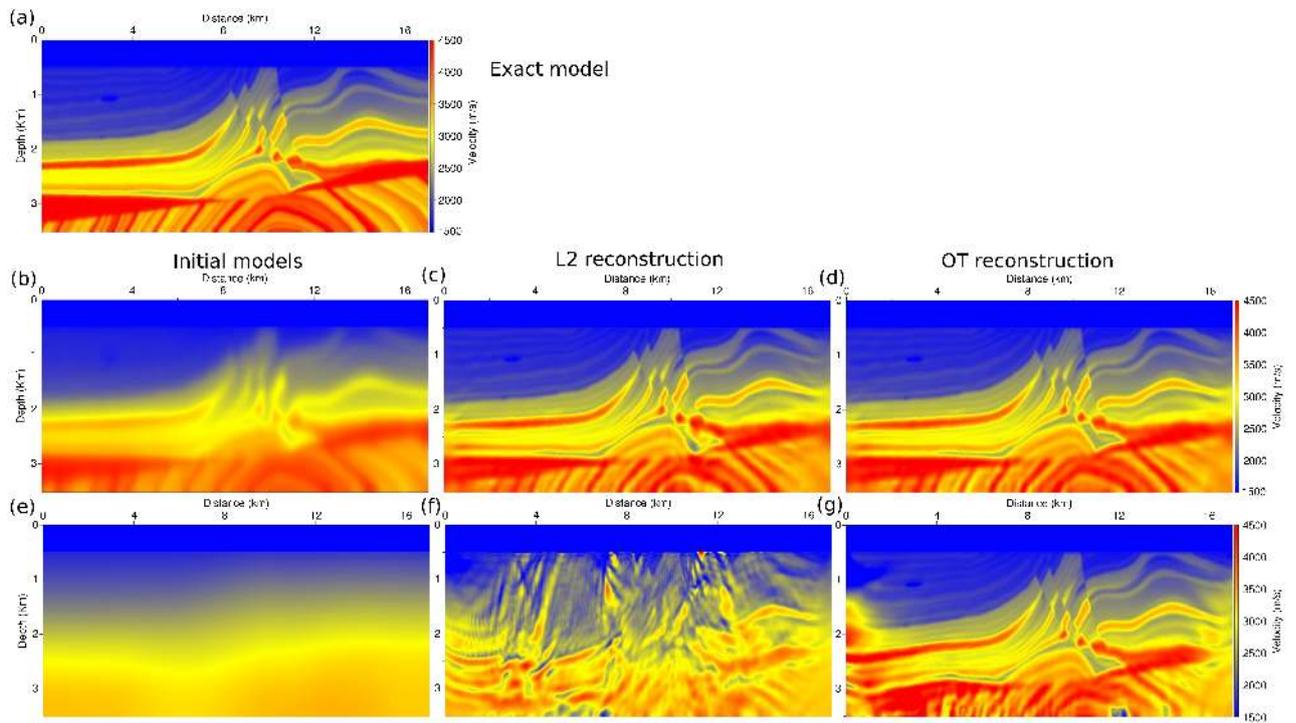


Figure 8. Marmousi 2 exact P-wave velocity model (a). Initial P-wave velocity models, computed from the exact model using a Gaussian filter with a correlation length of 250 m (b) and 2000 m (e). Corresponding P-wave velocity estimations with FWI using the L^2 misfit function (c),(f). Corresponding P-wave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

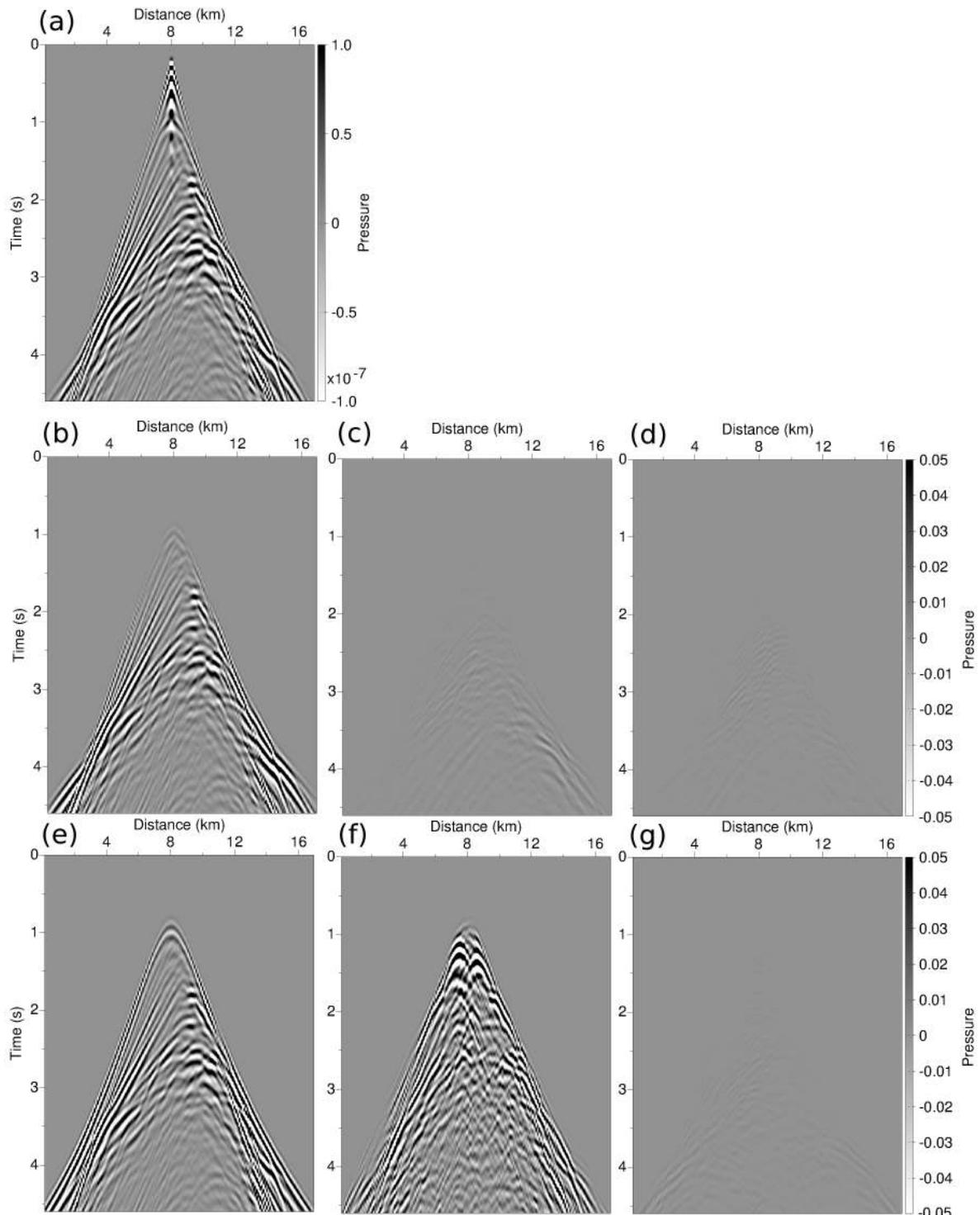


Figure 9. Marmousi 2 exact data for the shot-gather corresponding to the source position $x_S = 8$ km (a). Associated residuals in the initial P-wave velocity models (b),(e). Associated residuals in the P-wave velocity models estimated with FWI using the L^2 misfit function (c),(f). Associated residuals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

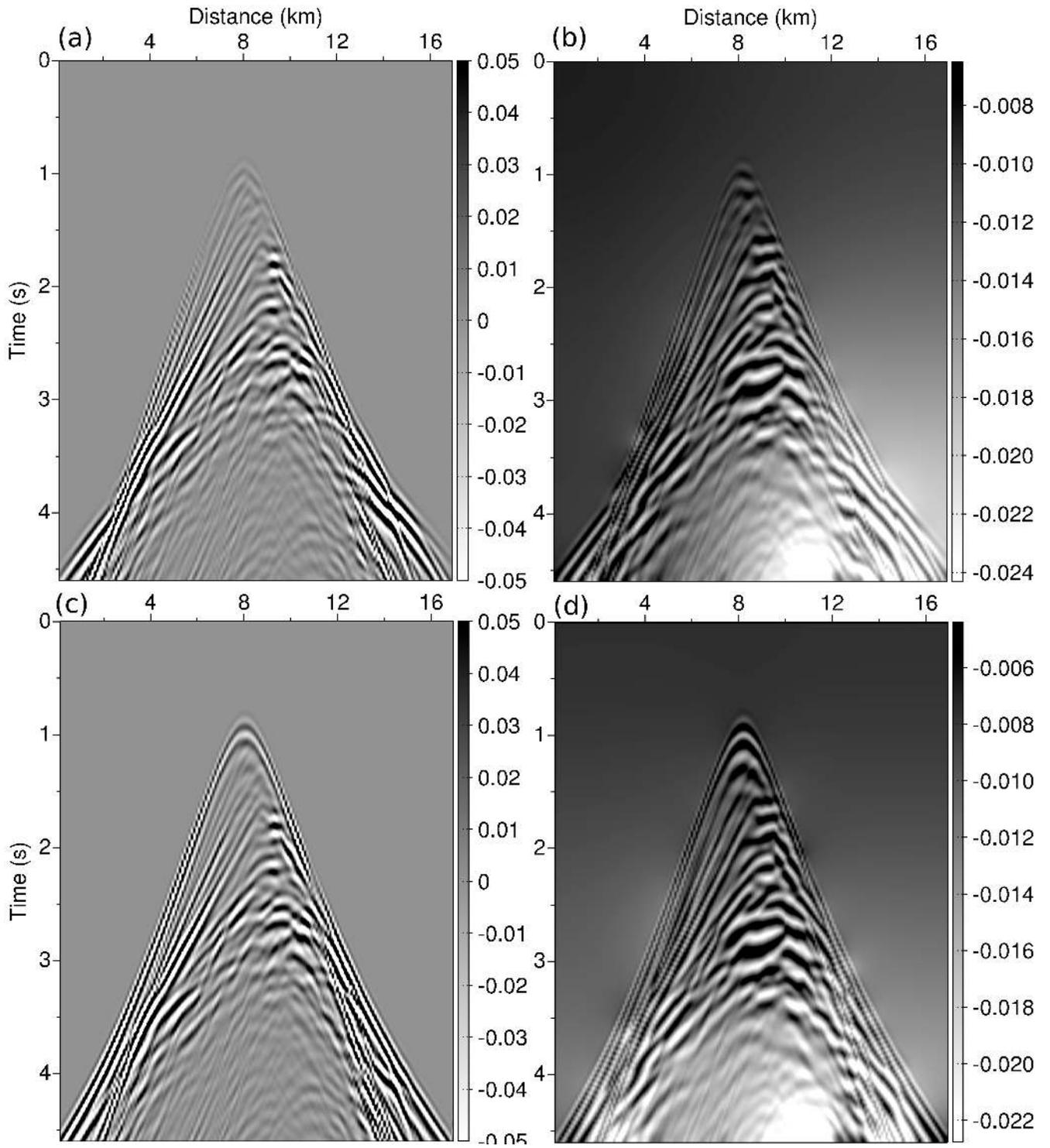


Figure 10. L^2 residuals in the initial model 1 (a) and 2 (c). Corresponding optimal transport residuals (b),(d).

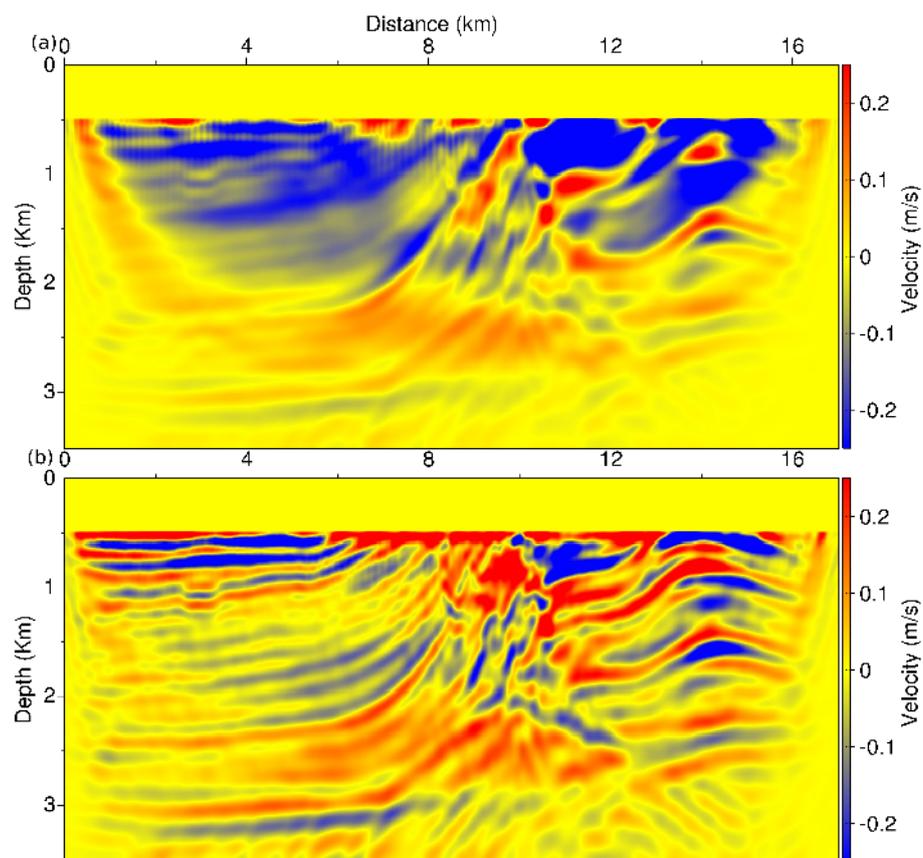


Figure 11. Initial descent direction (opposite of the gradient) in the initial model 3 using the L^2 distance (a) and the optimal transport distance (b).

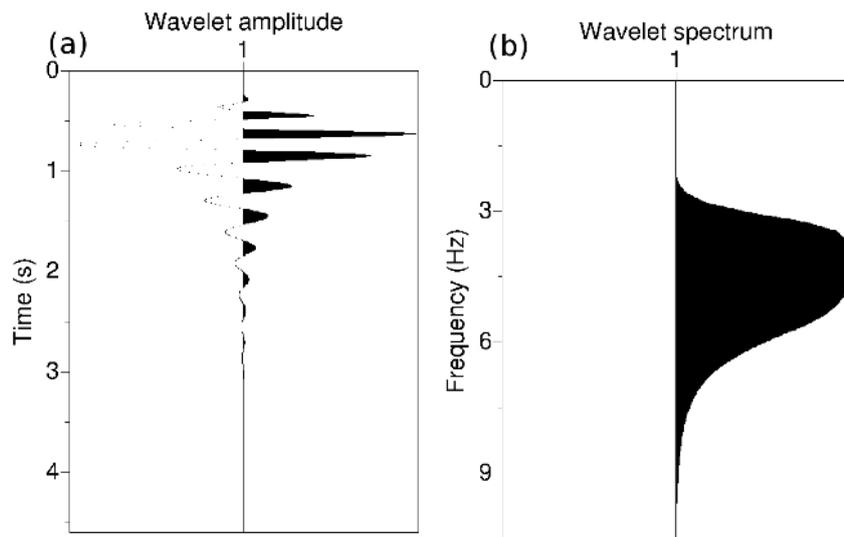


Figure 12. Source wavelet used to generate the synthetic dataset on the BP 2004 model (a). This source is obtained from a Ricker wavelet centered on 5 Hz. A whitening of its frequency content is performed before a low-pass and high-pass filter are applied, so that the corresponding spectrum spans an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

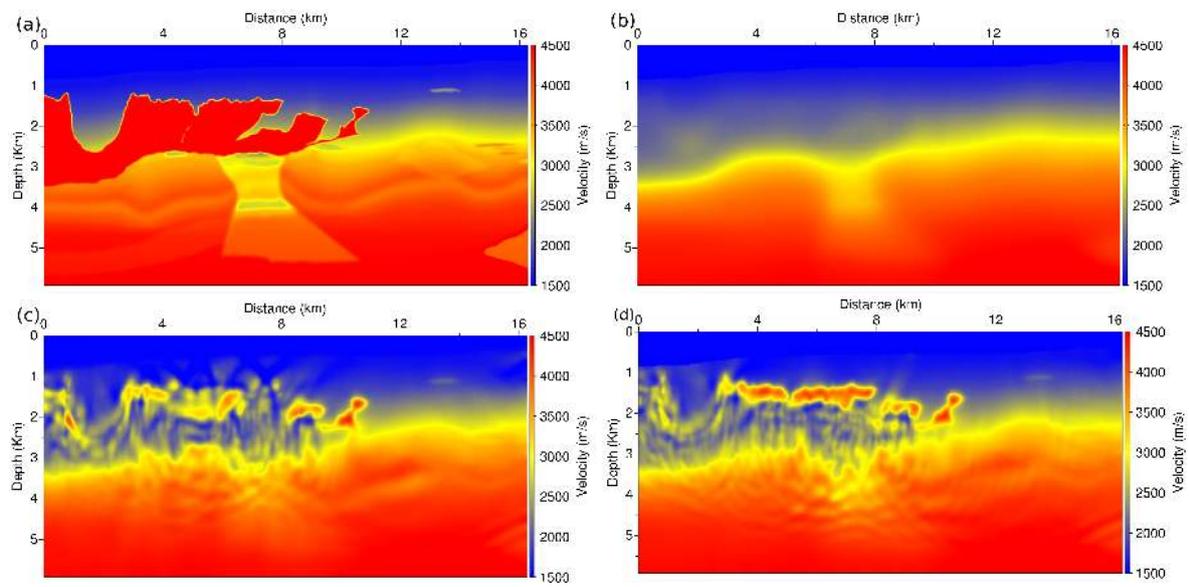


Figure 13. BP 2004 exact model (a) and initial model (b). P-wave velocity estimation with a standard L^2 norm on short-time window data (4.6 s) (c). The same with the optimal transport distance (d).

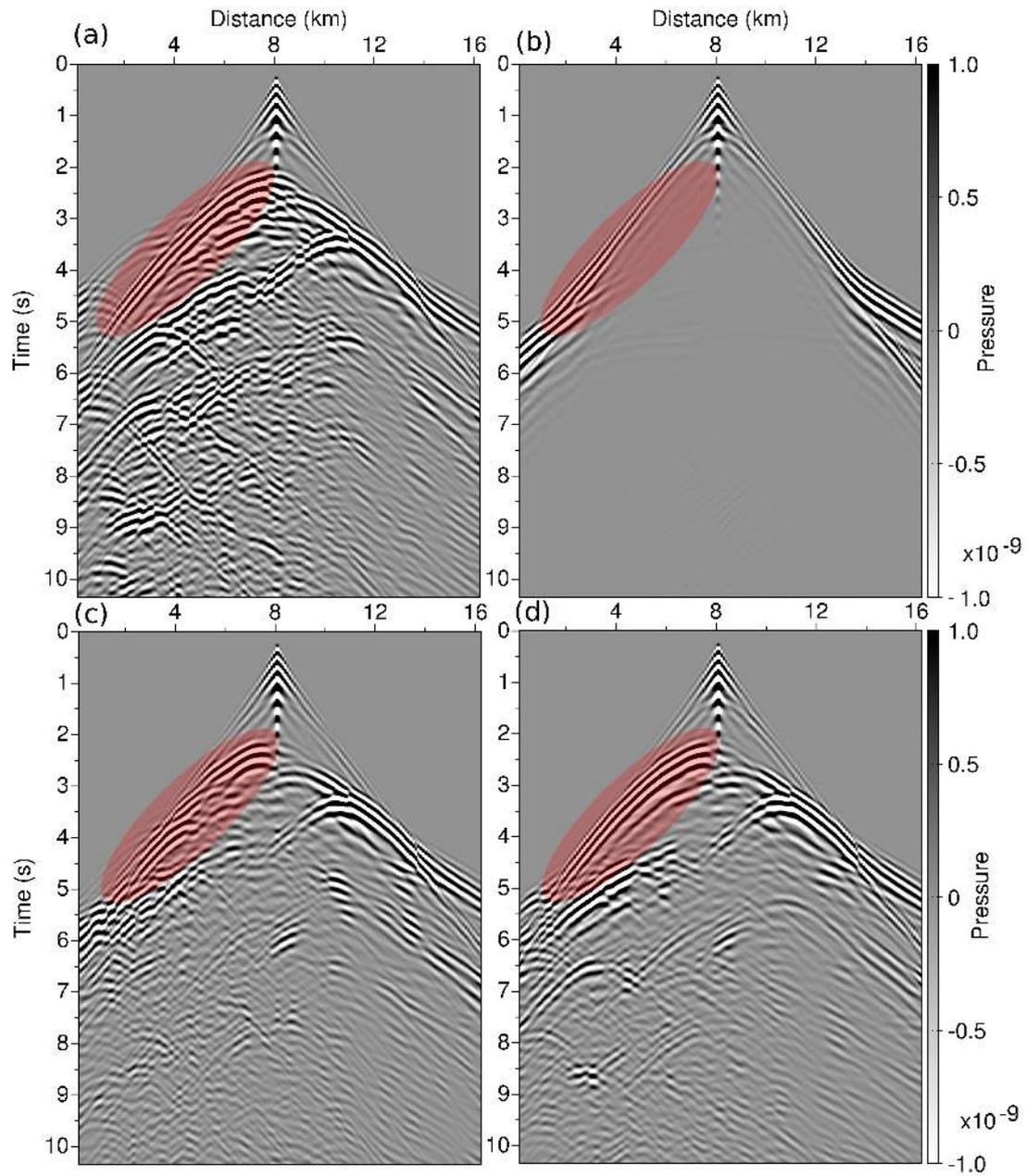


Figure 14. BP 2004 exact data (a) and initial data (b). Predicted data in the final model using a standard L^2 norm (c). Predicted data in the final model using the optimal transport distance using together with a layer stripping workflow (d). The red ellipses highlight the reflection on the salt roof. 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).

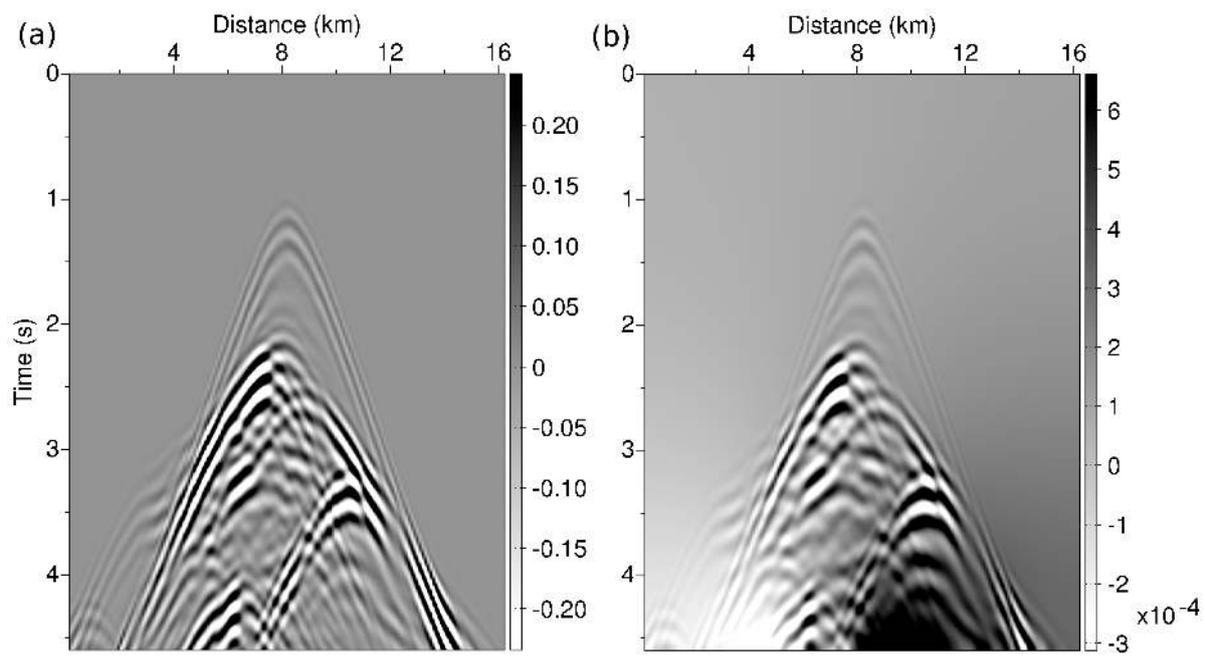


Figure 15. BP 2004 case study. L^2 residuals in the initial model (a). Optimal transport residuals in the initial model (b).

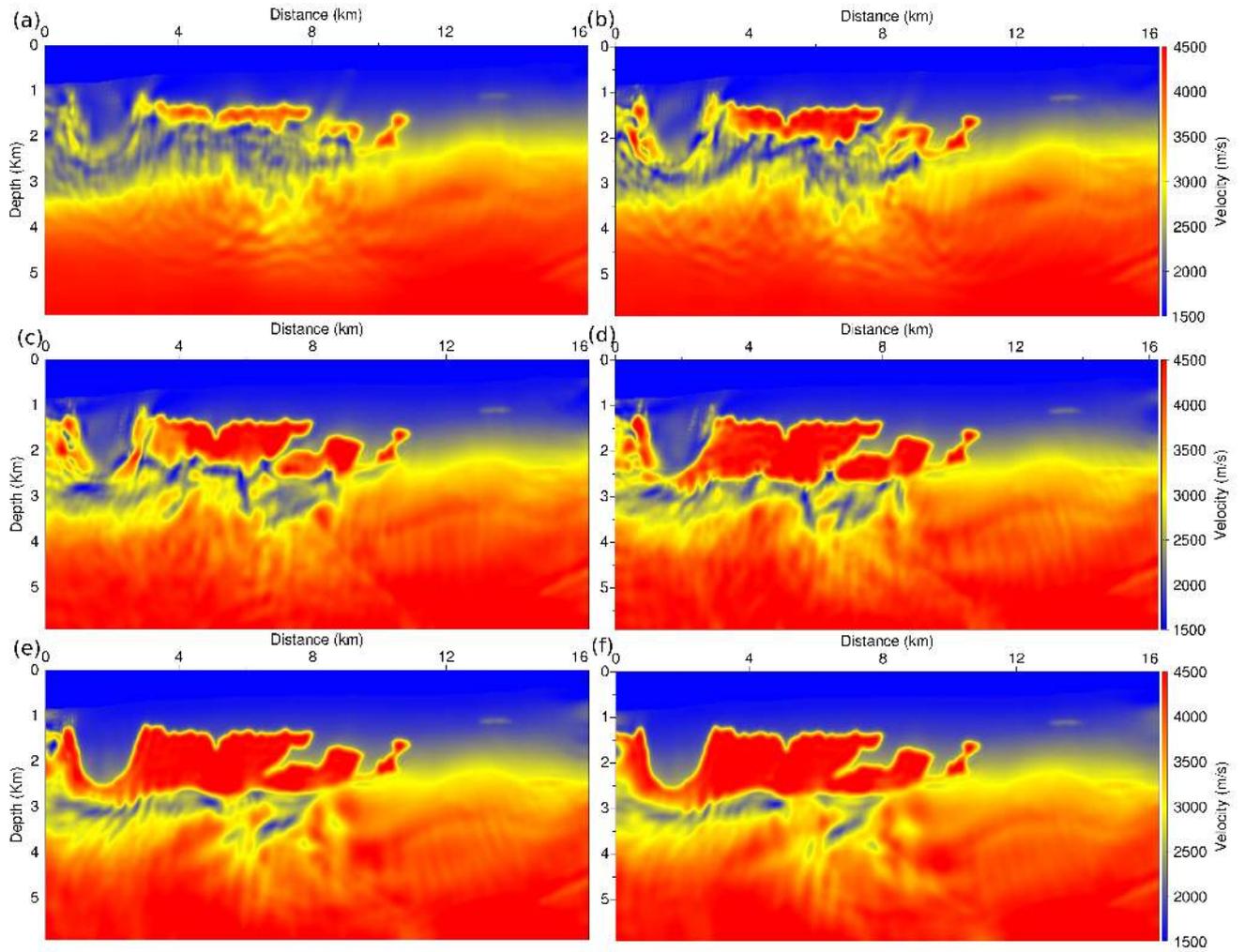


Figure 16. BP 2004 P-wave velocity estimation computed after the 1st(a), 3rd (b), 6th (c), 9th (d), 12th (e), and 15th (f) inversion using the optimal transport distance.

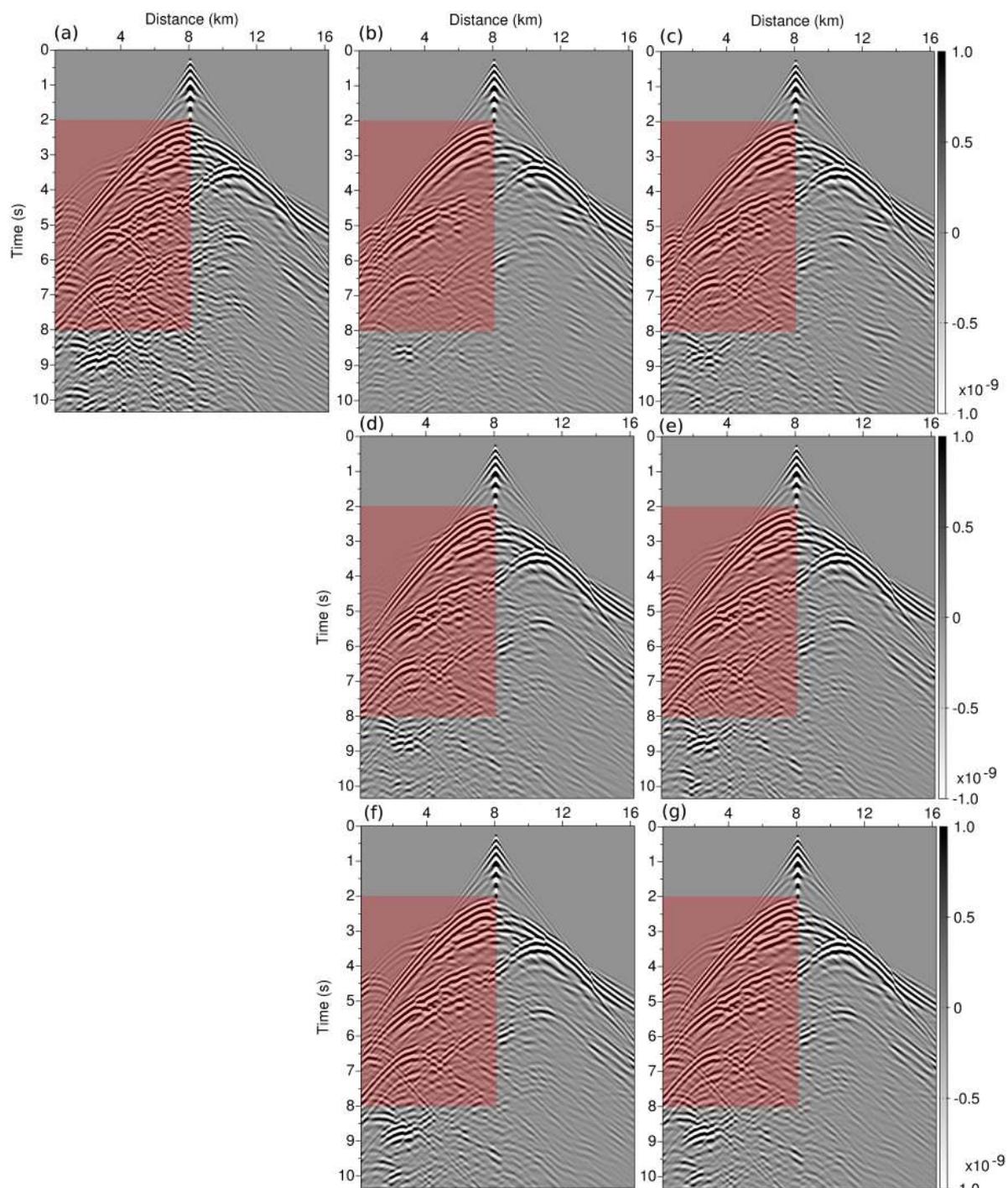


Figure 17. Synthetic data in the exact model (a) and in the intermediate models obtained with FWI using an optimal transport distance after the 1st(b), 3rd (c), 6th (d), 9th (e), 12th (f), and 15th (g) inversion. The red rectangles highlight the shot-gather zone associated with the diving waves traveling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

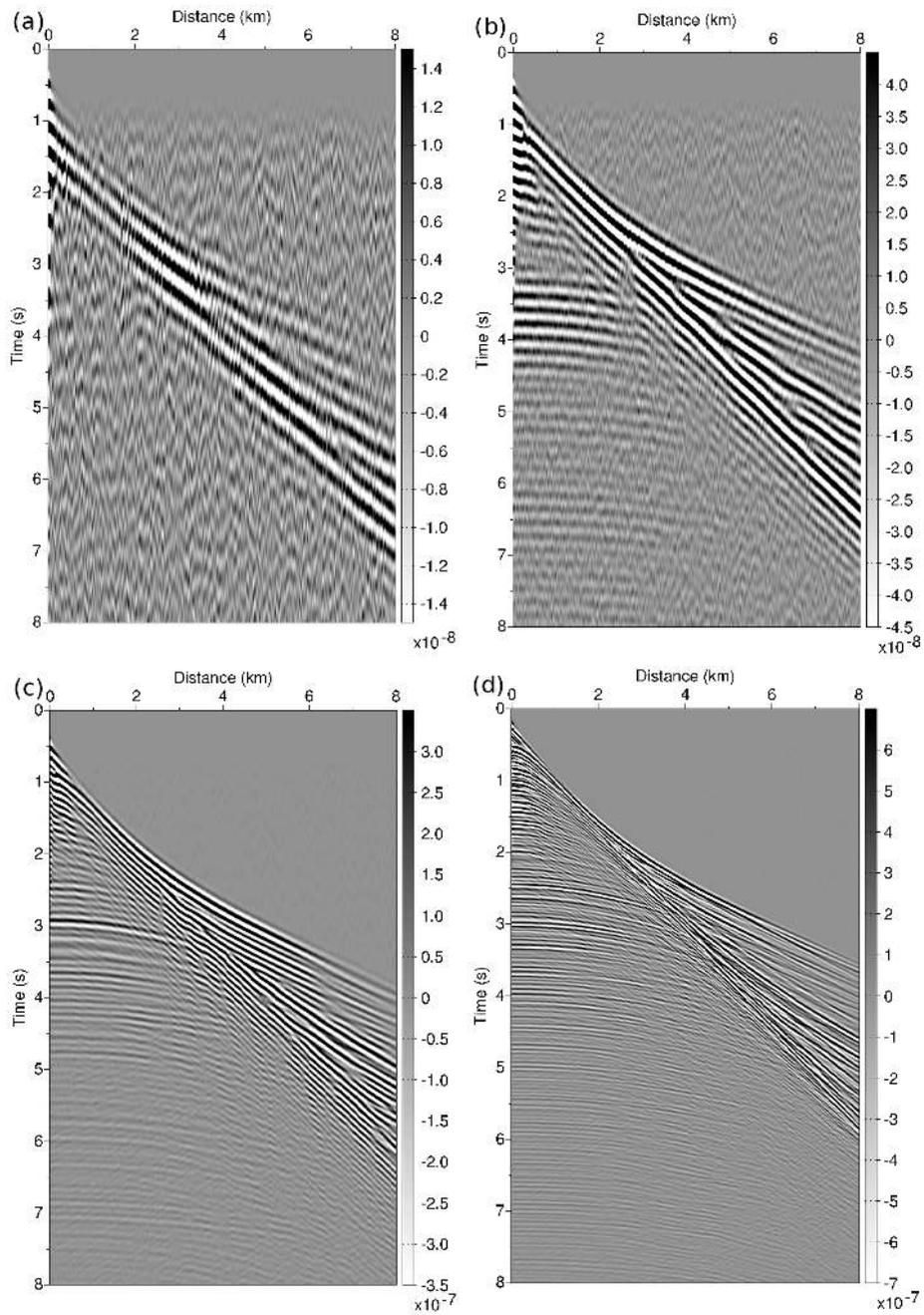


Figure 18. Chevron 2014 dataset. Common shot-gather for the source situated at $x = 0$ km for the frequency bands 1 (a), 5 (b), 10 (c), and 15 (d).

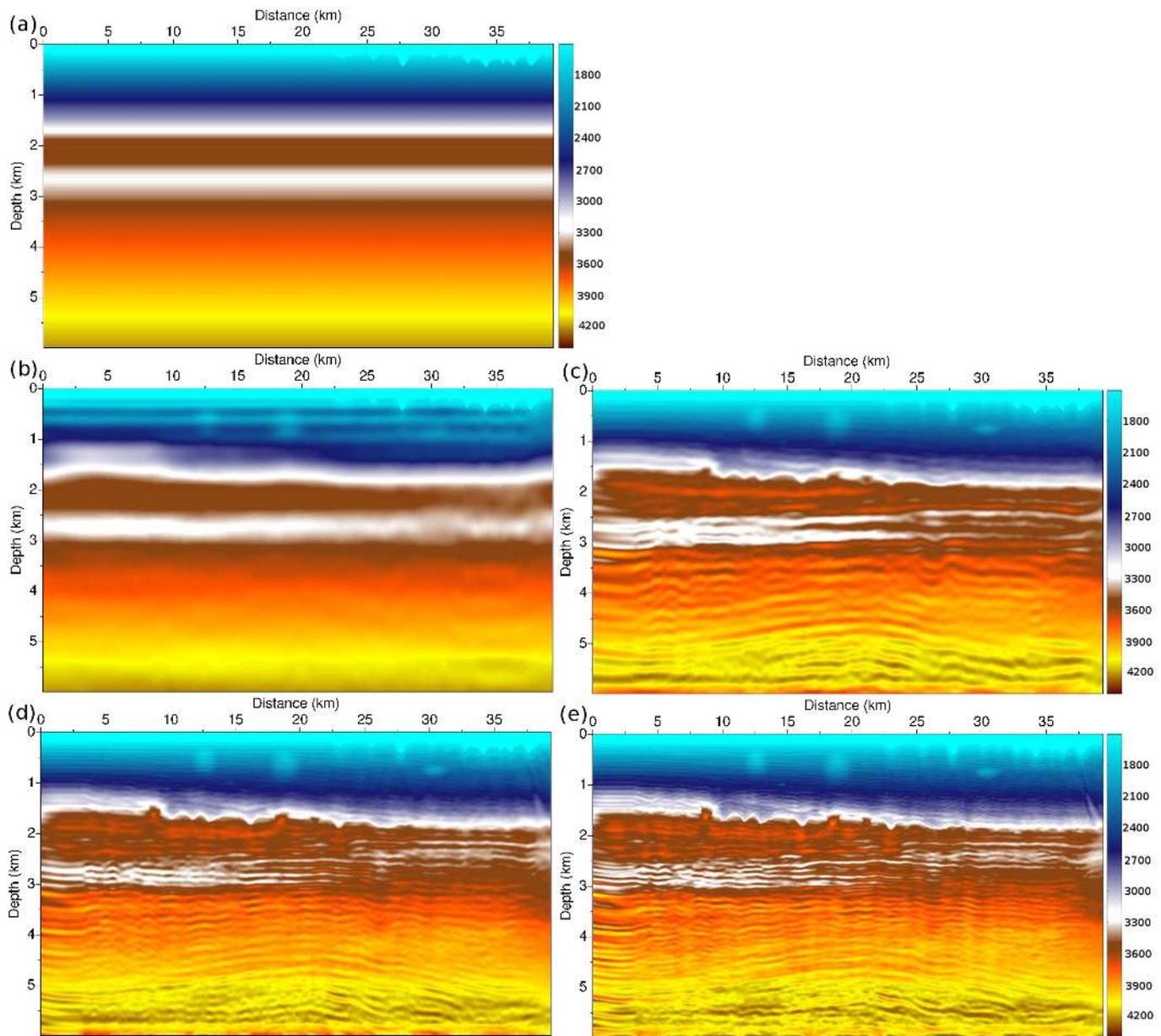


Figure 19. Chevron 2014 starting P-wave velocity model (a). Estimated P-wave velocity model at 4 Hz (b), 10 Hz (c), 16 Hz (d), 25 Hz (e).

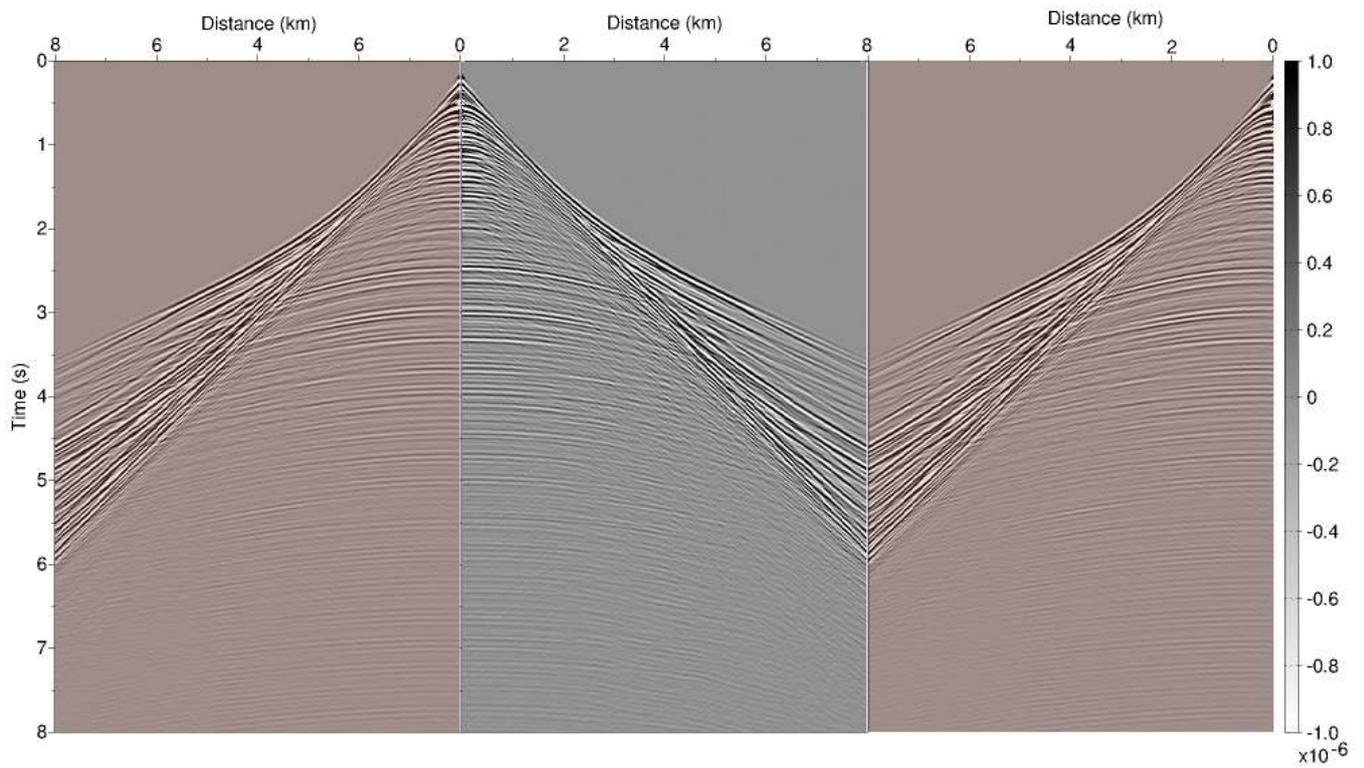


Figure 20. Exact common shot-gather for the left most source at 25 Hz, compared to the corresponding synthetic in the final model at 25 Hz (orange panels). The synthetic data is mirrored and placed on both sides of the real data to better compare the match of the different phases.

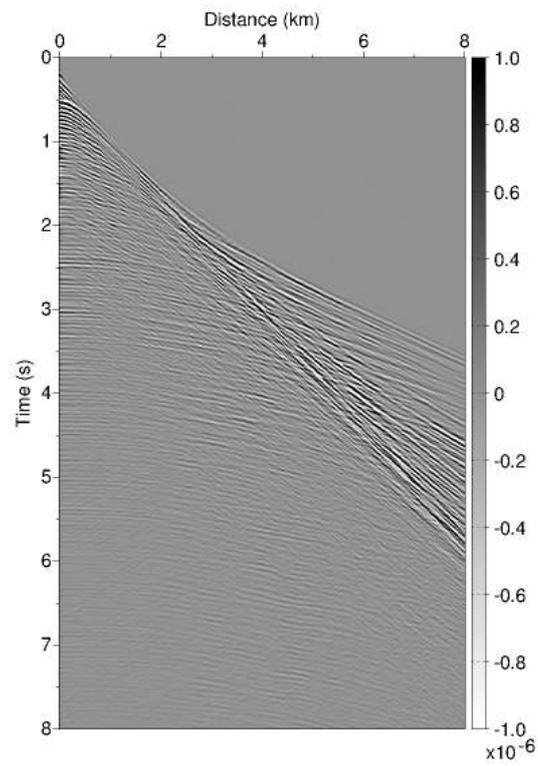


Figure 21. Residuals between the exact common shot-gather for the left most source at 25 Hz and the corresponding synthetic common shot-gather.

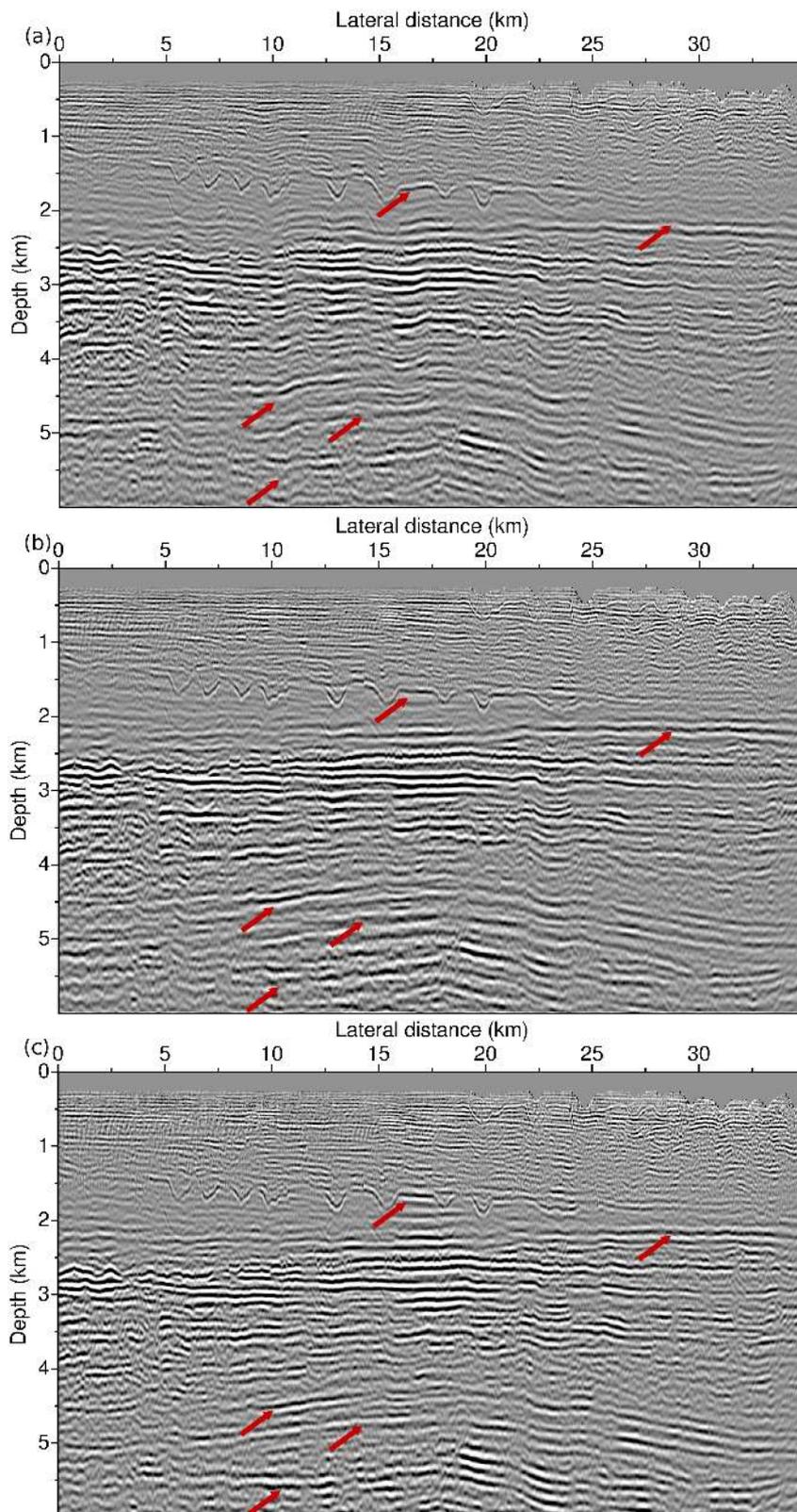


Figure 22. Migrated images in the initial model (a), in the model obtained at 10 Hz maximum frequency (b), in the model obtained at 16 Hz maximum frequency (c). Red arrows indicate identifiable improvements of the reconstruction of the reflectors and re-focusing of the energy. Improvements in the shallow 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.

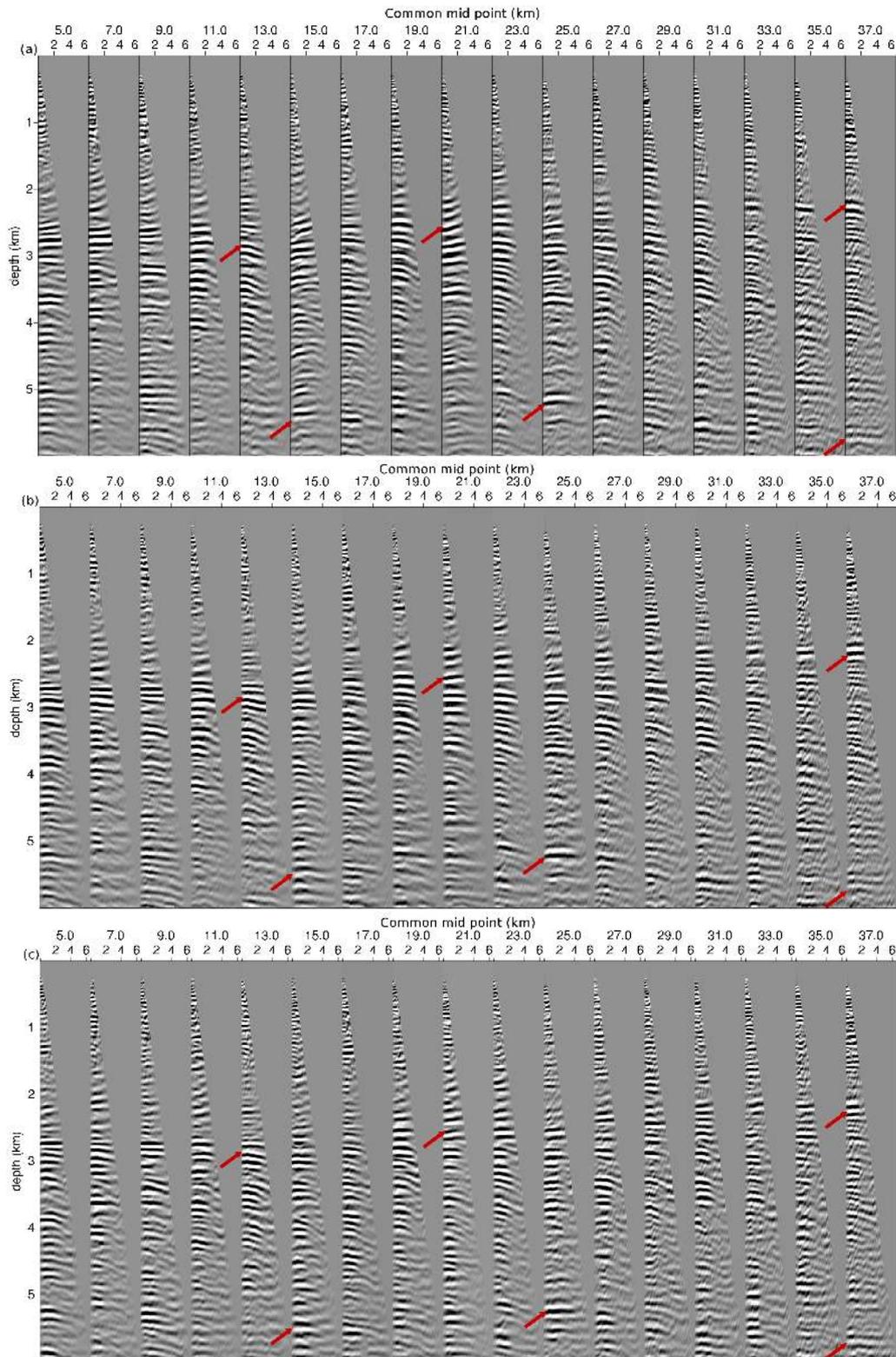


Figure 23. CIG in the initial model (a), in the model obtained at 10 Hz maximum frequency (b), in the model obtained at 16 Hz maximum frequency (c). Red arrows indicate identifiable improvement of the CIG continuity in the offset direction. As for the migrated images, improvements in the shallow part (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.

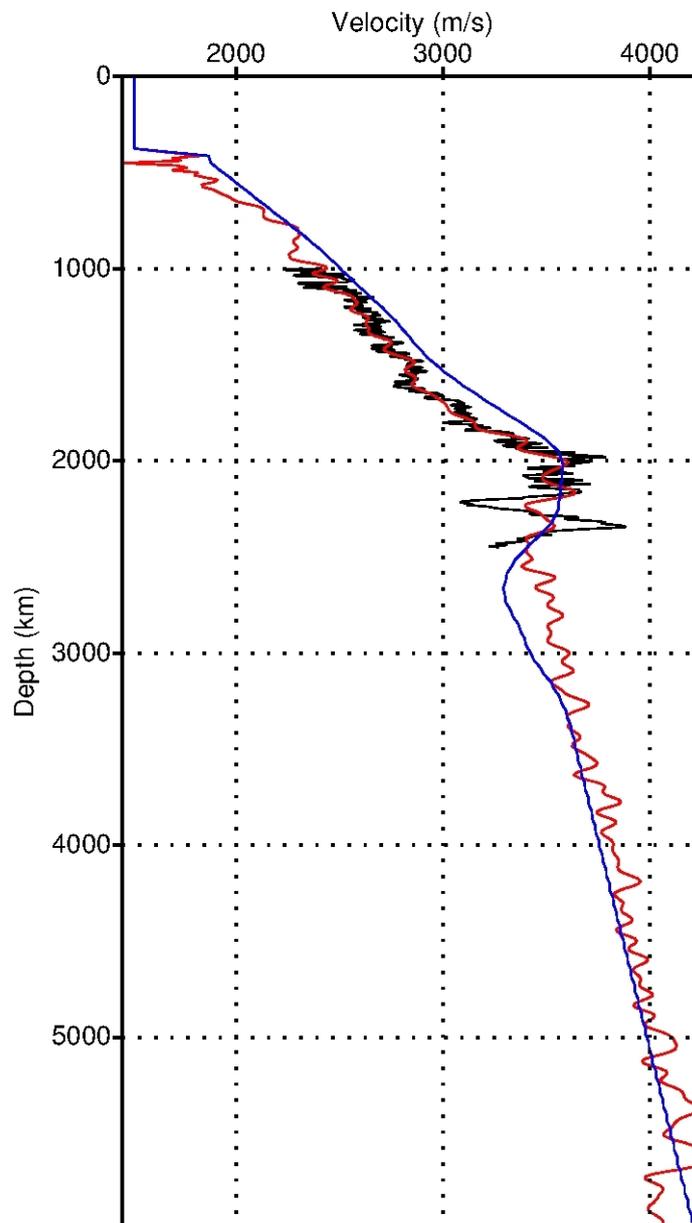


Figure 24. Vertical P -wave velocity log taken at $x = 39,375$ km. Initial model (blue), exact model (black), estimation at 25 Hz (red).

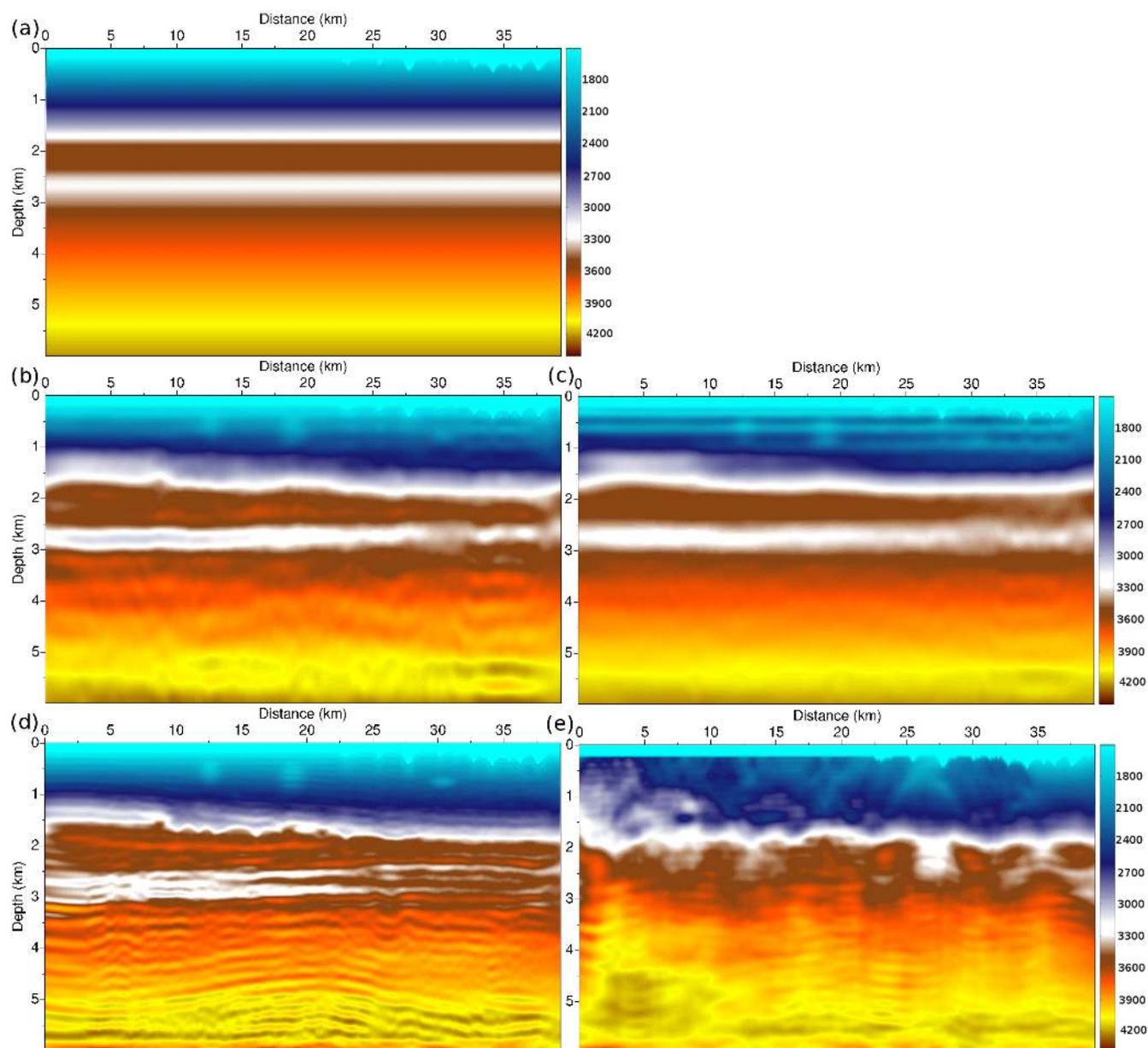


Figure 25. Chevron 2014 starting P-wave velocity model (a). Estimated P-wave velocity model at 4 Hz with the optimal transport distance (b), with the L^2 distance (c). Estimated P-wave velocity model at 10 Hz with the optimal transport distance (d), with the L^2 distance (e).

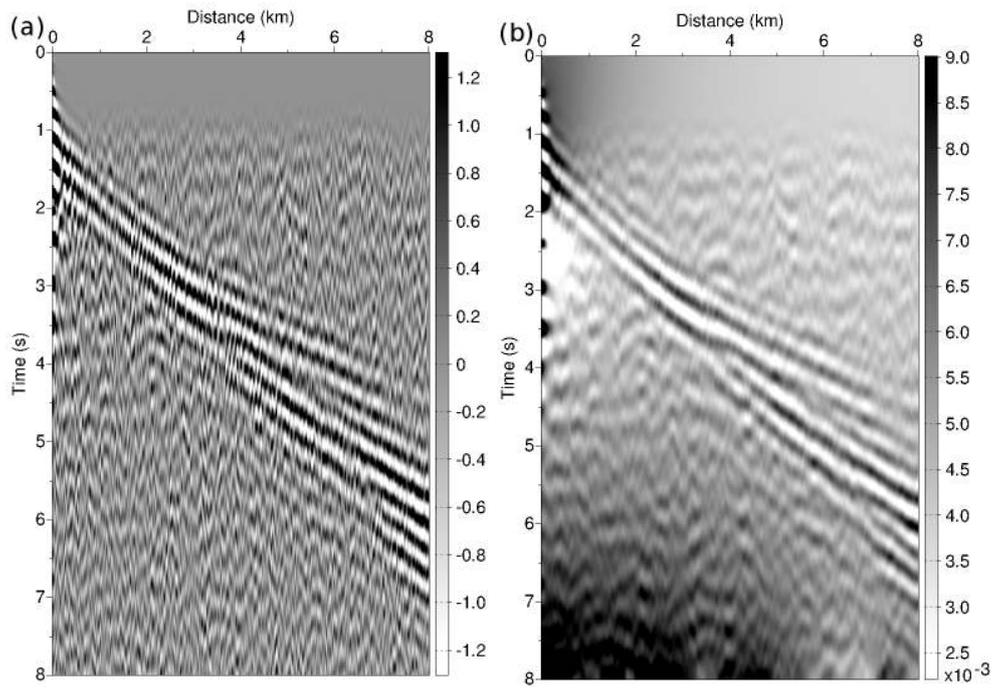


Figure 26. Residuals in the initial model for the first frequency band, using the L^2 norm misfit function (a), using the optimal transport distance (b).

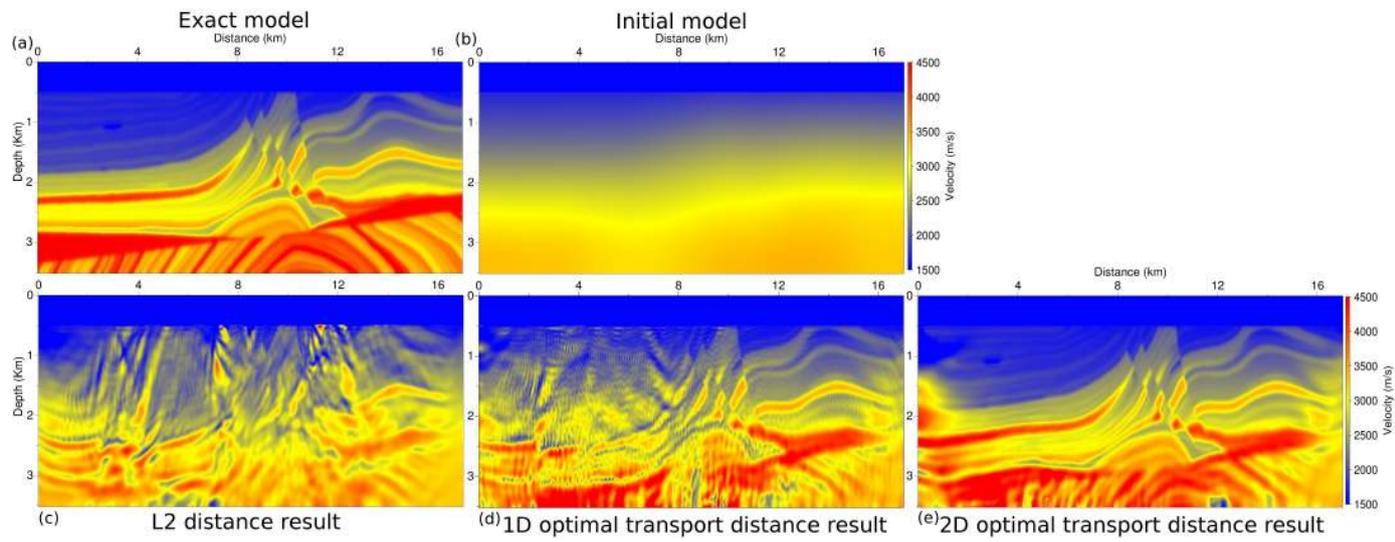


Figure 27. Exact Marmousi 2 P-wave velocity model (a). Initial model corresponding to the third initial model of Figure 8 (b). Reconstructed model using the L^2 distance (c), using 1D optimal transport distance (d), 2D optimal transport distance (e).