On the use of nonlinear anisotropic diffusion filters for seismic imaging using the full waveform

Ludovic Métivier, Romain Brossier

On the use of nonlinear anisotropic diffusion filters for seismic imaging using the full waveform

L. Métivier[1,2], **R. Brossier**[2]

[1] CNRS, Univ. Grenoble Alpes, LJK, F-38058 Grenoble, France
[2] Univ. Grenoble Alpes, ISTerre, F-38058 Grenoble, France

E-mail: ludovic.metivier@univ-grenoble-alpes.fr

January 2022

Abstract. Nonlinear anisotropic diffusion filters have been introduced in the field of image processing for image denoising and image restoration. They are based on the solution of partial differential equations involving a nonlinear anisotropic diffusion operator. From a mathematical point of view, these filters enjoy attractive properties, such as minimum-maximum principle, and an inherent decomposition of the images in different scales. We investigate in this study how these filters can be applied to help solving data-fitting inverse problems. We focus on seismic imaging using the full waveform, a well known nonlinear instance of such inverse problems. In this context, we show how the filters can be applied directly to the solution space, to enhance the structural coherence of the parameters representing the subsurface mechanical properties and accelerate the convergence. We also show how they can be applied to the seismic data itself. In the latter case, the method results in an original low-frequency data enhancement technique making it possible to stabilize the inversion process when started from an initial model away from the basin of attraction of the global minimizer. Numerical results on a 2D realistic synthetic full waveform inversion case study illustrate the interesting properties of both approaches.

Submitted to: *Inverse Problems*

1. Introduction

Among a wide variety of existing strategies dedicated to image denoising and image restoration, nonlinear anisotropic diffusion filters have been proposed at the end of the 1990s as an efficient partial differential equations (PDE) based technique. Compared to state-of-the-art techniques such as total variation regularization, denoising methods based on the solution of PDE inherits a natural theoretical framework, both at continuous and discrete levels, making it possible to derive specific properties for the filter such as stability, energy conservation, separation of scales (a notion developed as scale-space properties in the following), as well as efficient numerical schemes. Working in the framework of the solution of PDE provides also flexibility in terms of application of the method to fields beyond image processing. In this study, we are interested in the application of this technique in the field of high resolution seismic imaging, using a method named full waveform inversion (FWI).

FWI can be formulated as a data fitting inverse problem, where the observed data consist in the surface recording of mechanical waves propagating inside the Earth and the synthetic data is computed through the solution of a PDE describing the wave propagation in the subsurface. A distance between these observed and synthetic data is minimized through local optimization techniques to update iteratively the parameters of the PDE representing the subsurface mechanical properties.

Initially designed in the 80s (Lailly, 1983; Tarantola, 1984), FWI is now a mature technique, routinely applied at the Earth and regional scales (Fichtner et al., 2009; Tape et al., 2010; Lei et al., 2020; Górszczyk et al., 2021), as well as at the crustal scale (Plessix and Perkins, 2010; Stopin et al., 2014; Operto et al., 2015). These successful applications however depend on expertise in data processing and initial velocity model building. Mathematically, FWI remains an ill-posed inverse problem: the PDE depends nonlinearly on the parameters which are reconstructed, and both modeling and observation noise is to be accounted for.

Numerous strategies have thus been investigated, to make FWI a more stable and better posed inverse problem, with the objective to relax the conditions required for successful applications of FWI. Such investigations encompass misfit function modifications (van Leeuwen and Mulder, 2010; Bozdağ et al., 2011; Métivier et al., 2016; Warner and Guasch, 2016; Yang and Engquist, 2018; Métivier et al., 2019), to enhance the convexity of the misfit function, as well as "extension" strategies, consisting in rewriting the PDE-based inverse problem by integrating artificial degrees of freedom to reduce its nonlinearity (Symes, 2008; van Leeuwen and Herrmann, 2013; Huang et al., 2018; Aghamiry et al., 2020a).

Regularization, as in the general case of the solution of ill-posed inverse problems, is also a central technique in FWI. Such regularization can be seen as the injection of prior information on the solution, reducing the size of the solution space. This prior information can be related to known values of the reconstructed parameters in specific areas, for instance through upscaled sonic-logs extracted from wells for the velocity parameters. More generally, the prior information injected in the inverse problems is related to the smoothness/structure of the reconstructed parameters. By structure, we mean orientations/directions along which the variation of the reconstructed parameters is slow, while the variation perpendicular to this orientation is fast.

How to inject this prior information is a matter of choice. The most common consists in adding penalization terms (Tikhonov strategies, Tikhonov et al., 2013) measuring the distance of the model parameter to a prior model, or equal to the norm of its spatial derivatives to enforce smoothness (see for instance Asnaashari et al., 2013). Another well known technique is the application of specific smoothing filters to the model update (*i.e.* the gradient) at each iteration of the inversion. Among this category, the most widespread is the Gaussian filter, which can be made non-stationary to adapt it to the expected local resolution as is done for instance in Operto et al. (2006). The local resolution can be indeed estimated from diffraction tomography analysis (Devaney, 1984; Wu and Toksöz, 1987; Sirgue, 2003). When a prior information on the structure

is accessible, for instance from reflectivity images of the subsurface target, or a geological interpretation of the investigated zone, it can also be injected by adapting the smoothing operation so as to follow specific orientations, with the objective not to smooth across interfaces and preserve the imprint of the underlying structures. Directional Laplacian filtering has been proposed to implement such non-local oriented smoothing (Guitton et al., 2012). A PDE-based smoothing process based on Bessel's filter has also been recently proposed (Trinh et al., 2017a). Edge preserving smoothing through Total Variation (TV) regularization is also a conventional technique applied in FWI, with a special interest for the reconstruction of salt bodies in exploration case studies (Strong and Chan, 2003; Peters and Herrmann, 2017; Anagaw and Sacchi, 2018; Aghamiry et al., 2020b). The boundary of these structures is sharp, while the mechanical properties are almost constant within them, making TV regularization an appropriate tool for their reconstruction.

Note that smoothing the reconstructed parameters might be somehow counter-intuitive given the high resolution objective of FWI. However, the discretization required to represent the subsurface mechanical properties is driven by the solution of the PDE describing the wave propagation within the subsurface. A sufficiently small discretization mesh has to be used to ensure sufficiently low numerical dispersion. This requirement can make the mesh size below the resolution one can estimate in the diffraction tomography approximation. Behind this is the question how to parameterize adequately the model space for inversion. Instead of smoothing the model according to the resolution, projection of the parameters from a modeling space to an inversion space could be considered to avoid over-parameterization issues. Projection on a spline basis (Dierckx, 1993) has been for instance used in FWI (Barnier et al., 2019), and is common in tomography (Nolet, 2008). Another example, at a more theoretical level, is a parameterization on a basis of eigenvectors of a TV-based regularization operator (Grote et al., 2017). More recently, a re-parameterization based on equivalent media theory (homogenization) has also been proposed (Capdeville and Métivier, 2018). Homogenization provides smooth elastodynamics subsurface models which generate equivalent wavefields (up to a certain, controllable, accuracy) for the propagation of elastic waves in a given frequency band. One drawback of this approach is that a homogenized subsurface parameterization is intrinsically fully anisotropic, leading to 21 independent stiffness tensor coefficient for 3D elastodynamics inversion (Cupillard and Capdeville, 2018).

In this study, we are interested in an alternative regularization approach. This alternative approach is based on nonlinear anisotropic diffusion filters, which we will refer to as NADF in the following. We study how NADF can be used in the context of FWI. Such filters provide the possibility to enhance the structural coherence of a given image directly, without including it as a prior information. This is done through the definition of a nonlinear anisotropic diffusion operator, which is based on an eigendecomposition of the local structure tensor of the image. The diffusion operator depends on the current image, hence the nonlinearity of this filter.

Besides the use of NADF in the *model* space, we also consider the application of NADF in the *data* space. There are several reasons to support this idea. The first is the existence of noise which always contaminates the seismic data. Applying denoising filters is thus natural. The second is, as is detailed in the following, that NADF provides a natural hierarchy of scales, a property referred to as scale-space property in Weickert (1998). Exploiting such hierarchy is a conventional process for FWI applications which often rely on a hierarchical interpretation of the data, from low to high frequency (Bunks et al., 1995; Pratt, 1999), complemented with specific time/offset windowing (Shipp and Singh, 2002; Wang and Rao, 2009; Brossier et al., 2009). Third, we will see it connects also well with different attempts to enhance low-frequency components of the data to stabilize the FWI process (Li and Demanet, 2016; Sun and Demanet, 2020) and generalizes a Gaussian smoothing kernel technique promoted by Xue et al. (2016). Fourth, as is detailed in this study, the PDE-based formalism of this filtering technique makes it possible to derive, through the adjoint-state technique, an analytic formula for the gradient of a functional measuring the misfit between filtered observed and synthetic data. This would not be the case for a non-PDE based filtering technique.

The structure of the study is a follows. We first recall the main theoretical results regarding

NADF and introduce the concept of scale-space properties. Then, we present how we integrate them within the FWI algorithm. While it is relatively basic for the model space filtering, the modification induced by the data space filtering requires more care. The essence of the method is to apply the filter to both observed and calculated data, prior to evaluating the distance between both. This defines a generic misfit function modification, which can be applied to the conventional least-squares distance as well as to any other misfit functions. We show how the corresponding gradient of the misfit function can be calculated through the conventional adjoint state strategy, with an additional linearized diffusion equation to solve at each iteration of the process. This constitutes the main mathematical result of this study. Numerical illustrations of the proposed methods are presented next. After presenting schematic model space and data space examples we consider a realistic synthetic FWI experiment based on the Marmousi II model. This study is conducted so as to avoid inverse crime, with observed data computed on a fine (5 m) grid model and corrupted by Gaussian noise, and an inversion performed on a coarse grid (25 m). We show how the use of NADF in the model space can speed-up the convergence of the FWI process and enhance model structure in its reconstruction. When applied in the data space, NADF enhances the low frequency content of the data, making it possible to stabilize the inversion when starting from crude initial model, in combination with misfit modifications techniques. The combination of model space and data space regularization provides a nice enhancement of conventional FWI techniques. A discussion and a conclusion are given to finalize the study.

2. Diffusion filters in image processing

This section is intended to provide a summary of NADF and their main mathematical properties. It does not contain new material compared with what can be found in the reference book by Weickert (1998): the aim is to provide an overview of the main results and properties of interest for our application to seismic imaging. We refer the interested reader to Weickert (1998) for more details and a wider perspective on this filtering technique in the context of image processing.

2.1. Linear diffusion filter and its equivalence with Gaussian smoothing

First we recall a result stating the equivalence between linear diffusion filtering and Gaussian smoothing. This reminder is useful to set up the scene and to introduce general concepts and notations.

Let us consider an image $f(x) \in \mathcal{C}_b\left(\mathbb{R}^2\right)$ where $\mathcal{C}_b\left(\mathbb{R}^2\right)$ denotes the space of bounded and continuous functions on $\mathbb{R}^2$. The solution of the linear diffusion equation

$$\left\{\begin{array}{rcl} \partial_t u - \Delta u &=& 0 \\ u(x,0) &=& f(x), \end{array}\right. \tag{1}$$

where Δ is the Laplacian operator, is

$$\left\{\begin{array}{rcll} u(x,t) &=& f(x) & \text{for } t=0, \\ u(x,t) &=& \left(K_{\sqrt{2t}} * f\right)(x) & \text{for } t>0, \end{array}\right. \tag{2}$$

where $*$ denotes the 2D spatial convolution operation and, for $\sigma \in \mathbb{R}^+$, K_σ is the Gaussian kernel

$$K_\sigma(x) = \frac{1}{2\pi\sigma^2} \exp\left(\frac{-|x|^2}{2\sigma^2}\right). \tag{3}$$

In (3), $|.|$ is the conventional Euclidean norm on $\mathbb{R}^2$. Based on (2) and (3), given a correlation length $L \in \mathbb{R}^+$, smoothing $f(x)$ with an (isotropic) Gaussian filter K_L is equivalent to solve the linear diffusion process (1) with a final diffusion time $T = \frac{1}{2}L^2$.

The uniqueness of the solution is guaranteed if the following constraint is imposed

$$\exists\, (M,a) \in \left(\mathbb{R}^+\right)^2, \quad \forall t \in \mathbb{R}^+, \quad |u(x,t)| < M \exp(a|x|^2). \tag{4}$$

In addition, the solution depends continuously on the initial state $f(x)$ for the norm $\|.\|_{L^\infty(\mathbb{R}^2)}$, and it satisfies the minimum-maximum principle

$$\forall x \in \mathbb{R}^2, \quad \forall t \in \mathbb{R}^+, \quad \inf_{\mathbb{R}^2} f \leq u(x,t) \leq \sup_{\mathbb{R}^2} f. \tag{5}$$

Another important result is what is referred to as the "scale-space" property of the linear diffusion filter. The scale-space concept is very important in image processing, and is also meaningful for the application of diffusion filters in seismic imaging. We briefly sketch its definition below.

Definition 1. *A scale-space representation of an image $f(x) \in L^\infty(\mathbb{R})$ embeds it into a family of gradually simplified versions of it. If we denote by $\mathcal{F}_\tau$ a family of filters indexed by $\tau \in \mathbb{R}^+$, its scale-space representation corresponds to the ensemble $\{\mathcal{F}_\tau f | \tau \geq 0\}$, where $\mathcal{F}_\tau$ needs to satisfy the properties described below.*

Property 1. *Recursivity: a family of filters $\{\mathcal{F}_\tau f | \tau \geq 0\}$ is recursive if and only if*

$$\begin{array}{lcl} \mathcal{F}_0 f & = & f, \\ \mathcal{F}_{\tau+s} f & = & \mathcal{F}_\tau(\mathcal{F}_s f), \quad (\tau, s) \in \left(\mathbb{R}^+\right)^2. \end{array} \tag{6}$$

Property 2. *Smoothing properties/information reduction: no additional structures are introduced at a coarser representation which do not correspond to structures at a finer scale.*

There are multiple choices for implementing this property mathematically: no creation of level curves, non-enhancement of local extrema, decreasing number of local extrema for instance. See Weickert (1998) for more details.

Property 3. *Invariance: an image is a representative of an equivalence class depicting the same object. Two images of this class might differ by gray-level shifts, translation, rotation or affine mappings. The scale-space representation should be invariant to these transformations to focus on the analysis of the depicted object.*

It can be shown that the linear diffusion filter generates a scale space representation of an image. Introducing

$$\begin{array}{rccc} \mathcal{F}_\tau : & f & \to & \mathcal{F}_\tau(f) := u(.,\tau) \\ & L^\infty(\Omega) & \mapsto & H^1(\Omega) \end{array} \tag{7}$$

such that $u(x,t)$ is the solution of (1), $\{\mathcal{F}_\tau f | \tau \geq 0\}$ is a scale-space representation of f (Weickert, 1998).

Despite this attractive property, the linear diffusion filter suffers from important limitations. The most obvious is the underlying isotropic smoothing effect which destroys the image structure. Edges can also be dislocated from coarse to fine representation. This has been the original motivation to improve over linear diffusion filters. It is not our purpose to review the different generalizations which have been proposed, but it appears that one of the most advanced strategy is NADF, at the core of this study.

2.2. Nonlinear anisotropic diffusion filters

2.2.1. Design Let us consider a rectangular domain $\Omega = [0, a_1] \times [0, a_2]$, its frontier $\Gamma := \partial\Omega$, and an image $f(x) \in L^\infty(\Omega)$. NADF are defined by the solution of the following PDE

$$\left\{ \begin{array}{rcll} \partial_t u - \operatorname{div}\left(D(u)\nabla u\right) & = & 0, & \text{on } \Omega \times [0, +\infty], \\ u(x,0) & = & f(x), & \text{on } \Omega, \\ \langle D(u)\nabla u, n\rangle & = & 0, & \text{on } \Gamma \times [0, +\infty]. \end{array} \right. \tag{8}$$

In (8), the operator $D(u)$ is a second-order tensor (a matrix) named as diffusion tensor: $D(u) \in M_2(\mathbb{R})$, where $M_n(\mathbb{R})$ is the space of square real matrices of size $n \in \mathbb{N}$. Were $D(u)$ be a scalar, the system (8) would be referred to as "isotropic" nonlinear diffusion equation. The use of a diffusion tensor makes it possible to control the diffusion rate depending on the direction, hence the reference to anisotropy here. The nonlinearity of the tensor (dependence on $u(x,t)$) makes it possible

to adapt locally (in space) and progressively (in time) the tensor during the diffusion process.

Let us consider the structure tensor $S(u)$, defined by

$$S(u) \in M_2(\mathbb{R}), \quad S(u) := \nabla u \nabla u^T = \begin{pmatrix} |\partial_{x_1} u|^2 & (\partial_{x_1} u)(\partial_{x_2} u) \\ (\partial_{x_1} u)(\partial_{x_2} u) & |\partial_{x_2} u|^2 \end{pmatrix}. \tag{9}$$

The eigenvectors of $S(u)$ are respectively parallel and perpendicular to ∇u. The corresponding eigenvalues are $|\nabla u|^2$ and 0: they give the contrast in the corresponding eigendirections.

The information embedded in $S(u)$ is strictly local. In order to exploit it, it should be averaged over specific spatial scales $(\sigma, \rho) \in (\mathbb{R}^+)^2$. We thus consider

$$J_\rho(u) = K_\rho *. \left(\nabla u_\sigma \nabla u_\sigma^T\right), \quad u_\sigma := K_\sigma * u, \tag{10}$$

where $*.$ denotes a component-wise convolution. As mentioned in Weickert (1998), the pre-smoothing by K_σ removes oscillations smaller than $O(\sigma)$, and the parameter σ is referred to as the "noise scale". In turn, the scale ρ should reflect the characteristic window size over which the orientation is to be analyzed. It is referred to as the "integration scale".

Let us denote μ_1 and μ_2 the eigenvalues of $J_\rho(u_\sigma)$, such that $\mu_1 \geq \mu_2 \geq 0$. Thanks to the smoothing by K_ρ, they describe the average local contrast in the eigendirections. Denoting by v_1 and v_2 the corresponding eigenvectors, we observe that v_1 is the orientation with the highest value fluctuations, while v_2 corresponds to the direction with the smallest variations. In the following v_2 is thus referred to as the "coherence" direction. In the light of this interpretation, the eigenvalues μ_1 and μ_2 can be used to describe locally the image structure:

- $\mu_1 = \mu_2 = 0$ correspond to zones with constant values (no variations);
- $\mu_1 \gg \mu_2 = 0$ correspond to straight edges;
- $\mu_1 \geq \mu_2 \gg 0$ correspond to corners.

Finally, $(\mu_1 - \mu_2)^2$ is a measure of local coherence, which becomes large for anisotropic structures.

The diffusion tensor $D(u)$ used in the filter is computed from $J_\rho(u)$. Considering the eigenvalue decomposition

$$J_\rho(u) = P^T(u)\Sigma(u)P(u), \tag{11}$$

with

$$P(u) = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \quad \Sigma(u) = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix}, \tag{12}$$

the generic form for the diffusion tensor is

$$D(u) = P^T(u)\Lambda(u)P(u). \tag{13}$$

Depending on the choice of $\Lambda(u)$, different types of filter can be designed. In this study, we are interested in one instance of this filter, namely the coherence enhancing filter. For this filter, we have

$$\Lambda(u) = \begin{pmatrix} \alpha & 0 \\ 0 & h(\mu_1, \mu_2) \end{pmatrix}, \tag{14}$$

where $\alpha \in \mathbb{R}_*^+$ is a small constant and

$$h(\mu_1, \mu_2) = \begin{cases} \alpha \text{ if } \mu_1 = \mu_2, \\ \alpha + (1-\alpha)\exp\left(\frac{-C}{(\mu_1 - \mu_2)^{2m}}\right) \text{ else.} \end{cases} \tag{15}$$

In (15) $C \in \mathbb{R}_*^+$ and $m \in \mathbb{N}^*$ are additional constants designing the filter. In this study we set $\alpha = 10^{-5}, C = 10^{-8}$ and $m = 1$. We see that the function $h(\mu_1, \mu_2)$ increases rapidly to 1 as soon as the measure of local coherence $(\mu_1 - \mu_2)^2$ departs from 0. We can thus interpret the filter design as follows: as soon as a feature with local coherence is detected, the diffusion rate increases in the coherence direction v_2, while it remains weak in the orthogonal direction v_1. When the coherence is weak, the diffusion gets back to a small isotropic diffusion.

2.2.2. Well posedness and minimum-maximum principle The following result gives guarantees in terms of robustness of the NADF. The problem (8) can be shown to be well posed and obey minimum-maximum principle (under certain conditions to be satisfied by the diffusion tensor).

Theorem 1. *Let $C\left(0,\tau,L^2(\Omega)\right)$ be the space of continuous functions from $[0,\tau]$ to $L^2(\Omega)$, and $L^2(0,\tau,H^1(\Omega)$ the space of strongly measurable function from $[0,\tau]$ to $H^1(\Omega)$, with $H^1(\Omega)$ the conventional Sobolev space. We rewrite $D(u)$ as $D(u)=M\left(J_\rho(u)\right)$.*

For $M\in C^\infty\left(M_2\left(\mathbb{R}\right),M_2\left(\mathbb{R}\right)\right)$, such that $M(J)$ is symmetric for any symmetric matrix $J\in M_2\left(\mathbb{R}\right)$, and such that

$$\forall w\in L^\infty\left(\Omega,\mathbb{R}^2\right),\ \ |w(x)|<K,\ \ x\in\overline{\Omega}, \tag{16}$$

there exists a positive lower bound $\nu(K)$ for the eigenvalues of $D\left(J_\rho(w)\right)$.

Then, the solution of the system (8) exists and is unique, which satisfies

$$u\in C\left(0,\tau:L^2(\Omega)\right)\cap L^2(0,\tau;H^1(\Omega)),\ \ \partial_t u\in L^2(0,T;H^1(\Omega)). \tag{17}$$

In addition, $u\in C^\infty\left(\overline{\Omega}\times[0,\tau]\right)$, it depends continuously on f with respect to $\|.\|_{L^2(\Omega)}$, and it satisfies

$$\operatorname*{ess\,inf}_{\Omega} f\le u(x,t)\le \operatorname*{ess\,sup}_{\Omega} f \tag{18}$$

We refer the reader to Weickert (1998) and references therein for a complete proof of Theorem 1. The important point here is that by choosing $D(u)$ as a coherence enhancing diffusion operator such as the one described in the previous paragraph, the conditions of application of the theorem are satisfied.

2.2.3. Scale-space properties Besides well-posedness and minimum-maximum principle, the nonlinear anisotropic diffusion process benefits also from scale-space properties. As previously, we denote the filter family by $\mathcal{F}_\tau,\tau\ \in\mathbb{R}^+$, where

$$\begin{array}{rccc}\mathcal{F}_\tau: & f & \to & \mathcal{F}_\tau(f):=u(.,\tau)\\ & L^\infty(\Omega) & \mapsto & H^1(\Omega)\end{array} \tag{19}$$

such that $u(x,\tau)$ is now the solution of (8).

NADF satisfy the recursivity property (6). They also satisfy the following invariance properties

- constant shift invariance: $\mathcal{F}_\tau(0)=0,\ \ \forall C\in\mathbb{R},\ \ \mathcal{F}_\tau(f+C)=\mathcal{F}_\tau(f)+C$;
- reverse contrast invariance: $\mathcal{F}_\tau(-f)=-\mathcal{F}_\tau(f)$;
- conservation of average gray value: let $\mu:=\frac{1}{|\Omega|}\int_\Omega f(x)dx$ be the average gray value. Then we have
$$\frac{1}{|\Omega|}\int_\Omega \mathcal{F}_\tau(f)(x)dx=\mu; \tag{20}$$
- translation invariance: for $h\in\Omega$, define $(\tau_h f)(x)=f(x+h)$, then
$$\mathcal{F}_\tau(\tau_h f)=\tau_h(\mathcal{F}_\tau(f)); \tag{21}$$
- isometry invariance: for $R\in O_2\left(\mathbb{R}\right)$, where $O_2\left(\mathbb{R}\right)$ is the set of orthogonal matrices of rank 2, we have
$$\mathcal{F}_\tau(Rf)=R\mathcal{F}_\tau(f). \tag{22}$$

In terms of information reduction, it can be shown that NADF do not create new level curves, making it possible to trace back a structure from coarse scale to fine scale. A sufficient condition for this is that local extrema are not enhanced by the filter. It is proved in Weickert (1998) that this condition is satisfied, more particularly

Theorem 2. *Let u be a solution of (8). Let $\theta \in \mathbb{R}^+$ and $\xi \in \Omega$ a local extremum of $u(.,\theta)$ with non vanishing Hessian. Then*

$$\begin{cases} \partial_t u(\xi,\theta) < 0 & \text{if } \xi \text{ is a local maximum,} \\ \partial_t u(\xi,\theta) > 0 & \text{if } \xi \text{ is a local minimum.} \end{cases} \tag{23}$$

Finally, important long-term behavior properties of the filter can be demonstrated. In particular, the following theorem is proved in Weickert (1998).

Theorem 3. *For $u(x,t)$ solution of (8), the following functions are decreasing with $t \in [0,+\infty]$*

(i) $\|u(.,t)\|_{L^p(\Omega)}, \quad \forall p \geq 2;$

(ii) $\dfrac{1}{|\Omega|}\displaystyle\int_\Omega (u(x,t)-\mu)^{2n} dx, \quad n \in \mathbb{N}$, *with μ the average gray value;*

(iii) $\displaystyle\int_\Omega u(x,t)\ln u(x,t) dx, \quad$ *if* $\operatorname{ess\,inf}_\Omega f > 0.$

In short, these properties ensure that

(i) the energy of the solution is decreasing in time;

(ii) the solution converges to a constant in space solution equal to the average value μ of the initial data $f(x)$;

(iii) the entropy of the solution also decreases with time.

There properties are important in the perspective of application to seismic imaging, whether the filter is applied in the model space or in the data space. In particular, they bring the required robustness guarantees to apply these filters in a repeated way, on a variety of models and data, as will be necessary from the application of these filters in the context of FWI.

2.3. Discretization and implementation

In Weickert (1998), the derivation of a class of finite-difference discretization schemes which preserve all the above discussed properties at the discrete level is proposed. In particular, the existence and uniqueness of the solutions are guaranteed, as well as the scale-space properties (invariances, conservation of average value, long-term behavior and information reduction). It is shown that these properties are ensured as soon as the operator $\operatorname{div}(D(u)\nabla u)$ can be represented by a non-negative matrix. While it is straightforward in the isotropic case, it is more challenging due to the anisotropy matrix $D(u)$. A general result states how such a second-order accurate non-negative discretization can be found. An example with a 3x3 stencil is provided, which guarantees a non-negative discretization for spectral ratio (ratio between the largest and the smallest eigenvalue) smaller than a constant close to 6. This is the discretization we use in this study.

The coefficients of this scheme are obtained as follows. Let $x = (x_1, x_2) \in \Omega$. We consider a 2D spatial discretization of $u(x,t)$ with spatial step h_1 (respectively h_2) in the direction x_1 (respectively h_2). For a time step $t^n = n\Delta t$, we use the notation

$$u_{ij}^n := u(x_{1i}, x_{2j}, t^n) \tag{24}$$

For each i, j, n, the corresponding diffusion matrix $D(u_{ij}^n)$ is denoted by

$$D(u_{ij}^n) := \begin{pmatrix} a_{ij}^n & b_{ij}^n \\ c_{ij}^n & d_{ij}^n \end{pmatrix}. \tag{25}$$

There exists a mapping

$$\begin{array}{rccc} p: & (i,j) & \longrightarrow & p(i,j) \\ & \mathbb{N}\times\mathbb{N} & \longrightarrow & \mathbb{N} \end{array} \tag{26}$$

which makes it possible to go from a matrix representation of an image to a vector representation of it. The second-order finite-difference discretization of $\operatorname{div}(D(u^n)\nabla u^n)$ into $A(u^n)$ following the

3x3 stencil presented in (Weickert, 1998) obeys

$$\begin{cases}
A(u^n)_{p(i,j),p(i-1,j-1)} &= \frac{|b_{i-1,j-1}|+b_{i-1,j-1}}{4h_1h_2} + \frac{|b_{i,j}|+b_{i,j}}{4h_1h_2} \\
A(u^n)_{p(i,j),p(i-1,j)} &= \frac{a_{i-1,j}+a_{i,j}}{2h_1^2} - \frac{|b_{i-1,j}|+|b_{i,j}|}{2h_1h_2} \\
A(u^n)_{p(i,j),p(i-1,j+1)} &= \frac{|b_{i-1,j+1}|-b_{i-1,j+1}}{4h_1h_2} + \frac{|b_{i,j}|-b_{i,j}}{4h_1h_2} \\
A(u^n)_{p(i,j),p(i,j-1)} &= \frac{c_{i,j-1}|+c_{i,j}}{2h_2^2} - \frac{|b_{i,j-1}|+|b_{i,j}|}{2h_1h_2} \\
A(u^n)_{p(i,j),p(i,j)} &= -\frac{a_{i-1,j}+2a_{ij}+a_{i+1,j}}{2h_1^2} \\
&\quad -\frac{|b_{i-1,j+1}|-b_{i-1,j+1}+|b_{i+1,j+1}|+b_{i+1,j+1}}{4h_1h_2} \\
&\quad -\frac{|b_{i-1,j-1}|+b_{i-1,j-1}+|b_{i+1,j-1}|-b_{i+1,j-1}}{4h_1h_2} \\
&\quad +\frac{|b_{i-1,j}|+b_{i+1,j}|+b_{i,j-1}|+b_{i,j+1}|+2b_{i,j}|}{2h_1h_2} \\
&\quad -\frac{c_{i,j-1}+2c_{ij}+c_{i,j+1}}{2h_2^2} \\
A(u^n)_{p(i,j),p(i,j+1)} &= \frac{c_{i,j+1}|+c_{i,j}}{2h_2^2} - \frac{|b_{i,j+1}|+|b_{i,j}|}{2h_1h_2} \\
A(u^n)_{p(i,j),p(i+1,j-1)} &= \frac{|b_{i+1,j-1}|-b_{i+1,j-1}}{4h_1h_2} + \frac{|b_{i,j}|-b_{i,j}}{4h_1h_2} \\
A(u^n)_{p(i,j),p(i+1,j)} &= \frac{a_{i+1,j}+a_{i,j}}{2h_1^2} - \frac{|b_{i+1,j}|+|b_{i,j}|}{2h_1h_2} \\
A(u^n)_{p(i,j),p(i+1,j+1)} &= \frac{|b_{i+1,j+1}|+b_{i+1,j+1}}{4h_1h_2} + \frac{|b_{i,j}|+b_{i,j}}{4h_1h_2}
\end{cases} \tag{27}$$

The preservation of the continuous properties at the discrete level also depends on the time discretization scheme which is employed. A general second-order weighted semi-implicit time scheme is proposed in Weickert (1998) with a criterion on the time step Δt to be satisfied to preserve the properties of the filter. In this study, we exploit this scheme in the limit of a fully explicit time scheme to avoid any matrix inversion and keep a numerically efficient implementation. The resulting scheme can be written as

$$u^{n+1} = u^n + \Delta t A(u^n)u^n, \tag{28}$$

where Δt satisfies

$$\Delta t < \frac{1}{\max_i |A_{ii}(u^n)|}. \tag{29}$$

In (29), $A_{ii}(u^n)$ denotes the ith diagonal entry of $A(u^n)$. Thanks to the extremum principle satisfied by the scheme, a prior bound on $|A_{ii}(u^n)|$ can be obtained from the initial state $u(x,0) = f(x)$ (see Weickert, 1998, remark p. 105).

3. Application to full waveform inversion

3.1. Conventional full waveform inversion

Now we introduce the FWI problem, starting with the observed seismic data. Such data is generated by the recording of mechanical waves triggered by a seismic source, which can be a natural source of energy such as earthquakes or volcanoes (global and regional scale imaging), or a controlled source such as an airgun or a vibrating truck (exploration scale and near-surface scale imaging). In the context of marine acquisition, the receivers can be deployed in the sea along cables towed by a boat (streamer acquisition) or at the sea bottom (node acquisition). For land acquisition, the receivers are generally deployed at the Earth's surface. Depending on the context, the receivers record the pressure variation (hydrophones) and/or the displacement, velocities, acceleration components along different directions (geophones, nodes). In the following, such observed data will be denoted by

$$d_{obs,s}(x_r,t) \in \mathcal{L}^2(\Sigma_r \times [0,T]), \quad s = 1,\dots,N_s, \tag{30}$$

where $\Sigma_r \subset \mathbb{R}^{d-1}$ is the part of the surface on which the receivers are deployed, d is the dimension ($d = 2$ or $d = 3$), $T \in \mathbb{R}_*^+$ denotes the recording time and $N_s \in \mathbb{N}$ is the number of seismic sources.

The calculated data, which are to be compared with the observed data, are obtained through the modeling of mechanical waves within the subsurface. Such waves are modeled following the

linear elasticity approximation, which considers the propagation of pressure waves (P-waves), shear waves (S-waves), and surface waves (Rayleigh and Love waves). In specific contexts, such as marine acquisition data, it is however possible to focus only on the propagation of P-waves under the acoustic approximation. In the following we introduce a general wave propagation operator $A(m)$ such that the wave equation we consider is denoted by

$$A(m)w_s = b_s, \tag{31}$$

where $m(x) \in \mathcal{L}^2(\Omega)$ represents the subsurface mechanical parameters with $\Omega \subset \mathbb{R}^d$, $w_s(x,t) \in \mathcal{L}^2(\Omega \times [0,T])$ is the wavefield solution of this wave equation and $b_s(x,t) \in \mathcal{L}^2(\Omega \times [0,T])$ represents the seismic source term. In the following $m(x)$ will be referred to as the model parameter.

The calculated data $d_{cal,s}[m](x_r,t) \in \mathcal{L}^2(\Sigma_r \times [0,T])$ is defined for all $x_r \in \Sigma_r$ as

$$d_{cal,s}[m](x_r,t) = w_s[m](x_r,t), \tag{32}$$

where the bracket $[m]$ is a reminder of the dependency of $d_{cal,s}$ and w_s to the model parameter $m(x)$. In the following, we use a restriction operator R to denote the relationship between $d_{cal,s}$ and w_s, such that

$$\begin{array}{rccc} R: & w_s & \longrightarrow & Rw_s := d_{cal,s}. \\ & \mathcal{L}^2(\Omega \times [0,T]) & \longrightarrow & \mathcal{L}^2(\Sigma_r \times [0,T]). \end{array} \tag{33}$$

The operator R acts as a restriction of the wavefield space to the data space.

The general formulation for FWI is

$$\min_{m \in \mathcal{M}} f(m), \tag{34}$$

with $\mathcal{M}$ a general model space and

$$f(m) = \sum_{s=1}^{N_s} G(d_{cal,s}[m], d_{obs,s}), \tag{35}$$

where $G(.,.)$ is a positive function measuring the misfit between $d_{cal,s}$ and $d_{obs,s}$

$$\begin{array}{rccc} G: & (d_1,d_2) & \longrightarrow & G(d_1,d_2) \\ & \mathcal{L}^2(\Sigma_r \times [0,T]) \times \mathcal{L}^2(\Sigma_r \times [0,T]) & \longrightarrow & \mathbb{R}^+. \end{array} \tag{36}$$

The conventional choice for G is the least-squares misfit, such that

$$G(d_1,d_2) = \frac{1}{2} \int_{\Sigma_r} \int_0^T |d_1(x_r,t) - d_2(x_r,t)|^2 dx_r dt. \tag{37}$$

The solution of (34) is performed using local optimization methods. Starting from a model m_0, such method builds a sequence

$$m_{k+1} = m_k + \alpha_k \Delta m_k, \tag{38}$$

where $\alpha_k \in \mathbb{R}_*^+$ is a scaling parameter computed by linesearch, and Δm_k is a descent direction. In practice, we rely on quasi-Newton strategies, for which we have

$$\Delta m_k = -Q_k \nabla f(m_k), \tag{39}$$

where $\nabla f(m_k)$ is the gradient of the function $f(m)$ at m_k and Q_k is an approximation of the inverse Hessian of $f(m)$ at m_k denoted by $H(m_k)^{-1}$

$$Q_k \simeq H(m_k)^{-1} := \left(\nabla^2 f(m_k)\right)^{-1}. \tag{40}$$

We use the l-BFGS strategy to compute Q_k. It builds a low-rank approximation of the inverse Hessian from gradients computed during the l-previous iterations (Nocedal, 1980; Nocedal and

Wright, 2006).

From (38), (39) and (40), we see that implementing a FWI algorithm requires the ability to compute $f(m)$ and its gradient $\nabla f(m)$. The adjoint state strategy is usually employed (Plessix, 2006). Following this method, the gradient of the misfit function $f(m)$ is obtained as

$$\nabla f(m) = \sum_{s=1}^{N_s} \int_0^T \left(\frac{\partial A(m)}{\partial m} w_s[m] \right) (x,t) \lambda_s[m](x,t) dt, \tag{41}$$

where $\lambda_s[m]$ is the wavefield solution of the adjoint equation

$$A(m)^T \lambda_s = R^T \frac{\partial G}{d_{cal,s}}(d_{cal,s}, d_{obs,s}). \tag{42}$$

This well-known result has been derived in several studies; see for instance Métivier et al. (2016, 2019).

Equations (41) and (42) have a physical interpretation. The adjoint operator of the wave equation with an initial condition is the same wave equation with a final condition. Therefore the adjoint wavefield $\lambda_s(x,t)$ is computed by a reverse propagation in time of the source term $R^T \frac{\partial G}{d_{cal,s}}(d_{cal,s}, d_{obs,s})$ (the adjoint source). The adjoint of the restriction operator R^T acts as a lift from the data space to the wavefield space, yielding an adjoint source localized at the receiver positions.

In case of the least-squares misfit measurement (37), the adjoint source is simply

$$\frac{\partial G}{d_{cal,s}}(d_{cal,s}, d_{obs,s}) = d_{cal,s} - d_{obs,s}, \tag{43}$$

which is the difference between calculated and observed data, also known as the residual. A deeper physical interpretation of the gradient in FWI is provided in Operto et al. (2013) and Virieux et al. (2017) for instance.

3.2. Application of NADF to full waveform inversion

3.2.1. Model space regularization In this study, we consider two (possibly complementary) ways of applying NADF in the context of FWI. The first is the more straightforward. It consists in using it as a regularization/smoothing in the model space. This can be formalized as solving the problem

$$\min_{m \in \mathcal{M} \cap \mathcal{S}} f(m) \tag{44}$$

where $\mathcal{S}$ is the space of smooth models obtained after NADF is applied.

We implement it using the following strategy: the descent direction in the local optimization process is computed as

$$\Delta m_k = -Q_k \mathcal{F}_\tau(\nabla f(m_k)) \tag{45}$$

for a given diffusion time $\tau \in \mathbb{R}_*^+$ defined prior to the inversion. The descent direction is thus built from a filtered version of the gradient, and the model estimate m_k is built as the sum of filtered gradient directions, scaled by the matrix Q_k. The latter is an approximation of the inverse Hessian matrix built using l previous filtered gradient $\mathcal{F}_\tau(\nabla f(m_{k-l+1})), \ldots, \mathcal{F}_\tau(\nabla f(m_k))$.

We present in the numerical experiments in Section 4 how this gradient filtering strategy can simultaneously remove small scale oscillations due to the presence of noise in the data and incomplete illumination, and enhance the inherent structure of the subsurface model properties. Contrary to structure oriented smoothing such as the Bessel's smoothing filter presented in Trinh et al. (2017b), the information on the structure *does not need to be provided as a prior information.* The interest of the NADF is their ability to extract the structure by themselves, directly from the gradient.

3.2.2. Data space regularization The second strategy we consider consists in applying NADF to observed and synthetic data prior to their comparison through a given misfit function $G(d_{cal}, d_{obs})$ as in equation (36). The resulting FWI strategy can thus be formulated as

$$\begin{aligned} &\min_m f_\tau(m), \\ &f_\tau(m) := \sum_{s=1}^{N_s} G_\tau\left(d_{cal,s}[m], d_{obs,s}\right), \quad \tau \in \mathbb{R}_*^+. \end{aligned} \tag{46}$$

where

$$G_\tau(d_1, d_2) = G\left(\mathcal{F}_\tau(d_1), \mathcal{F}_\tau(d_2)\right). \tag{47}$$

Note that such kind of formulation has already been proposed by Xue et al. (2016), however with Gaussian convolution kernels playing the role of $\mathcal{F}_\tau$, *i.e.* linear diffusion filters instead of NADF. As has already been mentioned, such linear diffusion filters have interesting properties, but they are equivalent to low-pass filters. For seismic data, noise contaminates the data, which reduces the range of applicability of such filters to the frequency-band where the signal to noise ratio remains acceptable. As will be illustrated in the numerical section, NADF make it possible a low frequency enhancement by combining a low-pass filter and denoising properties.

Following the natural hierarchy associated with scale-space properties of the filter $\mathcal{F}_\tau$ one can then formulate a sequence of FWI problems with decreasing values of τ. This can be formalized introducing the operator

$$\begin{array}{rccc} \mathcal{I}: & \tau, m_0 & \to & \mathcal{I}(\tau, m_0) := \arg\min_m f_\tau(m), \text{ starting from } m_0 \\ & \mathbb{R}^+ \cup L^2(\Omega) & \mapsto & L^2(\Omega). \end{array} \tag{48}$$

The operator $\mathcal{I}(\tau, m_0)$ consists in solving the FWI problem (46) starting from the model m_0, given a local optimization scheme, such as steepest descent, nonlinear conjugate gradient, l-BFGS, or truncated Newton method strategy (Nocedal and Wright, 2006). In this study we rely on l-BFGS (Nocedal, 1980). The hierarchical (*i.e.* multi-scale) approach based on scale-space properties thus writes as follows. Starting from m_0, build a sequence m_p such that

$$m_{p+1} = \mathcal{I}(\tau_p, m_p) \tag{49}$$

with

$$\lim_{p \to +\infty} \tau_p = 0. \tag{50}$$

We shall note at this stage that the convergence of the strategy described from equations (48) to (50) depends on the assumption that no "bifurcation" occurs in the process, that is the solution of the minimization problem through the operator I is unique and depends continuously on the parameter τ. In practice, we do not have such guarantees. A sufficient condition is the non-singularity of the Hessian operator of the misfit function $f_\tau(m)$ for any τ but this is not easy to prove. Yet, in practice, we assume that this "no-bifurcation" assumption is satisfied, and that the steps in the decreasing sequence τ_p are sufficiently small to guarantee the stability of the process.

We basically have all the ingredients to implement this strategy except one. As mentioned in Section 3.1, we need to be able to compute the gradient of the misfit function $f_\tau(m)$. To this end, we prove the following theorem.

Theorem 4. *We introduce $X = (x_r, t) \in \mathcal{X} := \Sigma_r \times [0, T]$. For a given $\tau^* \in \mathbb{R}^+$, we define*

$$\mathcal{F}_{\tau^*}(d_{cal,s}) = \overline{u}_s(., \tau^*) \tag{51}$$

where $\overline{u}_s(X, \tau)$ is the solution of the system

$$\begin{cases} \partial_\tau u - \operatorname{div}\left(D(u)\nabla u\right) = 0, \quad on \;\; \mathcal{X} \times [0, \tau^*], \\ u(., 0) = d_{cal,s}, \quad on \;\; \mathcal{X}, \\ \langle D(u)\nabla u, n\rangle = 0, \quad on \;\; \partial\mathcal{X} \times [0, \tau^*]. \end{cases} \tag{52}$$

We also introduce $g_{1,s}$ *and* $g_{2,s}$ *as follows*

$$\begin{cases} g_{1,s} := \mathcal{F}_{\tau^*}(d_{cal,s}) \\ \\ g_{2,s} := \mathcal{F}_{\tau^*}(d_{obs,s}). \end{cases} \tag{53}$$

We have

$$\nabla f_{\tau^*}(m) = \sum_{s=1}^{N_s} \int_0^T \left(\frac{\partial A(m)}{\partial m} w_s[m] \right)(x,t) \lambda_s[m](x,t) dt, \tag{54}$$

where

$$A(m)^T \lambda_s = R^T \mu_s(x_r, t), \tag{55}$$

with $\mu_s(x_r,t)$ *defined as*

$$\mu_s(x_r,t) := \mu_s(X) := -\overline{\lambda}_{2,s}(X, \tau^*), \tag{56}$$

where $\overline{\lambda}_{2,s}$ *is the solution of the following linearized diffusion equation*

$$\begin{cases} \partial_\tau \lambda_2 - \operatorname{div}\left((D(\overline{u}_s) + M(\overline{u}_s)) \right) \nabla \lambda_2 = 0, \quad on \;\; \mathcal{X} \times [0, \tau^*] \\ \\ \lambda_2(.,0) = \dfrac{\partial G\left(g_{1,s}, g_{2,s}\right)}{\partial g_{1,s}}, \quad on \;\; \mathcal{X} \\ \\ \left\langle \left(D(\overline{u}_s) + M(\overline{u}_s)^T \right) \nabla \lambda_2, n \right\rangle = 0, \quad on \;\; \partial\mathcal{X} \times [0, \tau^*], \end{cases} \tag{57}$$

with $M(u)$ *the matrix such that*

$$\forall (u,v), \quad M(u)\nabla v = D'(u).v \nabla u. \tag{58}$$

Proof. According to the adjoint state strategy applied to the FWI problem, the gradient of $f_{\tau^*}(m)$ is given by equation (41) and (42), replacing $G(d_1,d_2)$ by $G_{\tau^*}(d_1,d_2)$, where G_{τ^*} is defined as in (47), *i.e.* the composition of $G(.,.)$ with the filter $\mathcal{F}_{\tau^*}$. The quantity we need to compute is thus the adjoint source associated with this new misfit function, which is given by $\frac{\partial G_{\tau^*}}{\partial d_1}$.

For the sake of simplicity we drop the index 1 for d_1 in what follows. We consider the case $N_s = 1$ as well, and drop the index s. To compute this partial derivative, we rely on the adjoint state strategy. Indeed, the functional $G_{\tau^*}(d,d_2)$ can be associated with a PDE-constrained problem, with the following Lagrangian functional

$$\begin{aligned} L(d,g,u,\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \;& G(g,g_2) + \\ & (g - u(.,\tau^*), \lambda_1)_{\mathcal{X}} + \\ & \left(\partial_t u - \operatorname{div}\left(D(u)\nabla u\right), \lambda_2\right)_{\mathcal{X}\times[0,\tau^*]} + \\ & (u(.,0) - d, \lambda_3)_{\mathcal{X}} + \\ & \left(\langle D(u)\nabla u, n\rangle, \lambda_4\right)_{\partial\mathcal{X}\times[0,\tau^*]}, \end{aligned} \tag{59}$$

where $(.,.)_{\mathcal{X}}, (.,.)_{\mathcal{X}\times[0,\tau^*]}, (.,.)_{\partial\mathcal{X}\times[0,\tau^*]}$ denote the conventional L^2 scalar product on the spaces in subscript, and $\lambda_i, \;\; i = 1, \ldots 4$ are adjoint variables associated with the constraints (52) and (53).

Let $\overline{u}[d](X,\tau)$ and $\overline{g}[d](X)$ solution of (52) and (53) (where the brackets are used as a reminder of the dependency to the initial condition d). We have

$$L(d, \overline{g}[d], \overline{u}[d], \lambda_1, \lambda_2, \lambda_3, \lambda_4) = G_{\tau^*}(d, d_2). \tag{60}$$

Therefore in this case, we can write (formally) that

$$\partial_d \left(L(d, \overline{g}[d], \overline{u}[d], \lambda_1, \lambda_2, \lambda_3, \lambda_4) \right) = \partial_d G_{\tau^*}(d, d_2). \tag{61}$$

The right hand side of (61) is the quantity we need to evaluate.

Developing the left hand side of (61) yields (again formally)

$$\frac{\partial L(d, \overline{g}, \overline{u}, \lambda_1, \lambda_2, \lambda_3)}{\partial d} + \frac{\partial L(d, \overline{g}, \overline{u}, \lambda_1, \lambda_2, \lambda_3)}{\partial g} \frac{\partial \overline{g}}{\partial d} + \frac{\partial L(d, \overline{g}, \overline{u}, \lambda_1, \lambda_2, \lambda_3)}{\partial u} \frac{\partial \overline{u}}{\partial d} = \partial_d G_{\tau^*}(d, d_2) \tag{62}$$

The essence of the adjoint state strategy is to determine $\overline{\lambda}_i[d],\ i = 1, \ldots, 4$, such that

$$\frac{\partial L(d, \overline{g}, \overline{u}, \overline{\lambda}_1, \overline{\lambda}_2, \overline{\lambda}_3, \overline{\lambda}_4)}{\partial g} = 0, \tag{63a}$$

$$\frac{\partial L(d, \overline{g}, \overline{u}, \overline{\lambda}_1, \overline{\lambda}_2, \overline{\lambda}_3, \overline{\lambda}_4)}{\partial u} = 0. \tag{63b}$$

In this case, we have

$$\partial_d G_{\tau^*}(d, d_2) = \frac{\partial L(d, \overline{g}, \overline{u}, \overline{\lambda}_1, \overline{\lambda}_2, \overline{\lambda}_3, \overline{\lambda}_4))}{\partial d} = -\overline{\lambda}_3[d]. \tag{64}$$

In the remainder of the proof, we develop (63a) and (63b) to explicit $\overline{\lambda}_i[d],\ i = 1, \ldots, 4$.

We start with (63a). We have

$$\frac{\partial L(d, \overline{g}, \overline{u}, \lambda_1, \lambda_2, \lambda_3)}{\partial g} = \frac{\partial G(\overline{g}, g_2)}{\partial g} + \lambda_1. \tag{65}$$

Therefore

$$\overline{\lambda}_1 = -\frac{\partial G(\overline{g}, g_2)}{\partial g}. \tag{66}$$

The development of (63b) is more involved. Consider v an increment of u. We have

$$\begin{aligned} L(d, g, u + v, \lambda_1, \lambda_2, \lambda_3) = \quad & G(g, g_2) + \\ & (g - (u + v)(., \tau^*), \lambda_1)_{\mathcal{X}} + \\ & (\partial_t (u + v) - \operatorname{div} \left(D(u + v) \nabla (u + v) \right), \lambda_2)_{\mathcal{X} \times [0, \tau^*]} + \\ & ((u + v)(., 0), \lambda_3)_{\mathcal{X}} + \\ & (\langle D(u + v) \nabla (u + v), n \rangle, \lambda_4)_{\partial \mathcal{X} \times [0, \tau^*]}. \end{aligned} \tag{67}$$

We focus on the nonlinear terms. First we have

$$\begin{aligned} (\operatorname{div} \left(D(u + v) \nabla (u + v) \right), \lambda_2)_{\mathcal{X} \times [0, \tau^*]} = \quad & (\operatorname{div} \left(D(u) \nabla u \right), \lambda_2)_{\mathcal{X} \times [0, \tau^*]} + \\ & (\operatorname{div} \left(D(u) \nabla v \right), \lambda_2)_{\mathcal{X} \times [0, \tau^*]} + \\ & (\operatorname{div} \left(D'(u).v \nabla u \right), \lambda_2)_{\mathcal{X} \times [0, \tau^*]} + O(\|v\|^2). \end{aligned} \tag{68}$$

Second we have

$$\begin{aligned} (\langle D(u + v) \nabla (u + v), n \rangle, \lambda_4)_{\partial \mathcal{X} \times [0, \tau^*]} = \quad & (\langle D(u) \nabla (u), n \rangle, \lambda_4)_{\partial \mathcal{X} \times [0, \tau^*]} + \\ & (\langle D(u) \nabla (v), n \rangle, \lambda_4)_{\partial \mathcal{X} \times [0, \tau^*]} + \\ & (\langle D'(u).v \nabla u, n \rangle, \lambda_4)_{\partial \mathcal{X} \times [0, \tau^*]} + O(\|v\|^2). \end{aligned} \tag{69}$$

Therefore

$$\begin{aligned}\frac{\partial L(d,g,u,\lambda_1,\lambda_2,\lambda_3,\lambda_4)}{\partial u}.v = \quad & (v(.,\tau^*),\lambda_1)_{\mathcal{X}} + (\partial_t v,\lambda_2)_{\mathcal{X}\times[0,\tau^*]} + (v(.,0),\lambda_3)_{\mathcal{X}} - \\ & (\operatorname{div}(D(u)\nabla v) + \operatorname{div}(D'(u).v\nabla u),\lambda_2)_{\mathcal{X}\times[0,\tau^*]} + \\ & (\langle D(u)\nabla v + D'(u).v\nabla u, n\rangle,\lambda_4)_{\partial\mathcal{X}\times[0,\tau^*]}.\end{aligned} \tag{70}$$

By two integration by parts, and using the symmetry of $D(u)$, we have

$$\begin{aligned}&(\operatorname{div}(D(u)\nabla v),\lambda_2)_{\mathcal{X}\times[0,\tau^*]} = \\ &(v,\operatorname{div}(D(u)\nabla\lambda_2))_{\mathcal{X}\times[0,\tau^*]} + (\langle D(u)\nabla v,n\rangle,\lambda_2)_{\partial\mathcal{X}\times[0,\tau^*]} - (v,\langle D(u)\nabla\lambda_2,n\rangle)_{\partial\mathcal{X}\times[0,\tau^*]}.\end{aligned} \tag{71}$$

In addition, we have

$$(\operatorname{div}(D'(u)v\nabla u),\lambda_2)_{\mathcal{X}\times[0,\tau^*]} = (\operatorname{div}(M(u)\nabla v),\lambda_2)_{\mathcal{X}\times[0,\tau^*]}, \tag{72}$$

where we do not explicit the matrix $M(u)$. Thus we have, again after two integration by part,

$$\begin{aligned}&(\operatorname{div}(D'(u)v\nabla u),\lambda_2)_{\mathcal{X}\times[0,\tau^*]} = \\ &\left(v,\operatorname{div}\left(M(u)^T\nabla\lambda_2\right)\right)_{\mathcal{X}\times[0,\tau^*]} + (\langle M(u)\nabla v,n\rangle,\lambda_2)_{\partial\mathcal{X}\times[0,\tau^*]} - \left(v,\left\langle M(u)^T\nabla\lambda_2,n\right\rangle\right)_{\partial\mathcal{X}\times[0,\tau^*]}.\end{aligned} \tag{73}$$

Finally, time integration by part also yields

$$\begin{aligned}&(\partial_t v,\lambda_2)_{\mathcal{X}\times[0,\tau^*]} = \\ &-(v,\partial_t\lambda_2)_{\mathcal{X}\times[0,\tau^*]} + (v(.,\tau^*),\lambda_2(.,\tau^*))_{\mathcal{X}} - (v(.,0),\lambda_2(.,0))_{\mathcal{X}}.\end{aligned} \tag{74}$$

Gathering (70), (71), (73) and (74), we obtain

$$\begin{aligned}\frac{\partial L(d,g,u,\lambda_1,\lambda_2,\lambda_3,\lambda_4)}{\partial u}.v = \quad & (v(.,\tau^*),\lambda_1+\lambda_2(.,\tau^*))_{\mathcal{X}} + (v(.,0),\lambda_3-\lambda_2(.,0))_{\mathcal{X}} - \\ & (\partial_t\lambda_2 + \operatorname{div}((D(u)+M(u)))\nabla\lambda_2, v)_{\mathcal{X}\times[0,\tau^*]} + \\ & (\langle(D(u)+M(u))\nabla v,n\rangle,\lambda_2+\lambda_4)_{\partial\mathcal{X}\times[0,\tau^*]} - \\ & \left(v,\left\langle\left(D(u)+M(u)^T\right)\nabla\lambda_2,n\right\rangle\right)_{\partial\mathcal{X}\times[0,\tau^*]}.\end{aligned} \tag{75}$$

The condition

$$\forall v,\quad \frac{\partial L(d,\overline{g},\overline{u},\overline{\lambda_1},\overline{\lambda_2},\overline{\lambda_3},\overline{\lambda_4})}{\partial u}.v = 0, \tag{76}$$

thus implies that

$$\begin{cases}\overline{\lambda_2}(.,\tau^*) = -\overline{\lambda_1}, \quad \text{on} \quad \mathcal{X} \\ -\partial_\tau\lambda_2 - \operatorname{div}((D(\overline{u})+M(\overline{u})))\nabla\lambda_2 = 0, \quad \text{on} \quad \mathcal{X}\times[0,\tau^*] \\ \left\langle\left(D(\overline{u})+M(\overline{u})^T\right)\nabla\lambda_2,n\right\rangle = 0, \quad \text{on} \quad \partial\mathcal{X}\times[0,\tau^*] \\ \overline{\lambda_3} = \overline{\lambda_2}(.,0), \quad \text{on} \quad \mathcal{X} \\ \overline{\lambda_4} = -\overline{\lambda_2}, \quad \text{on} \quad \partial\mathcal{X}\times[0,\tau^*].\end{cases} \tag{77}$$

We see that λ_2 is the solution of a linearized diffusion problem, backpropagating from a final condition. With a simple change of variable $\tau' = \tau^* - \tau$ we can change the system on λ_2 into a

prograde diffusion system from an initial condition. In addition, we can replace $\overline{\lambda_1}$ by its value following (66), to finally obtain

$$\begin{cases} \partial_\tau \lambda_2 - \operatorname{div}\left(\left(D(\overline{u}) + M(\overline{u})\right)\right) \nabla \lambda_2 = 0, \quad \text{on} \quad \mathcal{X} \times [0, \tau^*] \\ \overline{\lambda_2}(., 0) = \dfrac{\partial G(\overline{g}, g_2)}{\partial g}, \quad \text{on} \quad \mathcal{X} \\ \left\langle \left(D(\overline{u}) + M(\overline{u})^T\right) \nabla \lambda_2, n \right\rangle = 0, \quad \text{on} \quad \partial\mathcal{X} \times [0, \tau^*] \\ \overline{\lambda_3} = \overline{\lambda_2}(., \tau^*), \quad \text{on} \quad \mathcal{X} \\ \overline{\lambda_4} = -\overline{\lambda_2}, \quad \text{on} \quad \partial\mathcal{X} \times [0, \tau^*]. \end{cases} \tag{78}$$

From (64) we conclude that

$$\partial_d G^*_\tau(d, d_2) = -\overline{\lambda_2}(., \tau^*), \tag{79}$$

with $\overline{\lambda_2}(., \tau^*)$ solution of (52), which concludes the proof.

□

We see from theorem 4 that the the computation of the misfit function $\nabla f_{\tau^*}(m)$ and its gradient can be done through the following steps. First, both observed and calculated datasets are filtered with NADF. Second, a misfit measurement function $G(d_1, d_2)$ is applied to the filtered datasets. This yields the misfit function value. To build the gradient, following the adjoint state theory, the corresponding adjoint source is required. To compute it, one first evaluates the conventional adjoint source through the differentiation of the misfit measurement function $G(d_1, d_2)$ applied to the filtered data. This quantity is then re-filtered through the adjoint linear diffusion process to yield the final adjoint source. This new diffusion process is linearized around the field $u(X, \tau)$ which serves to filter the calculated data in the first step. The diffusion tensor is thus anisotropic and depends on the evolution variable τ but the diffusion process is linear as the diffusion tensor does not depend on λ.

This is summarized in algorithm 1. A naive implementation is presented: the filtering of the observed data could be done once and for all before the inversion so as to avoid applying the filter each time on the observed data. Also we do not discuss here how the incident wavefields $w_s[m]$ are handled to build the gradient: this is a well discussed topic in FWI, with alternatives proposed from storing it in memory to recomputing it from a final snapshot (see for instance Yang et al., 2016, for a review of these alternatives and the presentation of the CARFS strategy which we use in this study). Finally, let us note that neglecting the operator $M(u)^T$ in the adjoint diffusion system (57) makes it possible to use the same diffusion algorithm at each stage of the algorithm, which is appreciable in practice. This is the strategy which is used in this study. From our numerical experiments, it appears that neglecting this contribution to the construction of the adjoint source is not harmful to the convergence of the whole FWI strategy. A more careful analysis could be done however to quantify the error associated with neglecting this term in the linearized diffusion process.

4. Numerical experiments

All the experiments presented in this study are performed using our 2D acoustic second-order in time and fourth-order in space finite difference code TOYxDAC_TIME (Yang et al., 2018), designed for full waveform modeling and inversion. It is coupled with the SEISCOPE optimization library, which provides `FORTRAN` implementations of gradient-based and quasi-Newton optimization solvers (Métivier and Brossier, 2016). In this study we rely on the l-BFGS method to perform our inversion tests.

Algorithm 1: Algorithm for the computation of the misfit function $f_{\tau^*}(m)$ and its gradient $\nabla f_{\tau^*}(m)$.

Data: m, τ^*, $d_{obs,s}$, b_s, $\quad s = 1, \ldots, N_s$
Result: $f_{\tau^*}(m), \nabla f_{\tau^*}(m)$
// Initialization of misfit and gradient ;
$f_{\tau^*}(m) = 0$;
$\nabla f_{\tau^*}(m) = 0$;
// Main loop over the sources;
for $s = 1, \ldots, N_s$ **do**
  //Misfit function part;
  Compute $g_{2,s} = \mathcal{F}_{\tau^*}(d_{obs,s})$;
  Compute $d_{cal,s}[m]$ through the solution of (31) and (32) ;
  Compute $g_{1,s} = \mathcal{F}_{\tau^*}(d_{cal,s})$ and store $u_s(X, \tau)$ in memory $\forall \tau \in [0, \tau^*]$;
  Update the misfit function with contribution from source s
  $f_{\tau^*}(m) = f_{\tau^*}(m) + G\left(g_{1,s}, g_{2,s}\right)$;
  //Gradient part;
  $\mu_s = \frac{\partial G}{\partial g_1}\left(g_{1,s}, g_{2,s}\right)$;
  Filter μ_s through the diffusion equations (57) linearized around u_s;
  Compute the adjoint wavefield λ_s from equation (41);
  Update the gradient with contribution from source s (equation (54))
  $\nabla f_{\tau^*}(m) = \nabla f_{\tau^*}(m) + \int_0^T \left(\frac{\partial A(m)}{\partial m} w_s[m]\right)(x,t)\lambda_s[m](x,t)dt$);
end

4.1. Data generation

The numerical experiments we present in the following are based on the Marmousi II benchmark model (Martin et al., 2006). To avoid working in "inverse crime" settings, we generate our data using fine grid (5 m) P-wave velocity and density models, under the 2D acoustic variable density approximation. These models are obtained by upscaling the original P-wave velocity and density Marmousi II models defined on a 1.25 m grid. Perfectly matched layers (PML) Bérenger (1994) with 40 points thickness are used to decrease the energy of spurious reflections at the border of the numerical domains, on the left, right and bottom side. On top, a free surface condition is implemented. A Kaiser-windowed cardinal-sine interpolation (Hicks, 2002) is used to represent sources and receivers off the Cartesian grid. A Ricker source centered on 5 Hz, and filtered so as to remove energy below 2.5 Hz, is used to generate the data. An additive Gaussian noise, filtered in the frequency range of the data (0 to 12.5 Hz) is introduced, with a signal to noise ratio equal to 10. The acquisition contains 128 seismic shots (source positions) at 50 m depth in the water column, regularly spaced each 132 m in the horizontal direction, and 169 receivers, also at 50 m depth, regularly spaced each 100 m. The acquisition is done in a fixed-spread fashion: each receiver records all the sources.

The fine grid P-wave velocity and density models we use are presented in Figure 1. The corresponding seismic source and its amplitude spectrum are presented in Figure 2a and 2b, while two examples of seismic shot gathers for two different shot positions are presented in Figure 2c and 2d.

4.2. Inversion setup

The inversion is performed on a coarse, regular grid, with 25 m grid interval. Only the P-wave velocity model is reconstructed. We consider two different initial models presented in Figure 3. They are computed in two stages. First, the fine grid true velocity model is smoothed with an isotropic Gaussian filter with a correlation length equal to 250 m and 500 m respectively. The resulting velocity models are then downsampled from the fine to the coarse grid. The two

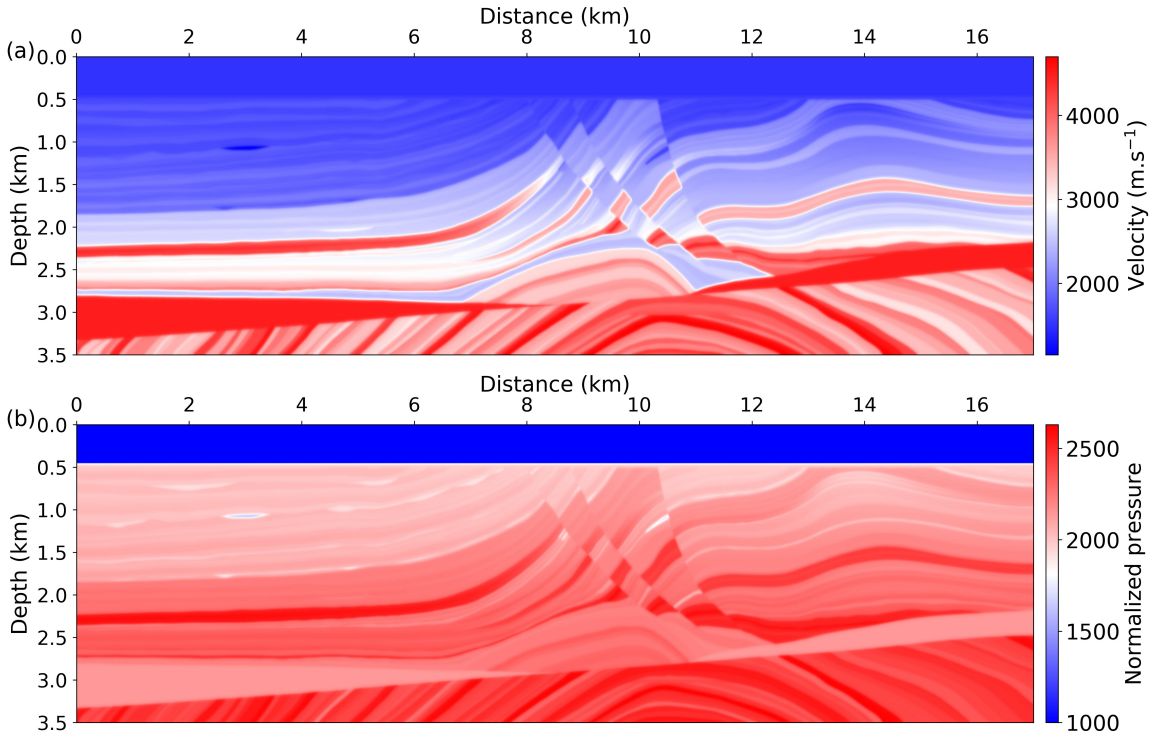

Figure 1. Fine grid P-wave velocity (a) and density (b) models used to generate the observed data. They are defined on a regular grid with a grid interval equal to 5 m.

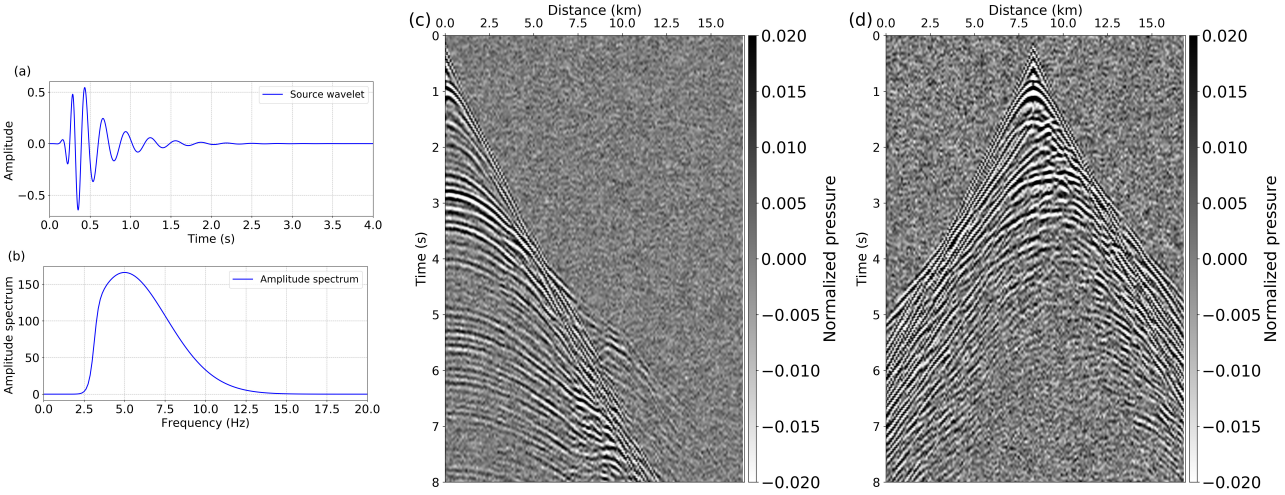

Figure 2. Source wavelet (a) and associated amplitude spectrum (b). The peak frequency is at 5 Hz, and the frequency band goes from 2.5 Hz to 12.5 Hz. Corresponding shot gathers computed in the fine grid P-wave velocity and density models (Fig. 1a and 1b) with the left most source $x_S = 0.05$ km (c), the source in the middle $x_S = 7.5$ km (d). A Gaussian noise filtered in the frequency band 2.5-12.5 Hz is added to the data.

corresponding initial density models are computed using the following Gardner's law

$$\rho(x) = 1741 \times \left(\frac{V_P}{1000}\right)^{0.25}. \tag{80}$$

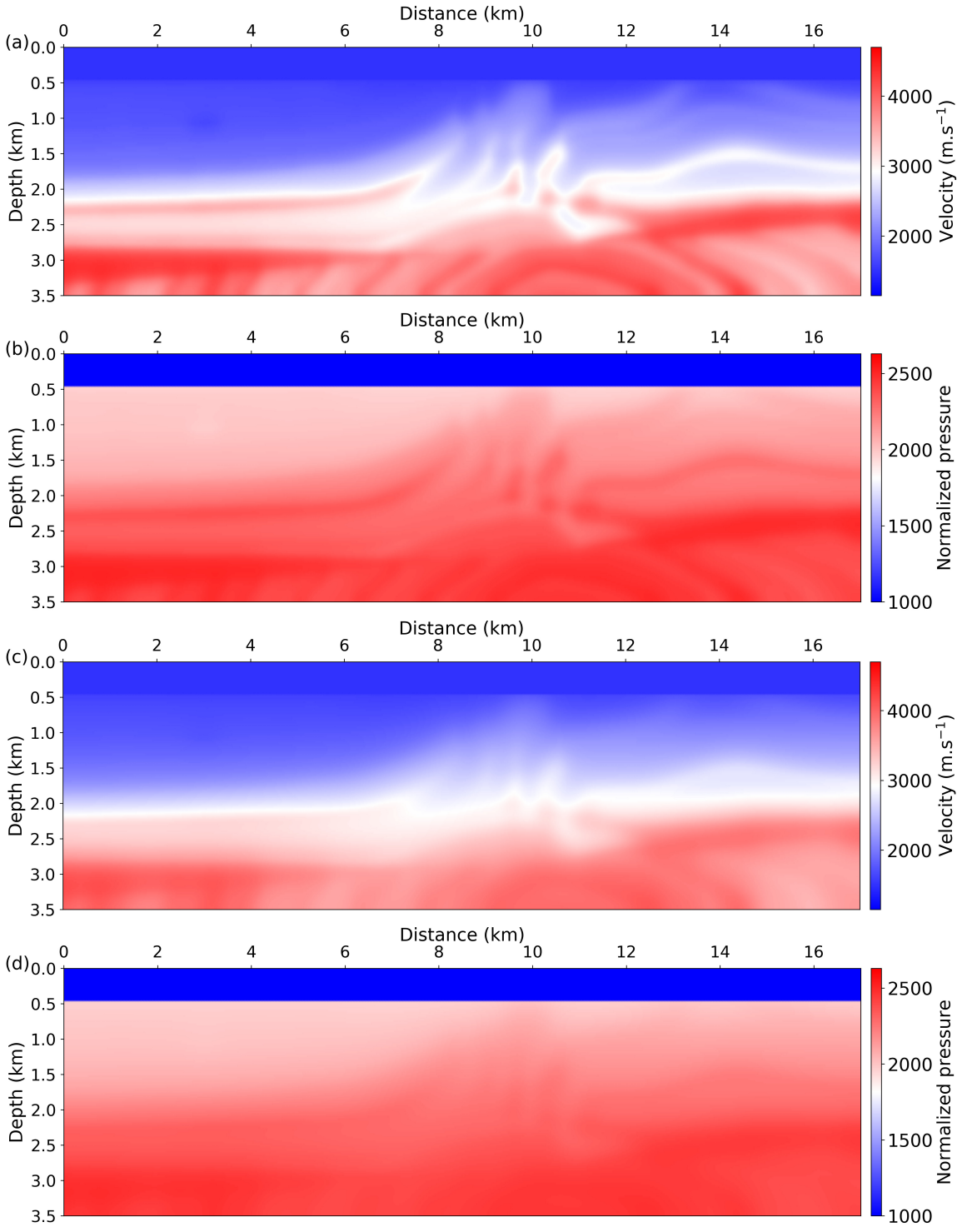

Figure 3. Initial velocity model 1 (a) and corresponding density model (b). Initial velocity model 2 (c) and corresponding density model (d).

From these two initial models, two source wavelets (time signatures) are computed. Computing these wavelets is equivalent to solve a linear deconvolution problem. We follow the frequency domain approach introduced by Pratt (1999). A taper is applied to make the wavelets causal, and remove oscillations after $t = 2$ s. The resulting wavelets and their normalized amplitude spectrum

are compared to the ones from the true wavelet in Figures 4a,b and 5a,b. As can be seen, the amplitude of the wavelet is slightly underestimated compared to the true wavelet, which is due to the inaccurate initial velocity and density models and the presence of noise in the data. The corresponding synthetic data for the two shot gathers presented in Figures 2 are presented in Figure 4c,d and 5c,d. The synthetic data is overlay in red/blue colors on the observed data presented in black and white. Some of the transmitted events are correctly predicted, especially in initial model 1. The reflected events are not predicted by any of the two models.

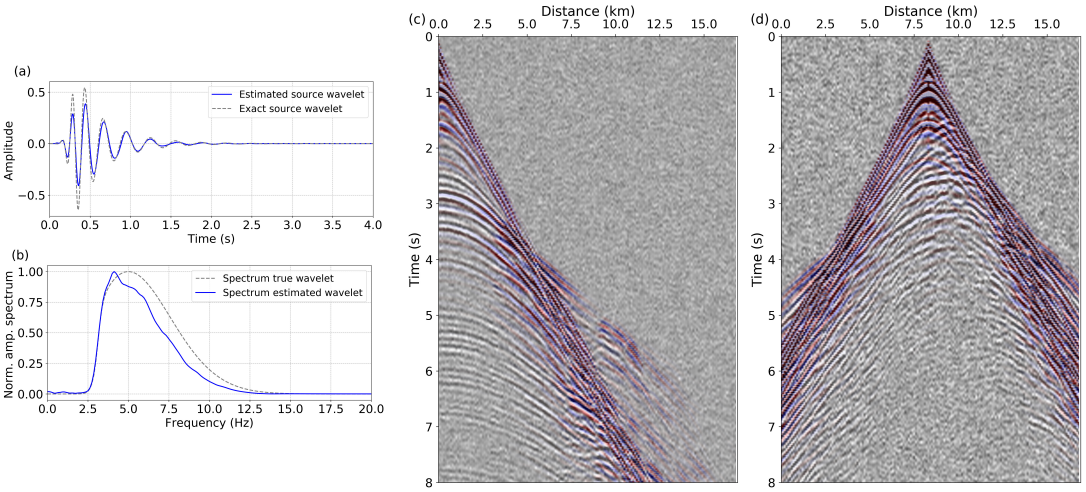

Figure 4. Estimated wavelet in model 1 (a) and corresponding normalized amplitude spectrum (b). Due to the inaccuracy of both velocity and density models, the wavelet amplitude is underestimated. Synthetic shot gather for $x_S = 0.05$ km (c) and $x_S = 8.5$ km (d) computed in model 1, overlay in red/blue colors on the corresponding shot gather in black/white.

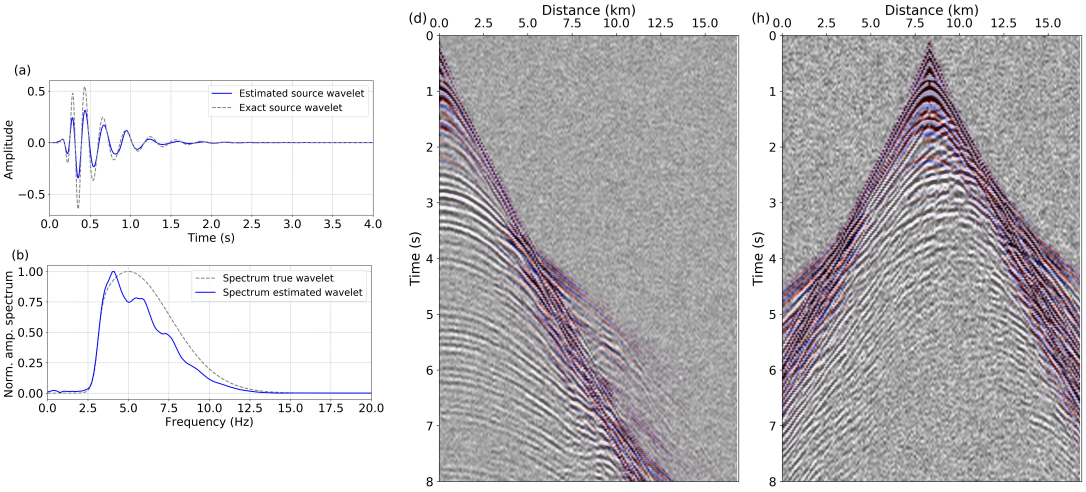

Figure 5. Same as Figure 4 using initial model 2.

4.3. Gradient and data smoothing

We start by analyzing NADF applied on the gradient computed in the initial model 1. The original gradient without smoothing is presented in Figure 6a1. To emphasize its regularity, we overlay its level curves in Figure 6a2, and we present its wavenumber spectrum in k_x and k_z directions in Figure 6a3. As can be seen, the original gradient exhibits some spatial irregularities and fast variations one could associate with imperfect subsurface illumination and noisy data. In Figures

6b,c,d, we present the same plots for the gradient obtained after the nonlinear anisotropic diffusion smoothing operator is applied, with diffusion times from $\tau^* = 2.5$, $\tau^* = 5$, and $\tau^* = 20$. The progressive simplification in the level curves and the reduction of the spread of the wavenumber spectrum is an illustration of the scale-space property of the filter. Interestingly, we can see how the main structures of the gradient are preserved through the smoothing, small scale structures being smoothed out progressively depending on the diffusion time.

For comparison, we present in Figure 6e the "conventional" smoothed gradient with a non-stationary isotropic Gaussian filter with correlation lengths adapted to an estimation of the local wavelength $\lambda(x, z)$. Here, we use

$$\lambda(x, z) = 0.4 \frac{V_P(x, z)}{f_0}, \tag{81}$$

with $V_P(x, z)$ described by model 1 and $f_0 = 5$ Hz, the central frequency of the source wavelet. As can be observed, such smoothing does not preserve the structure of the gradient. According to the velocity increase with depth, the deeper structures are also more severely affected.

We also present the effect of NADF on the data and illustrate its scale-space properties in Figure 7. The filter is applied to the shot gather from Figure 2c, with diffusion times equal to $\tau^* = 6.25$ (Fig. 7b), $\tau^* = 25$ (Fig. 7c) and $\tau^* = 50$ (Fig. 7d). The filter increases the coherency of the events in the time/offset panel. It also removes small scale noise oscillations to enhance larger scale oscillations. We will see in the full waveform inversion tests that this artificially structured noise is not detrimental to the quality of the inversion. The spectrum of the filtered data also reveals that the filter acts as a low-pass filter, enhancing the low frequency part of the data, and shifting the amplitude spectrum from high to low frequencies.

The latter effect is interesting. Enhancing the low-frequency content of the seismic data is natural in FWI to stabilize the early inversion before interpreting data with higher frequency content (Bunks et al., 1995). The difficulty is that the lowest frequency part of the data is often contaminated by noise, and standard linear filters are not able to separate the signal from the noise. The coherence enhancing filter applied here simultaneously plays the role of a low frequency pass filter and a denoising filter, making it possible to push the frequency content of the data towards the lower end of the spectrum without decreasing the signal to noise ratio.

4.4. Full waveform inversion results

4.4.1. Effect of model space regularization We first analyze the impact of NADF on FWI when applied in the model space. The initial P-wave velocity model (Fig.8a), is compared with the final models obtained with a Gaussian smoothing (Fig.8b) and NADF (Fig.8c). With Gaussian smoothing, the convergence of the l-BFGS algorithm is observed after 41 iterations. We compare both inversion for the same number of iterations. FWI combined with model space NADF is able to perform more iterations, but the additional iterations lead mainly to overfit the data. The final model estimation using NADF is more resolved at depth and exhibits clearer interfaces. This is expected, as the Gaussian smoothing tends to smooth out the structure from the model updates. The comparison of the convergence curves in terms of misfit function decrease (Fig.8d) and model error decrease (Fig.8e) shows how the use of NADF accelerates the convergence.

4.4.2. Effect of data space regularization and combination data/model regularization We next analyze the effect of NADF applied to the data. We proceed to a multi-scale inversion as described in equations 48 to 50 in 4 stages, with the diffusion times: $\tau^* = 25$, $\tau^* = 12.5$, $\tau^* = 6.25$, and $\tau^* = 0$ (the last stage corresponds to a conventional least-squares inversion without NADF). At each stage, the convergence of the l-BFGS minimization algorithm is observed after 32, 38, 49 and 44 iterations respectively. The models obtained after the first stage ($\tau^* = 25$) and the final one ($\tau^* = 0$) are presented in Figures 9c and 9d respectively. As can be seen, compared with the conventional result, already at the first stage of this workflow the estimated model is closer from the true model, especially regarding the estimation of the triangle shape low velocity zone

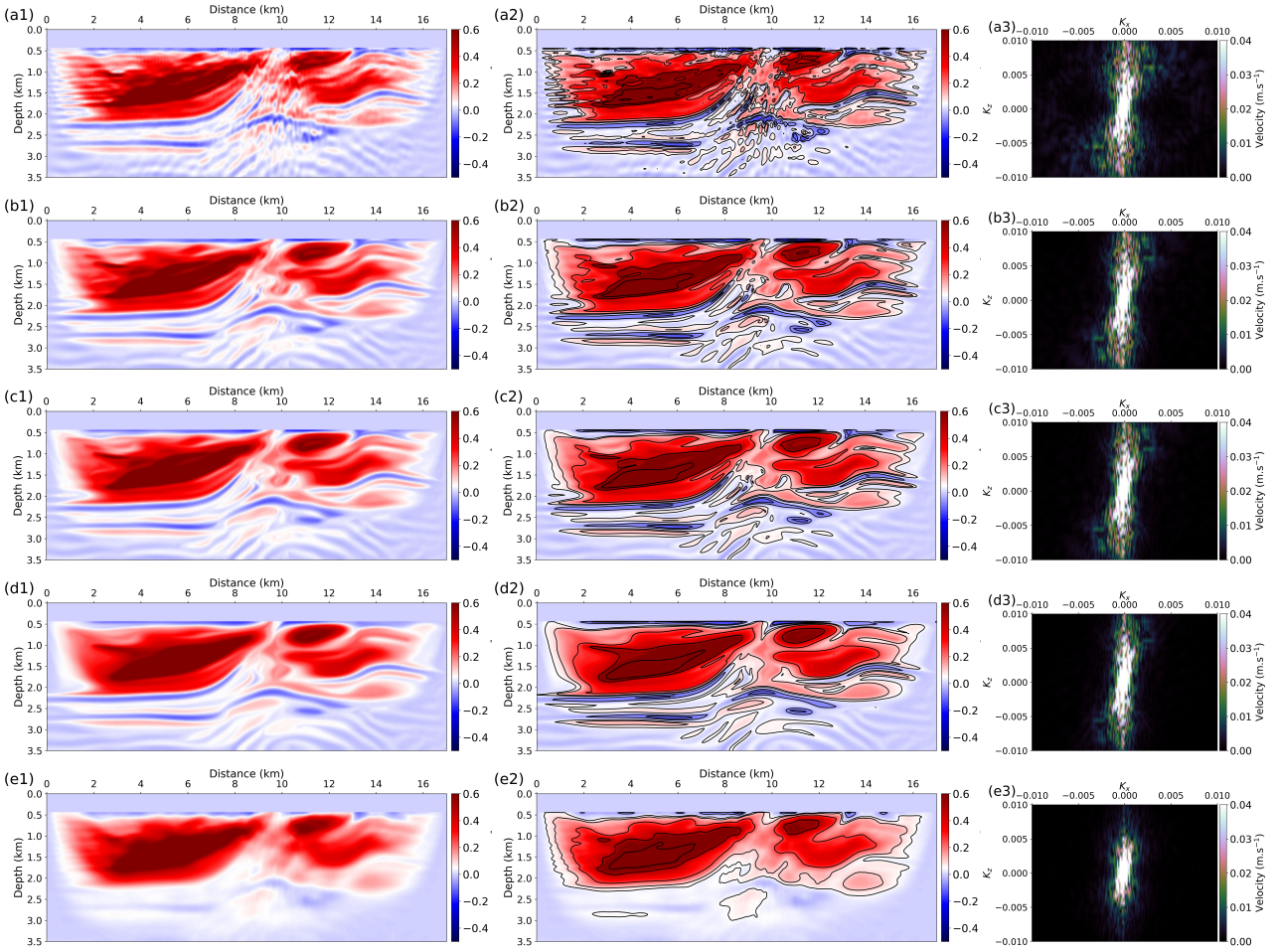

Figure 6. Scale space property of NADF and comparison with a nonstationary isotropic Gaussian filter. Gradient before filtering (a1-a3). Gradient after NADF with $\tau^* = 2.5$ (b1-b3), gradient after NADF with $\tau^* = 5$ (c1-c3) , gradient after NADF with $\tau^* = 20$ (d1-d3), gradient after nonstationary isotropic Gaussian smoothing (e1-e3). In column 1, the gradient in blue/red color scale are presented. In column 2, the level set are superimposed to emphasize the structural information embedded in the gradient. In column 3, the corresponding wavenumber spectrum is presented. One can see how increasing the diffusion time progressively smooth small scale structures while preserving larger scale structures. The spread of the wavenumber spectrum gradually decreases as the diffusion time increases. In comparison, the isotropic Gaussian smoothing destroys most of the small and large scale structures, especially at depth where the correlation length increases. The spread of the wavenumber spectrum is also significantly reduced.

in the middle of the model (at 2.5 km depth, between $x = 10$ and $x = 12$ km), which is one of the most complex area in the Marmousi II model (Martin et al., 2006). This area is incorrectly reconstructed by the conventional approach as the velocity is overestimated. Thanks to the multi-scale NADF approach, this area is better reconstructed, as well as deeper structures (Fig.9d).

The already good reconstruction obtained at the first stage of the multi-scale workflow prompts us to test a natural idea: combining both data space and model space regularization with NADF. The result obtained after 60 l-BFGS iterations is presented in Figure 9e. As can be seen, this combination provides an estimation similar to the one obtained after the multi-scale worfklow in less iterations.

The overall better reconstruction of the aforementioned triangle shape low velocity zone can

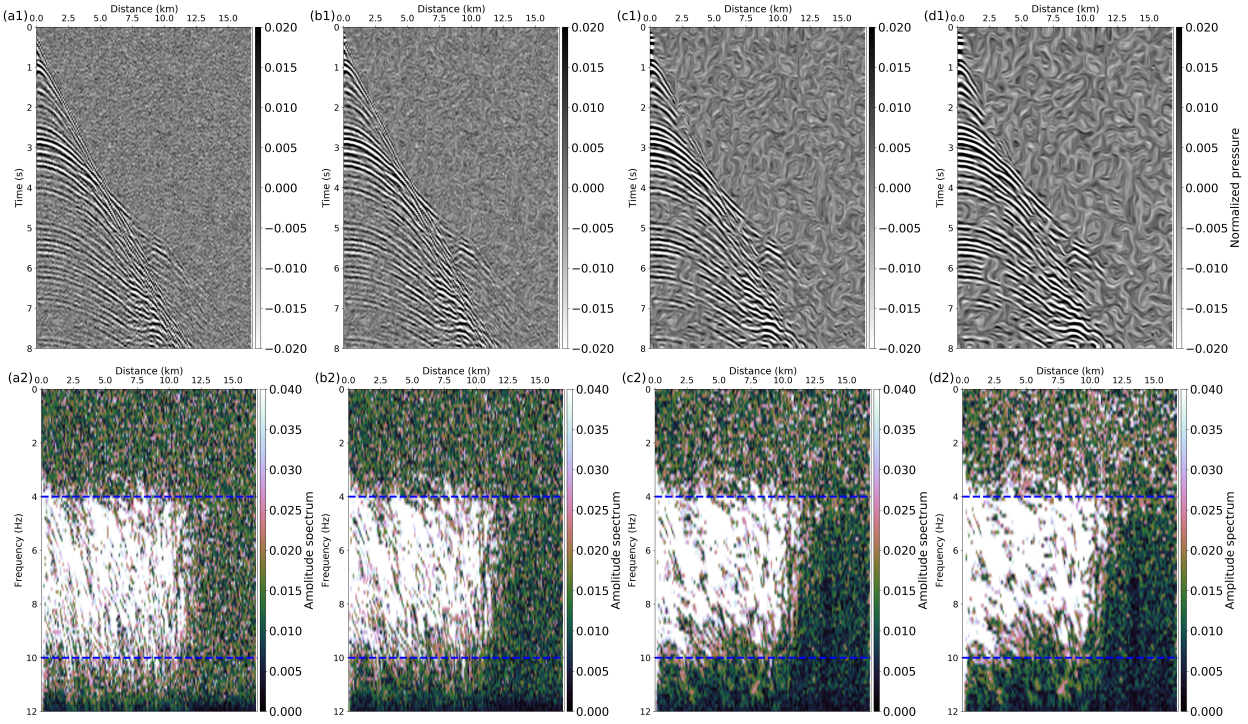

Figure 7. Scale space property of NADF illustrated on a shot gather. Shot gather before filtering (a1-a2), shot gather after NADF with $\tau^* = 6.25$ (b1-b2), shot gather after NADF with $\tau^* = 25$ (c1-c2), shot gather after NADF with $\tau^* = 50$ (d1-d2). On the first row we present the filtered shot gather, on the second row, the corresponding amplitude spectrum. The dashed blue lines indicating the 4-12 Hz window highlight how the frequency content slightly moves from high to low frequencies as the diffusion time increases.

be related to the lower frequency content of the filtered data using NADF. As an illustration of the effect of this data smoothing, we compare in Figure 10 the gradient at first iteration computed from raw data and no model smoothing (Fig.10c), from filtered data with diffusion time $\tau^* = 25$ and no model smoothing (Fig.10d), from filtered data with diffusion time $\tau^* = 25$ plus NADF model smoothing (Fig.10e). We can see already in Figure 10d how the enhancement of the low frequency content of the data, visible in the comparison between the conventional residuals (Fig.10a) and the residuals obtained when NADF is applied to the data (Fig.10b), translates into smoother and better delineated structures in the gradient. These are further enhanced by NADF applied directly in the model space in Figure 10e.

To further test the approach combining both data and model space regularization, we apply it starting from initial model 2. This is much more challenging, as can be seen in Figure 11b, which present the model estimated with a conventional FWI method. The convergence is observed after 72 iterations. This time, the final model is severely cycle skipped, especially in its left part. This is due to the too large kinematic error induced by the poorer initial model. In Figure 11c, we present the result obtained using a graph-space optimal transport (GSOT) distance, a strategy designed to mitigate such cycle skipping issues by designing a misfit function convex with respect to time shifts (Métivier et al., 2019). This time the convergence is observed after 23 iterations. The left part is more stably reconstructed, however the deeper part of the model remains inaccessible. Finally, in Figure 11d we present the result obtained by combining the GSOT approach with data and model space smoothing using NADF. The iterations are arbitrarily stopped after 100 iterations. One can see that the cycle skipping issues have been fixed, thanks to the combination of the GSOT approach and the low frequency enhancement provided by the low frequency enhancement. The model smoothing enables a sharp reconstruction of the structure and the interfaces.

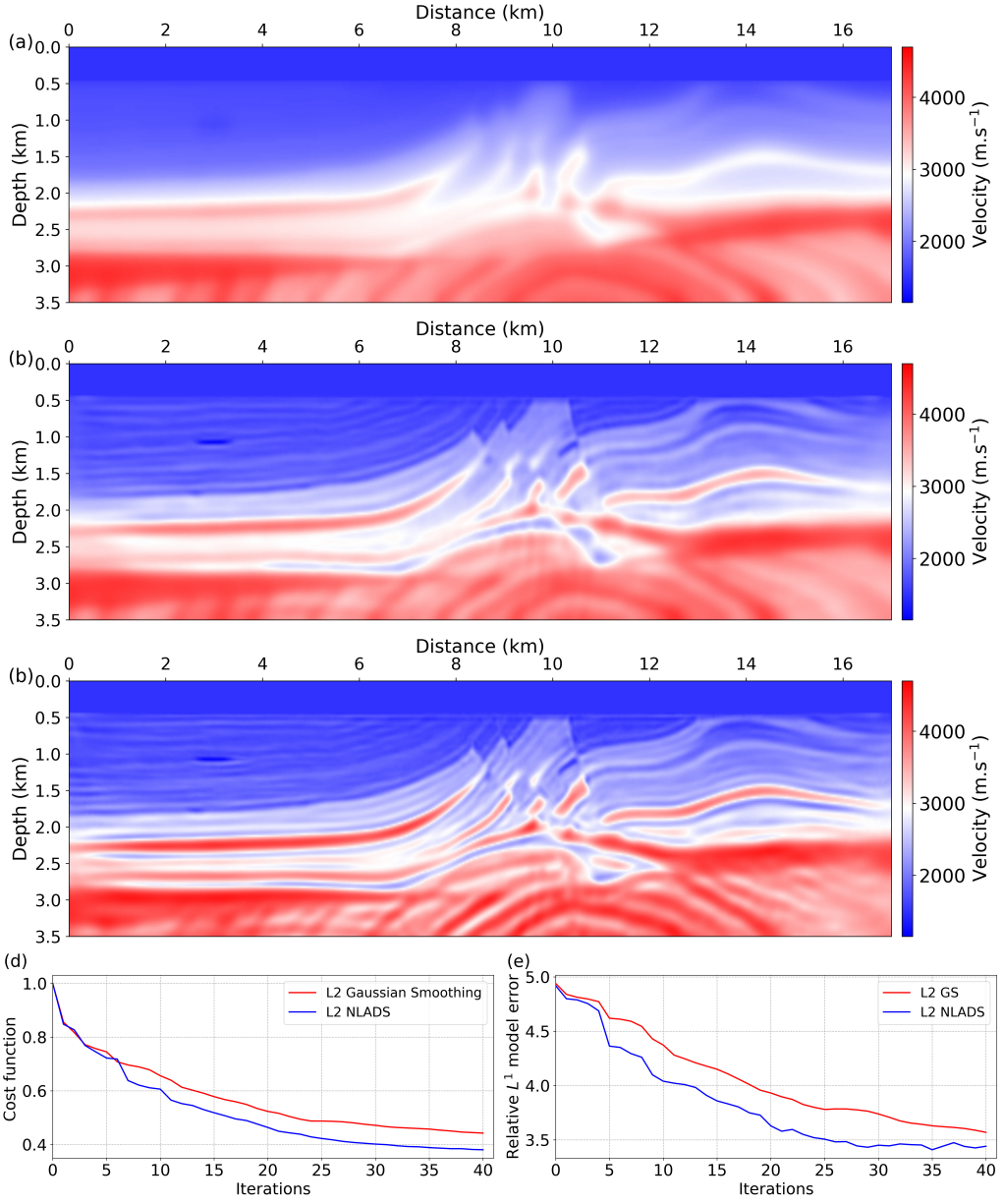

Figure 8. Initial model 1 (a), final model after 41 FWI iterations with a Gaussian smoothing (b), with NADF (c). Misfit function (d) and model error (e) history along the convergence path using Gaussian smoothing (solid red) and NADF (solid blue).

5. Discussion

NADF appears to be an interesting tool for FWI. When the FWI problem is well-posed, i.e. when it starts from an initial model in the valley of attraction of the global minimizer, applying NADF in the model space to the gradient avoids the undesirable effect of Gaussian smoothing which, at the same time, removes spurious oscillations but also destroys the main structure of the model. The behavior of the Gaussian smoothing appears symptomatic of a slowly convergent process: the seismic data contribute to enhance the subsurface model structure, while the model smoothing tends to remove this structure. This conflict between data and model space contributions results in an deterioration of the convergence rate. On the contrary, NADF make it possible to reconcile both contributions: NADF remove small scale oscillations due to noise and uneven illumination

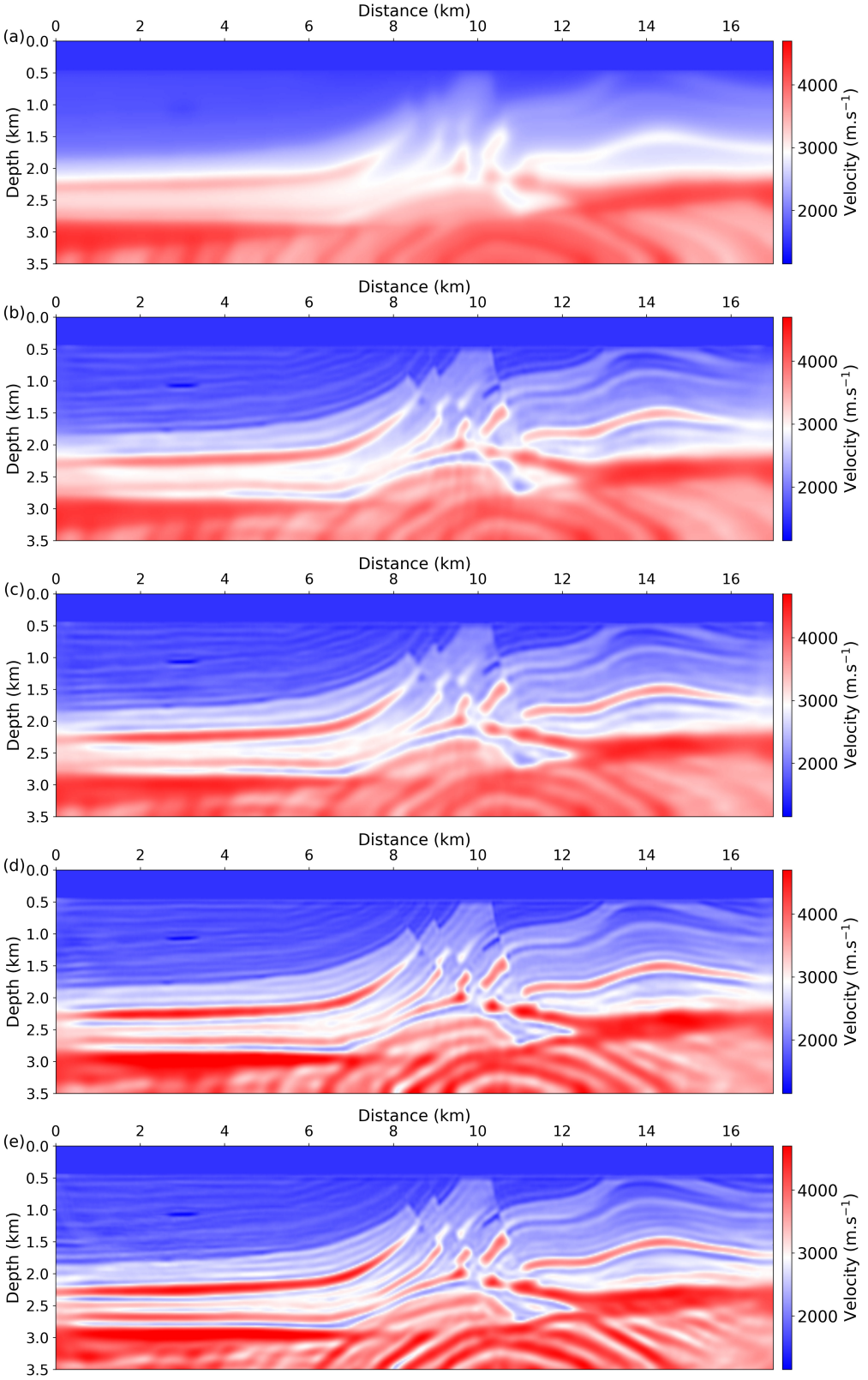

Figure 9. Initial model 1 (a), final model after 41 FWI iterations with a Gaussian smoothing (b), with a data smoothing $\tau^* = 25$ (c), with the multi-scale data smoothing strategy from $\tau^* = 25$ to $\tau^* = 0$ (d), combining data smoothing with $\tau^* = 25$ and NADF on the gradient (e).

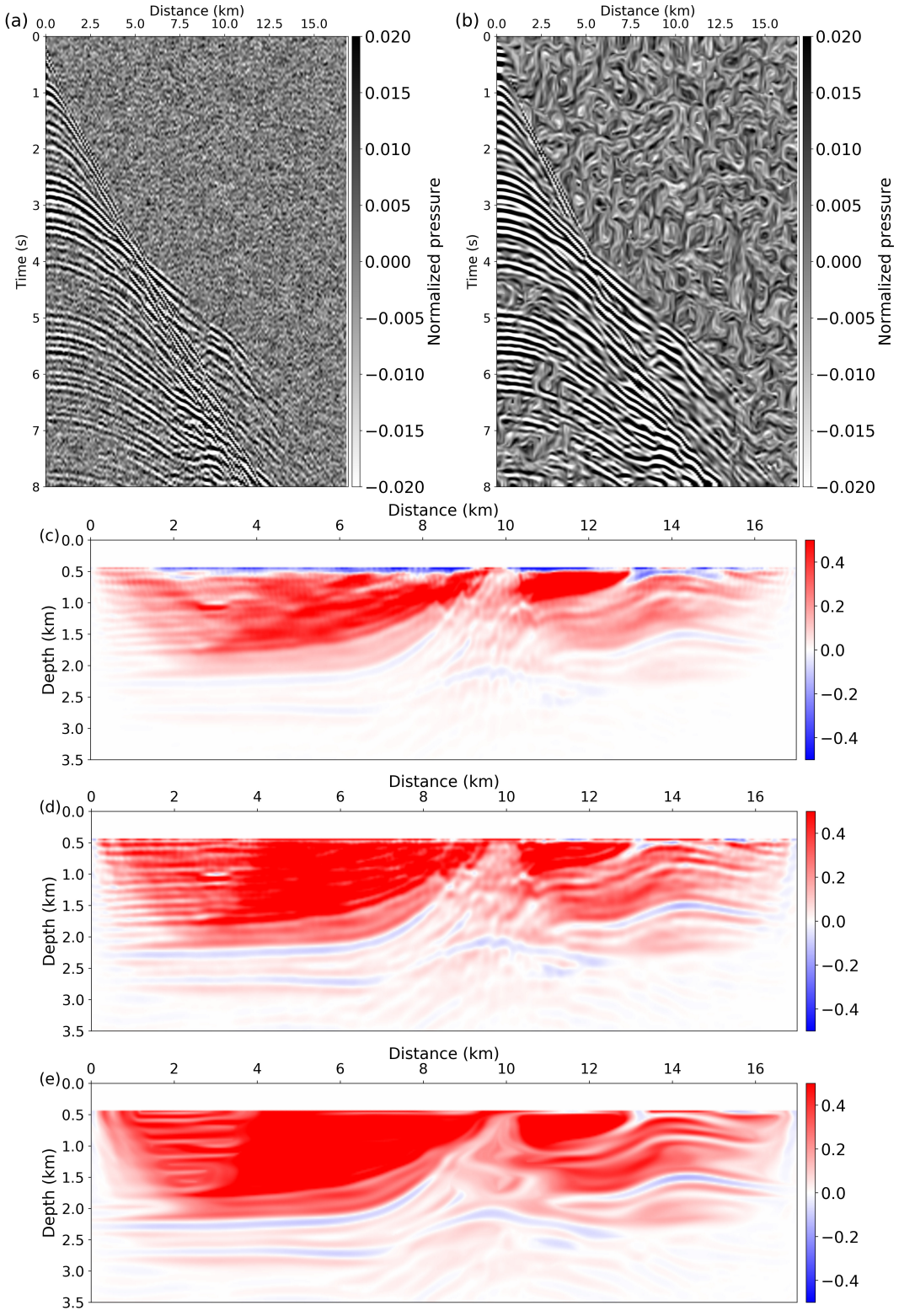

Figure 10. L^2 residuals (a), residuals corresponding to a data smoothing with $\tau^* = 25$ (b), conventional initial gradient (c), initial gradient with a data smoothing of $\tau^* = 25$ (d), initial gradient with data smoothing of $\tau^* = 25$ and an additional pass of NADF smoothing (e).

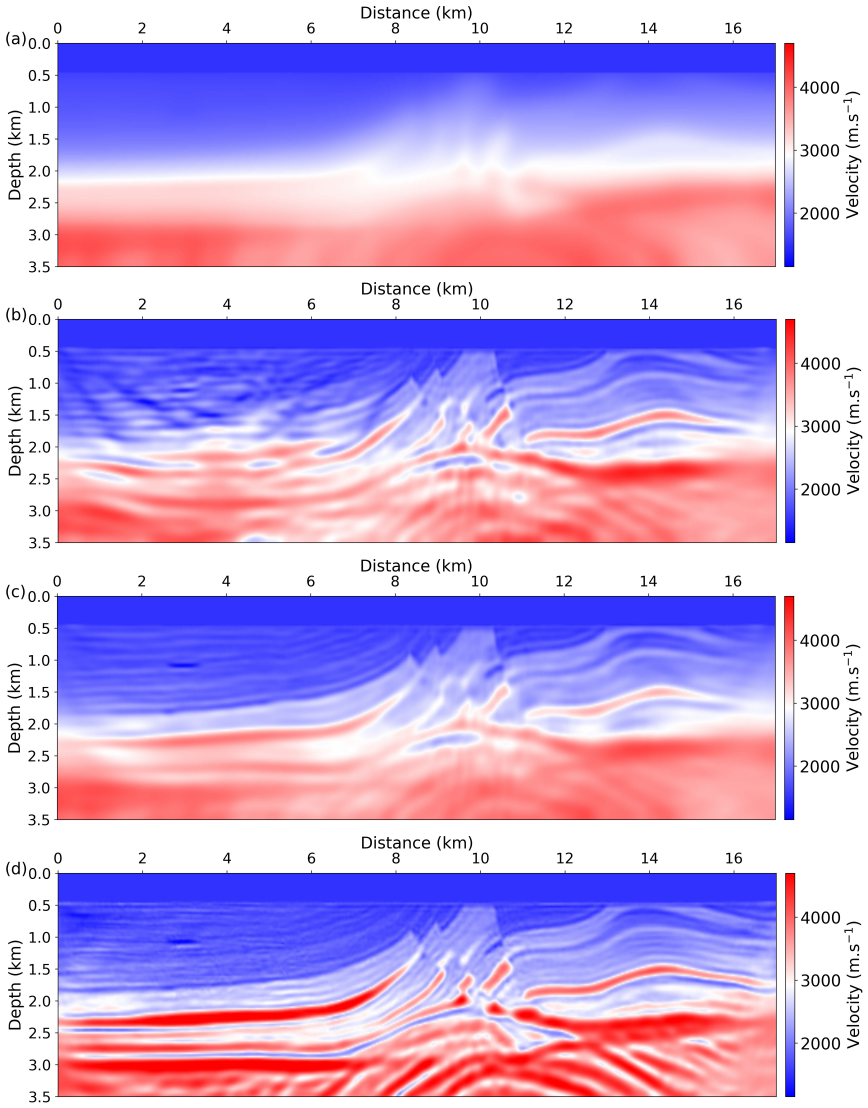

Figure 11. Initial model 2 (a), final model with a conventional FWI method (b), final model with a graph-space optimal transport misfit function (Métivier et al., 2019) (c), final model combining graph space optimal transport misfit function with data and model space smoothing through NADF (d).

while preserving the main structure. As a result, the convergence speed is improved. Of course, this has to be mitigated with situations where structures in the gradient would correspond to artifacts. In this case they would be enhanced instead of being removed. This issue is however related to the premise: NADF in the model space should be helpful to accelerate the convergence of already "well-posed" FWI problems.

Beyond traditional "low frequency" FWI, this suggests that NADF could be used fruitfully in the context of high resolution structural imaging using migration algorithms (reverse time or least-squares reverse time migration) or, as has been considered more recently, "high resolution FWI" (Shen et al., 2018). In this context, the corresponding inverse problems are relatively well posed. What remains challenging is the high computational cost of these methods as high frequency wave propagation problems have to be solved at each iteration (by high frequency we mean problems for which the waves have to be propagated over several hundreds of wavelengths in each direction of space). The convergence speed is thus crucial for these applications and such a structure

enhancing filter, amenable to accelerate the convergence, could therefore be particularly helpful. This calls for an extension of what is presented here to 3D NADF. Given the underlying PDE framework of NADF, this extension should be straightforward at the continuous level. Designing a finite-difference scheme which preserves the properties of NADF at the discrete level in 3D might be more delicate. This should be the matter of future investigation which should be inspired by the work from Fehrenbach and Mirebeau (2014); Mirebeau (2016) where finite-difference schemes adapted to the anisotropy of the diffusion operator are derived .

When applied in the data space, NADF seems to act as an interesting nonlinear low-pass filter, with the ability to enhance the low frequency content of the data without making the noise dominant. The scale-space property also provides a natural framework to design a hierarchical FWI scheme. Enhancing the low frequency part of the data is beneficial to FWI when the initial model lies outside the basin of attraction of the global minimum as it serves to stabilize the inversion. We would like however to mention that NADF alone on the data should not be enough to reduce cycle skipping issues in challenging settings such as the one we investigate here when starting from initial model 2. In this case, the low-frequency enhancing property of NADF applied on the data is not sufficient to make a conventional least-squares FWI converge toward a correct estimation of the subsurface velocity. Rather, it should be seen as a complementray strategy to other methods dedicated to mitigate cycle skipping in FWI, such as misift function modifications.

Indeed, from theorem 4, we see also that applying NADF in the data space can be combined with any misfit function modification strategy. In our numerical experiments, it seems in particular that it cooperates well with the GSOT strategy. It is illustrated in Métivier et al. (2019); Pladys et al. (2021); Górszczyk et al. (2021) that the GSOT strategy results in introducing high frequency components in the source of the adjoint field. Due to the computational cost of solving optimal transport problem on discrete point clouds, the GSOT strategy also considers each seismic trace independently, without taking into account the 2D coherency of the data in the time/receiver plane. Interestingly, when GSOT is composed with NADF, the resulting adjoint source corresponds to the standard GSOT adjoint source applied to the filtered data, *with an additional linearized diffusion filtering*. This additional filtering, as illustrated in Figure 12, removes the high frequency oscillations associated with GSOT, and also enhances the coherence of the adjoint source in the time/receiver plane. The composition of GSOT with NADF approach might thus be seen as a compensation for some of the current limitations of GSOT.

We would also like to mention at this stage, regarding practical applications, that the selection of the diffusion time τ^*, both in the data and model spaces is performed on a trial-and-error basis in this study. How to precisely and automatically design this parameter, and also the ones related to the pre-smoothing (noise scale σ and and coherence scale ρ in equation (10)) involved in the design of the diffusion operator, should be the matter of future inverstigation. While it is easy to design in the linear diffusion case through the equivalence between the final diffusion time and the coherent lenght of the underlying Gaussian kernel in equation (3), this is far more complicated in this nonlinear anisotropic settings.

Finally, another interest for the application of NADF in the data space could be related to data interpolation in the context of missing traces. It is usual in seismic surveys that some receivers do not respond correctly, either due to an incorrect coupling with the subsurface or simply a device breakdown. In such a case, the resulting seismograms present holes corresponding to seismic traces put to 0 for these defective receivers. Conventional data processing techniques consist in interpolating the data to recover the information from these missing traces. Thanks to its coherence enhancing feature, when NADF is applied to a seismogram presenting missing traces, it will simultaneously enhance the low frequency content of the data and play the role of a data interpolator. An example is provided in Figure 13. We consider the leftmost seismogram used in the numerical experiments, and set randomly 30 % of its traces to 0 (Fig.13a). In Figure 13b, we present the corresponding NADF version of this seismogram. It can be seen that the reflected events at short to medium offsets ($x < 5$ km) are relatively well interpolated.

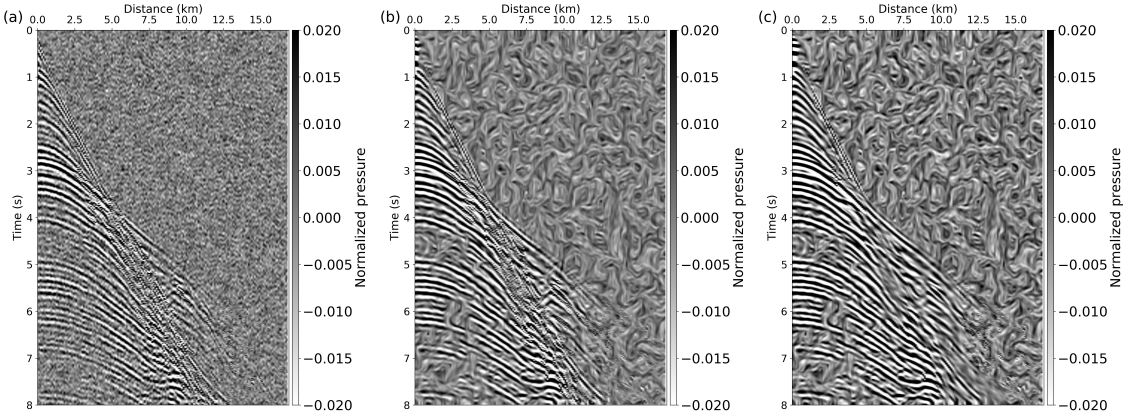

Figure 12. Comparison of residuals for the left most shot gather. Adjoint source associated with the GSOT strategy (a). Adjoint source associated with the GSOT strategy combined with NADF applied on the data with $\tau^* = 25$, *before* the solution of the linear diffusion process (b). Adjoint source associated with the GSOT strategy combined with NADF applied on the data, *after* the solution of the linear diffusion process (c). The residuals displayed in (b) correspond to the computation of $\mu_s = \frac{\partial G}{\partial g_1}(g_{1,s}, g_{2,s})$ at line 7 of algorithm 1. The residuals displayed in (c) correspond to μ_s after it has been filtered from the linearized diffusion process at line 8 of algorithm 1. Note that the GSOT approach inherently generates high frequency components in the residuals, visible in (a) and (b) (Métivier et al., 2019). The filtering procedure inherited from the coupling between GSOT and NADF in the data space filters out these high frequency components, as is visible in (c).

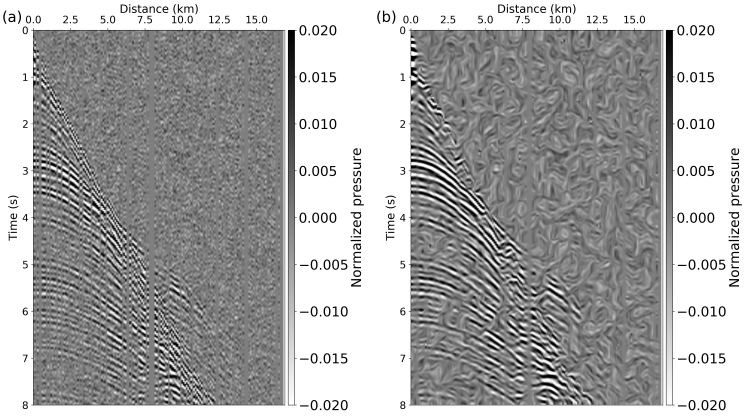

Figure 13. Data interpolation test. Seismogram with 30% of randomly selected traces set to 0 (a). Same seismogram after NADF filter (b). We see that the reflected events at short to medium offsets ($x < 5$ km) appear correctly interpolated, on top of the low frequency filter effect already detailed.

6. Conclusion

We present in this study how nonlinear diffusion filters, designed in the frame of image processing, can be applied to high resolution seismic imaging based on the full waveform. We first propose a short introduction to NADF, summarizing the main results and concepts presented in the reference book by Weickert (1998). Among them, minimum-maximum principle, scale-space properties, and the fact that they are based on PDE, make NADF an interesting tool for any data-fitting based inverse problem. The principle of coherence enhancing approach is to define a diffusion operator which depends on the local orientations of the image structure, which is accessed through the structure tensor of the image. Basically, the diffusion is set to be strong in the direction of small

variations and smaller in the direction of fast variations.

After introducing the FWI mathematical formalism, we show how NADF can be integrated as coherence enhancing filters both in the space of the subsurface mechanical parameters (model space) and in the space of the seismic recordings (data space). While the application of NADF in the model space is straightforward, its application in the data space requires more care. The key point is to be able to derive the gradient of the misfit function minimized through FWI. We prove that, following the conventional adjoint-state strategy, the corresponding adjoint source is obtained by the application of the conventional adjoint formula to the filtered data, and an additional filtering operation corresponding to a linearized version of the filter applied to the calculated data.

We present numerical illustrations of both model and data space strategies on a realistic synthetic case study. While model space filtering through NADF appears as an interesting tool to improve the convergence of FWI when the problem is well-posed, data space filtering appears as an interesting low-frequency enhancement strategy, making it possible to stabilize FWI convergence when the problem is ill-posed. The latter strategy is complementary with other investigated methods to overcome FWI ill-posedness such as misfit function modifications.

Based on the results presented here, field data applications and extension of this method to 3D model and data will be considered in the future.

Acknowledgements

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

References

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

Aghamiry, H., Gholami, A., and Operto, S. (2020b). Compound regularization of Full-Waveform Inversion for imaging piecewise media. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):1192–1204.

Anagaw, A. Y. and Sacchi, M. D. (2018). Edge-preserving smoothing for simultaneous-source full-waveform inversion model updates in high-contrast velocity models. *Geophysics*, 83(2):A33–A37.

Asnaashari, A., Brossier, R., Garambois, S., Audebert, F., Thore, P., and Virieux, J. (2013). Regularized seismic full waveform inversion with prior model information. *Geophysics*, 78(2):R25–R36.

Barnier, G., Biondi, E., and Clapp, R. (2019). Waveform inversion by model reduction using spline interpolation. In *SEG Technical Program Expanded Abstracts 2019*, pages 1400–1404. Society of Exploration Geophysicists.

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

Bozdağ, E., Trampert, J., and Tromp, J. (2011). Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophysical Journal International*, 185(2):845–870.

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

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

Capdeville, Y. and Métivier, L. (2018). Elastic full waveform inversion based on the homogenization method: theoretical framework and 2-d numerical illustrations. *Geophysical Journal International*, 213(2):1093–1112.

Cupillard, P. and Capdeville, Y. (2018). Non-periodic homogenization of 3-D elastic media for the seismic wave equation. *Geophysical Journal International*, 213(2):983–1001.

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

Dierckx, P. (1993). *Curve and surface fitting with splines*. Oxford Science Publications (Clarendon Press).

Fehrenbach, J. and Mirebeau, J.-M. (2014). Sparse Non-negative Stencils for Anisotropic Diffusion. *Journal of Mathematical Imaging and Vision*, 49(1):123–147.

Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H. P. (2009). Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods. *Geophysical Journal International*, 179(3):1703–1725.

Górszczyk, A., Brossier, R., and Métivier, L. (2021). Graph-space optimal transport concept for time-domain full-waveform inversion of ocean-bottom seismometer data: Nankai trough velocity structure reconstructed from a 1d model. *Journal of Geophysical Research: Solid Earth*, 126(5):e2020JB021504. e2020JB021504 2020JB021504.

Grote, M. J., Kray, M., and Nahum, U. (2017). Adaptive eigenspace method for inverse scattering problems in the frequency domain. *Inverse Problems*, 33(2):025006.

Guitton, A., Ayeni, G., and Díaz, E. (2012). Constrained full-waveform inversion by model reparameterization. *Geophysics*, 77(2):R117–R127.

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

Huang, G., Nammour, R., and Symes, W. W. (2018). Source-independent extended waveform inversion based on space-time source extension: Frequency-domain implementation. *Geophysics*, 83(5):R449–R461.

Lailly, P. (1983). The seismic inverse problem as a sequence of before stack migrations. In Bednar, R. and Weglein, editors, *Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia*, pages 206–220.

Lei, W., Ruan, Y., Bozdağ, E., Peter, D., Lefebvre, M., Komatitsch, D., Tromp, J., Hill, J., Podhorszki, N., and Pugmire, D. (2020). Global adjoint tomography—model glad-m25. *Geophysical Journal International*, 223(1):1–21.

Li, Y. E. and Demanet, L. (2016). Full-waveform inversion with extrapolated low-frequency data. *Geophysics*, 81(6):R339–R348.

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

Métivier, L. and Brossier, R. (2016). The SEISCOPE optimization toolbox: A large-scale nonlinear optimization library based on reverse communication. *Geophysics*, 81(2):F11–F25.

Métivier, L., Brossier, R., Mérigot, Q., and Oudet, E. (2019). A graph space optimal transport distance as a generalization of L^p distances: application to a seismic imaging inverse problem. *Inverse Problems*, 35(8):085001.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016). An optimal transport approach for seismic tomography: Application to 3D full waveform inversion. *Inverse Problems*, 32(11):115008.

Mirebeau, J.-M. (2016). Minimal stencils for discretizations of anisotropic pdes preserving causality or the maximum principle. *SIAM Journal on Numerical Analysis*, 54(3):1582–1611.

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

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, 2nd edition.

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

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

Operto, S., Miniussi, A., Brossier, R., Combe, L., Métivier, L., Monteiller, V., Ribodetti, A., and Virieux, J. (2015). Efficient 3-D frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: application to Valhall in the visco-acoustic vertical transverse isotropic approximation. *Geophysical Journal International*, 202(2):1362–1391.

Operto, S., Virieux, J., Dessa, J. X., and Pascal, G. (2006). Crustal imaging from multifold ocean bottom seismometers data by frequency-domain full-waveform tomography: application to the eastern Nankai trough. *Journal of Geophysical Research*, 111(B09306):doi:10.1029/2005JB003835.

Peters, B. and Herrmann, F. J. (2017). Constraints versus penalties for edge-preserving full-waveform inversion. *The Leading Edge*, 36(1):94–100.

Pladys, A., Brossier, R., Li, Y., and Métivier, L. (2021). On cycle-skipping and misfit function modification for full-wave inversion: Comparison of five recent approaches. *Geophysics*, 86(4):R563–R587.

Plessix, R. E. (2006). A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503.

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

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

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

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

Sirgue, L. (2003). *Inversion de la forme d'onde dans le domaine fréquentiel de données sismiques grand offset*. PhD thesis, Université Paris 11, France - Queen's University, Canada.

Stopin, A., Plessix, R.-E., and Al Abri, S. (2014). Multiparameter waveform inversion of a large wide-azimuth low-frequency land data set in Oman. *Geophysics*, 79(3):WA69–WA77.

Strong, D. and Chan, T. (2003). Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Problems*, 19:S165–S187.

Sun, H. and Demanet, L. (2020). Extrapolated full-waveform inversion with deep learning. *Geophysics*, 85(3):R275–R288.

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

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

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

Tikhonov, A., Goncharsky, A., Stepanov, V., and Yagola, A. (2013). *Numerical methods for the solution of ill-posed problems*. Springer Science & Business Media.

Trinh, P. T., Brossier, R., Métivier, L., Virieux, J., and Wellington, P. (2017a). Bessel smoothing filter for spectral element mesh. *Geophysical Journal International*, 209(3):1489–1512.

Trinh, P. T., Brossier, R., Métivier, L., Virieux, J., and Wellington, P. (2017b). Bessel smoothing filter for spectral element mesh. *Geophysical Journal International*, 209(3):1489–1512.

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

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

Virieux, J., Asnaashari, A., Brossier, R., Métivier, L., Ribodetti, A., and Zhou, W. (2017). An introduction to Full Waveform Inversion. In Grechka, V. and Wapenaar, K., editors, *Encyclopedia of Exploration Geophysics*, pages R1–1–R1–40. Society of Exploration Geophysics.

Wang, Y. and Rao, Y. (2009). Reflection seismic waveform tomography. *Journal of Geophysical Research*, 114(B3):1978–2012.

Warner, M. and Guasch, L. (2016). Adaptive waveform inversion: Theory. *Geophysics*, 81(6):R429–R445.

Weickert, J. (1998). *Anisotropic diffusion in image processing, Treubner Verlag, Stuttgart*. Treubner Verlag.

Wu, R. S. and Toksöz, M. N. (1987). Diffraction tomography and multisource holography applied to seismic imaging. *Geophysics*, 52:11–25.

Xue, Z., Alger, N., and Fomel, S. (2016). Full-waveform inversion using smoothing kernels. In *SEG Technical Program Expanded Abstracts 2016*, pages 1358–1363. Society of Exploration Geophysicists.

Yang, P., Brossier, R., Métivier, L., and Virieux, J. (2016). Wavefield reconstruction in attenuating media: A checkpointing-assisted reverse-forward simulation method. *Geophysics*, 81(6):R349–R362.

Yang, P., Brossier, R., Métivier, L., Virieux, J., and Zhou, W. (2018). A Time-Domain Preconditioned Truncated Newton Approach to Multiparameter Visco-acoustic Full Waveform Inversion. *SIAM Journal on Scientific Computing*, 40(4):B1101–B1130.

Yang, Y. and Engquist, B. (2018). Analysis of optimal transport and related misfit functions in full-waveform inversion. *Geophysics*, 83(1):A7–A12.