Geophysical Journal International

Geophys. J. Int. (2019) **216,** 130–147
Advance Access publication 2018 October 10
GJI Seismology

doi: 10.1093/gji/ggy419

An optimal transport approach to linearized inversion of receiver functions

N. Hedjazian,[1] T. Bodin[1] and L. Métivier[2]

[1] *Université de Lyon, UCBL, CNRS, LGL-TPE, 69622 Villeurbanne, France. E-mail:* navid.hedjazian@univ-lyon1.fr
[2] *Université Grenoble Alpes, CNRS, ISTerre, LJK, 38058 Grenoble, France*

Accepted 2018 October 9. Received 2018 October 1; in original form 2018 June 18

SUMMARY

Receiver function analysis is widely used to make quantitative inferences about the structure below a seismic station. As these observables are mainly sensitive to traveltimes of phases converted and reflected at seismic discontinuities, the resulting inverse problem is highly non-linear, the solution non-unique, and there are strong trade-offs between the depth of discontinuities and absolute velocities. To overcome this difficulty, we propose to measure the misfit between the predicted and observed data with an optimal transport distance instead of the conventional least-squares distance, a strategy that has shown its assets in the context of full waveform inversion. This approach views a seismogram as a distribution of 'mass'. The optimal transport distance between two waveforms is the minimal cost of transporting one waveform onto the other. We test the optimal transport approach on the inversion of a radial *P*-wave receiver function. We also show how it can be applied to measure the cross-convolution distance between the radial and vertical components, thus avoiding the need for deconvolution associated with the calculation of the receiver function. The resulting misfit function is minimized with a local optimization algorithm to constrain the receiver-side structure. The benefits of this methodology are studied in simple synthetic tests and with real data. In particular, we show that with its increased sensibility to time-shifts, the optimal transport distance reduces the number of local minima in the misfit function, which, in the case of a linearized inversion, significantly reduces the dependency to the starting model and results in a better convergence towards the solution model. A joint inversion of the *P*-wave receiver function and surface wave dispersion curves is performed at the Hyderabad station in India.

Key words: Asia; Inverse theory; Joint inversion; Body waves; Surface waves and free oscillations.

1 INTRODUCTION

Scattered teleseismic body waves contain information about the Earth's structure beneath the receiver. Seismologists use them to image discontinuities (i.e. sharp changes in wave velocities) from the crust to the upper mantle (for a complete review, see Bostock 2015). The *P*-to-*S* (Ps) converted phases observed in the *P*-wave coda are probably the most used information for this purpose. At low incidence angles, the majority of the *P*-wave energy arriving at a station is contained in the vertical component of the seismogram, whereas radial and transverse components include information about these converted phases. A straightforward procedure to unravel Earth's structural response consists in deconvolving the vertical from the radial component, thus removing the effect of the source time function and the instrumental response, then in stacking many deconvolved traces to expose the energy of the converted phases (Vinnik 1977; Langston 1979). Due to the noise and the limited bandwidth of the signals, the deconvolution operation is unstable and requires regularization (Langston 1979). The resulting time-series is called the receiver function (RF). It allows to identify the arrival of the different converted phases and the amount of converted energy. In the case of a single station, this information can be directly translated in terms of a 1-D layered structure with a simple analytic formula based on traveltimes of *P* and *S* waves in a 1-D Earth. However, the depth of the discontinuity can be inferred only at the expense of a trade-off with the absolute *S*-wave velocity (Ammon *et al.* 1990; Zhu & Kanamori 2000). Therefore, RFs (or directly the scattered energy) are usually migrated to depth with the use of a reference *S*-wave velocity model (e.g. Kosarev *et al.* 1999; Bostock 2015). The computational cost is low and, when a sufficiently dense array is available, this kind of method is able to image lateral or even 3-D structures (Ryberg & Weber 2000; Rondenay 2009; Cheng *et al.* 2017). However, two limitations are the necessary choice of a reference background velocity model and the difficulty to interpret multiple reflected phases. Also, migration schemes do not allow to quantify the seismic wave velocities.

1.1 Inversion of receiver functions

To obtain a more quantitative estimation of seismic properties below the station, the RF can be used as an observable for an inversion procedure. Here, we focus on the inversion of the radial component of P-wave RFs. A synthetic RF is generated for a particular candidate earth model and compared to the data. From there, the real Earth properties can be estimated with various techniques developed in conventional inverse problem theory. For example, early studies use the least-squares difference (referred to as L^2 distance hereafter) between predicted and observed RF as a misfit function and then minimize this misfit with an iterative linearized inversion procedure (Owens *et al.* 1984; Kind *et al.* 1995). The non-uniqueness of the solution resulting from the depth–velocity trade-off is usually addressed by introducing additional data such as surface-wave dispersion curves, thereby enhancing the sensitivity to the absolute S-wave velocity (Julià *et al.* 2000, 2003). This is now a well-established method which can be applied at large scale. The structure below an array of receivers is inferred by multiple 1-D inversions and eventually spatial interpolation between the resulting profile locations (Yoo *et al.* 2007; Sosa *et al.* 2014; Ward *et al.* 2014; Guo *et al.* 2015). While iterative least-squares procedures are quite popular, the RF depends non-linearly on seismic properties at depth, and the least-squares misfit function presents several local minima, leading to a strong dependence to the chosen initial model. These approaches also require some damping or smoothing to ensure stability.

In order to avoid getting trapped in local minima, an alternative is the use of global optimization techniques, mostly based on Monte Carlo sampling algorithms (Shibutani *et al.* 1996; Zhao *et al.* 1996; Vinnik *et al.* 2004). Furthermore, instead of giving a single best-fitting solution, they can be extended to sample the space of possible earth models and provide an ensemble of solutions that fit the data reasonably well. The problem can be formulated in a Bayesian framework, and algorithms performing importance sampling (e.g. Markov-Chain Monte Carlo) can be used to produce a model distribution proportional to the *a posteriori* probability distribution, which allows to quantify uncertainties on the estimated parameters (Sambridge 1999; Piana Agostinetti & Malinverno 2010; Shen *et al.* 2012; Dettmer *et al.* 2015). However, going towards more sophisticated methods of inference has also its drawbacks. While they extract the maximum information from the scattered teleseismic waves, Bayesian inversion methods based on sampling algorithms are computationally intensive, prone to convergence issues and difficult to implement. Studies have been up to date mainly limited to single station 1-D inversions (Piana Agostinetti & Malinverno 2010; Bodin *et al.* 2014; Dettmer *et al.* 2015). A computationally viable approach for moderate-size seismic networks is to relax some of the conditions in the Bayesian methodology, at the expense of a less complete uncertainty estimation (Shen *et al.* 2012; Kim *et al.* 2016).

In this work, we propose to address the issues of local minima and trade-offs by using an alternative misfit function. This function contains less local minima, and thus allows us to use a conventional linearized inversion, with minimal dependence on the initial model.

1.2 Optimal transport

The existence of local minima in the misfit function and the dependence of the result to the initial model are common to most applications of waveform tomography. This issue takes root in the definition of the misfit criterion. The L^2 distance is not suited to compare oscillatory signals such as seismograms because it performs only local point-to-point comparisons. It is highlighted in the phenomenon referred to as cycle skipping, when the predicted signal matches the data with one or several phase shifts. In applications where the high cost of solving the forward problem hinders the use of global optimization methods, this issue is particularly restricting. In this context, more suitable strategies to compare seismograms have been studied. For example, the multiscale frequency approach performs successive inversions using an increasing frequency content in the data (Bunks *et al.* 1995). Another recurring idea in seismology is to separate phase and amplitude information of the signal (Gee & Jordan 1992; Fichtner *et al.* 2008; Bozdağ *et al.* 2011), which enhances the linearity with respect to model parameters. To our knowledge, such alternative misfit strategies have not received much attention in RF inversion, although the RF possesses most properties of a seismogram.

Instead of the least-squares criterion, we propose to use a distance based on optimal transport to compare predicted and observed RFs. More specifically, we follow the same strategy as Métivier *et al.* (2016b) that has been successfully applied to time-domain full-waveform inversion (Métivier *et al.* 2016a). This approach is likely to be beneficial for RF inversion because it increases the sensitivity to time-shifts, and hence to traveltimes of converted and reflected phases.

The optimal transport problem consists of moving mass units from one set of locations to another with minimal effort. Recent advances in the numerical solution of this problem has enabled numerous applications in imaging methods (e.g. Lellmann *et al.* 2014). For our part, it is used to design new misfit functions to compare seismograms (Engquist & Froese 2014; Métivier *et al.* 2016b). The idea is to define the distance between two distributions as the solution to the optimal transport problem. Such a distance is called the Wasserstein distance. In this frame, a seismogram is considered as a discrete distribution of mass. The time axis is taken as equivalent to a spatial axis and the amplitude as the unit of mass. A transport plan is a sequence of displacements required to transport the mass units from one seismogram to another. A cost is associated with each displacement of a unit mass, with respect to some pre-defined distance on the time axis. For instance, Engquist & Froese (2014) chose the L^2 distance whereas Métivier *et al.* (2016b) use the L^1 distance. The cost associated with a transport plan is the sum, for all the required displacements, of the amount of mass displaced multiplied by the distance along which it is displaced. The Wasserstein distance is defined as the minimal cost over all possible transport plans. Defined this way, the misfit function provides a natural way to takes into account information in the time-shift and the amplitude difference between the two signals. In particular, it should be convex with respect to time-shifted patterns.

Several difficulties arise to compute the Wasserstein distance between seismograms. First, the two distributions are required to be positive. This is not the case for seismograms, where the amplitude can have an arbitrary sign. To solve this issue, Engquist & Froese (2014) consider the positive and negative parts of the seismogram separately, or add a constant to observed and predicted data to produce positive observables. Alternatively, Métivier *et al.* (2016b) work with a particular instance of the Wasserstein distance (the 1-Wasserstein distance), which can be easily extended to the comparison of non-positive signals. Another constraint is the conservation of mass between the two distributions. This is easily solved in the case of seismograms by demeaning the time signal. The third difficulty is associated with the computational cost of the solution of the transport problem. The formulation of Métivier *et al.* (2016b) is also interesting in this respect, as an algorithm with quasi-linear complexity (relative to the number of data in the time signal) is derived. This is very similar to the cost of computing the least-squares difference between RFs and the computation overhead is limited.

1.3 Outline

In this study, we construct a methodology to apply this optimal transport distance to RF inversion. The most simple implementation is to replace the L^2 distance by the Wasserstein distance. Its properties are compared to the least-squares misfit properties in the case of an inversion for a synthetic 1-D Earth structure. We show in particular that the inversion with the optimal transport distance is less sensitive to the starting model. In a second step, the methodology is applied to real waveforms recorded at the Hyderabad station in India. This is a well-studied station where the simple underlying structure satisfies the 1-D hypothesis. We can thus compare our results with similar studies in the region (Saul *et al.* 2000; Kumar *et al.* 2007; Kiselev *et al.* 2008; Julià *et al.* 2009; Oreshin *et al.* 2011; Bodin *et al.* 2014; Dettmer *et al.* 2015; Singh *et al.* 2015).

2 MISFIT FUNCTIONS FOR SCATTERED TELESEISMIC *P*-WAVE CODA

2.1 Least-squares distances

We denote the *P*-wave coda of a scattered teleseismic body wave on the radial and vertical components in the time domain as [$V(t)$, $R(t)$]. It can be considered as the convolution of the receiver-side structure impulse response [$v(t)$, $r(t)$] with an effective source time function $s_{\text{eff}}(t)$, that is, encompassing the effect of the source, path and instrument response:

$$\begin{aligned} R(t) &= r(t) * s_{\text{eff}}(t) + \epsilon_{\text{r}}(t) \\ V(t) &= v(t) * s_{\text{eff}}(t) + \epsilon_{\text{v}}(t), \end{aligned} \tag{1}$$

where ε_{r} an ε_{v} account for the noise on each component.

To construct the corresponding observed receiver function, $\text{RF}_{\text{obs}}(t)$, the effect of the effective source time function is removed with the deconvolution procedure of Langston (1979). The radial component R is divided by the vertical component V in the frequency domain with a water-level stabilization and a Gaussian filtering to remove high-frequency noise:

$$\text{RF}_{\text{obs}}(\omega) = \frac{R(\omega)V^*(\omega)}{\max\{V(\omega)V^*(\omega), \text{c}.|\text{V}|^2_{\max}\}} \exp\left(\frac{-\omega^2}{4a^2}\right), \tag{2}$$

where $|V|^2_{\max} = \max\limits_{\omega}(V(\omega)V^*(\omega))$ is the maximal value of V power spectra. In the following, we chose a water-level value $c = 0.005$ and a width factor for the Gaussian filter $a = 2$.

In order to invert the observed RF, the synthetic impulse response $r(t, \mathbf{m})$ and $v(t, \mathbf{m})$ for a given 1-D earth model $\mathbf{m}$ needs to be calculated. Those are obtained with a reflectivity Thomson–Haskell propagator-matrix method (Haskell 1962). The transmission response of a stack of isotropic, homogeneous layers is calculated in the Fourier domain for an incoming planar *P* wave, at a number of different frequencies. Surface displacements are obtained by an inverse Fourier transform. The predicted synthetic $\text{RF}_{\text{synth}}(t, \mathbf{m})$ receiver function for a given earth model $\mathbf{m}$ is calculated with the same deconvolution method, but using [$v(t, \mathbf{m})$, $r(t, \mathbf{m})$] instead of observed [$V(t)$, $R(t)$].

We study two possible misfit functions using the L^2 norm. We refer to the least-squares difference between observed and predicted RFs as ϕ_0. We also propose an alternative misfit which directly considers the *P*-wave coda on the vertical and radial components and computes the cross-convolution between the observed and predicted signals. This avoids the deconvolution and, thus, the use of two tuning parameters required to calculate a RF (Menke & Levin 2003; Bodin *et al.* 2014). This misfit function, referred to as ϕ_1, is defined as the least-squares distance between the two terms of the cross-convolution:

$$\phi_0(\mathbf{m}) = \| RF_{\text{synth}}(t, \mathbf{m}) - RF_{\text{obs}}(t) \|^2_{L^2}, \tag{3}$$

$$\phi_1(\mathbf{m}) = \| v(t, \mathbf{m}) * R(t) - r(t, \mathbf{m}) * V(t) \|^2_{L^2} . \tag{4}$$

2.2 Optimal transport distances

The solution of the optimal transport problem can be used to define a distance between two distributions. Let us consider two real-valued functions f, g defined on a subset X of $\mathbb{R}^n$. $f(x)$ is the amount of mass at position $x \in X$. As we will consider only 1-D seismograms, the dimension of the subset is $n = 1$, but we keep the formulation general.

In the classic formulation of the optimal transport problem, f and g need to be positive and to have the same total mass, that is, $\int f(x)\text{d}x = \int g(y)\text{d}y$. The ensemble $\mathcal{M}$ of mappings rearranging f into g is the ensemble of functions $T\colon\ X \rightarrow X$, called transport maps, verifying:

$$\forall A \subset X, \quad \int_{x \in A} g(x)\, \text{d}x = \int_{T(x) \in A} f(x)\text{d}x. \tag{5}$$

The aim is to find the transport map T requiring minimal effort. The displacement cost of a unit mass is here chosen as an L^p norm on $\mathbb{R}^n$. We can define the L^p Wasserstein distance as

$$W_p(f, g) = \left(\min_{T \in \mathcal{M}} \int_{x \in X} \| x - T(x) \|^p_{L^p} \; f(x)\text{d}x \right)^{\frac{1}{p}} . \tag{6}$$

The idea of applying this distance to seismic signals has first been promoted by Engquist & Froese (2014), using the W_2 Wasserstein metric. As seismograms can have an arbitrary sign, their formulation is based on a separation of the positive and negative part of the data. This choice ensures the convexity of the distance with respect to both time-shifts and variations in amplitude between the two time signals. Besides, it appears to be robust relative to noise. However, it is not straightforward to apply in the frame of inverse problems through local optimization, as the extraction of positive and negative parts of the data are not differentiable operations.

Alternatively, Métivier *et al.* (2016b) propose to use the W_1 Wasserstein metric. In this case, the cost associated with a displacement from x to y is the sum of absolute differences:

$$\| x - y \|_{L^1} = \sum_{i=1}^{n} |x_i - y_i|. \tag{7}$$

The main interest is the possibility to extend W_1 to the comparison of signed signals, through its associated dual form (e.g. Villani

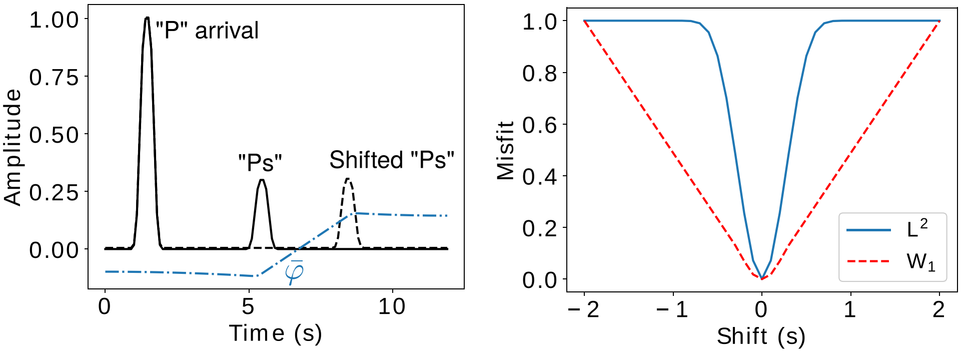

Figure 1. Comparison of the L^2- and W_1-based misfit functions for the two shifted converted phases on the left-hand panel. The blue dotted line shows $\bar{\varphi}$, the function φ achieving the maximum criterion of eq. (8). On the right-hand panel, the L^2 misfit (blue line) shows no dependence to the time-shift Δt. The W_1 misfit (dashed red) scales as $O(\Delta t)$.

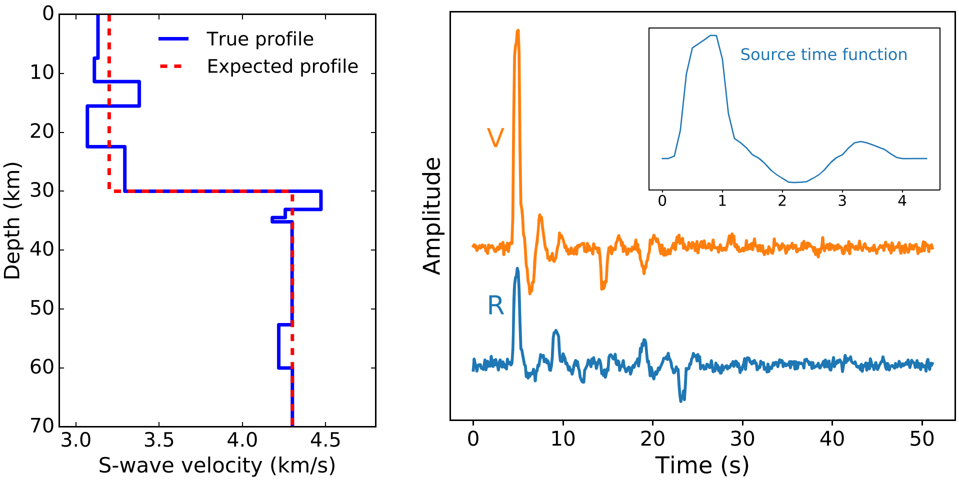

Figure 2. Synthetic true model and the corresponding synthetic waveform data. The true model corresponds to the V_S profile in blue solid line. The waveforms are constructed by convolving the structural response to the noisy source time function and adding white Gaussian noise. We try to invert these data for a two-parameter model (d_1, V_{S1}) similar to the dashed-red V_S profile.

2008). The corresponding formulation is

$$\begin{cases} W_1(f,g) = \max\limits_{\varphi \in \mathrm{Lip}_1(X)} \int_{x\in X} \varphi(x)\,(f(x)-g(x))\,\mathrm{d}x, \\ \mathrm{Lip}_1(X) = \{\varphi,\ \forall (x,y) \in X,\ |\varphi(x)-\varphi(y)| \leq \| x-y \|_{L^1}\}\,. \end{cases} \tag{8}$$

In this case, f and g do not require to be positive. We ensure the conservation of mass by imposing f and g to have zero mean.

A quasi-linear complexity algorithm is proposed in Métivier *et al.* (2016b) to solve the problem, for 1-D, 2-D and 3-D seismograms. The algorithm relies on the simplification of this problem to a linear programming problem with $O(N)$ constraints, where N is the number of discrete samples in the time signal.

To illustrate the advantages of W_1 compared to L^2, we show an example revealing their properties relative to a time-shift Δt between two simple signals. We mimic a hypothetical RF up to the first converted phase, where the Ps arrival phase is an energy pulse having a reduced amplitude and a time delay relative to the first arrival. It is compared with a second signal where the Ps pulse arrives with a varying delay. The shape of the L^2 and W_1 misfit functions with respect to the time-shift between both energy pulses is shown in Fig. 1. For a sufficiently large Δt (greater than the energy pulse width), the L^2 misfit function becomes independent of Δt. On the contrary, the W_1 misfit scales as $O(\Delta t)$ over the whole range of time-shifts. As a consequence, the W_1 misfit will enlarge the domain of convergence of local minimization algorithms towards the desired global minimum.

We use the W_1 metric to define two additional misfit functions ϕ_2 and ϕ_3. As with the L^2 norm, W_1 is applied both on the predicted and observed RFs and on the cross-convolution of predicted and observed waveforms:

$$\phi_2(\mathbf{m}) = W_1\left(\mathrm{RF}_{\mathrm{synth}}(t,\mathbf{m}),\ \mathrm{RF}_{\mathrm{obs}}(t)\right), \tag{9}$$

$$\phi_3(\mathbf{m}) = W_1\left(v(t,\mathbf{m}) * R(t),\ r(t,\mathbf{m}) * V(t)\right). \tag{10}$$

Note that each term in the misfit functions above have zero mean provided that R, V, r, v also have zero mean, which is a required hypothesis for the computation of the Wasserstein distance.

2.3 Inversion strategy

With a chosen misfit function ϕ, inference of Earth's seismic properties from observed waveforms is possible through an inversion procedure. For RFs, the non-linearity of the problem and the existence of local minima in the misfit function can be addressed by using global optimization such as genetic algorithms (Shibutani *et al.* 1996), simulated annealing (Zhao *et al.* 1996), or more

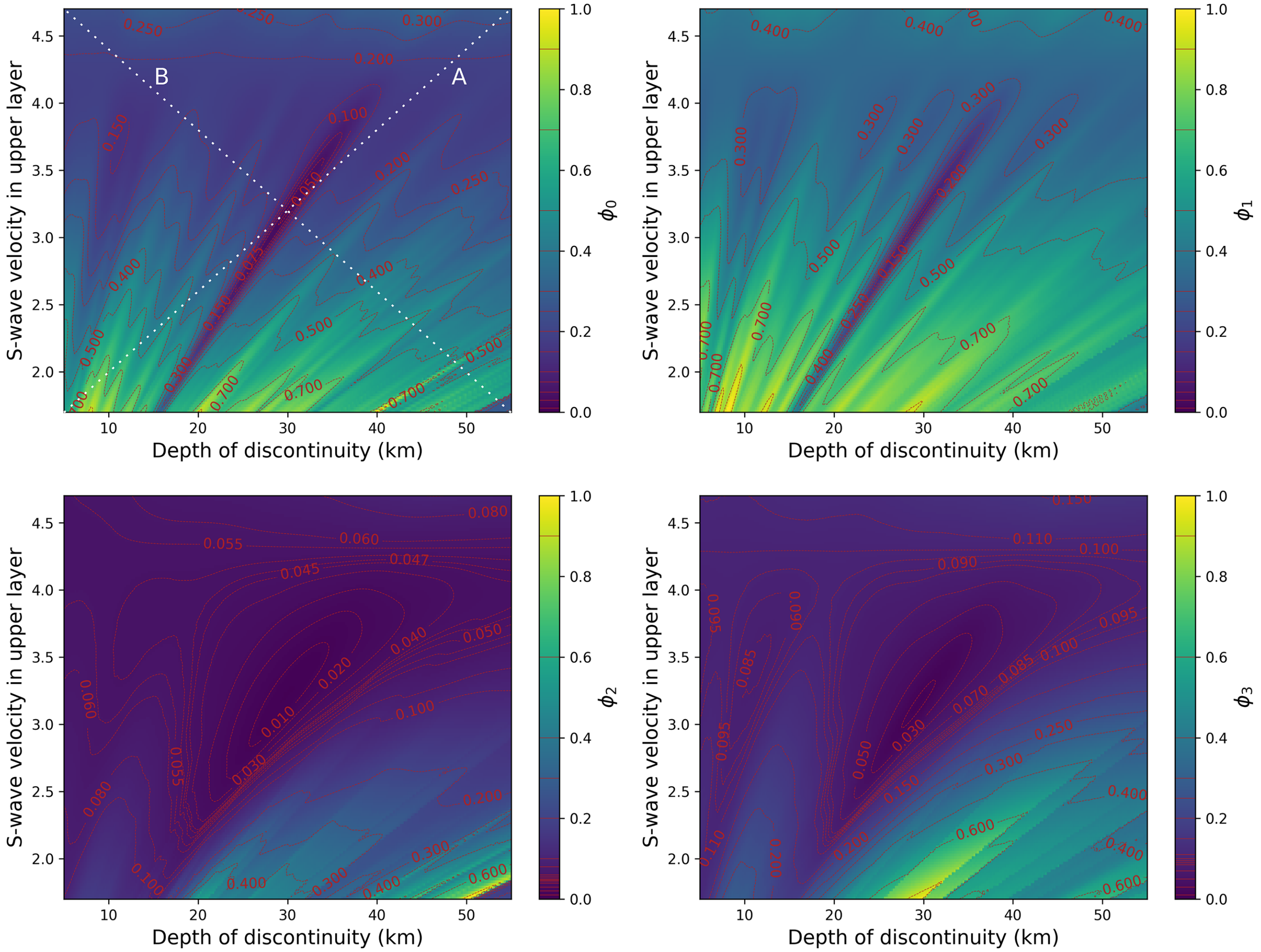

Figure 3. Representation of each misfit function as function of the discontinuity depth and the upper-layer S-wave velocity for the model of Fig. 2. ϕ_0 corresponds to the squared misfit between observed and predicted RFs, ϕ_1 to the cross-convolution distance between observed and predicted waveforms, ϕ_2 to the W_1 distance between RFs and ϕ_3 to the W_1 distance applied to the cross-convolution term. The expected model, with $d_1 = 30$ km and $V_{S1} = 3.2$ km s^{-1}, lies in the centre of each panel.

recently, probabilistic inversion methods based on Markov-Chain Monte Carlo (e.g. Piana Agostinetti & Malinverno 2010; Dettmer *et al.* 2015). These methods remain computationally expensive, as they are based on sampling approaches where the forward solution is calculated a large number of times. In this study, we show that changing the form of the misfit function reduces the number of local minima, thus allowing us to use gradient-based local optimization schemes.

We formulate the problem as an optimization problem, where the goal is to minimize the misfit $\phi(\mathbf{m})$ defining the distance between the data d_{obs} and some predicted observables $d_{\text{pred}} = g(\mathbf{m})$, where g is a non-linear function representing the forward model. As our intent is to compare the performance of different misfit functions, we use a general purpose algorithm that can be applied to all of them. The standard approach to locally minimize a non-linear misfit function is the quasi-Newton method. Starting with an initial model $\mathbf{m}_0$, it constructs a sequence of models $\mathbf{m}_i$ converging towards a local minimum of the misfit function. At step i, a descent direction $\Delta\mathbf{m}_i = -Q_i\nabla\phi(\mathbf{m}_i)$ is computed using information from the misfit function gradient $\nabla\phi(\mathbf{m}_i)$ and an approximation of the inverse of its Hessian Q_i. The descent step size α_i is chosen through a line search strategy (e.g. Bonnans *et al.* 2006; Nocedal & Wright 2006). The model update is computed as:

$$\mathbf{m}_{i+1} = \mathbf{m}_i + \alpha_i \Delta\mathbf{m}_i. \tag{11}$$

In this work, we calculate the gradient of the misfit function numerically with finite differences. The number of function evaluations will therefore increase linearly with the number of parameters in $\mathbf{m}$. Subsequently, the inverse Hessian is estimated with the l-BFGS algorithm. Specifically we use the implementation of Zhu *et al.* (1997) aimed at bound-constrained problems.

A more efficient method to evaluate the gradient of the misfit function is to derive $\phi(\mathbf{m})$ explicitly with respect to $\mathbf{m}$ by means of the chain rule. This strategy could be employed for each of the misfit functions ϕ_{i}. Here, we explicit the mathematical developments for the specific cases of the cross-convolution and the Wasserstein distance. We consider the modelling operator $g(\mathbf{m}) = [g_1(\mathbf{m}), g_2(\mathbf{m})] = [v(t, \mathbf{m}), r(t, \mathbf{m})]$ and define the cross-convolution as $w(\mathbf{m}, t) = v(t, \mathbf{m})*R(t) - r(t, \mathbf{m})*V(t)$. For the cross-convolution measure and

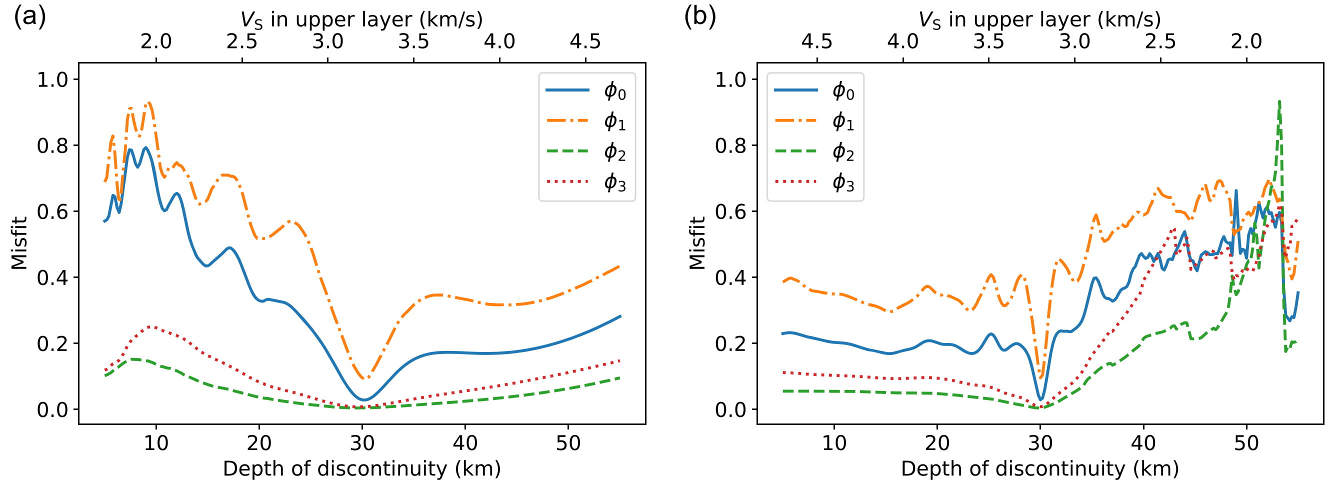

Figure 4. Cross-sections of each misfit function in the parameter space, along the directions denoted A and B in Fig. 3. The Wasserstein distance improves the convexity of the misfit function. Note, however, that the local minima along these directions are not necessarily local minima in the complete parameter space.

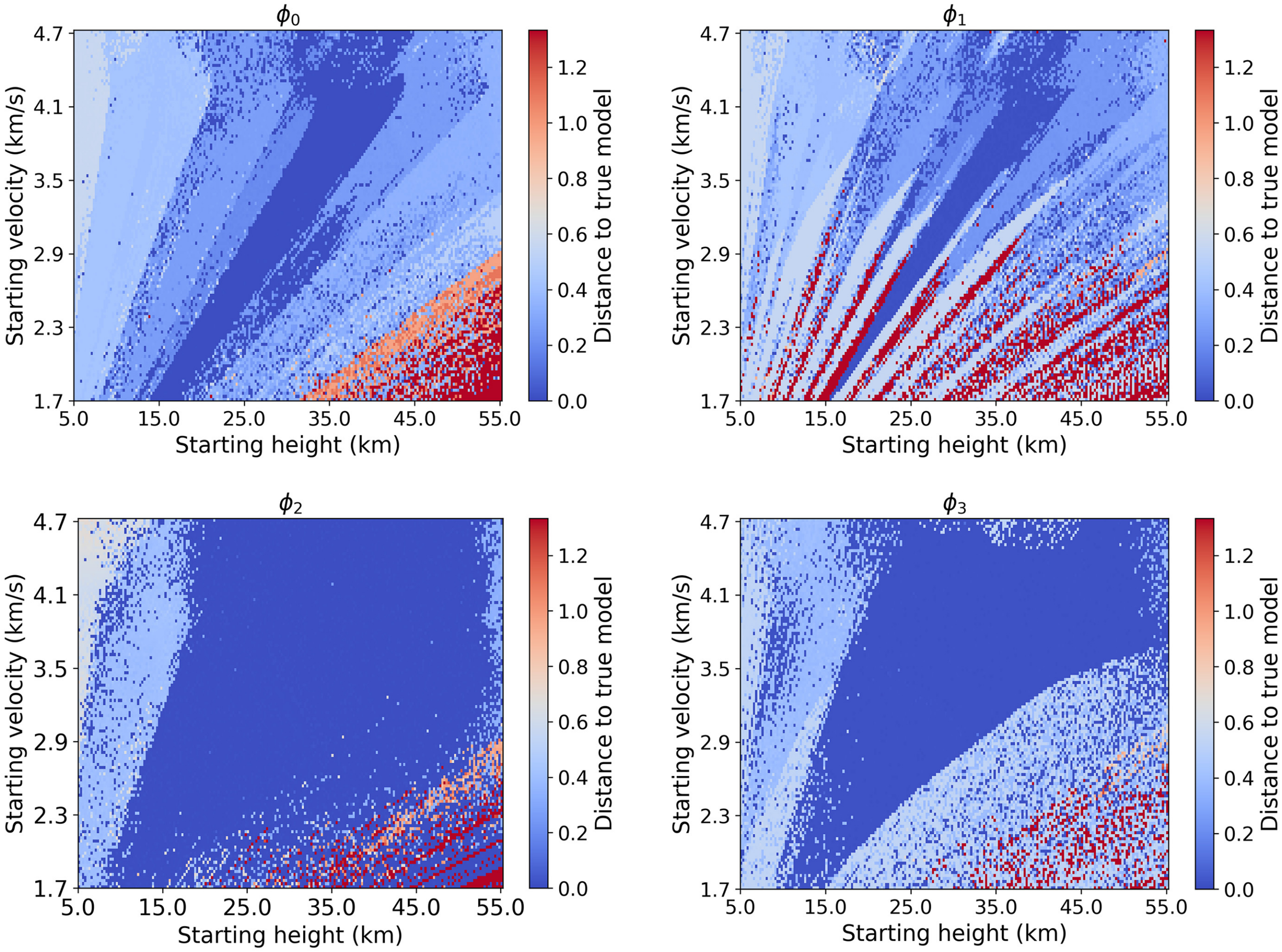

Figure 5. Attraction basin corresponding to each misfit function. For each couple of starting parameters $d_{1,\,i}$, $V_{\mathrm{S1,\,i}}$ on a regular grid, we undertake an inversion using the gradient-based local optimization method to obtain a final model. The colour in each pixel of the grid depicts the distance between the final model and the true model (calculated as a squared distance between V_{S} profiles). Hence the dark blue region corresponds to the starting models for which the inversion algorithm converges to the correct solution. W_1-based distances perform much better than L^2-based ones.

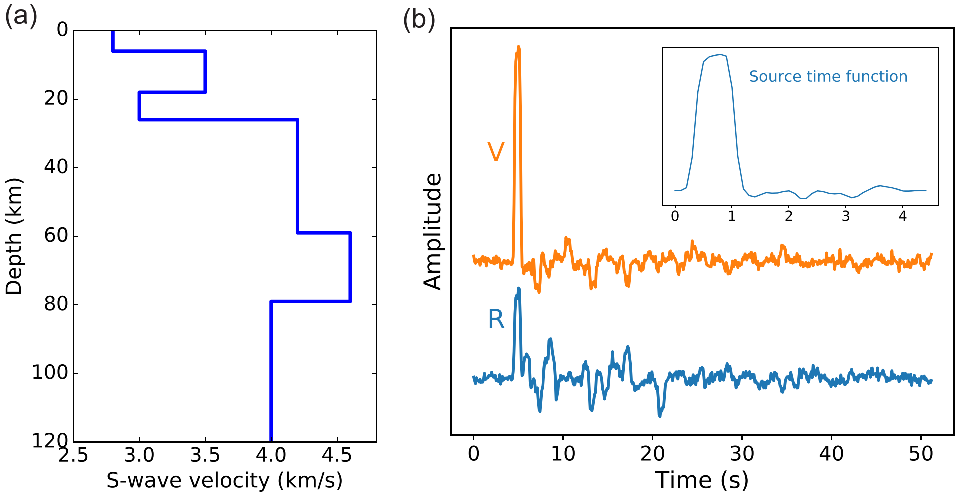

Figure 6. Synthetic six-layer model and its associated radial (blue) and vertical (orange) waveforms considered as observed data. Here, the waveforms are created by convolving the structural response with the source time function and by adding 3 per cent of Gaussian noise. Subsequently, we invert the data for a six-layer model, that is, for 12 parameters.

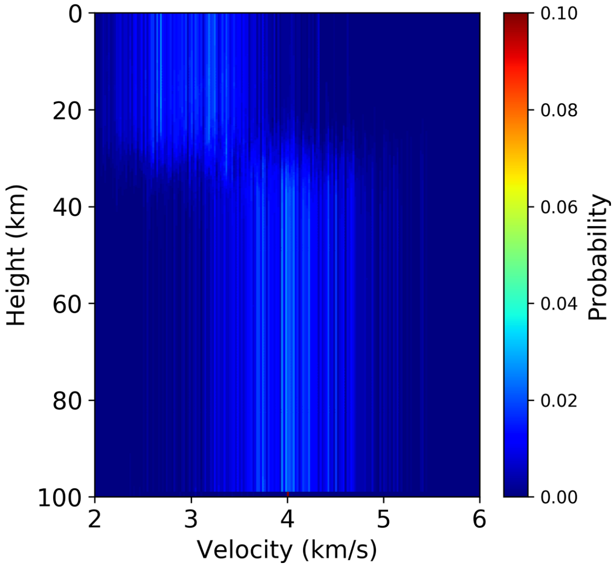

Figure 7. Ensemble of starting models for the inversion. We create 1000 random starting models following a distribution centred on a two-layer reference model with a strong discontinuity at 30 km depth.

the L^2 norm (which corresponds to ϕ_1), the gradient of the misfit function has been derived by Menke (2017):

$$\nabla\phi_1(\mathbf{m}) = 2\frac{\partial g_1}{\partial m}^T R \star w(\mathbf{m}) - 2\frac{\partial g_2}{\partial m}^T V \star w(\mathbf{m}), \tag{12}$$

where ^{T} denotes the adjoint operator and $\star$ denotes the cross-correlation operator. For the W_1 distance case, we have $\phi_3(\mathbf{m}) = \max_{\varphi\in\mathrm{Lip}_1(X)} \int_X \varphi w(\mathbf{m})$. We introduce

$$\bar{\varphi}(\mathbf{m}) = \underset{\varphi\in\mathrm{Lip}_1(X)}{\mathrm{argmax}} \int_X \varphi w(\mathbf{m}), \tag{13}$$

where the space X is here the time window of the seismogram. The function $\bar{\varphi}$ is the function φ that achieves the maximum criterion of eq. (8). The algorithm given by Métivier *et al.* (2016b) returns both the value of the W_1 distance and $\bar{\varphi}$, at no additional cost. The gradient of ϕ_3 is:

$$\nabla\phi_3(\mathbf{m}) = \frac{\partial g_1}{\partial m}^T R \star \bar{\varphi}(\mathbf{m}) - \frac{\partial g_2}{\partial m}^T V \star \bar{\varphi}(\mathbf{m}). \tag{14}$$

We give the details of the calculations in the Appendix. The function $\bar{\varphi}$ can be interpreted as the W_1 equivalent of the L^2 residual error. For example, in an adjoint state strategy, $\bar{\varphi}$ is the source term of the adjoint equation. The gradient of the modelling operator g, in the case of a reflectivity propagator-matrix method, can be obtained at low computational cost in comparison with finite differences of $\phi(\mathbf{m})$, for both RFs and waveforms (e.g. Randall 1994; Hu & Zhu 2017a,b).

We shall highlight here that the final model obtained after convergence of the quasi-Newton minimization will be an approximation of the true model, but no uncertainties will be available about this solution model. Uncertainties on inferred parameters can usually be assessed with a statistical formulation of the inverse problem. However, to our knowledge, the Wasserstein distance cannot be used to infer uncertainties on model parameters.

In waveform inversion, one usually make the hypothesis of additive noise: the statistical model of the data is $\mathbf{d} = g(\mathbf{m}) + \mathbf{e}$, where $\mathbf{e}$ is a random noise. The likelihood function is defined as $L(\mathbf{m}) = p(\mathbf{d} \mid \mathbf{m})$, and is related to the noise probability density (Kaipio & Somersalo 2006):

$$L(\mathbf{m}) = p(\mathbf{e} \mid \mathbf{m}) = p_{\mathrm{noise}}[\mathbf{d} - g(\mathbf{m})], \tag{15}$$

assuming that $\mathbf{m}$ and $\mathbf{e}$ are mutually independent. Therefore, the mathematical form of the log-likelihood function $\log L(\mathbf{m})$ (the Bayesian analogue of the misfit function) directly stems from the statistical model of the data and the assumed distribution of noise. For example, if $\mathbf{e}$ is the measurement error, the hypothesis of Gaussian and independent noise leads to a log-likelihood function defined as the least-squares distance between observed and predicted data.

It is uncertain which hypothesis on the statistical model and the noise distribution will result in a log-likelihood corresponding to the W_1 distance. This is by no means a limitation to use W_1 with global optimization methods. Simply, the final ensemble of models is not ensured to be proportional to the *a posteriori* probability distribution $p(\mathbf{m} \mid \mathbf{d})$. Similarly, as discussed by Dettmer *et al.* (2015), the exponential of the negative of cross-convolution misfit function

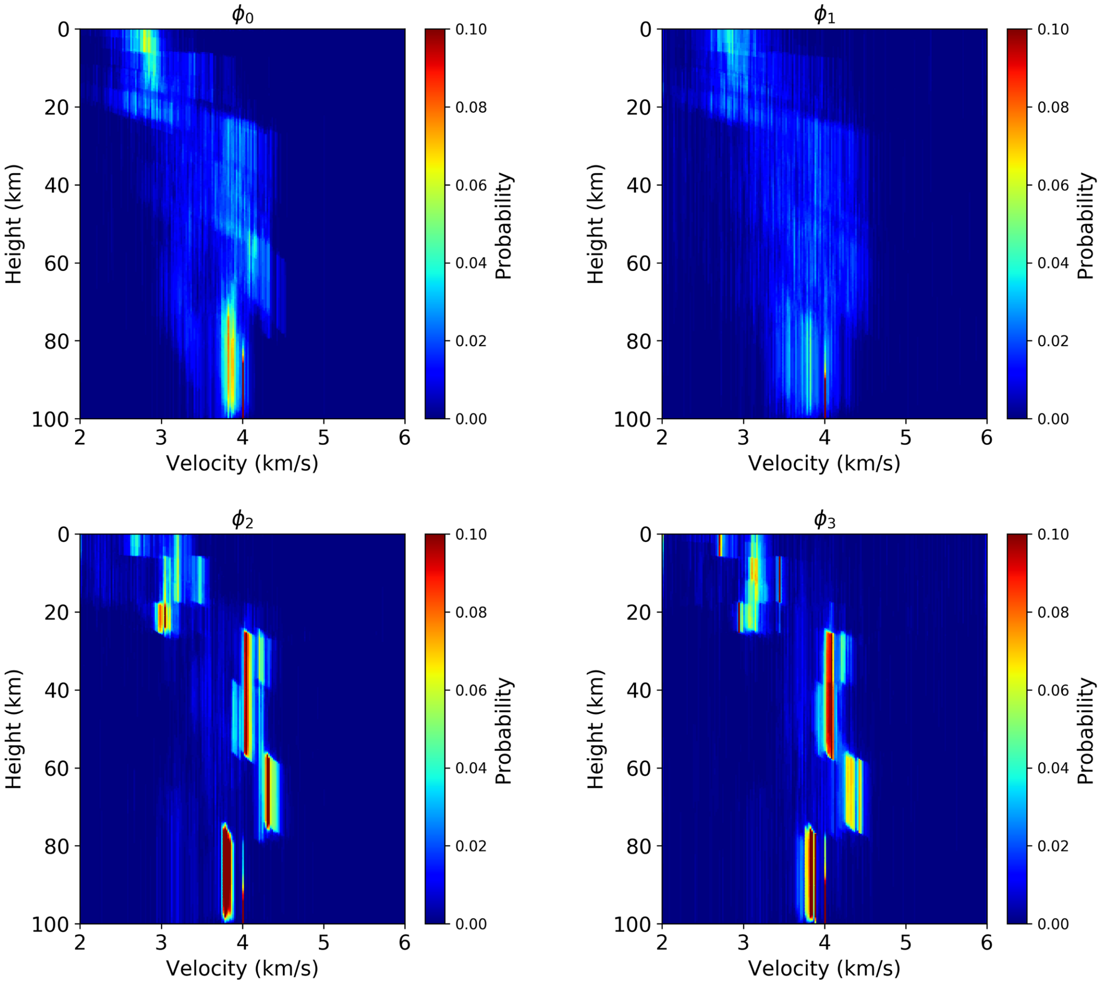

Figure 8. Resulting ensemble of solution models of a linearized inversion starting at different points in the model space, displayed as V_S profiles, for each misfit function ϕ_0 to ϕ_3. The distances ϕ_2 and ϕ_3, based on the Wasserstein distance, show the least dependence to the initial model and are able to recover more features from the true model.

ϕ_1, as used by Bodin *et al.* (2014) or Eilon *et al.* (2018), is not equal to $p(\mathbf{d} \mid \mathbf{m})$, and thus does not represent a proper likelihood function for uncertainty estimation.

3 SYNTHETIC TESTS

We construct synthetic 1-D horizontally layered Earth models where the seismic properties are isotropic and homogeneous in each layer. It is described as a vector of parameters $\mathbf{m}$ containing, for each layer, the *S*-wave velocity V_S and the depth of the lower interface (the last layer is a semi-infinite half-space). The ratio V_P/V_S is fixed, meaning that the *P*-wave velocity profile is always proportional to the V_S profile. Equally, the density of each layer is directly scaled from V_P using a simple analytic formula $\rho = 2.35 + 0.036(V_P - 3)^2$, as done in Tkalčić *et al.* (2006).

3.1 Recovering a simple two-layer model

In this first test, for each misfit function, we evaluate the ability of our quasi-Newton approach to recover a simple model with a single layer over a half-space. The two unknown parameters are the depth of the interface d_1 and the *S*-wave velocity of the upper layer V_{S1}. We construct a two-layer model having a discontinuity at 30 km depth. In order to introduce some realistic correlated noise in the seismic data, this *S*-wave velocity profile is perturbed in a 10-layered model (Fig. 2). Realistic waveform data are generated from eq. (1). The effective source time function is a Gaussian-filtered box-car function. To simulate the source-side effects and the instrumental response, it is subsequently convolved with an error transfer function T_{err} having unit spectral amplitude and a random phase between 0 and $\Pi/4$ at each frequency [following the methodology of Stähler & Sigloch (2016)]. Once the source is convolved with the structural response, we add 3 per cent of Gaussian noise (relative to the main P peak) to each waveform. The obtained (R_{obs}, V_{obs}) signals act as observed data for the inversion (Fig. 2).

Before comparing inversion results, we first show the shape of each misfit function with respect to the two model parameters. We compute the level of fit over a uniform grid for d_1 and V_{S1} while keeping the velocity of the half-space fixed at its true value (Fig. 3). The expected model, with $d_1 = 30\,\mathrm{km}$ and $V_{S1} = 3.2\,\mathrm{km\,s^{-1}}$, lies at the centre of each panel. The four misfit functions tested here feature some classical attributes related to the RF problem: they show a trade-off between d_1 and V_{S1} and present several local minima. However, the two misfit functions based on the optimal transport distance seem to bring about an improvement on these

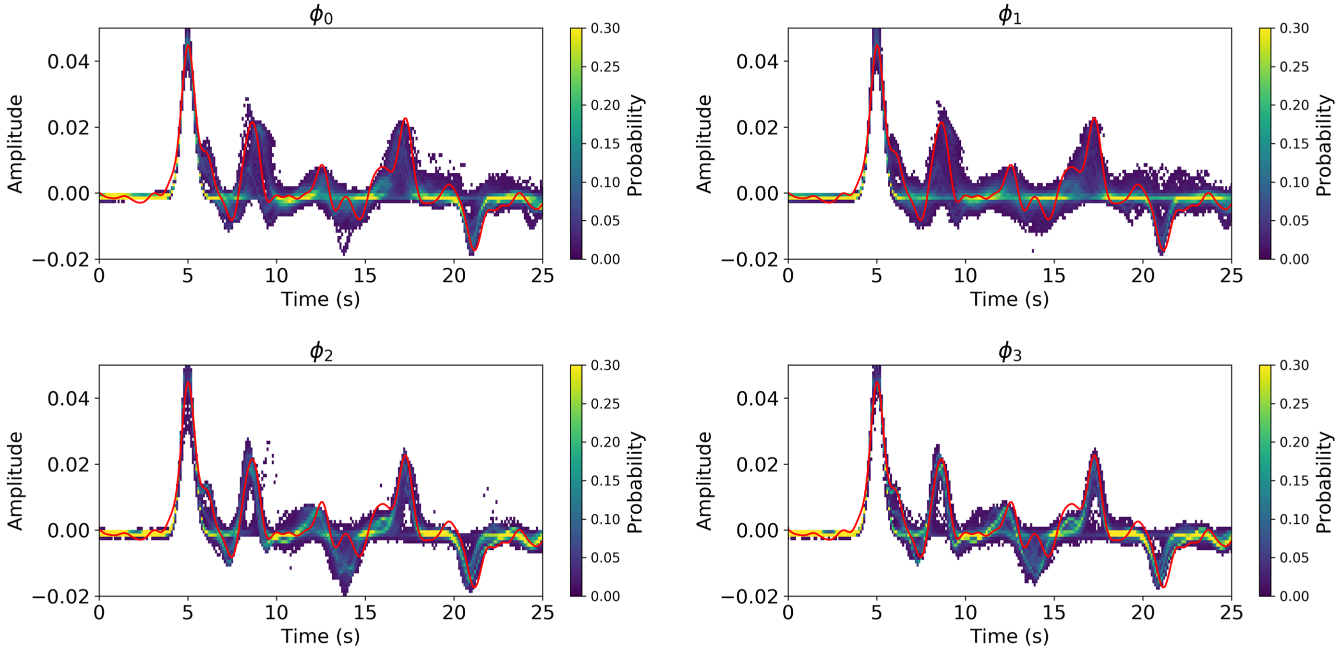

Figure 9. Ensemble of predicted receiver functions corresponding to the final models obtained in Fig. 8, calculated using a Gaussian width $a = 2$. The red line represents the true receiver function, obtained from the radial and vertical data components of Fig. 6.

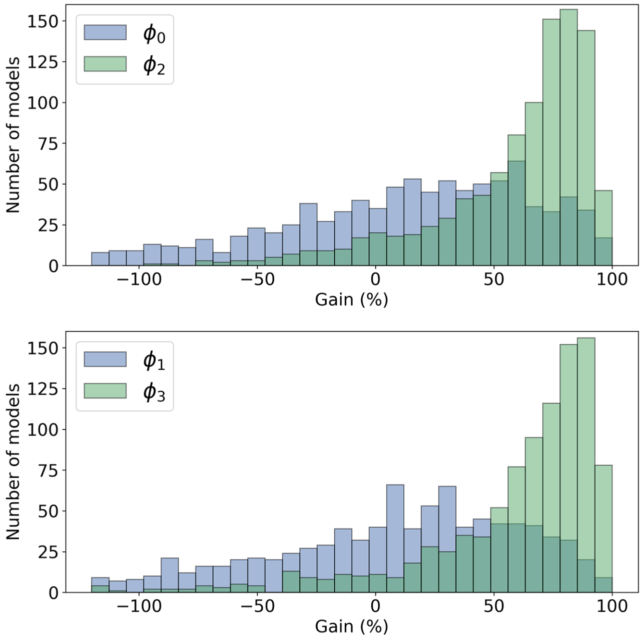

Figure 10. Measure of the minimization algorithm convergence for each inversion and each misfit function, defined as $100 \times \left(\parallel \mathbf{m}_\mathrm{i} - \mathbf{m}_\mathrm{true} \parallel^2 - \parallel \mathbf{m}_\mathrm{i} - \mathbf{m}_\mathrm{f} \parallel^2\right) / \left(\parallel \mathbf{m}_\mathrm{i} - \mathbf{m}_\mathrm{true} \parallel^2\right)$ (in per cent). Negative values indicate inversions where the final model is further away from the true model than the initial model. Values close to 100 per cent mean that the algorithm converges towards the correct answer.

two elements. At the global minimum, the local shape of the misfit function is described by its Hessian, and the Hessian can be interpreted in terms of local resolution analysis (Fichtner & Trampert 2011). Its diagonal elements characterize the local resolution of the model parameters, and its off-diagonal elements the trade-offs between the model parameters. The L^2-based misfit functions have the shape of a narrow canyon around the global minimum, corresponding to large off-diagonal elements in the Hessian. It reveals a significant depth/velocity trade-off but well-resolved parameters. On the contrary, the W_1-based misfit functions present a wider, more circular shape. In that respect, it reduces the trade-off between depth and velocity at the expense of less resolved parameters. Additionally, W_1 improves the convexity (i.e. it reduces the number of local minima) of the misfit function, as seen in the cross-sections of Fig. 4. As a consequence, it is more suitable for gradient-based minimization.

By further investigating the waveforms, we can explain two kinds of local minima inherent to the problem, which thus do not disappear when using an optimal transport distance. A flawed local minimum emerges when the PpPs phase of the predicted RF fits the Ps phase of the observed one. In Fig. 3, it can be seen by the narrow valley located at a value of d_1 roughly three times smaller than the true depth. A second type of flawed local minima is due to the finite length of the waveforms window. Indeed, when increasing the depth of the discontinuity or decreasing the velocity of the upper layer, some of the multiple reflected phases arrive after the end of the window and are not taken into account when calculating the misfit function anymore. This appears in Figs 3 and 4 through the abrupt jumps in value of the misfit functions, for a deep interface and a low S-wave velocity. It seems to be slightly more pronounced in the case of the optimal transport distance, probably because it considers a global rather than a point-to-point comparison of the signals. To alleviate this effect, we consider here waveforms that are longer than the required length to observe the Ps phase induced by the discontinuity. We observed that it could be further mitigated by tapering the signal at the end of the window. Anyhow, this issue is less substantial in real applications because the model space is larger and parameters vary on a narrower range during the search process.

To further quantify the improvement brought by the optimal transport distance to a standard linearized inversion process, we perform multiple inversions for the two model parameters using different

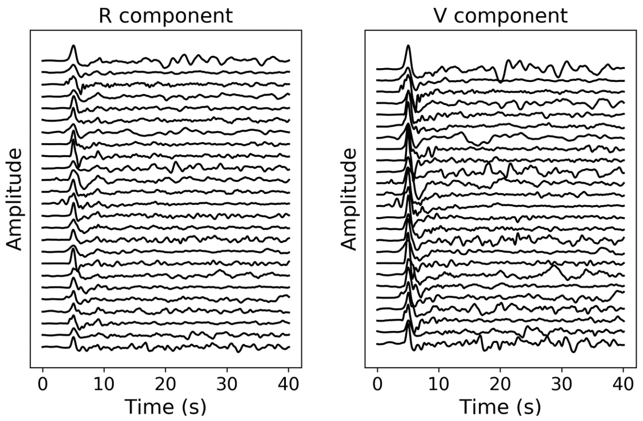

Figure 11. List of 25 waveforms used in the stacking process. They are lowpass filtered at 2 Hz and aligned using the time-shift determined by cross-correlation.

initial models $\mathbf{m}_i$, distributed on a regular grid. For each inversion, we calculate the distance between the final model and the true model as the least-squares difference between the two *S*-wave velocity profiles, then report it on the grid of $\mathbf{m}_i$ points as a colour map. Fig. 5 displays the results for each misfit function. The dark blue region corresponds to the ensemble of starting models from which our minimization algorithm succeeds in recovering the correct depth and velocity jump associated with the main discontinuity. The misfit functions using the Wasserstein distance clearly present a much wider attraction basin. On the contrary, the misfit functions based on L^2 norm recover the true model only if the starting model is close to the depth–velocity trade-off valley, meaning that the main phases of predicted and observed waveforms are already almost aligned.

3.2 Inverting for multiple parameters

We now test the quasi-Newton inversion with the four different objective functions on a problem with a larger number of model parameters and a more complicated structure. We consider the *S*-wave velocity profile depicted in Fig. 6. The synthetic data are generated from the true model using the same methodology as in the previous section. We invert for V_S and interface depths in a six-layer model, which corresponds to 12 unknown parameters (d_1, ..., d_6, V_{S1}, ..., V_{S6}). At depths >100 km the *S*-wave velocity is kept constant at $V_S = 4\,\mathrm{km\,s^{-1}}$. Here again, the minimization procedure requires a starting model $\mathbf{m}_\mathrm{i}$. To assess the dependence of each method to $\mathbf{m}_\mathrm{i}$, we generate a series of starting models using a random procedure. A reference initial model has one layer of 30 km thickness with $V_S = 3\,\mathrm{km\,s^{-1}}$ and one layer of 60 km thickness with $V_S = 4\,\mathrm{km\,s^{-1}}$ underneath. We generate an ensemble of starting models by perturbing the velocity and thickness of each layer of this reference model according to a Gaussian distribution with standard deviations of $0.5\,\mathrm{km\,s^{-1}}$ and 5 km, respectively. Each of the two layers are subsequently divided into three in order to get a six-layer model. The resulting ensemble of starting models, Fig. 7, displays a distribution centred around the reference model.

For every starting model, we use the quasi-Newton method described earlier to minimize the different types of misfit functions. During the minimization, the velocity parameters are constrained between 2 and $6\,\mathrm{km\,s^{-1}}$ and the depth of each interface has to be shallower than 100 km depth. We obtain an ensemble of final models for each misfit function, plotted as distributions in Fig. 8, whereas Fig. 9 displays the distribution of predicted RFs for each final model. Note here that the distribution of final models does not represent a Bayesian *a posteriori* probability distribution. Instead, the aim is to evaluate the sensitivity to the choice of a starting model in a linearized inversion scheme.

The ensembles of final models obtained with the L^2 misfit functions are much wider than for the optimal transport, and hence more dependent on the starting model. The ensemble clearly shows strong trade-offs between absolute velocities and depths of discontinuities, and hence lacks information about the absolute velocity profile. Fig. 9 shows they struggle to match most features of the data RF.

On the contrary, final models corresponding to W_1 misfit functions show very little sensitivity to the starting model and present much clearer interfaces. All four methods are more or less able to identify the main discontinuity at 25 km, but the transition is smoother for the L^2 misfit functions due to the non-resolved depth–velocity trade-off. Only the W_1 misfit functions, with their improved sensitivity to time-shifts, are able to robustly retrieve the other discontinuities. The deep discontinuities (at 60 and 80 km depth) are recovered, although at a shallower depth than the true one. It is however more uncertain for the shallowest discontinuity, because the main P peak partly obscures the corresponding Ps phase when calculating the RF (Fig. 9).

By improving the fit to observations, the inference process is supposed to approach the true model $\mathbf{m}_\mathrm{true}$, that is the final model $\mathbf{m}_\mathrm{f}$ should be closer to $\mathbf{m}_\mathrm{true}$ than $\mathbf{m}_\mathrm{i}$. This can be measured by looking at the quantity (in per cent):

$$100\,\frac{\| \mathbf{m}_\mathrm{i} - \mathbf{m}_\mathrm{true} \|^2 - \| \mathbf{m}_\mathrm{i} - \mathbf{m}_\mathrm{f} \|^2}{\| \mathbf{m}_\mathrm{i} - \mathbf{m}_\mathrm{true} \|^2}, \tag{16}$$

where $\| \|^2$ represents the squared difference between the *S*-wave velocity profiles corresponding to both models. A negative value means that the inference algorithm diverges, that is, the final model is further away from the true model than the initial model. A value close to 100 means that the algorithm converges towards the correct answer. This quantity, referred to as gain, is depicted for each misfit function and each initial model in Fig. 10. Again, the misfit functions ϕ_2 and ϕ_3 are the most successful in approaching the real model. The optimal transport distance is (slightly) more beneficial when applied on the cross-convolution of the waveforms, which reflects the loss of information that occurs when stabilizing the deconvolution.

4 APPLICATION TO REAL DATA: HYDERABAD STATION

The Hyderabad station (India) is located on the eastern part of the Dharwar craton that formed about 2.5 Gyr ago. Subject of many studies, the structure below the receiver is well known. A sharp Moho is identified at about 30–35 km depth and a mid-lithosphere discontinuity with a negative V_S velocity jump is visible in several studies (Kumar *et al.* 2007; Bodin *et al.* 2014; Dettmer *et al.* 2015). Studies including surface-wave data suggest a low-velocity zone at about 200 km depth (Mitra *et al.* 2006; Bodin *et al.* 2014; Dettmer *et al.* 2015). The layered structure in this region makes possible, at first order, to neglect the effect of interfaces' dip and anisotropy.

4.1 Data set

We consider the same data as Bodin *et al.* (2014), but use a pre-processing method more appropriate to waveform stacking. We follow a stacking strategy similar to Sippl *et al.* (2017) and described

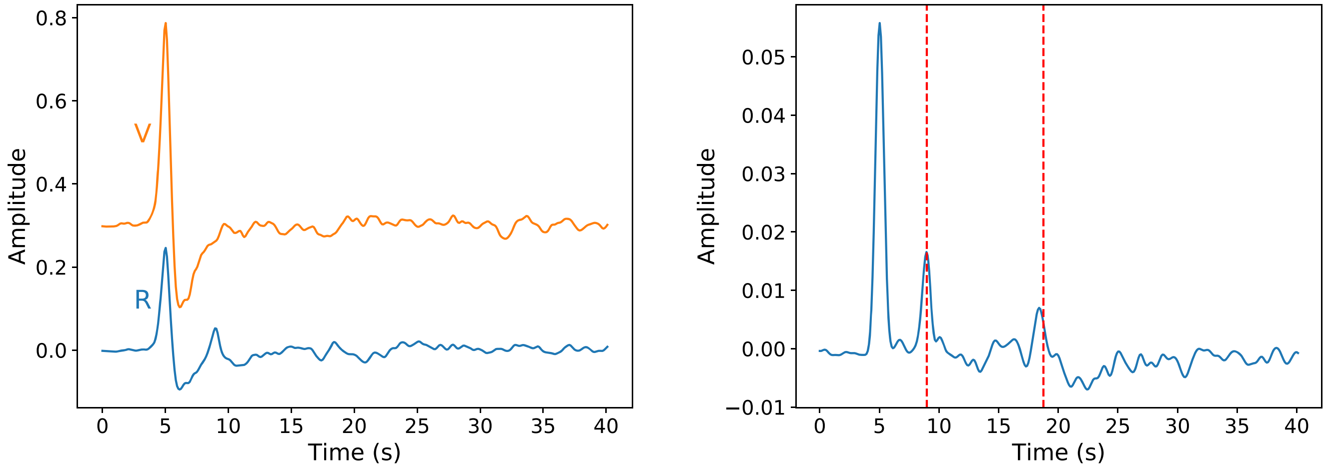

Figure 12. Stacked data from the Hyderabad station and corresponding receiver function, calculated with Gaussian filter width $a = 2$ and water-level $c = 0.0005$. On the left, the blue and orange lines represent the radial and vertical components, respectively. On the receiver function on the right, the red dotted lines show the theoretical arrival time of the Ps and PpPs phases converted at a ∼30 km depth Moho.

hereinafter. We select 562 events having back-azimuths between 265° and 315°, a slowness (ray parameter) range 0.6–0.8 s.km^{-1} and a sampling frequency of 10 Hz. At this stage, the data are low-pass filtered with a Butterworth filter having cut-off frequency of 2 Hz. After picking a time window of −10 to +15 s around the theoretical *P* arrival, the vertical components are cross-correlated to identify coherent traces. We select the largest population having a cross-correlation coefficient >0.85, which are subsequently aligned using the time-shift determined during the cross-correlation. To obtain a high-quality signal, we keep only those traces having a high signal-to-noise ratio, resulting in a stack of 25 waveforms displayed in Fig. 11. The final waveforms are time-windowed in the range −5 to +40 s around the first P peak. This ensures to image interfaces up to roughly 340 km depth. Whereas the frequency content of the waveforms extends up to 2 Hz, it is reduced up to 0.5 Hz for the data RF to stabilize the deconvolution. The obtained RF, Fig. 12, correlates well with the theoretical arrival of the Ps and PpPs phases for a 32 km depth Moho and $V_S = 3.5$ km s^{-1} in the upper layer, calculated using the formula of Zhu & Kanamori (2000). It shows several low-amplitude phases that could correspond to velocity jumps in the mid lithosphere.

4.2 Inversion of a receiver function

As a first step, we invert the data from the Hyderabad station with the same method as in the previous section. We invert for V_S and interface depths in an eight-layer model, repeating the process for various starting models. The reference starting model has a Moho at 30 km depth, its upper and lower *S*-wave velocities are 3.5 and 4.5 km s^{-1} up to 160 km depth. Below 160 km depth, the velocity profile is fixed at $V_S = 4.5$ km s^{-1} . To get a random distribution of starting models, the Moho depth, and both upper and lower velocities are perturbed according to a Gaussian distribution with standard deviations of 15 km and 0.3 km s^{-1}, respectively (Fig. 13). For each of the starting models generated that way, we minimize the misfit functions with the previously described algorithm. The velocity parameters are constrained between 2 and 6 km s^{-1} and the depth of each interface has to be shallower than 160 km depth. The distribution of final models for each misfit function is displayed in Fig. 14, the standard deviation of the V_S profiles shown in Fig. 15 characterizes their dependence to the starting models. The distribution of predicted RFs, Fig. 16, demonstrates the capability of each method to fit the data.

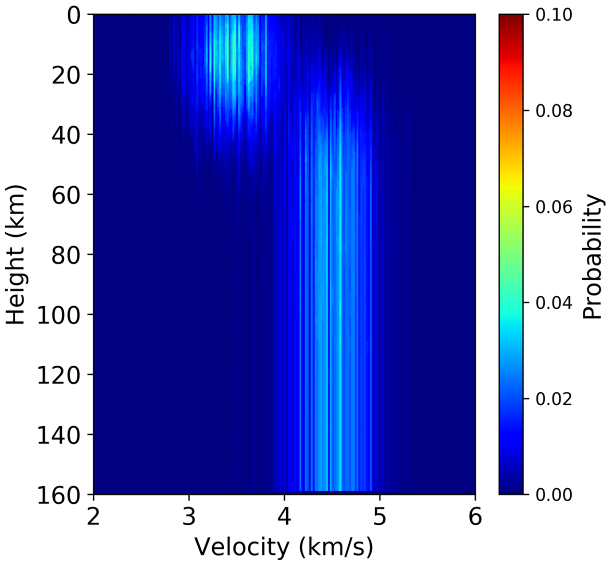

Figure 13. Ensemble of starting models for the receiver function inversion. We create 1000 random eight-layer starting models following a distribution centred on a two-layer model with a strong discontinuity at 30 km depth.

All methods are able to identify the general structure with a Moho at about 30 km depth, although they feature different values of absolute velocities. The best resolved features in the predicted data correspond to the most energetic phases, the Ps, PpPs and the PsPs/PpSs related the Moho. The methods using the optimal transport distance are more prone to see a shallower discontinuity at 10 km depth. Their final models distributions are narrower, particularly below the Moho, highlighting again their low dependency to the starting model (Fig. 15). The misfit function ϕ_3 gives the best results, both in stability relative to the starting model and in visually fitting the data. It is able to reconstruct more details of the RF (e.g. at 7 and 20 s, Fig. 16). The method using the misfit function ϕ_2, although it benefits from the stability of the Wasserstein distance, yields more variability in data prediction, probably because

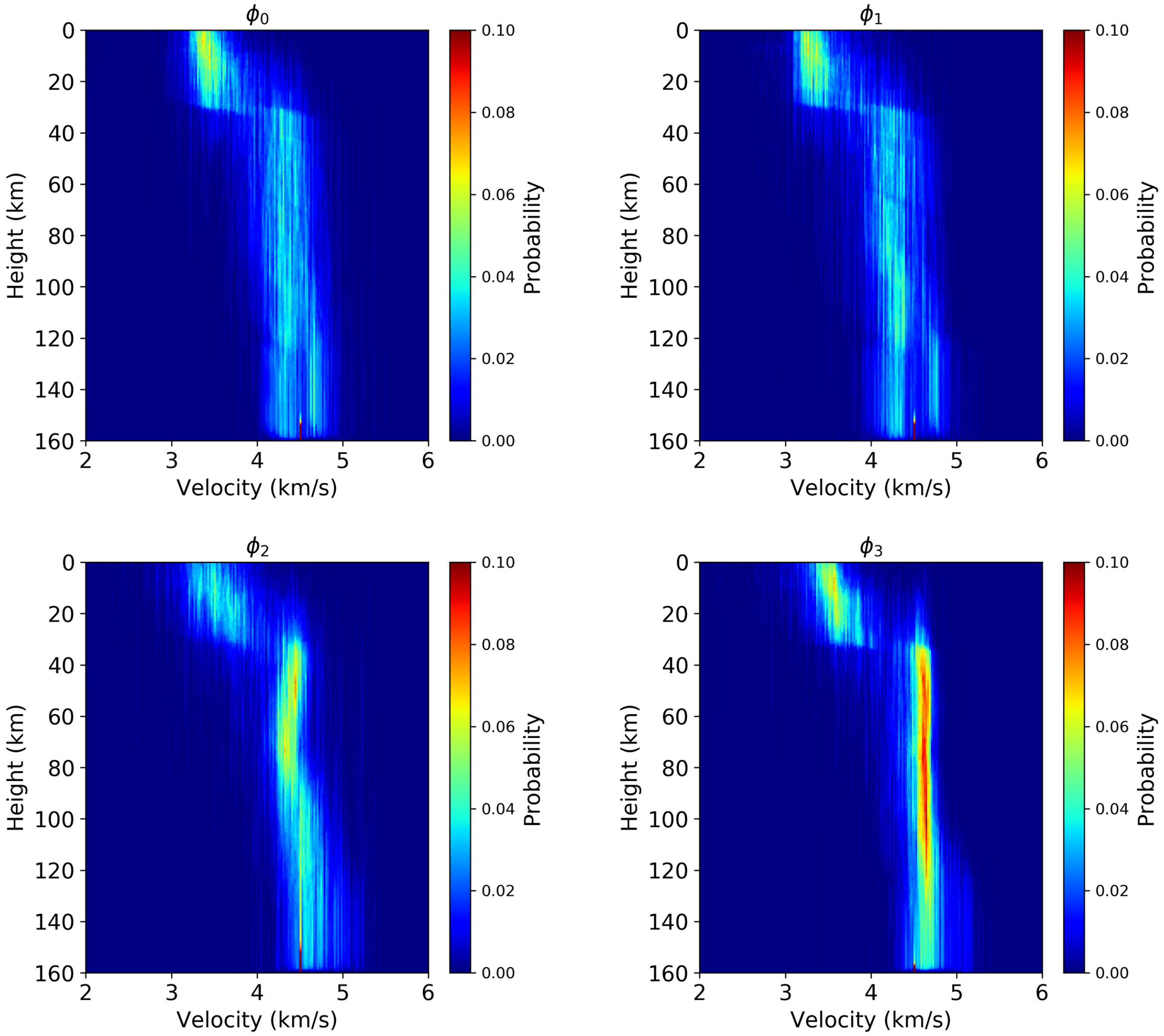

Figure 14. Resulting ensemble of models, displayed as V_S profiles, after the minimization of each misfit function for the Hyderabad station data (Fig. 11). All methods are able to recover a sharp Moho at about 30 km depth, but W_1-based methods are less dependent on the starting model.

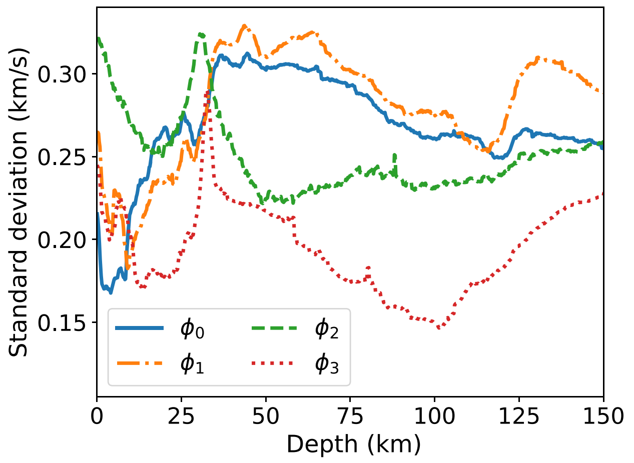

Figure 15. Standard deviation of the ensemble of final V_S profiles as a function of depth, for the resulting ensembles of models shown in Fig. 13.

the construction of the RFs requires regularization and filtering at high frequencies. Particularly, the shallow structure is not well resolved by this method (Fig. 15). The L^2-based resulting inversions exhibit more variability, but this does not seem to be reflected in the fit to the data displayed, Fig. 16. We deduce that, in this example, the L^2-based inversions have more difficulties to resolve the depth–velocity trade-off.

It may be noted that, for all misfit functions, the predicted RFs show little evidence of cycle skipping. The structure below the Hyderabad station being rather simple, a starting model including only the Moho is sufficient to make the local minimization algorithm converge. To verify this hypothesis, we repeat the experiment using a wider variety of starting models: each of the eight layers is now perturbed independently resulting in highly non-smooth models. In this case, we observe that a large number of final models obtained with L^2-based misfit functions do not converge to a meaningful solution, while W_1-based misfit functions are still able to consistently

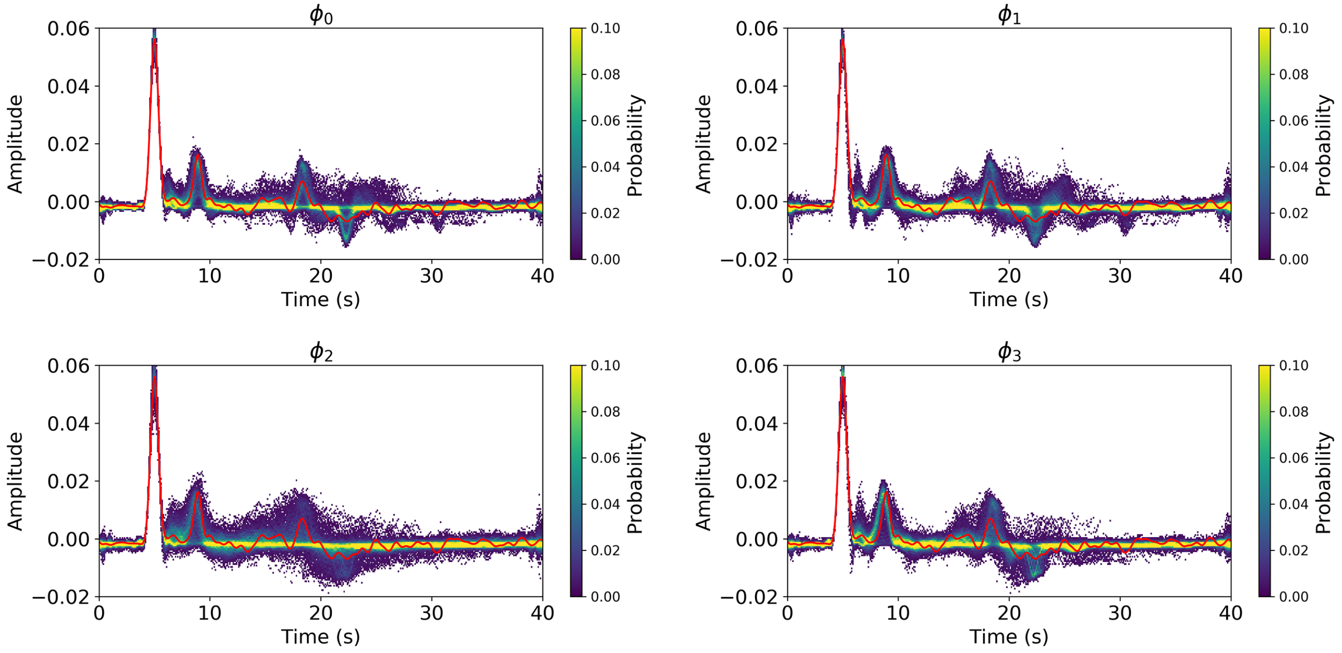

Figure 16. Distribution of the predicted receiver function for each of the final models presented Fig. 13, calculated with a Gaussian filter of width $a = 2$. The red line shows the Hyderabad station data RF.

recover the Moho (see the Supporting Information). From this experiment, we may conclude that using the Wasserstein distance is particularly useful when prior information on the structure below the station is not sufficient to build a good starting model.

4.3 Joint inversion of receiver function and surface wave dispersion curve

In many studies, the scattered *P*-wave coda is combined with longer period observations to overcome the depth–velocity trade-off afflicting the inversion of a RF alone. One complementary source of information are surface wave dispersion curves, which are sensitive to the absolute *S*-wave velocity. In this section, we develop a 1-D inversion strategy and parametrization similar to Julià *et al.* (2000), but that allows the use of the optimal transport distance and apply it to the Hyderabad station. On top of the *P*-wave coda of Fig. 11, we consider fundamental mode phase velocity measurements of Rayleigh waves in the range of periods 25–150 s given by Ekström (2011). To compute synthetic dispersion curves from a given 1-D earth model, we use the normal-mode summation method of Saito (1988).

The synthetic earth models are constructed in a similar manner as the ones presented earlier, except that the depths of the interfaces are now fixed and we invert only for the *S*-wave velocities. The model structure consists of 41 layers divided in the following manner: 16 layers of 2.5 km thickness up to 40 km depth, 8 layers of 5 km thickness up to 80 km depth and finally 17 layers of 10 km thickness up to 250 km depth. Below 250 km, the *S*-wave velocity profile is kept fixed and roughly similar to PREM. The model is parametrized as a 41-D vector **m** containing the *S*-wave velocity values for each layer. The waveforms predicted from a model **m** are compared to the observed data using the misfit functions ϕ_0 to ϕ_3, while the surface wave dispersion curves are compared with the least-squares misfit criteria. We denote $c(T)$ the Rayleigh phase velocity at period T. The misfit between observed and predicted phase velocities is

$$\phi_{\mathrm{SWD}} = \sum_T [c_{\mathrm{obs}}(T) - c_{\mathrm{pred}}(T)]^2. \tag{17}$$

Due to the high number of layers, a smoothing constraint might be required to limit rapid variations of velocity with depth (e.g. Julià *et al.* 2000). This can be achieved by minimizing the sum of velocity differences between adjacent layers, $\mid \mathbf{D}V_{\mathrm{S}} \mid = \sum_{i=1}^{41} \mid V_{\mathrm{S},i+1} - V_{\mathrm{S},i} \mid$. This regularization term, called total variation, has the property to preserve contrasts in model parameters by promoting piece-wise constant models (e.g. Strong & Chan 2003; Loris *et al.* 2010), thus is appropriate to recover seismic wave velocity discontinuities. However, in the form presented here, it is non-differentiable. This is not an issue in our strategy, where the gradient of the misfit function is calculated with finite differences. For an explicit differentiation of the misfit function, an adaptation of the total variation criteria is required (Vogel 2002).

The final misfit function to jointly invert RFs and surface wave dispersion curves thus contains three terms. As in any joint minimization problems, it requires to choose parameters to determine the relative weight of each data set and the amount smoothing. For each of the functions ϕ_i, we define:

$$\tilde{\phi}_i = \alpha_i\, \phi_i + \beta\, \phi_{\mathrm{SWD}} + \theta\ \mid \mathbf{D}V_{\mathrm{S}} \mid, \tag{18}$$

where θ is the smoothing parameter and is chosen (after a few trials and errors) as 0.03 for all inversions. Several methods exist to make the optimal choice for θ. As our purpose is to compare different misfit functions, it will not be attempted here. The four different distances can result in very different ranges of values, they need to be equalized in order to provide comparable results. α_i and β are normalization factors for each misfit term. α_i is chosen as $1/\phi_i(\mathbf{m}_{\mathrm{ref}})$, ensuring that the residual misfits associated with the *P*-wave coda are normalized to 1 at a reference model $\mathbf{m}_{\text{ref}}$, independently of the function ϕ_i used. Similarly, β is chosen as $1/\phi_{\text{SWD}}(\mathbf{m}_{\text{ref}})$. More refined choices can take into account the ratio between the number of samples in the RF and the number of periods in the surface wave dispersion curve, or consider the variance of each data point (e.g. Julià *et al.* 2000).

Our reference model is a two-layer model with $V_S = 4.0\,\text{km s}^{-1}$ up to 40 km and $V_S = 4.5\,\text{km s}^{-1}$ up to 250 km. It is used to calculate the values of α_i and β. From this reference model, we construct 100 different starting models by randomly choosing the depth of the discontinuity between 15 and 45 km depth and by perturbing the *S*-wave velocity of each layer in the same manner as outlined above. The resulting distribution of starting models is shown in the Supporting Information. For each of the starting models, the inverse problem is then solved by minimizing the misfit function $\tilde{\phi}_i$.

The resulting final models for the four misfit functions, Fig. 17, present little variability. The parameter space is much better constrained after introducing surface wave dispersion curves and adding a smoothing factor. A large part of the dependency to the starting model is due to the depth–velocity trade-off. The introduction of surface wave dispersion curves in the data significantly reduces this problem even for the L^2-based misfit functions (e.g. Julià *et al.* 2000). Here, there are few discrepancies in the final profiles depending on the misfit function. They also reproduce most features of the data RF quite well (Fig. 18).

Our velocity profiles particularly agree with the one obtained by Julià *et al.* (2009). We identify a sharp Moho at 35 km depth, corresponding to the Ps phase observed at 9 s in the data RF in Fig. 12, and also a shallower discontinuity separating upper and lower crust at 10 km depth whose Ps phase (at ~7 s) is much less energetic. Additionally, all methods show evidence of a low-velocity zone between 70 and 80 km depth, which is required to explain the sequence of negative and positive phases at 13–15 s in the data RF. The negative converted phase corresponding to 70 km depth could match a mid-lithosphere discontinuity seen in other stations in the region (Kumar *et al.* 2013). The positive V_S discontinuity at about 80 km depth, sometimes called Hales discontinuity, has been observed by Saul *et al.* (2000), Oreshin *et al.* (2011) and Dettmer *et al.* (2015). These studies also resolve a negative V_S discontinuity at 110 km, which is however not visible in our profiles. Several limitations of our methodology can explain this absence. In the data RF, the associated negative Ps phase may be hidden by the multiples of the Moho. Also, because our stacking method does not account for moveout corrections, the discontinuities below 100 km depth become difficult to resolve. Finally, the results from Oreshin *et al.* (2011) and Dettmer *et al.* (2015) suggest that the discontinuity is associated with changes in the V_P/V_S ratio, which is kept constant in our parametrization. This mantle discontinuity is best resolved using Sp RFs and has sometimes been interpreted as the lithosphere-asthenosphere boundary (LAB; Kumar *et al.* 2007, 2013). Instead, we observe a diffuse LAB between 150 and 200 km depth mostly constrained by the surface wave dispersion data. The methods using the W_1 distance are able to recover a sharper negative discontinuity at 150 km depth, which corresponds to a negative phase at 22 s in the data RF. This is, however, obscured by the PsPs phase of the Moho. Eventually, we are able to recover the main discontinuities of the crust and shallow lithosphere beneath Hyderabad station, but the resolution potential of the data used here does not enable to robustly infer the structure below 100 km depth.

5 DISCUSSION

Optimal transport provides a natural way to compare distributions. Several strategies have been elaborated to adapt the Wasserstein distance to the comparison of seismograms and increase the sensitivity of the misfit function to time-shifts (Engquist & Froese 2014; Métivier *et al.* 2016a). Here, we demonstrate that the W_1 distance can also be applied to radial P-RFs and has benefits compared to the L^2 distance. The main advantage is the improvement in convexity with respect to the model parameters and consequently the reduced dependency to the starting model in the case of a linearized inversion. In particular, we obtained a misfit function less subject to the depth/velocity trade-off. Furthermore, we propose to apply the W_1 distance to the cross-convolution misfit introduced by Menke & Levin (2003). This method does not require deconvolution, filtering or regularization, thus preventing any loss of information. In all of our experiments, it exhibits slightly better performance than the W_1 distance between RFs. However, it requires to identify events having similar source time function, which limits its application to stations comprising large number of data. We also set up a methodology to jointly invert RFs and surface wave dispersion curves inspired by Julià *et al.* (2000). In the case of the relatively simple 1-D structure such as beneath Hyderabad station, the dependency of the obtained parameters to the starting model is mostly due to the depth–velocity trade-off. The introduction of surface wave dispersion curves, which allow to resolve the isotropic *S*-wave velocity, is a successful approach to alleviate this issue. The resulting velocity profiles are thus, to a few minor discrepancies, similar for L^2 and W_1-based misfit functions.

Alternative, more convex misft functions can have drawbacks relative to the standard least-squares misfit. These include increased computation cost or difficulty of implementation, the need to window the data in a specific manner, higher sensitivity to noise, or loss in resolution. Below, we discuss these potential limitations in the case of the W_1-based misfit function used here.

On the first point, although the solution to the optimal transport problem is an iterative process, the algorithm developed by Métivier *et al.* (2016b) has a linear complexity. The additional cost compared to the L^2 distance is almost transparent in the RF application, as the majority of the computational power is allocated to the forward model. We outline that the W_1 distance is particularly easy to implement in any pre-existing inversion strategy. The solution does not require the manual tuning of any parameter and also returns the associated "residual" term $\bar{\varphi}$, required to calculate the misfit gradient or used in adjoint-state strategies, at no additional cost.

Second, as it performs a global comparison between two waveforms, the optimal transport distance is more strongly affected by the finite size of the seismogram window. For example, when increasing the depth of an interface (or reducing the layer wave velocity) in the parameter space, it may happen that a converted phase goes out of the predicted data window which drastically reduces the value of the misfit function and creates a barrier for the minimization algorithm (see Fig. 3 for example). The use of a global comparison method needs to go together with a choice of a sufficiently large number of samples in the RF. Furthermore, our experiments suggest that the W_1 distance is not particularly sensitive to noise.

The Wasserstein-based misfit function has a flatter shape around the global minimum, which could be seen as a loss of resolution in estimating the model parameters. If this is of concern, a solution is to use the model obtained with the Wasserstein distance as a starting point for a new minimization using the L^2 distance. Our experiments on synthetic tests suggest that, in most cases, this does not

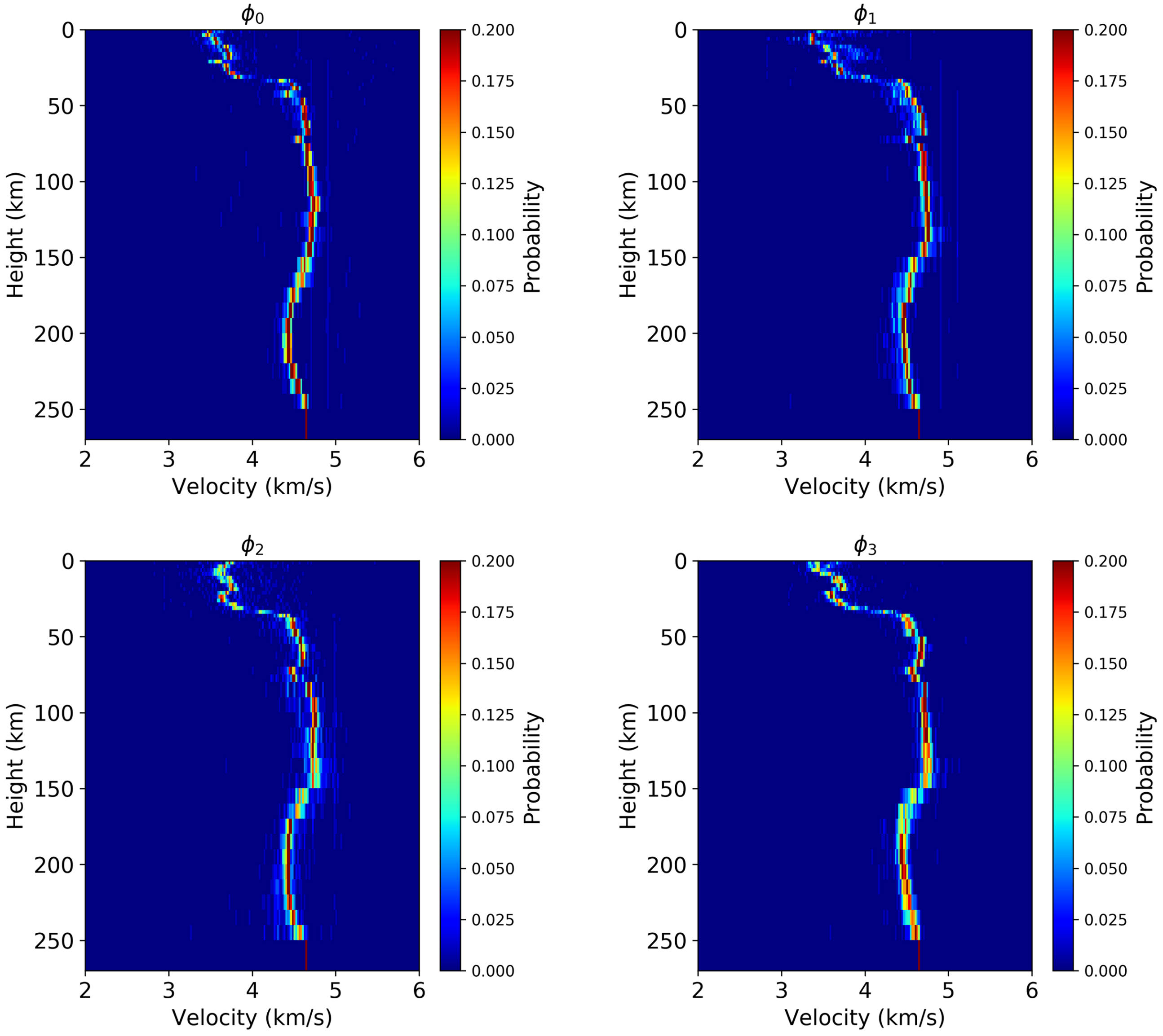

Figure 17. Resulting ensemble of models for each misfit function $\tilde{\phi_0}$ to $\tilde{\phi_4}$ in the joint inversion of RF and surface wave dispersion curves. The introduction of dispersion curves and smoothing greatly reduces the dependency to the starting model, as all inversions seem to converge towards the same minimum.

provide any improvement towards the true model (see an example in the Supporting Information). This is because the misfit function is multimodal. The final model obtained with the W_1 distance is not necessary in the valley of the L^2 global minimum. For example, the predicted seismograms and the data can still be cycle-skipped.

Although this study only deals with a gradient-based minimization method for the resolution of the inverse problem, optimal transport can also be applied with global minimization algorithms. As such, it provides alternatives to misfit functions based on dynamic time warping (e.g. Berndt & Clifford 1994), which has been investigated to resolve an anisotropic structure from radial and transverse RFs (Wirth *et al.* 2016). As the misfit functions using the Wasserstein distance present few local minima, they are likely to improve the exploration rate of the parameter space with Monte Carlo sampling techniques. But, contrary to the least-squares distance, we outlined that the W_1 distance cannot be related to the level of data uncertainties. If posterior uncertainties are sought, it might help to find good starting models for Bayesian methods.

The inversion of P-RFs for an Earth's velocity profile is a nonlinear problem where the misfit function can present multiple local minima. One of the most appropriate strategy is to jointly invert RFs with surface wave data to constrain the absolute *S*-wave velocity profile. In this context, the benefit of linearized inversion schemes is the possibility to quickly perform multiple 1-D inversions on large seismic arrays. This has resulted in numerous studies mapping the continental crust and lithosphere structure (e.g. Julià *et al.* 2003, 2009; Dugda *et al.* 2007; Xu *et al.* 2007; Yoo *et al.* 2007; Tokam *et al.* 2010; Sosa *et al.* 2014; Ward *et al.* 2014). We suggest that the W_1 optimal transport distance can improve the stability of such inversions with a small implementation difficulty. A reduced dependency to the starting model can alleviate the manual calibration of the inversion and makes easier its application to many stations. In particular, it will be beneficial in the case of complex structures where the signals are subject to cycle skipping, or when there is no good *a priori* model available. Eventually, the methodology could be extended to the inversion of other types of RFs, as

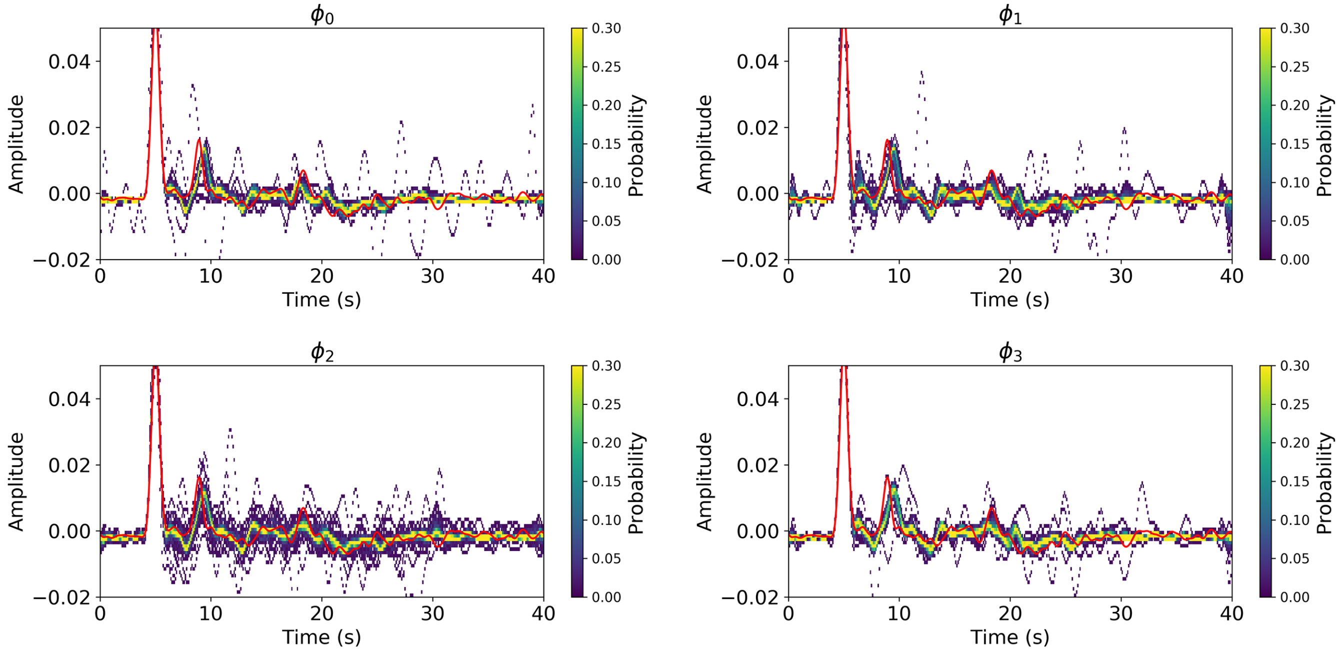

Figure 18. Distribution of the predicted receiver functions for each final model in the joint inversion of RF and dispersion curves, calculated with a Gaussian filter of width $a = 2$. The red line shows the data RF from the Hyderabad station.

long as the null mean hypothesis is preserved. If the method is of interest to the community, the theoretical developments presented in this study will make the optimal transport approach to RF inversion computationally competitive to be applied on large seismic arrays.

ACKNOWLEDGEMENTS

We thank Yann Capdeville for discussions at a meeting funded by the ANR (ANR project ANR-16-CE31-0022-01), which have initiated this research. This work was funded by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 716542. We are grateful to the associated editor, Jan Dettmer and an anonymous reviewer, whose comments helped to clarify the manuscript.

REFERENCES

Ammon, C.J., Randall, G.E. & Zandt, G., 1990. On the nonuniqueness of receiver function inversions, *J. geophys. Res.,* **95**(B10), 15 303–15 318.

Berndt, D.J. & Clifford, J., 1994. Using dynamic time warping to find patterns in time series, in *Proc. of KDD Workshop,* **10,** Seattle, WA, p. 359.

Bodin, T., Yuan, H. & Romanowicz, B., 2014. Inversion of receiver functions without deconvolution—application to the Indian craton, *Geophys. J. Int.,* **196**(2), 1025–1033.

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

Bostock, M., 2015. Theory and observations—seismology and the structure of the Earth: teleseismic body-wave scattering and receiver-side structure, in *Treatise on Geophysics (Second Edition),* Schubert, Gerald, pp. 253–275, Elsevier.

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

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

Cheng, C., Bodin, T., Tauzin, B. & Allen, R.M., 2017. Cascadia subduction slab heterogeneity revealed by three-dimensional receiver function Kirchhoff migration, *Geophys. Res. Lett.,* **44**(2), 694–701.

Dettmer, J., Dosso, S.E., Bodin, T., Stipčević, J. & Cummins, P.R., 2015. Direct-seismogram inversion for receiver-side structure with uncertain source–time functions, *Geophys. J. Int.,* **203**(2), 1373–1387.

Dugda, M.T., Nyblade, A.A. & Julia, J., 2007. Thin lithosphere beneath the Ethiopian Plateau revealed by a joint inversion of Rayleigh wave group velocities and receiver functions, *J. geophys. Res.,* **112**(B8).

Eilon, Z., Fischer, K.M. & Dalton, C.A., 2018. An adaptive Bayesian inversion for upper-mantle structure using surface waves and scattered body waves, *Geophys. J. Int.,* **214**(1), 232–253.

Ekström, G., 2011. A global model of Love and Rayleigh surface wave dispersion and anisotropy, 25-250 s, *Geophys. J. Int.,* **187**(3), 1668–1686.

Engquist, B. & Froese, B.D., 2014. Application of the Wasserstein metric to seismic signals, *Commun. Math. Sci.,* **12**(5), 979–988.

Fichtner, A. & Trampert, J., 2011. Resolution analysis in full waveform inversion, *Geophys. J. Int.,* **187**(3), 1604–1624.

Fichtner, A., Kennett, B.L., Igel, H. & Bunge, H.-P., 2008. Theoretical background for continental-and global-scale full-waveform inversion in the time–frequency domain, *Geophys. J. Int.,* **175**(2), 665–685.

Gee, L.S. & Jordan, T.H., 1992. Generalized seismological data functionals, *Geophys. J. Int.,* **111**(2), 363–390.

Guo, Z., *et al.*, 2015. High resolution 3-D crustal structure beneath NE China from joint inversion of ambient noise and receiver functions using NECESSArray data, *Earth planet. Sci. Lett.,* **416,** 1–11.

Haskell, N.A., 1962. Crustal reflection of plane P and SV waves, *J. geophys. Res.,* **67**(12), 4751–4768.

Hu, S. & Zhu, L., 2017a. Calculation of differential seismograms using analytic partial derivatives—I: teleseismic receiver functions, *Geophys. J. Int.,* **210**(2), 887–897.

Hu, S. & Zhu, L., 2017b. Calculation of differential seismograms using analytic partial derivatives—II: regional waveforms, *Geophys. J. Int.,* **212**(1), 390–399.

Julià, J., Ammon, C., Herrmann, R. & Correig, A.M., 2000. Joint inversion of receiver function and surface wave dispersion observations, *Geophys. J. Int.,* **143**(1), 99–112.

Julià, J., Ammon, C.J. & Herrmann, R.B., 2003. Lithospheric structure of the Arabian Shield from the joint inversion of receiver functions and surface-wave group velocities, *Tectonophysics,* **371**(1), 1–21.

Julià, J., Jagadeesh, S., Rai, S. & Owens, T., 2009. Deep crustal structure of the Indian shield from joint inversion of *P* wave receiver functions and Rayleigh wave group velocities: implications for Precambrian crustal evolution, *J. geophys. Res.,* **114**(B10).

Kaipio, J. & Somersalo, E., 2006. *Statistical and Computational Inverse Problems,* Vol. **160,** Springer Science & Business Media.

Kim, S., Dettmer, J., Rhie, J. & Tkalčić, H., 2016. Highly efficient Bayesian joint inversion for receiver-based data and its application to lithospheric structure beneath the southern Korean Peninsula, *Geophys. J. Int.,* **206**(1), 328–344.

Kind, R., Kosarev, G. & Petersen, N., 1995. Receiver functions at the stations of the German Regional Seismic Network (GRSN), *Geophys. J. Int.,* **121**(1), 191–202.

Kiselev, S., Vinnik, L., Oreshin, S., Gupta, S., Rai, S., Singh, A., Kumar, M.R. & Mohan, G., 2008. Lithosphere of the Dharwar craton by joint inversion of P and S receiver functions, *Geophys. J. Int.,* **173**(3), 1106–1118.

Kosarev, G., Kind, R., Sobolev, S., Yuan, X., Hanka, W. & Oreshin, S., 1999. Seismic evidence for a detached Indian lithospheric mantle beneath Tibet, *Science,* **283**(5406), 1306–1309.

Kumar, P., Yuan, X., Kumar, M.R., Kind, R., Li, X. & Chadha, R., 2007. The rapid drift of the Indian tectonic plate, *Nature,* **449**(7164), 894–897.

Kumar, P., Ravi Kumar, M., Srijayanthi, G., Arora, K., Srinagesh, D., Chadha, R. & Sen, M.K., 2013. Imaging the lithosphere-asthenosphere boundary of the Indian plate using converted wave techniques, *J. geophys. Res.,* **118**(10), 5307–5319.

Langston, C.A., 1979. Structure under Mount Rainier, Washington, inferred from teleseismic body waves, *J. geophys. Res.,* **84**(B9), 4749–4762.

Lellmann, J., Lorenz, D.A., Schonlieb, C. & Valkonen, T., 2014. Imaging with Kantorovich–Rubinstein discrepancy, *SIAM J. Imaging Sci.,* **7**(4), 2833–2859.

Loris, I., Douma, H., Nolet, G., Daubechies, I. & Regone, C., 2010. Nonlinear regularization techniques for seismic tomography, *J. Comput. Phys.,* **229**(3), 890–905.

Menke, W., 2017. Sensitivity kernels for the cross-convolution measure, *Bull. seism. Soc. Am.,* **107**(5), 2213–2224.

Menke, W. & Levin, V., 2003. The cross-convolution method for interpreting SKS splitting observations, with application to one and two-layer anisotropic earth models, *Geophys. J. Int.,* **154**(2), 379–392.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E. & Virieux, J., 2016a. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion, *Geophys. J. Int.,* **205**(1), 345–377.

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

Mitra, S., Priestley, K., Gaur, V. & Rai, S., 2006. Shear-wave structure of the south Indian lithosphere from Rayleigh wave phase-velocity measurements, *Bull. Seism. Soc. Am.,* **96**(4A), 1551–1559.

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

Oreshin, S., Vinnik, L., Kiselev, S., Rai, S., Prakasam, K. & Treussov, A., 2011. Deep seismic structure of the Indian shield, western Himalaya, Ladakh and Tibet, *Earth planet. Sci. Lett.,* **307**(3-4), 415–429.

Owens, T.J., Zandt, G. & Taylor, S.R., 1984. Seismic evidence for an ancient rift beneath the Cumberland Plateau, Tennessee: a detailed analysis of broadband teleseismic P waveforms, *J. geophys. Res.,* **89**(B9), 7783–7795.

Piana Agostinetti, N. & Malinverno, A., 2010. Receiver function inversion by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.,* **181**(2), 858–872.

Randall, G.E., 1994. Efficient calculation of complete differential seismograms for laterally homogeneous earth models, *Geophys. J. Int.,* **118**(1), 245–254.

Rondenay, S., 2009. Upper mantle imaging with array recordings of converted and scattered teleseismic waves, *Surv. Geophys.,* **30**(4-5), 377–405.

Ryberg, T. & Weber, M., 2000. Receiver function arrays: a reflection seismic approach, *Geophys. J. Int.,* **141**(1), 1–11.

Saito, M., 1988. DISPER80: a subroutine package for calculation of seismic normal-mode solution, in *Seismological Algorithm: Computational Methods and Computer Programs,* pp. 293–319, ed. Doornbos, D.J., Academic Press.

Sambridge, M., 1999. Geophysical inversion with a neighbourhood algorithm—II. Appraising the ensemble, *Geophys. J. Int.,* **138**(3), 727–746.

Saul, J., Kumar, M.R. & Sarkar, D., 2000. Lithospheric and upper mantle structure of the Indian shield, from teleseismic receiver functions, *Geophys. Res. Lett.,* **27**(16), 2357–2360.

Shen, W., Ritzwoller, M.H., Schulte-Pelkum, V. & Lin, F.-C., 2012. Joint inversion of surface wave dispersion and receiver functions: a bayesian monte-carlo approach, *Geophys. J. Int.,* **192**(2), 807–836.

Shibutani, T., Sambridge, M. & Kennett, B., 1996. Genetic algorithm inversion for receiver functions with application to crust and uppermost mantle structure beneath eastern Australia, *Geophys. Res. Lett.,* **23**(14), 1829–1832.

Singh, A., Singh, C. & Kennett, B.L.N., 2015. A review of crust and upper mantle structure beneath the Indian subcontinent, *Tectonophysics,* **644,** 1–21.

Sippl, C., Kumar, A. & Dettmer, J., 2017. A cross-correlation-based approach to direct seismogram stacking for receiver-side structural inversion, *Bull. seism. Soc. Am.,* **107,** 1545–1550.

Sosa, A., Thompson, L., Velasco, A.A., Romero, R. & Herrmann, R.B., 2014. 3-D structure of the Rio Grande Rift from 1-D constrained joint inversion of receiver functions and surface wave dispersion, *Earth planet. Sci. Lett.,* **402,** 127–137.

Stähler, S.C. & Sigloch, K., 2016. Fully probabilistic seismic source inversion-part 2: modelling errors and station covariances, *Solid Earth,* **7**(6), 1521.

Strong, D. & Chan, T., 2003. Edge-preserving and scale-dependent properties of total variation regularization, *Inverse Probl.,* **19**(6), S165.

Tkalčić, H., Pasyanos, M.E., Rodgers, A.J., Gök, R., Walter, W. & Al-Amri, A., 2006. A multistep approach for joint modeling of surface wave dispersion and teleseismic receiver functions: implications for lithospheric structure of the Arabian Peninsula, *J. geophys. Res.,* **111**(B11).

Tokam, A.-P.K., Tabod, C.T., Nyblade, A.A., Juli, J., Wiens, D.A. & Pasyanos, M.E., 2010. Structure of the crust beneath Cameroon, West Africa, from the joint inversion of Rayleigh wave group velocities and receiver functions, *Geophys. J. Int.,* **183**(2), 1061–1076.

Villani, C., 2008. *Optimal Transport: Old and New,* Vol. **338,** Springer Science & Business Media.

Vinnik, L., 1977. Detection of waves converted from P to SV in the mantle, *Phys. Earth planet. Inter.,* **15**(1), 39–45.

Vinnik, L.P., Reigber, C., Aleshin, I.M., Kosarev, G.L., Kaban, M.K., Oreshin, S.I. & Roecker, S.W., 2004. Receiver function tomography of the central Tien Shan, *Earth planet. Sci. Lett.,* **225**(1), 131–146.

Vogel, C.R., 2002. *Computational Methods for Inverse Problems,* Vol. **23,** SIAM.

Ward, K.M., Zandt, G., Beck, S.L., Christensen, D.H. & McFarlin, H., 2014. Seismic imaging of the magmatic underpinnings beneath the Altiplano-Puna volcanic complex from the joint inversion of surface wave dispersion and receiver functions, *Earth planet. Sci. Lett.,* **404,** 43–53.

Wirth, E.A., Long, M.D. & Moriarty, J.C., 2016. A Markov chain Monte Carlo with Gibbs sampling approach to anisotropic receiver function forward modeling, *Geophys. J. Int.,* **208**(1), 10–23.

Xu, L., Rondenay, S. & van der Hilst, R.D., 2007. Structure of the crust beneath the southeastern Tibetan plateau from teleseismic receiver functions, *Phys. Earth planet. Inter.,* **165**(3-4), 176–193.

Yoo, H., Herrmann, R., Cho, K. & Lee, K., 2007. Imaging the three-dimensional crust of the Korean Peninsula by joint inversion of surface-wave dispersion and teleseismic receiver functions, *Bull. seism. Soc. Am.,* **97**(3), 1002–1011.

Zhao, L.-S., Sen, M.K., Stoffa, P. & Frohlich, C., 1996. Application of very fast simulated annealing to the determination of the crustal structure beneath Tibet, *Geophys. J. Int.,* **125**(2), 355–370.

Zhu, C., Byrd, R.H., Lu, P. & Nocedal, J., 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization, *ACM Trans. Math. Softw.,* **23**(4), 550–560.

Zhu, L. & Kanamori, H., 2000. Moho depth variation in southern California from teleseismic receiver functions, *J. geophys. Res.,* **105**(B2), 2969–2980.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Figure S1. Ensemble of non-smooth starting models for the receiver function inversion. In this scenario, we start again from the reference model described in Section 4.2 of the main text, and subsequently perturb all eight layers independently following a Gaussian distribution of standard deviation 0.3 km s^{-1}.

Figure S2. Resulting ensemble of models for each misfit function ϕ_0 to ϕ_3, displayed as V_S profiles. The results correspond to the ensemble of non-smooth starting models presented in Fig. S1. A substantial number of final models obtained with L_2-based misfits diverge towards the bounds of the parameter space ($V_S = 2$ km s^{-1} and $V_S = 6$ km s^{-1}). On the other hand, inversions using W_1 distance converge more consistently towards meaningful solutions because the misfit function is more convex.

Figure S3. Ensemble of starting models for the joint inversion of the receiver function and the dispersion curve. The corresponding ensemble of final models is presented in the main text.

Figure S4. Ensemble of dispersion curve predictions for the final models obtained in the joint inversion at the Hyderabad station. The red line shows the dispersion curve used as data.

Figure S5. Resulting ensemble of models after a two-step procedure of Hyderabad station data inversion. Using the ensemble of starting models presented in Fig. 13, a first step minimizes the W1-based misfit functions ϕ_2 and ϕ_3 to obtain the models displayed in Fig. 14. These models are used as starting models for a supplementary minimization, this time with L_2-based misfit functions ϕ_0 and ϕ_1, respectively. The obtained ensemble of final models corresponding to RF-based misfits ϕ_2 and ϕ_0 is displayed on the left, the ensemble of models using cross-convolution-based misfits ϕ_3 and ϕ_1 is displayed on the right.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX: GRADIENT OF THE MISFIT FUNCTIONS

A1 Cross-convolution and the least-squares norm

For a function $\phi(\mathbf{m})$ and an infinitesimal perturbation in the parameter space dm, the gradient $\nabla\phi$ is defined by the following relation:

$$\phi(\mathbf{m} + dm) = \phi(\mathbf{m}) + < \nabla\phi(\mathbf{m}),\ dm >_M + O(\|dm\|_M^2), \qquad \text{(A1)}$$

where $<.,.>_M$ and $\|.\|_M$ are the scalar product and associated norm defined in the model space. Similarly, we denote $<.,.>_X$ and $\|.\|_X$ the scalar product and norm in the time window space.

To calculate the gradient of $\phi_1(\mathbf{m})$ with respect to a perturbation dm, we expand:

$$\phi_1(\mathbf{m} + dm) = \|v(\mathbf{m} + dm) * R - r(\mathbf{m} + dm) * V\|_X^2 \qquad \text{(A2)}$$

$$= \phi_1(\mathbf{m}) + 2 < \left(\frac{\partial g_1}{\partial m} dm\right) * R - \left(\frac{\partial g_2}{\partial m} dm\right) * V,\ w(\mathbf{m}) >_X + O(\|dm\|_M^2). \qquad \text{(A3)}$$

Using the property of the adjoint of a convolution $<u{*}v, w> = <v, u{\star}w>$, we have:

$$\phi_1(\mathbf{m} + dm) = \phi_1(\mathbf{m}) + 2 < dm,\ \frac{\partial g_1}{\partial m}^T R \star w(\mathbf{m}) - \frac{\partial g_2}{\partial m}^T V \star w(\mathbf{m}) >_M + O(\|dm\|_M^2). \qquad \text{(A4)}$$

Identifying the expression of the gradient between eqs (A1) and (A4) leads to the expression of eq. (12).

A2 Cross-convolution and the Wasserstein distance

We define the function $f(\mathbf{m}, \varphi)$ as

$$f(\mathbf{m}, \varphi) = \int_{t\in X} \varphi(t) w(\mathbf{m}, t)\mathrm{d}t = < \varphi,\ w(\mathbf{m}) >_X \qquad \text{(A5)}$$

such that $\phi_3(\mathbf{m}) = f(\mathbf{m}, \bar{\varphi})$. The gradient of $\phi_3(\mathbf{m})$ can be expressed as a function of f through the chain rule:

$$\frac{\partial\phi_3}{\partial m} = \frac{\partial f}{\partial m} + \frac{\partial f}{\partial\varphi}\frac{\partial\bar{\varphi}}{\partial m}. \qquad \text{(A6)}$$

By definition of $\bar{\varphi}$ as the maximum of f (see eq. 13), the partial derivatives of f with respect to φ at this point vanish:

$$\frac{\partial\phi_3}{\partial m}(\mathbf{m}) = \frac{\partial f}{\partial m}[\mathbf{m}, \bar{\varphi}(\mathbf{m})]. \qquad \text{(A7)}$$

We now expand f with respect to a perturbation dm:

$$f(\mathbf{m}{+}dm, \varphi) = < \varphi,\ v(\mathbf{m}{+}dm) * R{-}r(\mathbf{m}{+}dm) * V >_X \qquad \text{(A8)}$$

$$= f(\mathbf{m}, \varphi) + < \varphi,\ \left(\frac{\partial g_1}{\partial m} dm\right) * R - \left(\frac{\partial g_2}{\partial m} dm\right) * V >_X + O(\|dm\|_M^2) \qquad \text{(A9)}$$

$$= f(\mathbf{m}, \varphi) + < \frac{\partial g_1}{\partial m}^T * R \star \varphi - \frac{\partial g_2}{\partial m}^T * V \star \varphi,\ dm >_M + O(\|dm\|_M^2). \qquad \text{(A10)}$$

Again, using the expression of the gradient in eq. (A1) , we have:

$$\frac{\partial f}{\partial m} = \frac{\partial g_1}{\partial m}^T * R \star \varphi - \frac{\partial g_2}{\partial m}^T * V \star \varphi. \qquad \text{(A11)}$$

Now using eq. (A7), we obtain the gradient of the misfit function ϕ_3 as seen eq. (14).