Accelerated full-waveform inversion using dynamic mini-batches

Geophysical Journal International

Geophys. J. Int. (2020) **221,** 1427–1438
Advance Access publication 2020 February 21
GJI Seismology

doi: 10.1093/gji/ggaa079

Accelerated full-waveform inversion using dynamic mini-batches

Dirk Philip van Herwaarden,[1] Christian Boehm,[1] Michael Afanasiev,[1] Solvi Thrastarson,[1] Lion Krischer,[1] Jeannot Trampert[2] and Andreas Fichtner[1]

[1] *Department of Earth Sciences, Institute of Geophysics, ETH Zurich, 8092 Zurich, Switzerland. E-mail: dirkphilip.vanherwaarden@erdw.ethz.ch*
[2] *Department of Earth Sciences, Utrecht University, Princetonlaan 8a, 3584 CB Utrecht, The Netherlands*

Accepted 2020 February 12. Received 2019 September 13; in original form 2020 February 3

SUMMARY

We present an accelerated full-waveform inversion based on dynamic mini-batch optimization, which naturally exploits redundancies in observed data from different sources. The method rests on the selection of quasi-random subsets (mini-batches) of sources, used to approximate the misfit and the gradient of the complete data set. The size of the mini-batch is dynamically controlled by the desired quality of the gradient approximation. Within each mini-batch, redundancy is minimized by selecting sources with the largest angular differences between their respective gradients, and spatial coverage is maximized by selecting candidate events with Mitchell's best-candidate algorithm. Information from sources not included in a specific mini-batch is incorporated into each gradient calculation through a quasi-Newton approximation of the Hessian, and a consistent misfit measure is achieved through the inclusion of a control group of sources. By design, the dynamic mini-batch approach has several main advantages: (1) The use of mini-batches with adaptive size ensures that an optimally small number of sources is used in each iteration, thus potentially leading to significant computational savings; (2) curvature information is accumulated and exploited during the inversion, using a randomized quasi-Newton method; (3) new data can be incorporated without the need to re-invert the complete data set, thereby enabling an evolutionary mode of full-waveform inversion. We illustrate our method using synthetic and real-data inversions for upper-mantle structure beneath the African Plate. In these specific examples, the dynamic mini-batch approach requires around 20 per cent of the computational resources in order to achieve data and model misfits that are comparable to those achieved by a standard full-waveform inversion where all sources are used in each iteration.

Key words: Inverse theory; Waveform inversion; Computational seismology; Seismic tomography.

1 INTRODUCTION

Seismic tomography has seen significant progress since the early applications by Aki & Lee (1976), Aki *et al.* (1977) and Dziewonski *et al.* (1977). Increasing data availability and increases in computational power have opened the doors to ever more sophisticated methodologies, including, for instance, finite-frequency tomography (e.g. Yomogida 1992; Dahlen *et al.* 2000; Friederich 2003; Montelli *et al.* 2004; Yoshizawa & Kennett 2005), joint inversions of body and surface wave data (e.g. Ritsema *et al.* 1999, 2011; Chang *et al.* 2010; Koelemeijer *et al.* 2017) or fully probabilistic approaches (e.g. Devilee *et al.* 1999; Trampert *et al.* 2004; Bodin & Sambridge 2009; Mosca *et al.* 2012). During the last decade it has become computationally feasible to perform regional-scale (e.g. Chen *et al.* 2007; Fichtner *et al.* 2009; Tape *et al.* 2010; Rickers *et al.* 2013; Simute *et al.* 2016) and global-scale seismic tomography (e.g. French & Romanowicz 2014; Bozdag *et al.* 2016; Fichtner *et al.* 2018) using full-waveform inversion (FWI), conceptualized already in the late 1970s and early 1980s (Bamberger *et al.* 1977, 1982; Lailly 1983; Tarantola 1984).

Although FWI is able to account for the full physics of wave propagation and recover detailed Earth structure, it is not being applied as widely as more traditional techniques, such as ray-based traveltime tomography (e.g. Grand *et al.* 1997; Gorbatov & Kennett 2003; Lebedev & van der Hilst 2008). This partly stems from various practical challenges, such as the handling of numerous different file formats related to forward and adjoint simulations, the processing of the observed waveform data, the quantification of multifrequency waveform misfits, the computation of adjoint sources, and the often tedious and impractical interaction with remote supercomputing systems. The Obspy toolkit (Krischer *et al.* 2015b) made a significant contribution to simplify these tasks.

Additionally, the sheer number of files involved in FWI led to performance degradation of file systems on high-performance computing (HPC) systems, which motivated the development of the adaptable seismic data format (Krischer *et al.* 2016). Furthermore, the automation of previously manual tasks such as window picking (Maggi *et al.* 2009; Krischer *et al.* 2015a) helped to make the workflow faster and more robust. All of these developments have led to a point where it has become increasingly feasible to perform FWI almost automatically using HPC resources (Krischer *et al.* 2018). Despite these improvements, significant challenges remain.

First, and foremost, the computational requirements of FWI are substantial, as the algorithm requires at least two numerical simulations of the wave equation for each source at each iteration. Previous steps to mitigate this problem include the use of graphical processing units (Rietmann *et al.* 2012; Gokhberg & Fichtner 2016), simultaneous sources (e.g. Capdeville *et al.* 2005; Krebs *et al.* 2009; Moghaddam *et al.* 2013; Tromp & Bachmann 2019), coupling with computationally less intense methods (e.g. Capdeville *et al.* 2003; Monteiller *et al.* 2012; Masson & Romanowicz 2016, 2017; Capdeville & Métivier 2018), the acoustic approximation (e.g. Alkhalifah 2000; Operto *et al.* 2013; Cance & Capdeville 2015), using wavefield adapted meshes (van Driel *et al.* 2020; Thrastarson *et al.* 2020) and coarsening the numerical mesh with the help of homogenization (e.g. Capdeville *et al.* 2013). Most commonly, however, the number of sources, for example, earthquakes or explosions, is limited at the expense of tomographic resolution.

Second, and related to the first issue, efficient methods must be found to evolve FWI earth models over time, as new data become available. Pioneering evolutionary earth models using approximate and computationally inexpensive forward simulators are already operational (Debayle *et al.* 2016). However, to be practical, an evolutionary approach for FWI must (1) avoid costly re-inversions of all previously considered data and (2) harness second-derivative information from previous iterations using Limited-memory-Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) -type optimization techniques (Nocedal 1980; Liu & Nocedal 1989; Nocedal & Wright 1999).

Here, we propose an optimization and inversion framework, which aims to address these challenges. It is based on a stochastic gradient descent (SGD) method, using dynamic mini-batches, and it incorporates curvature information through the L-BFGS approximation of the Hessian and its inverse. Similar methods have become popular in machine-learning (Bottou 2010; Byrd *et al.* 2011, 2016; Masters & Luschi 2018). In pure SGD, the randomly chosen updates can improve optimization for non-convex objective functions, as it allows the optimization to escape from saddle points or local minima (Ge *et al.* 2015) by quasi-randomly changing the misfit topography with each view on the data. However, this comes at the cost of slow convergence. Mini-batch gradient descent methods promise to retain the beneficial properties of pure SGD methods for non-convex optimization, while at the same time enabling efficient convergence through the exploitation of redundancies in the data set. Optimization based on mini-batch methods has been shown to reduce computational costs in exploration geophysics (van Leeuwen & Herrmann 2013; Fabien-Ouellet *et al.* 2017; Yang *et al.* 2018; Matharu & Sacchi 2019) and has been applied in medical imaging as well (Boehm *et al.* 2018). To our knowledge, these methods have not found their way to seismological problems. In this contribution, we present a framework for mini-batch optimization with adaptive batch size that is particularly well suited for specific challenges faced in seismology. These include, for instance, the large heterogeneity between the number of receivers per source and uneven source–receiver distributions. In addition, we show that the method offers significant benefits by allowing for the flexible integration of new data, faster model convergence, constant window updating, as well as enabling the use of large data sets without increasing iteration costs.

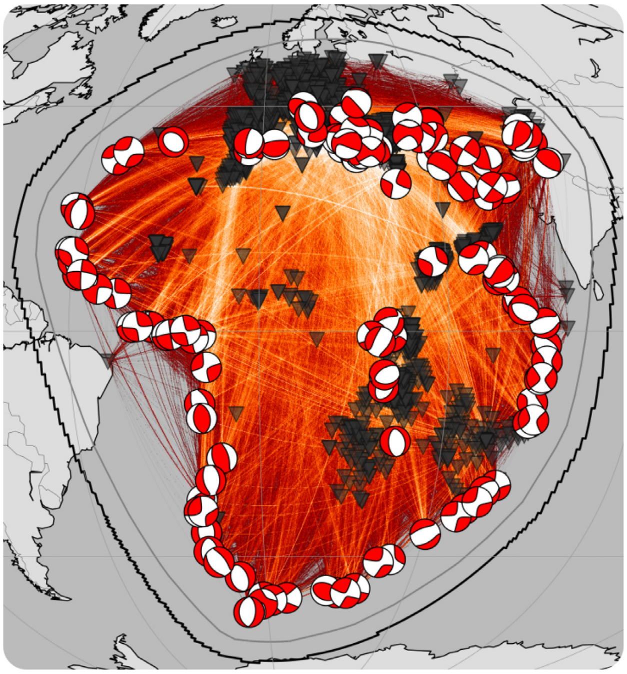


Figure 1. Surface ray coverage. Earthquake locations and mechanisms are indicated by beach balls and stations by black triangles. Bright colours represent a relatively higher density of rays. In total 125 earthquakes were recorded at 2648 unique stations. The complete data set has 51 865 source–receiver pairs. The black line marks the edge of the computational domain and the grey line marks the start of the absorbing boundary region.

This manuscript is structured as follows: In Section 2, we briefly introduce the study region and the data set that we later use to illustrate our developments. Subsequently, in Section 3, we present the dynamic mini-batch optimization for FWI, the corresponding inversion workflow, as well as the modelling tools used in the examples. In Section 4, we show synthetic and real-data inversions, illustrating that our approach converges and that it does so at significantly lower computational cost than mono-batch FWI, where all sources are used in each iteration. Since the focus of this work is on methodological improvements, the interpretation of the final model is deferred to a future publication.

2 EXAMPLE APPLICATION

Throughout this paper, we illustrate our developments using a data set that covers the African Plate. Coverage, summarized in Fig. 1, is uneven and partly sparse, thus making it a suitable test region for the proposed approach. We select 125 earthquakes from the Global Centroid Moment Tensor (GCMT) Catalog (Ekström *et al.* 2012). Magnitudes range from 5.5 to 6.7. This range empirically provides a good signal in the frequency band of interest while minimizing finite-source effects (Vallée 2013). Recordings from AfricaArray and the Network of Autonomously Recording Seismographs (NARS) were complemented with data that are publicly available through the International Federation of Digital Seismograph Networks (FDSN) (Romanowicz & Dziewonski 1986) Web Services. The earthquakes were recorded at 2648 unique stations. Since not each station was online for each earthquake, the data set contains 51 865 unique source–receiver pairs. The considered period band is 65–120 s.

This source–receiver geometry is used for both synthetic and real-data inversions, the results of which are described in more detail in Sections 4.1 and 4.2, respectively. For the real-data inversion, the first generation of the collaborative seismic earth model (Fichtner *et al.* 2018) is used as initial model. In the synthetic inversions, the initial model is Preliminary reference Earth model (PREM) (Dziewoński & Anderson 1981).

While our method is applicable in combination with arbitrary misfit functions, the example inversions employ time- and frequency-dependent phase misfits (Fichtner *et al.* 2008). This misfit measure does not require the isolation of specific phases and largely ignores less reliable amplitude information. To balance sensitivity across regions, we employ a station weighting scheme that empirically leads to faster convergence. For this, misfits at station location $\mathbf{x}_r$ are multiplied by the factor

$$W_r = \left(\sum_{i=1, i \neq r}^{n} \frac{1}{|\mathbf{x}_i - \mathbf{x}_r|} \right)^{-1}, \tag{1}$$

where n is the total number of all other station locations $\mathbf{x}_i$. For a review of station weighting methods in the context of regional to global scale FWI, the reader is referred to Ruan *et al.* (2019).

3 STOCHASTIC GRADIENT DESCENT WITH DYNAMIC MINI-BATCHES

In this section, we describe all steps from forward modelling to the inversion framework. While our method is independent of the particular numerical wave propagation solver, we perform all forward and adjoint modelling using the Salvus software suite (Afanasiev *et al.* 2019). This is a fully 3-D implementation of the spectral-element method (Seriani & Priolo 1994; Faccioli *et al.* 1996, 1997; Komatitsch & Vilotte 1998), which has the advantages of implicitly satisfied free-surface boundary conditions and geometric flexibility.

3.1 Optimization scheme

The goal of most deterministic inverse problems is to find a model that explains data within their observational errors. In seismology, this is commonly done by defining and minimizing a misfit function $\chi(\mathbf{m})$ that quantifies differences between observed seismograms and synthetic seismograms computed for earth model $\mathbf{m}$. The misfit function is composed of individual misfits, each corresponding to one of N sources. To facilitate a varying amount of sources, we define $\chi(\mathbf{m})$ as a sample average, where each sample i is the misfit contribution $\chi_i(\mathbf{m})$ from a single source and forward simulation, that is,

$$\chi(\mathbf{m}) = \frac{1}{N} \sum_{i=1}^{N} \chi_i(\mathbf{m}). \tag{2}$$

The individual misfits $\chi_i(\mathbf{m})$ themselves contain the misfits of all seismic traces related to that source. The function χ may be approximated by a randomly or systematically chosen mini-batch B, which is a subset of all N sources and normalized by the number of sources, $|B|$, in the batch,

$$\chi(\mathbf{m}) \approx \chi_B(\mathbf{m}) = \frac{1}{|B|} \sum_{i \in B} \chi_i(\mathbf{m}). \tag{3}$$

As described in the following paragraphs, the composition of the batches varies per iteration in such a way that information from the entire data set is still incorporated during the optimization procedure. Iterative optimization methods subsequently update the model $\mathbf{m}_k$ in the kth iteration as

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \mathbf{s}_k, \tag{4}$$

where the update $\mathbf{s}_k$ can be computed with various optimization methods using the mini-batch B_k of the kth iteration. Here, we propose a mini-batch variant of the trust-region method (Nocedal & Wright 1999; Conn *et al.* 2000), which determines the update $\mathbf{s}_k$ by solving the trust-region subproblem. For this, we consider a quadratic approximation q_k of the misfit function χ, using Taylor's expansion around the current model $\mathbf{m}_k$,

$$\min q_k(\mathbf{s}) = \chi_{B_k}(\mathbf{m}_k) + \nabla \chi_{B_k}(\mathbf{m}_k)^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H}_k \mathbf{s}$$
$$\text{subject to} \quad \|\mathbf{s}\| \leq \Delta_k. \tag{5}$$

Here and in contrast to line-search methods, the model update is computed in a single step, without the separation into a search direction and a step length. The trust-region radius Δ_k limits the maximum distance between two consecutive models and automatically adapts to the quality of the approximation during the iterations. In eq. (5), we replaced the misfit χ and its gradient $\nabla\chi$ by the mini-batch approximations χ_{B_k} and $\nabla\chi_{B_k}$, respectively. Furthermore, $\mathbf{H}_k$ is the L-BFGS approximation of the Hessian, computed from previously calculated mini-batch gradients and model updates.

Since misfits and gradients are not computed for all sources in each iteration, it is not possible to check that the total misfit decreases in each iteration, that is, that $\chi(\mathbf{m}_{k+1}) < \chi(\mathbf{m}_k)$ for all $k > 0$. A common mitigation in conventional stochastic gradient methods is to use diminishing step sizes (Nemirovski *et al.* 2009; Byrd *et al.* 2016). However, this has two major drawbacks: (1) With an increasing number of iterations, the model updates become very small and (2) the updates do not utilize any previously accumulated curvature information, which may lead to slow convergence. Since the composition of the mini-batches changes (quasi-randomly) from one iteration to the next, additional simulations would be required to ensure that the mini-batch approximation of the misfit has decreased, meaning that

$$\chi_{B_{k+1}}(\mathbf{m}_{k+1}) < \chi_{B_k}(\mathbf{m}_k). \tag{6}$$

For these reasons, we define a control group $C_k \subset B_k$, consisting of a subset of sources of the current mini-batch. This control group remains in the mini-batch of the subsequent iteration, that is, $C_k \subset B_{k+1}$. As explained in Section 3.2, the events in the control group are selected dynamically, and therefore we generally do not have $C_k = C_{k+1}$. This concept is visualized in Fig. 2.

The purpose of the control group is to accept or reject proposed model updates and to steer the sizes of the mini-batch, the control group itself and the trust-region radius. Using the L-BFGS approximation of the inverse Hessian, we solve eq. (5) approximately using the dogleg method (Nocedal & Wright 1999) and obtain a trial model update $\mathbf{s}_k$. This only requires a few vector products, but neither storing a matrix nor solving a linear system. Next, we compute the misfit reduction of the mini-batch sample average that is

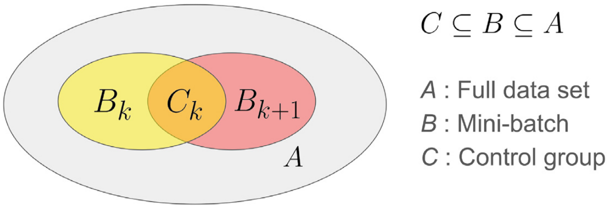


Figure 2. Schematic representation of the mini-batch approach. Mini-batches B_k for iteration k are a subset of the complete data set A, and the control group C_k is a subset of the current mini-batch. The mini-batch for the next iteration, B_{k+1}, consists of events that were chosen as control group events from the latest mini-batch B_k as well as other events that are quasi-randomly chosen from the full data set.

predicted by the quadratic model q_k for the trial model $\mathbf{s}_k$:

$$\rho^{B_k}_{\text{pred.}} = q_k(\mathbf{s}_k) - \chi_{B_k}(\mathbf{m}_k)\,. \tag{7}$$

If $\rho^{B_k}_{\text{pred.}} \geq 0$, the curvature information extracted from the previous batches is inconsistent, and the L-BFGS approximation of the inverse Hessian is not positive definite. In this case, the size of the mini-batch needs to be increased or the curvature information reset. Next, we compute the predicted misfit reduction for the control group

$$\rho^{C_k}_{\text{pred.}} = \nabla\chi_{C_k}(\mathbf{m}_k)^T\mathbf{s}_k + \frac{1}{2}\mathbf{s}_k^T\mathbf{H}_k\mathbf{s}_k\,, \tag{8}$$

which corresponds to the quadratic model obtained from using not only the gradients from the control group but also all available curvature information. If $\rho^{C_k}_{\text{pred.}} \geq 0$, the curvature information between the control group and the entire batch is inconsistent, and more events of the current batch need to be added to the control group. Otherwise, if $\rho^{C_k}_{\text{pred.}} < 0$, we continue with computing the misfits of the control group events for the trial model, $\mathbf{m}_k + \mathbf{s}_k$, to obtain the actual misfit reduction for the control group

$$\rho^{C_k}_{\text{act.}} = \chi_{C_k}(\mathbf{m}_k + \mathbf{s}_k) - \chi_{C_k}(\mathbf{m}_k)\,. \tag{9}$$

If $\rho^{C_k}_{\text{act.}} < 0$, the model update $\mathbf{s}_k$ is accepted, and we proceed with the next iteration. Otherwise, we need to repeat the previous steps with a smaller trust-region radius to improve the quality of the quadratic approximation q_k.

As a final step, we update the trust-region radius based on the ratio of actual and predicted reductions of the control group misfit, $\rho^{C_k}_{\text{act.}}/\rho^{C_k}_{\text{pred.}}$. This is a standard procedure in trust-region methods (Conn *et al.* 2000), except that we only consider events from the control group to compute this ratio. If the ratio is significantly smaller than 1, the approximation of the quadratic approximation q_k was poor, and we decrease the trust-region radius Δ_k for the next iteration. Otherwise, we may increase Δ_k to allow for larger model updates in the following iteration. This procedure is identical to algorithm 4.1 of Nocedal & Wright (1999), except that the trust-region radius is halved when the ratio is smaller than 0.25.

The advantages of this strategy are threefold: First, the composition of the mini-batch is fully dynamic and allows us to interchange sources, as well as measurement time windows, for computing misfits in every iteration. Second, we retain curvature information using the L-BFGS approximation of the inverse Hessian, and thus we can use curvature information to influence the computed descent direction. Third, with the help of the control group and the trust-region framework, we ensure convergence without the need for additional simulations to evaluate misfits or gradients for particular events.

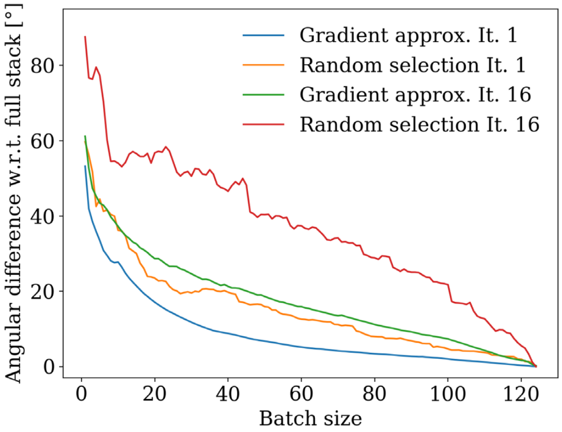


Figure 3. Angle between the gradient for the complete data set and mini-batch gradient approximations or random event selection for variable batch sizes, ranging from 1 to 125 (all sources). At iteration 1, the gradient is dominated by few prominent features, meaning that there is significant redundancy in the data set. Therefore, a smaller amount of sources is able to provide a gradient that is close to the complete gradient. As the model converges (towards iteration 16), gradients tend to contain more short-wavelength structure, and each individual source becomes more important to further improve the model. Note that the gradient approximation algorithm results in significantly smaller batch size without increasing the angular difference. The gradient approximation is made for the gradient with respect to the model parameters v_{sv}, v_{sh} and v_{pv}.

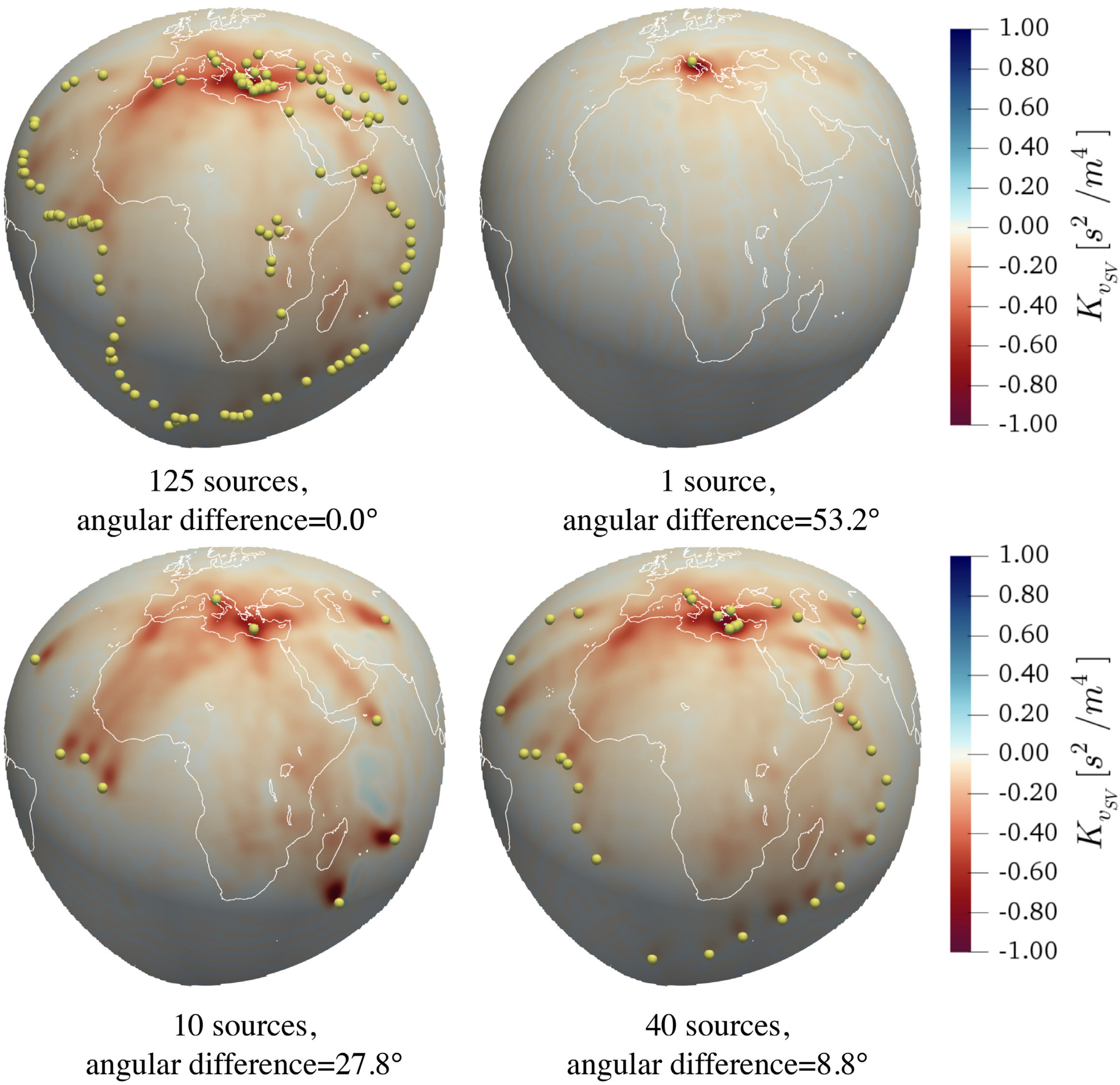
The actual rules to determine the sizes and compositions of the mini-batches and control groups will be explained in the following section.

3.2 Selection of mini-batch and control group sources

In principle, a variety of strategies could be employed to select sources for both the mini-batches and the control group. Our specific approach rests on the observation that uneven coverage in regional to global tomography often causes a small number of sources to dominate both the misfit and the gradient. This effect is related to the variation in the number and quality of data recordings for each event. Therefore, it is possible to approximate the gradient for the complete data set using a smaller subset of sources.

The selection of sources for the mini-batch B is a multistage process, starting with Mitchell's Best-Candidate algorithm (Mitchell 1991) to select sources that have not been used in previous iterations. For this, the source furthest away from the already selected sources is added to the mini-batch. The first source in the initial mini-batch is chosen randomly. This approach ensures that all available sources are incorporated quickly, while homogenizing spatial coverage in each iteration. As the iterative inversion progresses, the number of events in the mini-batches, needed to approximate the complete gradient, typically increases. This is illustrated in Fig. 3, which shows the angular difference between the complete gradient and mini-batch gradients for variable mini-batch sizes at iterations 1 and 16. For comparison with the previously discussed gradient approximation method, we also include a comparison with simple random event selection. Note that with the gradient approximation algorithm, significantly fewer sources are required for an equally good approximation.

Control group events are selected by attempting to approximate the gradient of the mini-batch. This is done iteratively, by removing one source at a time from the mini-batch and measuring the angle



125 sources,
angular difference=0.0°

1 source,
angular difference=53.2°

10 sources,
angular difference=27.8°

40 sources,
angular difference=8.8°

Figure 4. Examples of the mini-batch gradient approximation. Shown are spherical slices at 100 km depth through the normalized sensitivity kernels $K_{v_{sv}}$ with respect to the *SV* velocity, v_{sv}, computed for the initial model of the real-data inversion. Source locations are indicated by the yellow circles. The dominant feature in the full gradient for all 125 sources (top left) suggests that an increase in v_{sv} is required in the Eastern Mediterranean to reduce the misfit. The gradients shown in the top right, bottom left and right respectively are the approximations with 1, 10 or 40 sources with corresponding 53.2°, 27.8° and 8.8° angular differences. As the number of sources that are used for the approximation increases, the quality improves. It is however import to note that each of the shown approximations lie within 90° of the full gradient, and, would therefore still provide a direction of descent for the full problem.

between the gradients of the reduced test control group and the complete mini-batch. The source that results in the smallest angular change between the test control group gradient and the mini-batch gradient is removed. Essentially, we remove those sources from the current batch that have the smallest influence on the search direction. This iterative process can be described as choosing the source

$$s_{i+1} = \arg\max_{s\in B_i} \left\{ \frac{\nabla\chi^T_{C_i/\{s\}}\nabla\chi_B}{\|\nabla\chi_{C_i/\{s\}}\|\cdot\|\nabla\chi_B\|} \right\}, \qquad (10)$$

in order to update the control group as $C_{i+1} = C_i/s_{i+1}$. The initial test control group C_0 is equal to the complete mini-batch B. This process can then either be terminated once the control group reaches a predefined minimum size or when the angular difference between the mini-batch gradient and the control group gradient reaches a certain threshold value. Examples of mini-batch gradient approximations are shown in Fig. 4. The complete *SV* velocity gradient, computed for all 125 sources with respect to the initial model, is dominated by a negative contribution in the Eastern Mediterranean. This feature is preserved when the gradient is computed with fewer sources. We define the size of the subsequent mini-batch to be twice the size of the control group to ensure that we continue to sample from the remaining events.

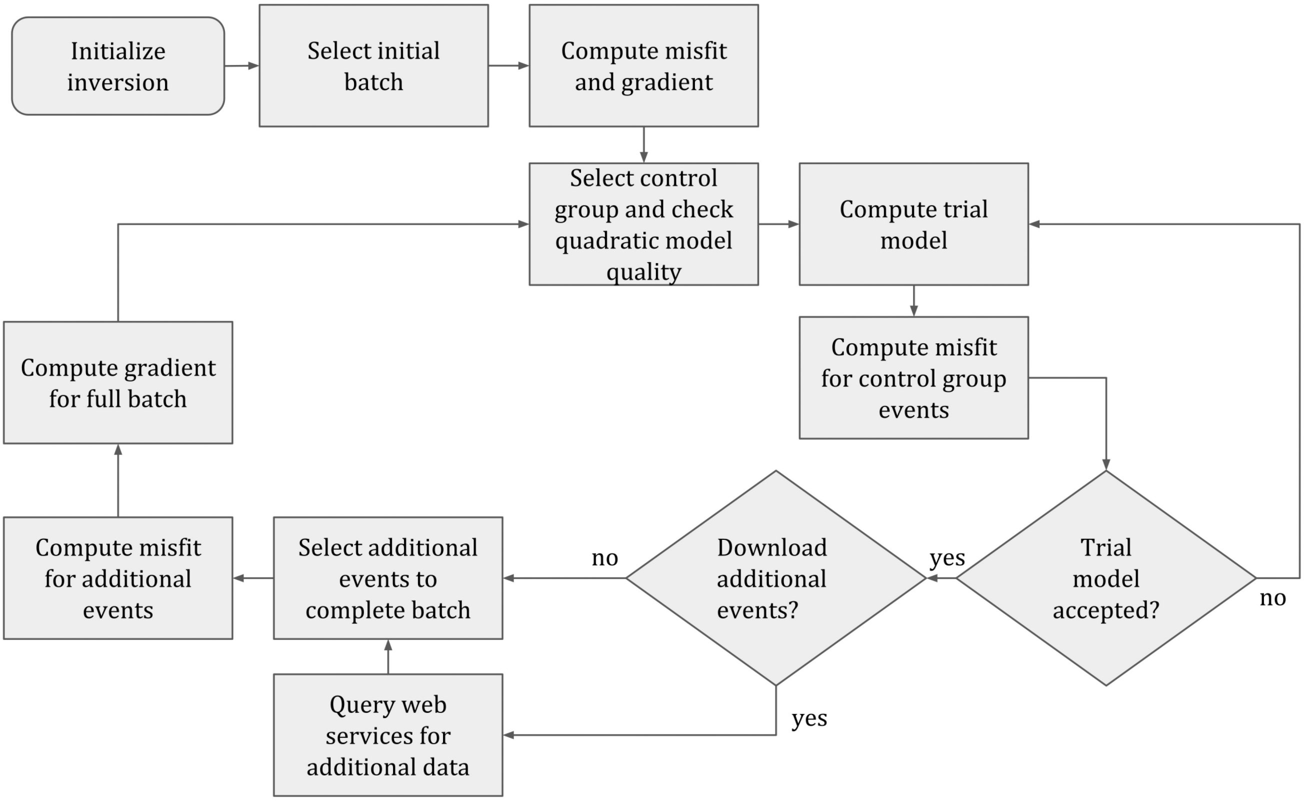

Initialize inversion
Select initial batch
Compute misfit and gradient
Select control group and check quadratic model quality
Compute trial model
Compute misfit for control group events
Trial model accepted?
no
yes
Download additional events?
no
yes
Select additional events to complete batch
Query web services for additional data
Compute misfit for additional events
Compute gradient for full batch

Figure 5. Flowchart summarizing the inversion procedure. The above workflow allows us to easily incorporate new data each time a new model is accepted. Additionally, measurement windows can be re-selected each time a source is not included in the control group. This naturally enables the use of a dynamically changing objective function that comprises an increasing number of measurements.

Additionally, we store the removal order $(i + 1)/|B|$, which provides information on which sources had a large effect on the batch gradient and which sources had less influence. This information is updated every time a source is removed from a batch and then used to assign probabilities for picking previously used sources randomly to complete subsequent mini-batches. The probability is determined by the removal order and normalized such that the total probability equals 1. This process constitutes a dynamic optimal experimental design problem. Typically, the model converges first in regions where the gradient is dominant during the first few iterations. As a consequence, the gradient in those regions decreases, and events constraining other parts of the domain obtain a relatively higher likelihood of being selected for the subsequent mini-batches.

3.3 Workflow

The flowchart in Fig. 5 summarizes the inversion procedure. In the initialization stage, the complete data set is assembled, and an initial set of sources for the first mini-batch is selected using Mitchell's best candidate algorithm. Subsequently, synthetic seismograms, misfits and gradients are computed. Using the mini-batch gradients, the first control group is selected. With the gradient information available, a model update can be computed together with the control group misfits. When the model is accepted at this stage, new sources are selected to complete the next mini-batch. Misfits and gradients are then again computed for the mini-batch, which now contains both the newly added sources and the control group sources carried over from the previous mini-batch. With the new gradients available, the next control group is selected for the next iteration.

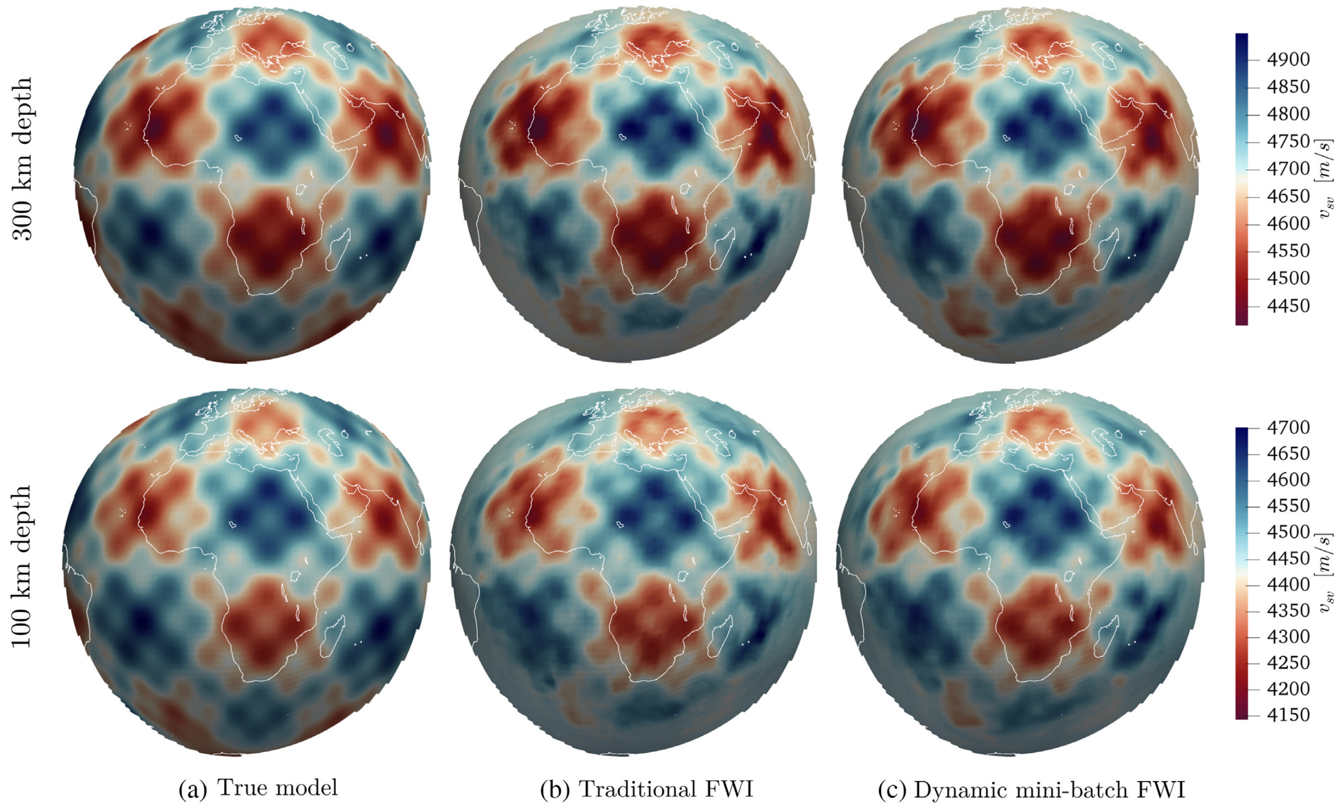
Measurement windows for which misfits are computed can be reselected for the newly added sources. This continued window reselection allows us to increase the number of measurements as the model improves, thereby increasingly avoiding cycle skips.

4 SYNTHETIC AND REAL-DATA INVERSIONS

To illustrate the proposed optimization scheme, we present two inversion examples, using the scenario described in Section 2. A synthetic inversion allows us to quantify the quality of the recovery in relation to the computational costs. In the subsequent real-data inversion, we compare the dynamic mini-batch approach to a more traditional mono-batch inversion with L-BFGS optimization, where all data are used in each iteration. All gradients are smoothed to prevent subwavelength structure from entering the model by effectively convolving the gradients with a Gaussian filter with a standard deviation of 150 km. Such an effective convolution is efficiently implemented through the numerical solution of the diffusion equation (Afanasiev *et al.* 2018).

4.1 Synthetic inversion

For the synthetic inversion, we set the minimum control group size to three sources, in order to begin with a reasonable coverage of the study region. Additionally, we set the maximum allowable angle between the control group and the mini-batch gradient to 22.5°, and



(a) True model (b) Traditional FWI (c) Dynamic mini-batch FWI

Figure 6. Comparison of synthetic mono-batch and dynamic mini-batch FWI. (a) *SV* velocity, v_{sv}, in the target model, used to compute artificial waveform data. (b) Reconstructed v_{sv} after 21 mono-batch iterations with all 125 sources. (c) Reconstructed v_{sv} after 61 dynamic mini-batch iterations. Both methods produce similarly good reconstructions in areas with sufficient coverage. However, the total number of simulations, forward plus adjoint, is around 5300 for the mono-batch FWI and 1300 for the dynamic mini-batch FWI.

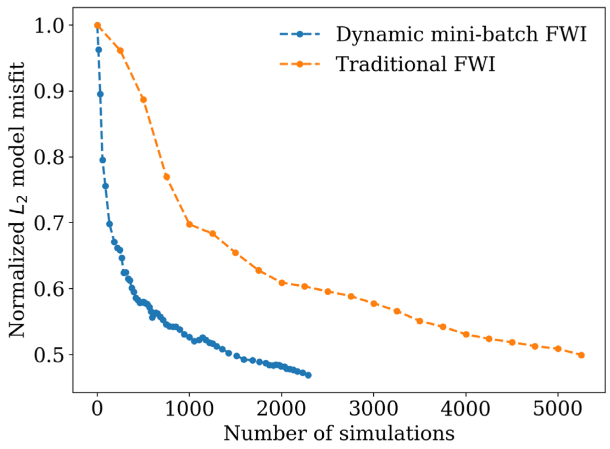


Figure 7. Normalized L_2 model misfit versus the combined number of forward and adjoint simulations. Each dot represents an iteration. For a given number of iterations the mono-batch FWI retrieves a better model. However, since the computational cost is directly tied to the number of simulations, the dynamic mini-batch inversion recovers a model of similar quality at a substantially lower computational cost. The model misfit does not converge to zero because the coverage is insufficient to constrain the entire domain.

define the mini-batch size for the subsequent iteration to be twice the previous control group size. These values were empirically found to produce good results, and may vary when other data sets are considered.

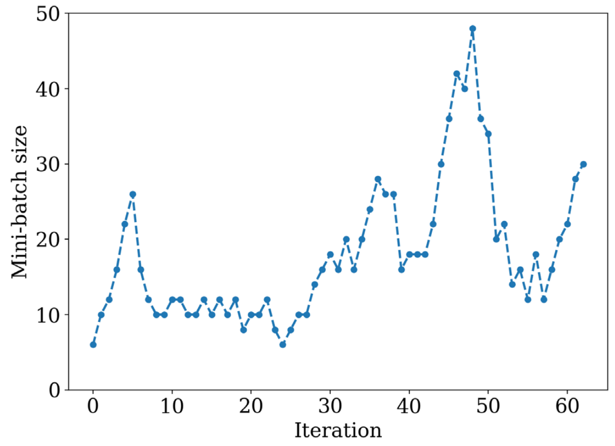


Figure 8. Development of the mini-batch sizes over the course of 61 dynamic mini-batch iterations. Note the general trend to increase the mini-batch size. This is to be expected as individual sources start to contribute more unique information as the model improves. The subsequent mini-batch size is always twice the size of the last iteration's control group. Peaks occur when the degree of redundancy in the batch decreases. In this case, each event is detected to contribute unique information, and the algorithm attempts to include as many of these unique directions as possible. If a few sources dominate the search direction, or events point in a similar direction, the batch size is shrunk.

The earth model is radially anisotropic and parametrized in terms of density ρ, the wave speeds v_{pv}, v_{ph}, v_{sv} and v_{sh}, and the dimensionless parameter η. The target model, used to compute artificial

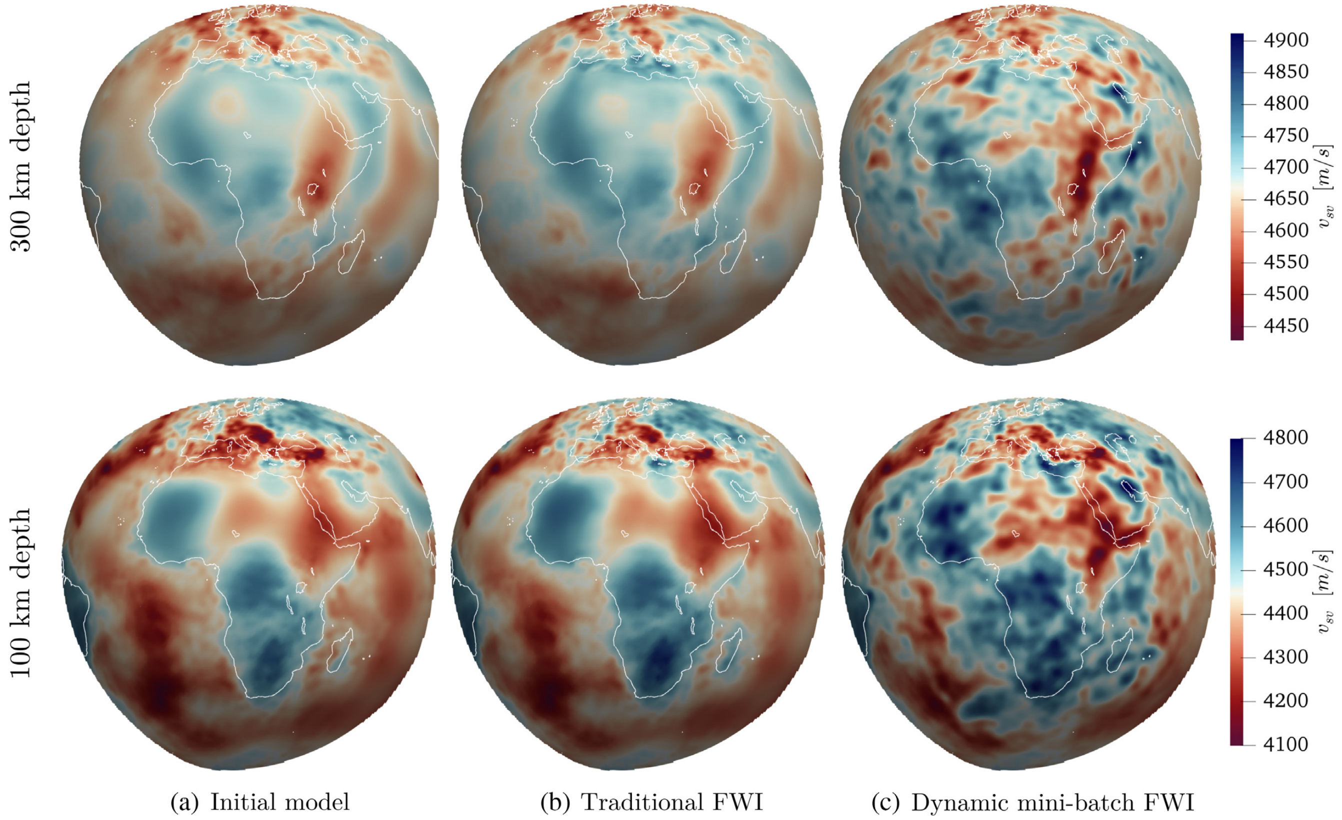


Figure 9. Mono-batch and dynamic mini-batch inversions with similar computational cost, corresponding to around 750 simulations, forward plus adjoint. (a) Initial v_{sv} distribution (Fichtner *et al.* 2018). (b) Mono-batch inversion result after 3 iterations, each including all 125 sources. (c) Dynamic mini-batch inversion after 40 iterations, with batch sizes around 10. The total data misfit reduction for the mono-batch FWI is 21.36 per cent with respect to the initial model misfit. For the dynamic mini-batch FWI, it is 50.99 per cent.

data, contains equal perturbations of all velocities with respect to PREM (Dziewoński & Anderson 1981), which also serves as initial model. To assess recovery on different scales, short wavelength anomalies with a 2 per cent perturbation are placed on top of long wavelength anomalies with a 5 per cent perturbation. This is illustrated in Fig. 6(a). During the inversion, we enforce $v_{\mathrm{pv}} = v_{\mathrm{ph}}$ and $\eta = 1$ because we do not expect our long-period data set to resolve *P*-wave anisotropy.

Figs 6(b) and (c) show the recovered v_{sv} models after 21 iterations of a mono-batch FWI and 61 iterations using the dynamic mini-batch approach, respectively. While both methods recover the velocity anomalies in the regions with sufficient coverage, the latter required a significantly smaller number of forward and adjoint simulations, and therefore less computational resources. This is quantified in Fig. 7, which shows model misfit in terms of the L_2 norm of the model parameter residuals, normalized by the L_2 norm of the initial model. To achieve a similar model misfit reduction, the dynamic mini-batch inversion requires only around 25 per cent of the simulations needed by the mono-batch FWI where all sources are used in each iteration.

Fig. 8 shows the adaptive sizes of the mini-batches as a function of iteration number. Since the mini-batch size was tied to the allowable angular difference, it automatically changed during the inversion. Note the general trend of increasing mini-batch size. This effect is to be expected as we gradually need more sources to approximate the complete gradient, as indicated before in Fig. 3.

4.2 Real-data inversion

To demonstrate the practical applicability of our method, we perform a real-data FWI using the setup described in Section 2, which is identical to the previously presented synthetic inversion. We compare two cases, the ‘traditional’ approach where all data is used for each model update, and the dynamic mini-batch approach. Since the focus of this contribution is on the method and not on the model, we defer a geologic interpretation to a later publication.

To prepare for the inversion, we processed data using ObsPy (Megies *et al.* 2011; Krischer *et al.* 2015b). This included linear detrending, removal of the instrument response, and bandpass filtering to 65–120 s period. This period band allows us to keep the computational requirements relatively low, while avoiding cycle-skipping problems.

To compare mono-batch FWI with the dynamic mini-batch approach, we first contrast inversion results where the total number of forward and adjoint simulations is similar, around 750. In the mono-batch approach, 750 simulations correspond to three iterations, each requiring a forward and an adjoint run for each of the 125 sources. The final model in Fig. 9(b) still closely resembles the initial model (Fichtner *et al.* 2018) in Fig. 9(a). For the nearly identical computational cost, the dynamic mini-batch approach was able to perform 40 iterations with mini-batch sizes around 10. The resulting model, displayed in Fig. 9(c), contains substantially more detail than the initial model and the mono-batch inversion result. While a detailed resolution analysis is beyond the scope of this

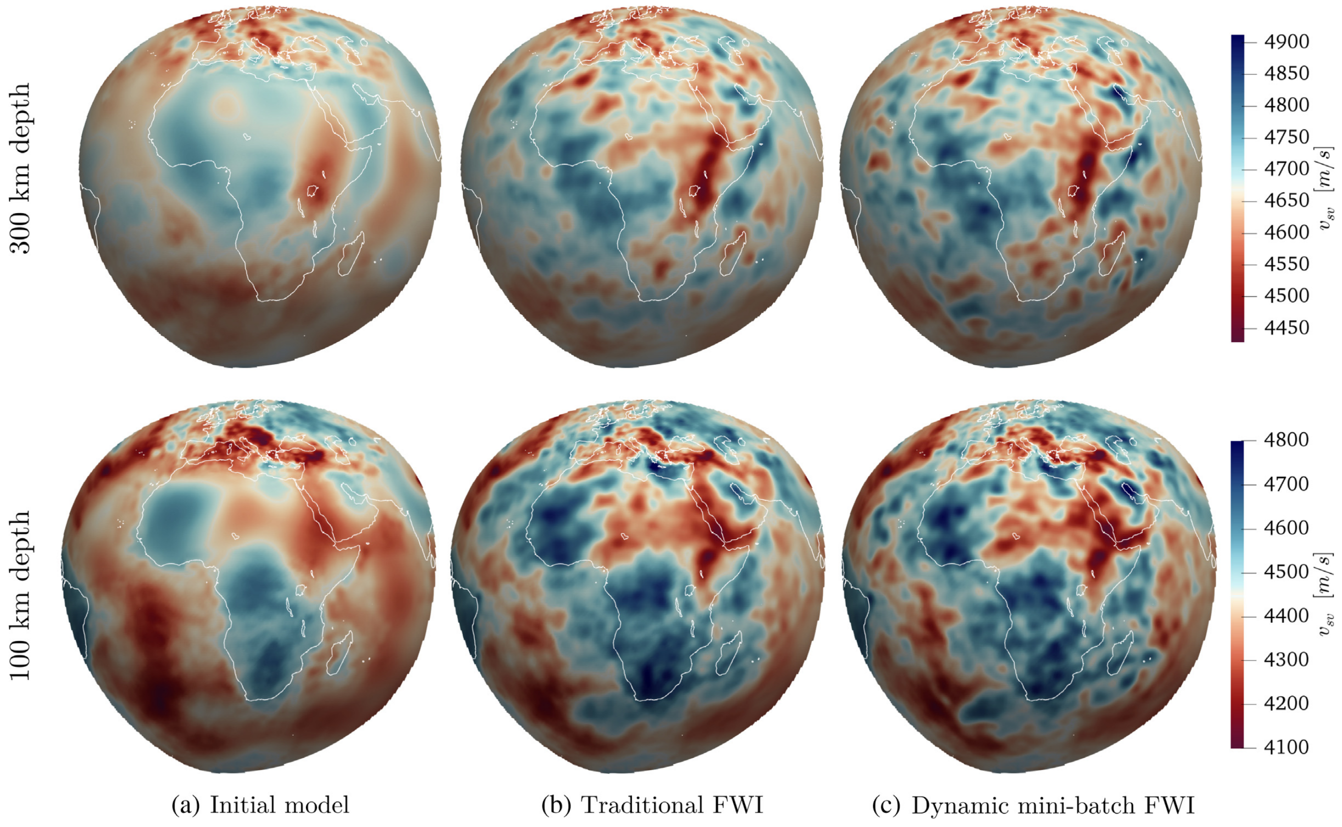


Figure 10. Mono-batch and dynamic mini-batch inversions with similar data misfit reduction. (a) Initial v_{sv} distribution (Fichtner *et al.* 2018). (b) Distribution of v_{sv} after mono-batch inversion with 16 iterations, corresponding to 4000 simulations, forward and adjoint combined. (c) Distribution of v_{sv} after 40 dynamic mini-batch inversions, corresponding to around 800 simulations. The total data misfit reductions are comparable, 48.23 per cent for the mono-batch inversion, and 50.99 per cent for the dynamic mini-batch inversion.

work, we note that the data misfit reduction relative to the misfit for the initial model is 21.36 per cent for the mono-batch approach. In the dynamic mini-batch inversion, the misfit is reduced by 50.99 per cent.

Fig. 9 motivates the question if more similar results could be found if the mono-batch inversion was continued further. In fact, after additional 13 mono-batch iterations, the misfit is reduced by 48.23 per cent, closely approaching the 50.99 per cent of the dynamic mini-batch inversion. Also, the corresponding earth models, displayed in Fig. 10, are visually more similar. However, to achieve this result with the mono-batch FWI, we required a total of 16 × 250 = 4000 simulations, compared to 750 for the dynamic mini-batch version. Thus the dynamic mini-batch inversion only required 19 per cent of the computational costs.

5 DISCUSSION

We presented a quasi-random mini-batch optimization technique with adaptive batch size and its application to full seismic waveform inversion. In the following paragraphs, we will discuss the extent to which the method meets the goals formulated in Section 1, namely (1) the reduction of computational cost and (2) the easy integration of new data without the need to re-invert the complete data set. Furthermore, we will discuss other advantages and limitations of the method, as well as its relation to previous work.

5.1 Computational efficiency

As shown in both the synthetic and real-data examples, the dynamic mini-batch approach converges significantly faster than the mono-batch inversion. However, we note that the relative convergence of two very different methods is not easy to compare. In this context, we note that the two real-data inversion results in Fig. 10 are not exactly identical for various reasons. Most importantly, batch gradients merely approximate the gradient of the complete data set. The quality of the approximation depends on the size of the mini-batches, but also on the amount of noise in the data. One may argue that mini-batch sizes should be increased to ensure a close approximation of the minimum. On the other hand, we found that a limited batch size, in fact, helps to avoid over-fitting because random noise is harder to fit by being inconsistent between different subsets of events. Thus, as in any real-data application, careful preliminary inversion experiments are required to ensure that meaningful results can be obtained at optimally low cost.

Our approach fundamentally rests on the presence of redundancies in the data set. This makes it particularly suitable for seismological applications based on earthquake data. Earthquake hypocentres tend to cluster, thereby naturally introducing redundancy that the dynamic mini-batch approach can exploit. This also implies that the approach may have less benefits for source–receiver configurations, where redundancies are minimized, for example, with optimal experimental design (Curtis 1999; Martiartu *et al.* 2017; Maurer *et al.* 2017). More generally, the dynamic mini-batch optimization must be considered in the framework of the no-free-lunch theorem (Wolpert & Macready 1997; Mosegaard 2012), loosely stating that the efficiency of any method does not rest within the method itself but in its application to suitable problems.

5.2 Evolutionary FWI

The dynamic mini-batch approach has the advantage of being 'evolutionary' on two levels: First of all, it naturally allows us to incorporate new data without having to re-invert the whole data set, which is done in evolutionary inversions based on computationally less expensive forward problem approximations (Debayle *et al.* 2016). This ability rests on the interpretation of the complete data set in terms of both sources that have already acted and sources that will act in the future. In this sense, adding data from new sources, for example, from quasi-randomly occurring earthquakes and seismic array deployments, simply corresponds to the selection of mini-batch members that had not been selected before.

Second, each time a source enters a mini-batch without being part of the control group, we adapt the measurement windows. This allows us to add new measurement windows and to extend existing ones, as the waveform fit gradually improves during the inversion.

5.3 Limitations

One of the obvious disadvantages of our approach is that it adds complexity to the already complex FWI workflow. Additionally, the size of the mini-batch is effectively determined by the desired quality of the gradient approximation, which is a tuning parameter. Setting this tuning parameter appropriately requires intuition for the problem. If the batch size is taken too small, this might lead to slow convergence. If it is taken too large, one may miss some of the performance benefits. Furthermore, since misfits are not evaluated for the entire data set at each iteration, the mini-batch approach does not provide a misfit reduction curve, commonly used to assess convergence. Instead, the convergence curve can only be approximated by successive mini-batch misfits.

Another drawback is the potential imprint of so-called 'stochastic noise' in the model, where the contributions from gradients with respect to individual sources can easily be recognized in the model. This effect is especially evident in the early phases of the inversion and is likely to be stronger when using higher frequency data and their narrower associated Fresnel zones. Using larger batches and/or performing more iterations helps to mitigate this effect. The synthetic example shows that for a given number of iterations, mono-batch FWI always produces a model that is closer to the true model (see Fig. 7). The dominant cost in FWI, however, is not the total number of iterations, but the total number of wavefield simulations required. In this metric, the stochastic mini-batch method presented in this paper outperforms mono-batch FWI, even though a higher total number of total iterations are required to mitigate the stochastic noise.

5.4 Comparison to similar approaches

Other ideas have been proposed to accelerate FWI, such as source-stacking (Capdeville *et al.* 2005; Krebs *et al.* 2009; Romanowicz *et al.* 2019). Although source-stacking theoretically provides significant computational savings, we are unaware of any real data applications. Missing data, artefacts in gradients due to cross-talk between forward and adjoint wavefields and workflow complexity may have contributed to this. Recent developments (Tromp & Bachmann 2019) have shown promising results to address the first two challenges by utilizing superpositions of monochromatic sources. We think it is likely that both methods carry value, depending on the nature of the problem. It is important to note that they are not mutually exclusive, that is, there is nothing that prevents one from using source-stacks within the framework of a mini-batch inversion.

6 CONCLUSIONS

We presented a novel FWI approach that can lead to significant computational savings, consistently accumulates and exploits curvature information, and enables an evolutionary mode where new data can be incorporated without re-inverting the complete data set.

The method is based on a variant of stochastic gradient descent, specifically adapted to applications in seismic tomography. Quasi-random mini-batches of sources, for example earthquakes, are used to approximate the misfit and the gradient for the complete data set. The size of the mini-batches is dynamic and mostly controlled by the desired quality of the gradient approximation. Furthermore, members of a mini-batch are chosen to (1) homogenize spatial coverage and (2) exploit redundancies in the data set.

Our synthetic and real-data inversions for upper-mantle structure beneath the African Plate indicate that the dynamic mini-batch approach requires around 20 per cent of the computational resources in order to achieve data and model misfits that are comparable to those achieved by a standard FWI where all sources are used in each iteration. Naturally, these numbers will depend on the specific application and in particular on the extent to which a given data set is redundant.

ACKNOWLEDGEMENTS

We gratefully acknowledge editor Carl Tape, as well as Gian Matharu, Yder Masson and an anonymous reviewer for their constructive comments that helped to improve the manuscript. This work was supported by the European Research Council (ERC) under the EU's Horizon 2020 programme (grant no. 714069). We also gratefully acknowledge support from the Swiss National Supercomputing Centre (CSCS) in the form of computing time grants c13 and s868. In addition, we would like to acknowledge Martin van Driel for his help with generating the meshes used for the simulations.

REFERENCES

Afanasiev, M., Boehm, C., van Driel, M., Krischer, L. & Fichtner, A., 2018, Flexible high-performance multiphysics waveform modeling on unstructured spectral-element meshes, in *SEG Technical Program Expanded Abstracts 2018,* pp. 4035–4039.

Afanasiev, M., Boehm, C., van Driel, M., Krischer, L., Rietmann, M., May, D.A., Knepley, M.G. & Fichtner, A., 2019. Modular and flexible spectral-element waveform modelling in two and three dimensions, *Geophys. J. Int.,* **216**(3), 1675–1692.

Aki, K. & Lee, W.H.K., 1976. Determination of three-dimensional velocity anomalies under a seismic array using first *P* arrival times from local earthquakes: 1. A homogeneous initial model, *J. geophys. Res.,* **81**(23), 4381–4399.

Aki, K., Christoffersson, A. & Husebye, E.S., 1977. Determination of the three-dimensional seismic structure of the lithosphere, *J. geophys. Res.,* **82**(2), 277–296.

Alkhalifah, T., 2000. An acoustic wave equation for anisotropic media, *Geophysics,* **65,** 1239–1250.

Bamberger, A., Chavent, G. & Lailly, P., 1977. Une application de la théorie du contrôle à un problème inverse sismique, *Ann. Geophys.,* **33,** 183–200.

Bamberger, A., Chavent, G., Hemons, C. & Lailly, P., 1982. Inversion of normal incidence seismograms, *Geophysics,* **47,** 757–770.

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *Geophys. J. Int.,* **178,** 1411–1436.

Boehm, C., Martiartu, N.K., Vinard, N., Balic, I.J. & Fichtner, A., 2018. Time-domain spectral-element ultrasound waveform tomography using a stochastic quasi-Newton method, in *Proceedings of SPIE: Medical Imaging 2018: Ultrasonic Imaging and Tomography,* Vol. 10580, p. 105800H, International Society for Optics and Photonics.

Bottou, L., 2010. Large-Scale Machine Learning with Stochastic Gradient Descent, in *Proceedings of COMPSTAT'2010,* pp. 177–186, Physica-Verlag HD, Heidelberg.

Bozdag, E., Peter, D., Lefebvre, M., Komatitsch, D., Tromp, J., Hill, J., Podhorszki, N. & Pugmire, D., 2016. Global adjoint tomography: first-generation model, *Geophys. J. Int.,* **207**(3), 1739–1766.

Byrd, R., Chin, G., Neveitt, W. & Nocedal, J., 2011. On the use of stochastic Hessian information in optimization methods for machine learning, *SIAM J. Opt.,* **21**(3), 977–995.

Byrd, R., Hansen, S., Nocedal, J. & Singer, Y., 2016. A stochastic quasi-Newton method for large-scale optimization, *SIAM J. Opt.,* **26**(2), 1008–1031.

Cance, P. & Capdeville, Y., 2015. Validity of the acoustic approximation for elastic waves in heterogeneous media, *Geophysics,* **80,** T161–T173.

Capdeville, Y. & Métivier, L., 2018. Elastic full waveform inversion based on the homogenization method: theoretical framework and 2-D numerical illustrations, *Geophys. J. Int.,* **213**(2), 1093–1112.

Capdeville, Y., Chaljub, E., Vilotte, J.-P. & Montagner, J.-P., 2003. Coupling the spectral element method with a modal solution for elastic wave propagation in global earth models, *Geophys. J. Int.,* **152,** 34–66.

Capdeville, Y., Gung, Y. & Romanowicz, B., 2005. Towards global earth tomography using the spectral element method: a technique based on source stacking, *Geophys. J. Int.,* **162,** 541–554.

Capdeville, Y., Stutzmann, E., Montagner, J.-P. & Wang, N., 2013. Residual homogenization for seismic forward and inverse problems in layered media, *Geophys. J. Int.,* **194,** 470–487.

Chang, S.-J. *et al.*, 2010. Joint inversion for three-dimensional S velocity mantle structure along the Tethyan margin, *J. geophys. Res.,* **115,** doi:10.1029/2009JB007204.

Chen, P., Zhao, L. & Jordan, T.H., 2007. Full 3D tomography for the crustal structure of the Los Angeles region, *Bull. seism. Soc. Am.,* **97,** 1094–1120.

Conn, A.R., Gould, N.I.M. & Toint, P.L., 2000. *Trust Region Methods,* SIAM.

Curtis, A., 1999. Optimal experiment design: cross-borehole tomographic examples, *Geophys. J. Int.,* **136**(3), 637–650.

Dahlen, F.A., Hung, S.-H. & Nolet, G., 2000. Fréchet kernels for finite-frequency traveltimes—I. Theory, *Geophys. J. Int.,* **141**(1), 157–174.

Debayle, E., Dubuffet, F. & Durand, S., 2016. An automatically updated *S*-wave model of the upper mantle and the depth extent of azimuthal anisotropy, *Geophys. Res. Lett.,* **43,** 674–682.

Devilee, R.J.R., Curtis, A. & Roy-Chowdhury, K., 1999. An efficient, probabilistic neural network approach to solving inverse problems: inverting surface wave velocities for Eurasian crustal thickness, *J. geophys. Res.,* **104,** 28 841–28 857.

Dziewoński, A.M. & Anderson, D.L., 1981. Preliminary reference Earth model, *Phys. Earth planet. Inter.,* **25,** 297–356.

Dziewonski, A.M., Hager, B.H. & O'Connell, R.J., 1977. Large-scale heterogeneities in the lower mantle, *J. geophys. Res.,* **82**(2), 239–255.

Ekström, G., Nettles, M. & Dziewonski, A.M., 2012. The global CMT project 2004-2010: centroid moment tensors for 13,017 earthquakes, *Phys. Earth planet. Inter.,* **200–201,** 1–9.

Fabien-Ouellet, G., Gloaguen, E. & Giroux, B., 2017. A stochastic L-BFGS approach for full-waveform inversion, in *SEG Technical Program Expanded Abstracts 2017,* pp. 1622–1627.

Faccioli, E., Maggio, F., Quarteroni, A. & Tagliani, A., 1996. Spectral-domain decomposition methods for the solution of acoustic and elastic wave equations, *Geophysics,* **61**(4), 1160–1174.

Faccioli, E., Maggio, F., Paolucci, R. & Quarteroni, A., 1997. 2D and 3D elastic wave propagation by a pseudospectral domain decomposition method, *J. Seismol.,* **1,** 237–251.

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

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

Fichtner, A. *et al.*, 2018. The Collaborative Seismic Earth Model: Generation 1, *Geophys. Res. Lett.,* **45**(9), 4007–4016.

French, S.W. & Romanowicz, B.A., 2014. Whole-mantle radially anisotropic shear velocity structure from spectral-element waveform tomography, *Geophys. J. Int.,* **199**(3), 1303–1327.

Friederich, W., 2003. The *S*-velocity structure of the East Asian mantle from inversion of shear and surface waveforms, *Geophys. J. Int.,* **153,** 88–102.

Ge, R., Huang, F., Jin, C. & Yuan, Y., 2015. Escaping from saddle points—online stochastic gradient for tensor decomposition, *CoRR,* abs/1503.02101.

Gokhberg, A. & Fichtner, A., 2016. Full-waveform inversion on heterogeneous HPC systems, *Comput. Geosci.,* **89,** 260–268.

Gorbatov, A. & Kennett, B.L.N., 2003. Joint bulk-sound and shear tomography for Western Pacific subduction zones, *Earth planet. Sci. Lett.,* **210,** 527–543.

Grand, S., VanDerHilst, R. & Widiyantoro, S., 1997. Global seismic tomography: a snapshot of convection in the earth, *Geol. Soc. Am. Today,* **7**(4), 1–7.

Koelemeijer, P., Deuss, A. & Ritsema, J., 2017. Density structure of Earth's lowermost mantle from Stoneley mode splitting observations, *Nat. Commun.,* **8,** 15241, doi:10.1038/ncomms15241.

Komatitsch, D. & Vilotte, J.P., 1998. The spectral element method: an effective tool to simulate the seismic response of 2D and 3D geological structures, *Bull. seism. Soc. Am.,* **88,** 368–392.

Krebs, J.R., Anderson, J.E., Hinkley, D., Neelamani, R., Lee, S., Baumstein, A. & Lacasse, M.-D., 2009. Fast full-wavefield seismic inversion using encoded sources, *Geophysics,* **74**(6), WCC177–WCC188.

Krischer, L., Fichtner, A., Zukauskaite, S. & Igel, H., 2015a. Large-scale seismic inversion framework, *Seismol. Res. Lett.,* **86**(4), 1198.

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C. & Wassermann, J., 2015b. ObsPy: a bridge for seismology into the scientific Python ecosystem, *Comput. Sci. Discovery,* **8**(1), 014003.

Krischer, L. *et al.*, 2016. An adaptable seismic data format, *Geophys. J. Int.,* **207**(2), 1003–1011.

Krischer, L., Fichtner, A., Boehm, C. & Igel, H., 2018. Automated large-scale full seismic waveform inversion for North America and the North Atlantic, *J. geophys. Res.,* **123**(7), 5902–5928.

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

Lebedev, S. & van der Hilst, R.D., 2008. Global upper-mantle tomography with the automated multimode inversion of surface and *S*-wave forms, *Geophys. J. Int.,* **173,** 505–518.

Liu, D.C. & Nocedal, J., 1989. On the limited-memory BFGS method for large-scale optimisation, *Math. Program.,* **45,** 503–528.

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

Martiartu, N.K., Boehm, C., Vinard, N., Balic, I.J. & Fichtner, A., 2017. Optimal experimental design to position transducers in ultrasound breast imaging, *Proc. SPIE,* **10139,** 129–143.

Masson, Y. & Romanowicz, B., 2016. Fast computation of synthetic seismograms within a medium containing remote localized perturbations: a numerical solution to the scattering problem, *Geophys. J. Int.,* **208**(2), 674–692.

Masson, Y. & Romanowicz, B., 2017. Box tomography: localized imaging of remote targets buried in an unknown medium, a step forward for understanding key structures in the deep Earth, *Geophys. J. Int.,* **211**(1), 141–163.

Masters, D. & Luschi, C., 2018. Revisiting Small Batch Training for Deep Neural Networks, *CoRR,* abs/1804.07612.

Matharu, G. & Sacchi, M., 2019. A subsampled truncated-Newton method for multiparameter full-waveform inversion, *Geophysics,* **84**(3), R333–R340.

Maurer, H., Nuber, A., Martiartu, N.K., Reiser, F., Boehm, C., Manukyan, E., Schmelzbach, C. & Fichtner, A., 2017. Chapter one—optimized experimental design in the context of seismic full waveform inversion and seismic waveform imaging, in *Advances in Geophysics,* Vol. 58, pp. 1–45, ed. Nielsen, L., Elsevier.

Megies, T., Beyreuther, M., Barsch, R., Krischer, L. & Wassermann, J., 2011. ObsPy—what can it do for data centers and observatories?, *Ann. Geophys.,* **54,** 47–58.

Mitchell, D.P., 1991. Spectrally optimal sampling for distribution ray tracing, *SIGGRAPH Comput. Graph.,* **25**(4), 157–164.

Moghaddam, P., Keers, H., Herrmann, F. & Mulder, W., 2013. A new optimization approach for source-encoding full-waveform inversion, *Geophysics,* **78,** 125–132.

Monteiller, V., Chevrot, S., Komatitsch, D. & Fuji, N., 2012. A hybrid method to compute short-period synthetic seismograms of teleseismic body waves in a 3-D regional model, *Geophys. J. Int.,* **192,** 230–247.

Montelli, R., Nolet, G., Masters, G., Dahlen, F.A. & Hung, S.-H., 2004. Global *P* and *PP* traveltime tomography: rays versus waves, *Geophys. J. Int.,* **158**(2), 637–654.

Mosca, I., Cobden, L., Deuss, A., Ritsema, J. & Trampert, J., 2012. Seismic and mineralogical structures of the lower mantle from probabilistic tomography, *J. geophys. Res.,* **117,** doi:10.1029/2011JB008851.

Mosegaard, K., 2012. *Limits to Nonlinear Inversion,* Springer, pp. 11–21.

Nemirovski, A., Juditsky, A., Lan, G. & Shapiro, A., 2009. Robust stochastic approximation approach to stochastic programming, *SIAM J. Opt.,* **19**(4), 1574–1609.

Nocedal, J., 1980. Updating quasi-Newton matrices with limited storage, *Math. Comput.,* **35,** 773–782.

Nocedal, J. & Wright, S.J., 1999. *Numerical Optimization,* Springer.

Operto, S., Prieux, V., Virieux, J. & Brossier, R., 2013. Multiparameter full waveform inversion of multicomponent ocean-bottom-cable data from the Valhall field. Part 1: imaging compressional wave speed, density and attenuation, *Geophys. J. Int.,* **194**(3), 1640–1664.

Rickers, F., Fichtner, A. & Trampert, J., 2013. The Iceland—Jan Mayen plume system and its impact on mantle dynamics in the North Atlantic region: evidence from full-waveform inversion, *Earth planet. Sci. Lett.,* **367,** 39–51.

Rietmann, M. *et al.*, 2012. Forward and adjoint simulations of seismic wave propagation on emerging large-scale GPU architectures, in *SC '12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis,* Salt Lake City, UT, pp. 1–11.

Ritsema, J.,Heijst,H.J.v. & Woodhouse, J.H., 1999. Complex shear wave velocity structure imaged beneath Africa and Iceland, *Science,* **286**(5446), 1925–1928.

Ritsema, J., Deuss, A., van Heijst, H.J. & Woodhouse, J.H., 2011. S40rts: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltime and normal-mode splitting function measurements, *Geophys. J. Int.,* **184**(3), 1223–1236.

Romanowicz, B., Chen, L.-W. & French, S.W., 2019. Accelerating full waveform inversion via source stacking and cross-correlations, *Geophys. J. Int.,* **220**(1), 308–322.

Romanowicz, B.A. & Dziewonski, A.M., 1986. Toward a federation of broadband seismic networks, *EOS, Trans. Am. geophys. Un.,* **67**(25), 541–542.

Ruan, Y., Lei, W., Modrak, R., Orsvuran, R., Bozdag, E. & Tromp, J., 2019. Balancing unevenly distributed data in seismic tomography: a global adjoint tomography example, *Geophys. J. Int.,* **219**(2), 1225–1236.

Seriani, G. & Priolo, E., 1994. Spectral element method for acoustic wave simulation in heterogeneous media, *Finite Elem. Anal. Des.,* **16**(3-4), 337–348.

Simute, S., Steptoe, H., Cobden, L., Gokhberg, A. & Fichtner, A., 2016. Full-waveform inversion of the Japanese islands region, *J. geophys. Res.,* **121**(5), 3722–3741.

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

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

Thrastarson, S., van Driel, M., Krischer, L., Boehm, C., Afanasiev, M., van Herwaarden, D.-P. & Fichtner, A., 2020. Accelerating numerical wave propagation by wavefield-adapted meshes, Part II: Full-waveform inversion, *Geophys. J. Int*., doi:10.1093/gji/ggaa065.

Trampert, J., Deschamps, F., Resovsky, J. & Yuen, D., 2004. Probabilistic tomography maps chemical heterogeneities throughout the lower mantle, *Science,* **306,** 853–856.

Tromp, J. & Bachmann, E., 2019. Source encoding for adjoint tomography, *Geophys. J. Int.,* **218**(3), 2019–2044.

Vallée, M., 2013. Source time function properties indicate a strain drop independent of earthquake depth and magnitude, *Nat. Commun.,* **4**(1), 2606, doi:10.1038/ncomms3606.

van Driel, M., Boehm, C., Krischer, L. & Afanasiev, M., 2020. Accelerating numerical wave propagation by wavefield-adapted meshes, Part I: Forward and adjoint modelling, *Geophys. J. Int*., doi:10.1093/gji/ggaa058.

van Leeuwen, T. & Herrmann, F.J., 2013. Fast waveform inversion without source-encoding, *Geophys. Prospect.,* **61**(s1), 10–19.

Wolpert, D.H. & Macready, W.G., 1997. No free lunch theorems for optimization, *IEEE Trans. Evolutionary Comput.,* **1,** 67–82.

Yang, H., Jia, J., Wu, B. & Gao, J., 2018. Mini-batch optimized full waveform inversion with geological constrained gradient filtering, *J. Appl. Geophys.,* **152,** 9–16.

Yomogida, K., 1992. Fresnel zone inversion for lateral heterogeneities in the Earth, *Pure appl. Geophys.,* **138**(3), 391–406.

Yoshizawa, K. & Kennett, B.L.N., 2005. Sensitivity kernels for finite-frequency surface waves, *Geophys. J. Int.,* **162**(3), 910–926.