Global seismic tomography with sparsity constraints: Comparison with smoothing and damping regularization

Jean Charlety, Sergey Voronin, Guust Nolet, Ignace Loris, Frederik Simons, Karin Sigloch, Ingrid Daubechies

JOURNAL OF GEOPHYSICAL RESEARCH: SOLID EARTH, VOL. 118, 4887–4899, doi:10.1002/jgrb.50326, 2013

Global seismic tomography with sparsity constraints: Comparison with smoothing and damping regularization

Jean Charléty,[1,2] Sergey Voronin,[1] Guust Nolet,[1] Ignace Loris,[3] Frederik J. Simons,[4] Karin Sigloch,[5] and Ingrid C. Daubechies[6]

Received 11 December 2012; revised 24 July 2013; accepted 31 July 2013; published 17 September 2013.

[1] We present a realistic application of an inversion scheme for global seismic tomography that uses as prior information the sparsity of a solution, defined as having few nonzero coefficients under the action of a linear transformation. In this paper, the sparsifying transform is a wavelet transform. We use an accelerated iterative soft-thresholding algorithm for a regularization strategy, which produces sparse models in the wavelet domain. The approach and scheme we present may be of use for preserving sharp edges in a tomographic reconstruction and minimizing the number of features in the solution warranted by the data. The method is tested on a data set of time delays for finite-frequency tomography using the USArray network, the first application in global seismic tomography to real data. The approach presented should also be suitable for other imaging problems. From a comparison with a more traditional inversion using damping and smoothing constraints, we show that (1) we generally retrieve similar features, (2) fewer nonzero coefficients under a properly chosen representation (such as wavelets) are needed to explain the data at the same level of root-mean-square misfit, (3) the model is sparse or compressible in the wavelet domain, and (4) we do not need to construct a heterogeneous mesh to capture the available resolution.

Citation: Charléty, J., S. Voronin, G. Nolet, I. Loris, F. J. Simons, K. Sigloch, and I. C. Daubechies (2013), Global seismic tomography with sparsity constraints: Comparison with smoothing and damping regularization, *J. Geophys. Res. Solid Earth*, *118*, 4887–4899, doi:10.1002/jgrb.50326.

1. Introduction

[2] In order to increase the resolution of tomographic images, we seek improvements in the way that the information contained in seismograms is used. *Dahlen et al.* [2000] introduced the use of frequency-dependent body wave data (delay time and/or amplitude) and derived kernels with which finite-frequency effects are taken into account. *Chevrot and Zhao* [2007] showed that the parameterization of the model is of crucial importance to benefit from the spatial structure of a finite-frequency sensitivity kernel. These kernels show variations in sensitivity over small spatial length scales that potentially allow for a solution with better resolution than offered by ray theory. However, in order to exploit the enhanced sensitivity, it is essential that the details of its structure are not smoothed away by a parameterization that allows only for long wavelengths, so we will have no choice but to overparameterize. Another reason for overparameterization of a tomographic model is that with the very high station density obtained today in networks like USArray and its flexible component, or HiNet in Japan, a resolution of the order of better than 50 km is often within reach in global tomography, if only locally and for shallow structure.

[3] In a tomographic inverse problem, we generally encounter the following phenomenon: The system to be solved is underdetermined; that is, for linear problems, the sensitivity matrix has more columns than rows and we need to solve for more unknowns than there are data. On the right-hand side of the problem, the data are noisy, and the singular values of the matrix decrease rapidly toward zero. Generally speaking, the matrices encountered in this setting are not well conditioned. Since the problem is underdetermined, constraints on the solution are generally added to impose uniqueness of the solution. Due to the noise in the right-hand side and the ill-conditioning of the matrix, it is necessary to use regularization for the solution of the linear system. The choice of regularizing constraints and the

[1]Géoazur, Centre National de la Recherche Scientifique (UMR 6526), Observatoire de la Côte d'Azur, Université de Nice Sophia-Antipolis, Valbonne, France.
[2]Now at IFP Energies nouvelles, Rueil-Malmaison, France.
[3]Département de Mathématique, Université Libre de Bruxelles, Brussels, Belgium.
[4]Department of Geosciences, Princeton University, Princeton, New Jersey, USA.
[5]Geophysics, Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München, Munich, Germany.
[6]Department of Mathematics, Duke University, Durham, North Carolina, USA.

Corresponding author: J. Charléty, IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, FR-92852 Rueil-Malmaison CEDEX, France. (jean.charlety@ifpen.fr)

2169-9313/13/10.1002/jgrb.50326

utilized algorithm are crucial to the characteristics of the obtained solution. In this paper, we adopt a framework for dealing with these problems using sparsity-constrained optimization applied in the wavelet model domain and apply it to actual experimental data.

[4] The poor conditioning of the matrix means that relatively few singular values are of sufficient magnitude, in comparison to the maximal rank for a matrix of that size. A simple regularization method is a truncated singular-value decomposition (SVD) whereby we compute the solution using only the singular vectors of the matrix that correspond to the largest singular values. However, real-world tomographic systems are commonly too large to allow for such an approach, since computing the SVD is expensive.

[5] An often-chosen option is smoothing by applying an ℓ_2 norm constraint on the solution (or, alternatively, on its gradient or Laplacian). The resulting quadratic problem is easily solved by means of a linear system or via augmented least-squares [*Paige and Saunders*, 1982], but has the disadvantage of smoothing away sharp boundaries (e.g., of a subducting slab in global terrestrial tomography). One way to alleviate the effects of overparameterization in large systems is to combine adjacent voxels into larger voxels, as was first done by *Abers and Roecker* [1991]. One can use ray coverage within each voxel as a guideline for combining voxels, but the procedure is not unique, and there is no guarantee that a combined voxel is actually resolved, nor is this system in a straightforward way related to SVD.

[6] In this paper we obtain regularized solutions by imposing an ℓ_1 norm constraint on the wavelet representation of the model following on the work in seismic tomography by *Loris et al.* [2007, 2010]. Our approach shares many conceptual similarities with a variety of methods in other seismological settings, such as those proposed by *Li et al.* [1996], *Lin and Herrmann* [2007], and *Herrmann and Hennenfent* [2008], and considered for other imaging and signal processing problems by *Figueiredo et al.* [2007], *Vonesch and Unser* [2008], and numerous other authors such as reviewed by *Bruckstein et al.* [2009]. In these papers, the sparsity-seeking behavior of the ℓ_1 norm is discussed in detail.

[7] If a sparsity constraint can be imposed on the solution via a penalty function, based on some prior knowledge, a resolution that seemingly exceeds fundamental limitations can be obtained. In a Bayesian interpretation, the extra information supplied can perhaps even be gleaned adaptively during the course of the experiment [*Haupt and Nowak*, 2012]. *Szameit et al.* [2012] exploit the knowledge of the sparsity of an object in a known basis to reconstruct features much smaller than the classical diffraction limit. Under some special circumstances, "compressed sensing" enables the realization of sub-Nyquist sampling [*Davenport et al.*, 2012; *Herrmann et al.*, 2012]. Since wavelets represent identifiable structures localized in space, the ℓ_1 norm minimization tends to satisfy our demand that the tomographic solution does not have more "structure" than warranted by the data. It can also conserve sharp boundaries in the solution if these are imposed by the data.

[8] It is important to realize that every form of regularization represents a subjective choice of one model over infinitely many that satisfy the data to the same degree. Annihilating null-space components by a simple damping of the model norm leads to erratic solutions unless the data coverage is without any major gaps [*VanDecar and Snieder*, 1994]. To bridge gaps left by imperfect data coverage, smoothing has been, until now, the preferred method in global seismic tomography—whether implemented by limiting the number of coefficients in a spherical harmonic expansion [*Ritsema et al.*, 2011] or by damping the ℓ_2 norm of the gradient or the Laplacian of the model.

[9] Although both methods use components of the null-space to bridge gaps, the ℓ_1 norm regularization offers an alternative to smoothing in the sense of an ℓ_2 penalty on the Laplacian. Whether it is "better" than smoothing depends on the situation at hand, and probably also on the subjective preference of the geophysicist. If information on the sharpness of discontinuities is available and can be used as a formal constraint in a Bayesian sense, the ability to preserve sharp boundaries may very well give ℓ_1 norm regularization an advantage over smoothing. At the same time, the coefficient thresholding that we employ should guard us against the inclusion of null-space components that are not strongly warranted by the data—and furthermore, such wavelet-basis components will wield an influence that is strictly localized in the model space.

[10] In this paper we quantify the differences between either option in terms of solution quality, computation speed, and algorithmic complexity. For a theoretical justification that goes beyond arguing for subjective preferences on the part of the user, we point to the early work by *Donoho* [1995] which contrasts the linear filtering of global, operator-dependent eigenfunctions of classical regularization techniques that include truncated SVD solutions, with the nonlinear approach of thresholding model coefficients in a wavelet basis—see our section 2. The crucial difference lies in our emphasis on representing the model, i.e., the object to be recovered, rather than the operator of the inverse problem itself, and in the space-scale localization and thus sparsifying nature of the wavelet basis acting on the model.

[11] Up to now, to our knowledge, the wavelet basis has been used in global seismic tomography in two different ways. The first uses the resolvability properties of wavelets [*Chiao and Kuo*, 2001] and the second their compressibility [*Chevrot and Zhao*, 2007; *Chevrot et al.*, 2012]. The first method benefits from the spatial variation of the resolution that can be achieved within the wavelet basis. The second uses the property that in the wavelet basis, the model can be sparser and therefore, in this basis, the model can be compressed. Here we discuss an inversion scheme that benefits from both properties, resolvability and compressibility, and that flexibly handles spatial variations in model resolution. In areas where the allowable model resolution is better than others, by being able to utilize information locally, our method flexibly handles such situations. Our present paper follows the philosophy of the earlier work by *Simons et al.* [2011], who developed a wavelet basis on the cubed Earth and described (but did not actually test in three-dimensional space) an inversion algorithm. We implemented their wavelet transformation by extension to three dimensions and thereby applied their proposed method in its full complexity, using a large set of real data, including the effects of source-correction terms. The novelty of our contribution lies in the implementation, for realistically sized applications in global seismic tomography, of methods

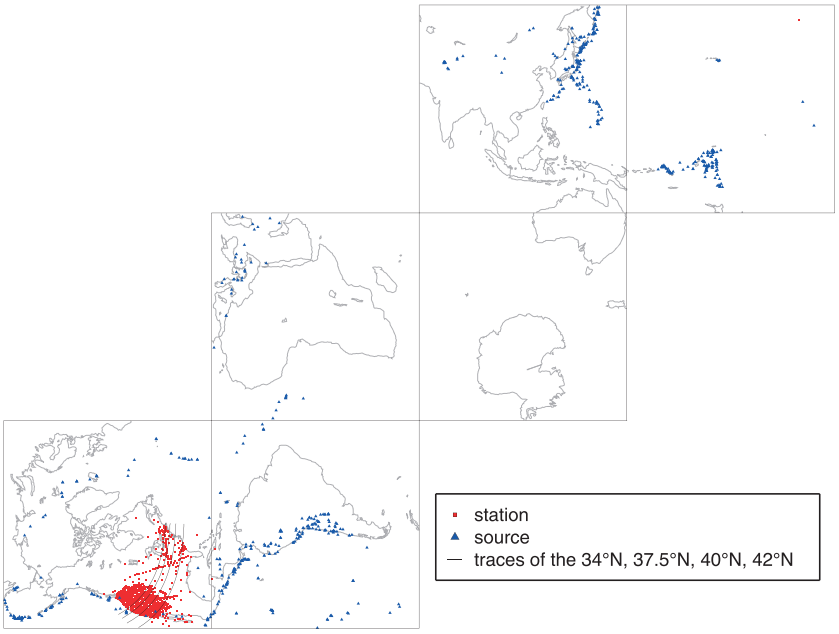

Figure 1. Location of the sources (blue triangles) and stations (red squares) used in our inversion experiment. The latitudes marked correspond to the model cross sections presented in Figures 6–9.

that are related to those that have been espoused in particular by the exploration-seismological community [e.g., *Li et al.*, 1996; *Tikhotsky and Achauer*, 2008; *Gholami and Siahkoohi*, 2010; *Tikhotskii et al.*, 2011, *Li et al.*, 2012].

[12] We first present the wavelet parameterization and the ℓ_1 norm-regularized least-squares inversion method (section 2). Subsequently, we determine the correct regularization parameter by constructing an L-curve (section 5.1). In section 5.2 we show the final preferred velocity model and compare it with the result of *Sigloch et al.* [2008], which was obtained with a more commonly used ℓ_2 inversion scheme using imposed, isotropic smoothing and damping in a tetrahedral voxel parameterization whose mesh size increased with depth. Using ℓ_1 regularization is a sensible choice in the wavelet model domain, as the transform is spatially localized and naturally sparsifying. One could envisage replicating the ℓ_2-regularizing schemes in the wavelet domain also, but since our wavelet transform is almost norm preserving, models obtained with ℓ_2 regularization are virtually unchanged under a change of parameterization from tetrahedra to our wavelets. This invariance explains our choice to use Sigloch's output model rather than regenerating it from the primary data sets.

2. Method

[13] We use the "cubed Earth," that is, the cubed sphere representation of *Ronchi et al.* [1996], a wavelet transform, and the Fast Iterative Soft-Thresholding Algorithm (FISTA) of *Beck and Teboulle* [2009] to invert the data in the new parameterization.

[14] The details of the parameterization or wavelet transformation can be found in *Simons et al.* [2011]. Briefly, the Earth is parameterized with six chunks that divide the surface of a sphere (see Figure 1). Each chunk, for this study, is sampled by 128 × 128 voxels in the angular directions. In depth, there are 37 layers (Table 1) distributed unevenly by subsampling an original division of 128 layers of equal thickness (Figure 2). Therefore, the model consists of $6 \times 128 \times 128 \times 37$ (3,637,248) voxels or unknowns. This number has to be compared to the 92,175 unknowns or grid points of the tetrahedral parameterization used by *Sigloch* [2008], *Sigloch et al.* [2008], and *Tian et al.* [2009] for models of the mantle under USArray or 19,279 grid points for the global model of *Montelli et al.* [2006].

[15] In contrast to these earlier studies, the size of voxels is smaller at the core-mantle boundary (CMB) than near the surface. The angular discretization remains constant irrespective of the radius. As it decreases with depth, the ratio of the size of a voxel at the surface to one in the lowermost layer is around 2. The linear horizontal size of a voxel at the surface is around 80 km and thus 40 km at the CMB. The resulting voxels are close to cubic in shape near the surface, but their height of 90 km remains constant. In comparison, for the parameterization used by *Sigloch et al.* [2008], the face length is around 200 km in the upper mantle under unconstrained regions (Pacific, Asia, for example)

Table 1. Radius of the Layer Boundaries for the 37 Layers of the Model Domain

Layer	Radius (km)	Layer	Radius (km)	Layer	Radius (km)	Layer	Radius (km)
1	3481.4	11	4294.1	21	5197.1	31	5964.7
2	3526.6	12	4384.4	22	5287.4	32	6009.8
3	3571.7	13	4474.7	23	5377.7	33	6100.1
4	3662.0	14	4565.0	24	5468.0	34	6190.4
5	3752.3	15	4655.3	25	5558.3	35	6280.7
6	3842.6	16	4745.6	26	5648.6	36	6325.9
7	3932.9	17	4835.9	27	5693.8	37	6348.4
8	4023.2	18	4926.2	28	5738.9	38	6371.0
9	4113.5	19	5016.5	29	5829.2		
10	4203.8	20	5106.8	30	5919.5		

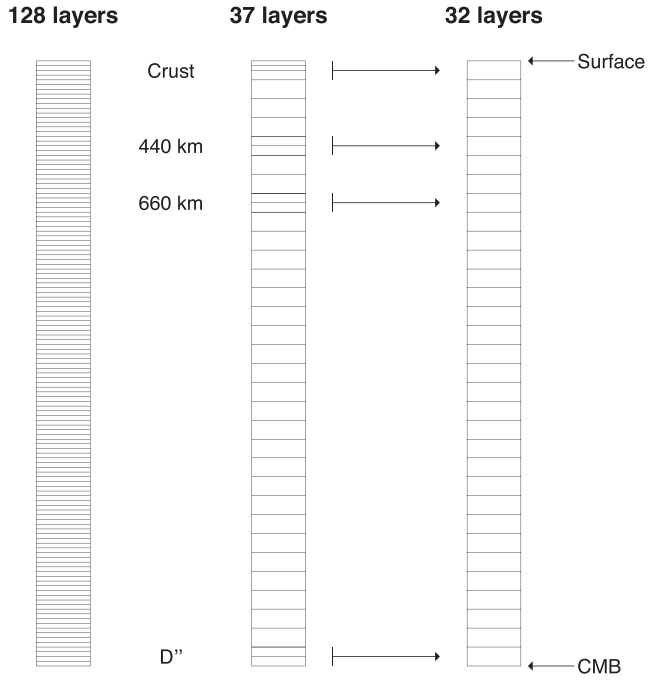

Figure 2. Illustration of the mapping of the 128 divisions to the 37 layers and of the 37 layers to the 32-component vector required for the wavelet transform.

but is around 70 km under USArray with a linear decrease to 600 km at the center of the Earth.

[16] The cubed Earth representation is chosen because it allows us to define a Cartesian coordinate system on which we can use various families of wavelets developed in other fields (for example, image processing). Terrestrial heterogeneities must be well represented in the chosen wavelet basis; there are many different wavelets to choose from to ensure that this is the case.

[17] The separable wavelet transformation defines a multiresolution basis in the combined three dimensions. *Simons et al.* [2011] tested a number of wavelets, starting with orthogonal *Daubechies* [1988] wavelets of varying filter lengths, namely D2 (Haar), D4, and D6. In those cases, the resolved model suffered from artifacts because the wavelets, except for the Haar wavelets, are not (anti)symmetric. Random-shift methods [*Figueiredo et al.*, 2007; *Vonesch and Unser*, 2008] might have been called to action, but these would involve recalculating our matrix A, which is costly. In the end, we sacrificed orthogonality for symmetry, as is often done in image processing, by using the *Cohen et al.* [1992] biorthogonal CDF wavelet family. In the angular directions, the chosen wavelet basis is CDF 4–2 [see *Simons et al.*, 2011, Figure 4]. In the radial dimension, we retained the Haar wavelet family. Since the latter basis encompasses "layers" of constant value, it implicitly allows for abrupt discontinuities. Also, as there are only 37 layers in our model domain, using smoother wavelets (with more vanishing moments, corresponding to longer filters) would be more difficult to implement across many decomposition scales. We would simply have too few scales available to attain the asymptotic regime in which using wavelets with more vanishing moments, appropriate for the second-order tomographic operator, would make a material difference.

[18] The algorithm used for the wavelet transform uses a discretized model with 2^n elements in each spatial direction. We have 128 voxels in the two horizontal (x, y) directions in the cubed Earth parameterization; in the z direction, we could adopt 32, 64, or 128 elements. However, we must explicitly represent the depths of discontinuities and possibly crustal layering. We chose 32 elements but felt compelled to transform parts of the model using a smaller voxel spacing in the z direction (e.g., near known boundaries in the Earth such as the 410, 660 km, and core-mantle discontinuities). Figure 2 illustrates the mapping of the 128 divisions to an initial set of 37 layers from which the final 32-element radial vector is constructed. The construction shown in Figure 2 combines the thinner layers by averaging in two stages. Starting from a 128-layer division, in a first step, we create 37 layers, most of them composed of four original layers except near major discontinuities and the surface. These 37 are subsequently reduced to $32 = 2^5$ but the differences between adjacent thin layers are stored so that the reverse transformation is possible. Thus, the transformation from 37 to 32 layers is invertible and mimics a lifting-scheme algorithm for the wavelet transform using only means and differences.

[19] The wavelet transform reorganizes the information into a set of details appearing at different resolutions, or levels. Multiresolution analysis consists of successively projecting the signal onto subspaces in a series of increasingly coarser approximations. Given a sequence of increasing resolutions $(r_j)_{j\in\mathcal{J}}$, the details of the information at resolution r_j are defined as the difference of information between its approximation at the resolution r_j and its approximation at the lower resolution r_{j-1} [*Mallat*, 2008]. A variety of algorithms is available to carry out the transforms efficiently [*Strang and Nguyen*, 1997; *Jensen and la Cour-Harbo*, 2001]; for the short filter lengths that we use, computation speeds do not vary appreciably between algorithms [*Sweldens*, 1996].

2.1. Toward a Sparse Model

[20] We proceed to describe the background to our approach. We start at the linear system:

$$Am = b, \tag{1}$$

where $A \in \mathbb{R}^{N\times M}$ contains the kernels, $b \in \mathbb{R}^N$ the data, and m is the "true" but unknown model to be estimated. In our problem, N, the number of data, is smaller than M, the number of unknowns, making the system underdetermined. Moreover, the matrix A is generally ill-conditioned and only a noisy version of the data b is measured. The classical way to deal with the noise in b is to minimize the term $\|Am-b\|_2^2$ in the inversion. It is, however, well understood that additional constraints must be added to the linear system (1) to account for the system being underdetermined and to obtain a reasonably bounded solution owing to the ill-conditioning of the matrix.

[21] We thus look for a solution $\bar{m}$ that minimizes the functional:

$$\bar{m} = \arg\min_m \left\{\mathcal{F}(m) = \|Am-b\|_2^2 + \mathcal{R}(m)\right\}, \tag{2}$$

where $\mathcal{R}(m)$ is some penalty function. In classical tomography, $\mathcal{R}(m)$ is equal to $\lambda\|m\|_2^2$, or more generally $\lambda\|\Phi m\|_2^2$,

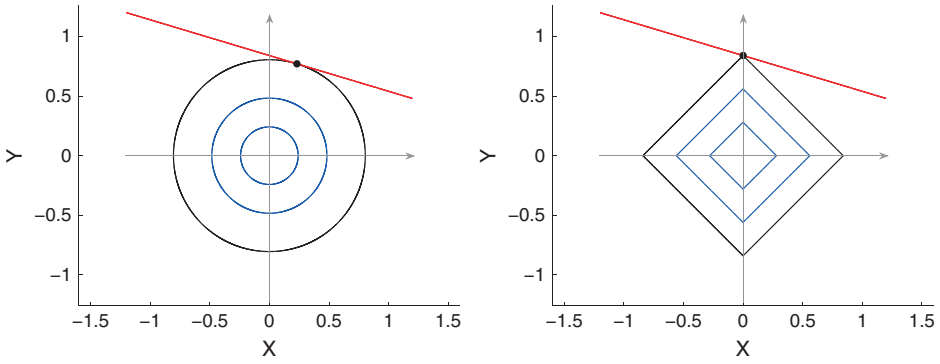

Figure 3. Comparison of ℓ_2 norm and ℓ_1 norm minimization in the two-dimensional (x, y) plane. The solution is defined by the black dot at the intersection of the black (for the quantity to be minimized) and the red (for the linear constraint) curves. (left) $\min(|x|^2 + |y|^2)$, different contours shown in blue, subject to $ax + by = c$, in red. (right) $\min(|x| + |y|)$ subject to $ax + by = c$. The ℓ_1 norm minimization generally produces a sparse solution: In this case, one of the components is exactly zero.

where Φ is a linear operator such as the gradient or Laplacian used to impose model smoothness. The advantage of this approach, commonly known as Tikhonov regularization, lies in its simplicity. When $\Phi = I$ and $\mathcal{R}(m) = \lambda\|m\|_2^2$, the solution to the quadratic problem is given in terms of the linear system:

$$\bar{m} = \arg\min_m \left\{\|Am - b\|_2^2 + \lambda\|m\|_2^2\right\} \iff (A^TA + \lambda I)\bar{m} = A^Tb. \quad (3)$$

One way to implement Tikhonov regularization is via solving the augmented least-squares problem:

$$\bar{m} = \arg\min_m \left\| \begin{bmatrix} A \\ \sqrt{\lambda}\Phi \end{bmatrix} m - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2, \quad (4)$$

usually by standard least-squares solvers such as LSQR [*Paige and Saunders*, 1982]. The desired effect of the regularization is to filter the contributions from the small singular values of the matrix A to the solution. Using the SVD $A = U\Sigma V^T$, we can write the solution by substitution in equation (4) for the case $\Phi = I$ as

$$\bar{m} = \arg\min_m \left\| \begin{bmatrix} U\Sigma V^T \\ \sqrt{\lambda}I \end{bmatrix} m - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2 = V\mathrm{diag}\left(\frac{\sigma_i}{\sigma_i^2 + \lambda}\right) U^Tb, \quad (5)$$

and the effect of this regularization is to replace the small singular values σ_i by $\sigma_i/(\sigma_i^2 + \lambda)$, which prevents those smaller than $\sqrt{\lambda}$ from dominating the solution.

[22] Introducing a two-norm (ℓ_2) constraint on the solution, however, is not the only way to realize the benefits of regularization. If other assumptions on the model can be made, different constraints can be used. The main hypothesis we make in this paper is that the unknown model is sparse in a particular basis.

[23] Wavelets provide a way to sparsely represent a (geophysical) model using functions at different scales, which allows the representation and analysis [e.g., *Herrmann and Bernabé*, 2004] of different features corresponding to different wavelengths (or sharpness). *Simons et al.* [2011] show that the seismic velocity model of *Montelli et al.* [2006] or that of *Ritsema et al.* [2011] can be efficiently represented in a "four-tap" orthogonal *Daubechies* [1988] wavelet basis (D4) using few nonzero expansion coefficients. In other words, these models can be considered sparse in this wavelet basis. For this paper, we assume that this is also true with respect to the biorthogonal CDF 4–2 basis which we use for the model that we wish to reconstruct. Physically, this means we assume that heterogeneity in the Earth is organized in identifiable entities that are well modeled with wavelets [*Piromallo et al.*, 2001]. Consequently, we suppose that tomographic models have a sparse representation in the wavelet basis and that this new type of constraint can be used to find an estimate of velocity variations within the Earth.

[24] Mathematically, forcing a sparsity constraint on the model norm in the wavelet domain implies that we would like to introduce a penalty on the term Wm, where W is to represent the action of the forward wavelet transform. Instead of imposing a quadratic penalty on Wm, which would not give a sparse solution, we could consider the so-called ℓ_0 measure which counts the number of nonzero components of a vector. The ℓ_0 measure is, however, not a norm and highly nonconvex (which means it has many local minima) making it difficult to work with numerically, especially in the case of ill-conditioned matrices. Moreover, direct minimization of the ℓ_0 measure would be combinatorially difficult. For these reasons and based on an abundance of results from the compressive-sensing literature [see, for example, *Candès and Wakin*, 2008; *Donoho*, 2006], we will use instead the closest convex norm to the ℓ_0 measure, the ℓ_1 norm, which is simply the sum of the absolute values:

$$\|Wm\|_1 = \sum_{i=1}^{M} |(Wm)_i|. \quad (6)$$

Under certain conditions on the matrix A in equation (1), both the ℓ_0 measure and the ℓ_1 norm penalties return identical results (see Figure 3 for a simple example and a comparison with the ℓ_2 norm). As parts of the globe are not illuminated by the kernel coverage (leading to near-zero columns in the matrix A), we know that these conditions may not be satisfied. However, in such a case, the ℓ_1 norm penalty approach still yields stable sparse solutions. *Daubechies et al.* [2004] highlight the regularizing action (small changes in the data do not lead to high variance in the reconstructed models) and prove the convergence of such schemes as will

be discussed below. Thus, the minimization problem that we wish to solve is now given by

$$\bar{w} = \arg\min_{w} \left\{ \|AW^{-1}w - b\|_2^2 + 2\tau\|w\|_1 \right\} \quad \text{with} \quad \bar{m} = W^{-1}\bar{w}, \tag{7}$$

[25] where we have identified the wavelet coefficients $w = Wm$ (the forward wavelet transform of the model) and $m = W^{-1}w$ (the inverse wavelet transform of the forward transformed model). Remember that our transform is not orthogonal; hence, we write the inverse and not the transpose of the operator W. We have replaced the regularization parameter λ with 2τ for convenience in writing down the algorithm for its minimization. Equation (7) forms the basis of the algorithm that we consider in this paper. If—for any reason—one preferred to keep the number of coefficients low at particular length scales and not others, equation (7) can easily be adapted by weighting each scale differently.

2.2. Algorithm

[26] We now discuss the approach we use to solve (7). The main advantage of the one-norm ℓ_1 penalty $\|\cdot\|_1$ is its convexity. Indeed, $\|AW^{-1}w - b\|_2^2 + 2\tau\|w\|_1$ is convex and is globally minimized for the conditions:

$$\begin{array}{rcll} \left[(AW^{-1})^T(b - AW^{-1}\bar{w})\right]_i & = & \tau\,\mathrm{sgn}(\bar{w}_i)\ , \ \forall i & \text{with} \quad \bar{w}_i \neq 0, \\ |\left[(AW^{-1})^T(b - AW^{-1}\bar{w})\right]_i| & \leq & \tau\ , \ \forall i & \text{with} \quad \bar{w}_i = 0. \end{array} \tag{8}$$

Such conditions, however, appear much more complicated than the linear system that arose in the case of ℓ_2 norm regularization, although we may immediately observe that $\bar{w}_i = 0 \quad \forall i$ when $\tau > \|(AW^{-1})^T b\|_\infty$. This gives us an upper bound on the choice of τ. Beyond this fact, however, we cannot make efficient use of these conditions directly. Fortunately, simple algorithms for the above minimization exist. They are based on the soft-thresholding function:

$$\mathcal{S}_\tau(u) \equiv \begin{cases} u - \tau & u > \tau \\ 0 & |u| \leq \tau\ , \\ u + \tau & u < -\tau \end{cases} \tag{9}$$

as defined by *Donoho and Johnstone* [1994]. The simple use of the majorization-minimization approach [*Daubechies et al.*, 2004] then yields the following straightforward scheme for solving equation (7) starting with any initial estimate m^0 for the model:

$$w^0 = Wm^0, \tag{10}$$

$$w^{n+1} = \mathcal{S}_\tau\left(w^n + (AW^{-1})^T b - (AW^{-1})^T(AW^{-1})w^n\right), \tag{11}$$

$$m^{n+1} = W^{-1}w^{n+1}. \tag{12}$$

[27] The above scheme, known as ISTA, short for Iterative Soft-Thresholding Algorithm, converges for any initial guess with the condition that $\|AW^{-1}\|_2 < 1$, which can be accomplished by rescaling the matrix A, the data b, and the penalty parameter. The parameter τ is chosen to be smaller than $\|(AW^{-1})^T b\|_\infty$. When τ is large, the convergence is faster and the solution is sparser in the wavelet domain. It typically is also slightly smaller—motivating some to opt for a debiasing step by a final ℓ_2 projection on the support of the result [*Mallat*, 2008]. We ran tests with matrices of similar conditioning as our A and with synthetically generated models, but we did not observe significant differences in terms of mean-square error when using debiasing after the ℓ_1 solve. In principle, τ might be chosen differently for different types of coefficients, allowing us to vary the degree of sparsity, for example, between detail and approximation coefficients. In practice, this simple scheme is known to converge considerably more slowly than a similar accelerated (F for Fast) scheme known as FISTA [*Beck and Teboulle*, 2009], in which the soft-thresholding operation is applied to a linear combination of the two previous iterates, so that for $n = 1, \ldots,$

$$w^0 = 0, \tag{13}$$

$$w^{n+1} = T\left(w^n + \frac{t_n - 1}{t_{n+1}}(w^n - w^{n-1})\right), \tag{14}$$

$$T(x) = \mathcal{S}_\tau(x + V^T b - V^T V x), \tag{15}$$

with $V = AW^{-1}$ and t_n a sequence of numbers defined by $t_{n+1} = (1 + \sqrt{1 + 4t_n^2})/2$, and $t_1 = 1$. The FISTA has the same computational complexity as ISTA but a substantially faster rate of convergence. The computational requirements at each iteration are to perform matrix-vector multiplications with the matrix AW^{-1} and its transpose $W^{-T}A^T$. This requires the existence of the inverse wavelet transform W^{-1} and the inverse-transpose transform W^{-T}, the dual of the forward transform W.

[28] In short, while there are many alternatives [e.g., *Figueiredo et al.*, 2007; *van den Berg and Friedlander*, 2008], our predilection for FISTA also keeps the number of matrix-vector multiplications low, and with a suitable choice of step lengths (and apart from the empirical lowering of the penalty parameter), we enjoy a guaranteed convergence of the algorithm.

3. Incorporating Corrections

[29] In seismic tomography, there is generally a trade-off between velocity model perturbations and corrections for the published origin times of the earthquakes, directivity effects that influence the cross correlations used to measure delays, and the coordinates of the hypocenter—each of which influences the data. Therefore, we include in our inversion corrections, u, for the locations and origin times of the earthquakes. Hereby, we follow established practice, while noting that there is recent work by *Aravkin and van Leeuwen* [2012] on the theoretical justification of solving for such and other "nuisance parameters" in a tomographic context. This part of the inversion must not be constrained in the same way as the model, and therefore, the system of equations that we solve approximately is

$$\begin{bmatrix} AW^{-1} & C \\ 0 & D \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{16}$$

The submatrix C contains the correction coefficients applied to the subvector u, while D is a diagonal matrix of damping terms. In our experiment, the vector u will be composed of about 2000 elements and as such, is substantially smaller than the vector w which will count over 3 million elements. To enforce the sparsity of w as a way to regularize the problem, we now seek to solve the problem

$$[\bar{w}, \bar{u}] = \arg\min_{w,u} \left\| \begin{bmatrix} AW^{-1} & C \\ 0 & D \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix} - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2 + 2\tau\|w\|_1. \tag{17}$$

Letting $V = AW^{-1}$, the two-parameter functional that we minimize is

$$F(w, u) = \|Vw + Cu - b\|_2^2 + \|Du\|_2^2 + 2\tau\|w\|_1. \tag{18}$$

To minimize the above functional with respect to both variables, we use alternate minimization: We update w and u independently for some initial guesses w^0 and u^0, one after the other in the iteration. Holding $u = u^n$ constant we can write

$$w^{n+1} = \arg\min_w \left(\|Vw - k\|_2^2 + 2\tau \|w\|_1 \right) \text{ with } k = b - Cu^n. \quad (19)$$

This update proceeds iteratively, through the accelerated FISTA scheme. Once w^n has been updated to w^{n+1}, we hold $w = w^{n+1}$ constant to update u^n:

$$u^{n+1} = \arg\min_u \left(\|Cu - l\|_2^2 + \|Du\|_2^2 \right) \text{ with } l = b - Vw^{n+1}. \quad (20)$$

The latter equation is quadratic, with u^{n+1} satisfying the linear system:

$$\left(C^T C + D^T D\right) u^{n+1} = C^T \left(b - Vw^{n+1}\right), \quad (21)$$

where $m^{n+1} = W^{-1} w^{n+1}$ represents the tomographic model and u^{n+1} the correction terms.

[30] The whole algorithm is given in pseudocode as

Algorithm 1: Alternate Minimization Algorithm

Input :The matrices A, C, and D and the right-hand side vector b. The inverse wavelet transform W^{-1} and the inverse-transpose transform W^{-T}. The regularization parameter $\tau < \|(AW^{-1})^T b\|_\infty$ and a maximum number of iterations $n_{\max}$.

Output :An estimate of the regularized model solution $\bar{m}$.

$m^0 = (0, \ldots, 0)^T$;
$u^0 = (0, \ldots, 0)^T$;
for $n = 0, 1, \ldots, n_{\max}$ **do**
  $k = b - Cu^n$;
  Solve with FISTA:
    $w^{n+1} = \arg\min_w \|AW^{-1}w - k\|_2^2 + 2\tau\|w\|_1$;
  Solve the small linear system:
    $(C^T C + D^T D)u^{n+1} = C^T(b - Vw^{n+1})$;
  $m^{n+1} = W^{-1} w^{n+1}$;
end
$\bar{w} = w^{n_{\max}}$;
$\bar{m} = m^{n_{\max}}$;

[31] For the inversion, two important technical aspects have to be addressed: the choice of the threshold, τ, and the maximum level of the wavelet decomposition. The optimal level of wavelet decomposition depends on the structure of the model and the data sensitivity to it. To choose τ, we use a continuation scheme that starts with the zero vector at τ just below $\|(AW^{-1})^T b\|_\infty$. At this value of τ, $w = 0$ is a good initial guess for the wavelet-transformed solution, since the optimality conditions of the ℓ_1 functional (8) indicate that the solution vanishes for values exceeding τ. Thus, the convergence at the initial point is expected to be fast. We then proceed to use the obtained solution, lowering τ at every iteration. We stop this procedure when the obtained solution satisfies the desired mean-square error, although we may elect different stopping criteria.

[32] In the following section, tests with real data are presented. The aim of these tests is to appreciate the behavior of the algorithm and the role of the regularization parameter τ. Only after determining an optimal threshold and decomposition level can we perform an inversion on real data. The maximum level of decomposition for the wavelet transform is set to 2 in our experiments. By implication, the thresholding of the wavelet coefficients is only applied to structures of size $2^2 = 4$ times the filter length or smaller. In subsequent sections we will provide physical length scales (in angular degrees across the surface or in kilometers at depth) for these and other values of the wavelet scales, where appropriate.

4. Data

[33] For this study, we use the cross-correlation delay times from the database built by *Sigloch* [2008] in whose published work more details can be found. This is only a subset of the data used by *Sigloch et al.* [2008], who also included amplitude variations in their inversion. Also excluded are a small number of regional International Seismological Centre (ISC) delay times that only influence shallow structure. Source and receiver locations are shown in Figure 1. The receivers are those of the USArray experiment in its early stage; that is, they are concentrated in the western part of the U.S. The locations of the sources provide a suitable azimuthal sampling for this region.

[34] Our data set is composed of $N = 430{,}554$ P wave delay times. These are estimated in eight different frequency bands whose central periods are 30, 21.2, 15, 10.6, 7.5, 5.3, 3.7, and 2.7 s. The frequency bands are tapered with a Gabor function that minimizes sidelobes in the kernels, which renders them spatially relatively compact.

5. Results

5.1. Threshold Determination and the Pareto Curve

[35] The first parameter to establish is the regularization parameter, $\lambda = 2\tau$. The role and influence of that parameter were determined by computing several inversions with different values for a two-level wavelet decomposition. Each inversion was run for a limited identical number of 566 iterations, chosen for practicality. While this number of iterations did not necessarily lead to complete convergence, nevertheless, an acceptable solution was found in all cases.

[36] We computed a reduced chi-square statistic $\chi^2_{\text{red}} = \chi^2/N$, the number of nonzero coefficients of the model in the wavelet domain, and the ℓ_1 norm of the model in the wavelet domain. This information is used to construct a Pareto curve whose implication and significance are discussed for ℓ_1-regularized least-squares problems by *Hennenfent et al.* [2008] and *van den Berg and Friedlander* [2008], to name a few.

[37] As expected, the larger the threshold, the more difficult it is to reach a good fit to the data (Figure 4 and Table 2); for a large threshold, the number of removed wavelet coefficients is too large to find a model that satisfies the data. On the other hand, if the threshold is too low, the fit to the data is easily assured and some coefficients only have a negligible impact on explaining the data.

[38] In our experiments, the χ^2_{red} value bottoms out to about 0.39. This value is also found by *Sigloch* [2008,

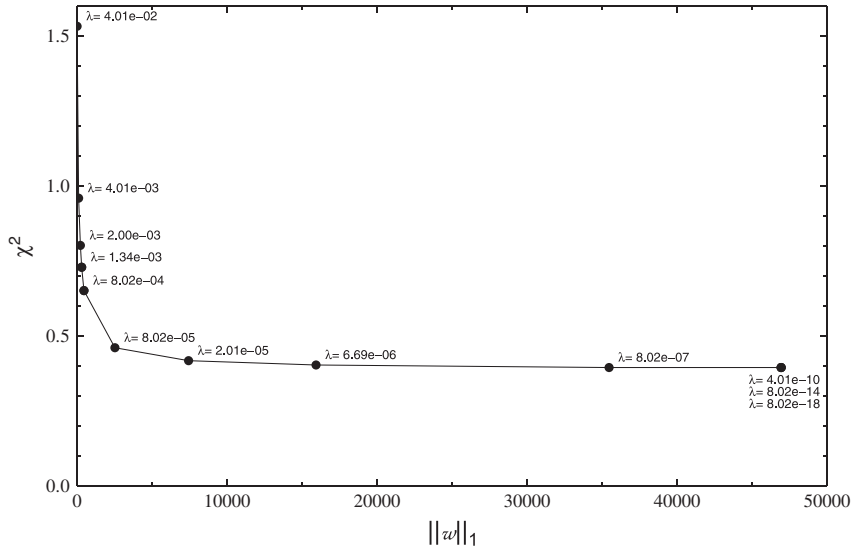

Figure 4. Trade-off L-curve for the FISTA inversion scheme. The value of the threshold λ is shown at each point.

Table 2-1, p. 96, Chapter 2] using a more classical inversion with a tetrahedral parameterization. If the data errors in the direct problem were correctly estimated, the χ^2_{red} value is expected to reach 1. However, as stated in *Sigloch* [2008, Chapter 2, p. 98, Figures 2–4], the initial error estimates in the delay times may have been too pessimistic, by about 30%. We chose the value of the threshold located at the corner of the L-curve [*Hansen*, 1992] shown in Figure 4: the regularization parameter $\lambda \approx 8.10^{-5}$ and $\chi^2_{\text{red}} = 0.46$. This value of the regularization parameter yields a good balance between a small residual norm $\|A\bar{m}-b\|$ and a small solution $\|\bar{w}\|_1$.

[39] For the chosen value of the regularization parameter, Figure 5 shows the evolution of the data misfit and model norms during one inversion whose maximum number of iterations was limited to 700. Color represents the iteration number. The curve decreases monotonically. Convergence is almost fully achieved after 300 iterations. During the last 400 iterations, the evolution of the two norms is much slower. The final value of χ^2_{red} is similar to the one obtained with 566 iterations: 0.4590 (for 700 iterations) compared to 0.4609 (for 566 iterations). The final model has 5289 nonzero coefficients, a value to be compared with the 3,637,248 degrees of freedom, equivalent to a relative sparsity of 0.14%. Should we take into account that our stations are located mainly in North America (Figure 1), and were we to compute sparsity only over one of the six chunks of the cubed Earth and only for the upper 1500 km of our model, the percentage would still be 1.5%. Either way, the hypothesis that a sparse representation in the wavelet domain exists for the given parameterization is confirmed, since the models have a very small number of nonzero wavelet coefficients. Using those, we are able to explain the data down to the estimated noise level.

Table 2. Value of the ℓ_1 Norm of the Final Model in the Wavelet Space, the Reduced Chi-Square Statistic χ^2_{red}, the Corresponding Threshold λ, and the Number of Nonzero Wavelet Coefficients

$\|w_n\|_1$	χ^2_{red}	λ	Nonzeros
4	1.53	4×10^{-02}	3
110	0.96	4×10^{-03}	104
219	0.80	2×10^{-03}	217
309	0.73	1×10^{-03}	321
462	0.65	8×10^{-03}	539
2,521	0.46	8×10^{-05}	5,838
7,435	0.42	2×10^{-05}	39,951
15,941	0.40	7×10^{-06}	128,955
35,481	0.39	8×10^{-07}	404,161
46,944	0.39	4×10^{-10}	1,444,608
46,962	0.39	8×10^{-14}	1,641,522
46,962	0.39	8×10^{-18}	1,643,240

5.2. Velocity Model

[40] Our preferred velocity model is shown in Figures 6–9, where it is compared to that obtained by

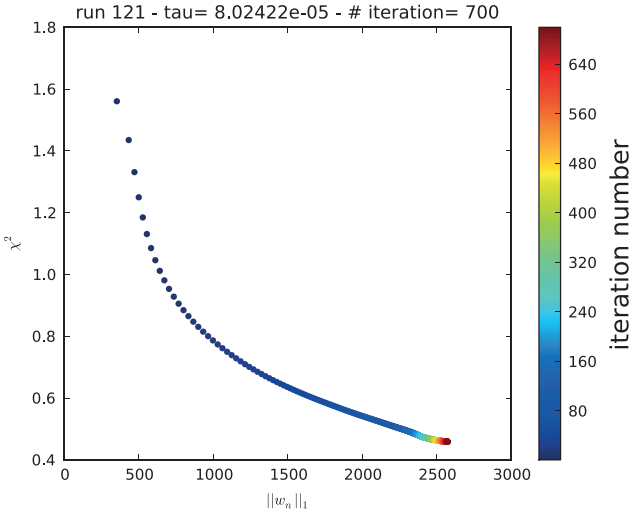

Figure 5. Evolution of the reduced chi-square misfit and the ℓ_1 norm of the model in the wavelet domain with the iteration number color-coded, for one inversion.

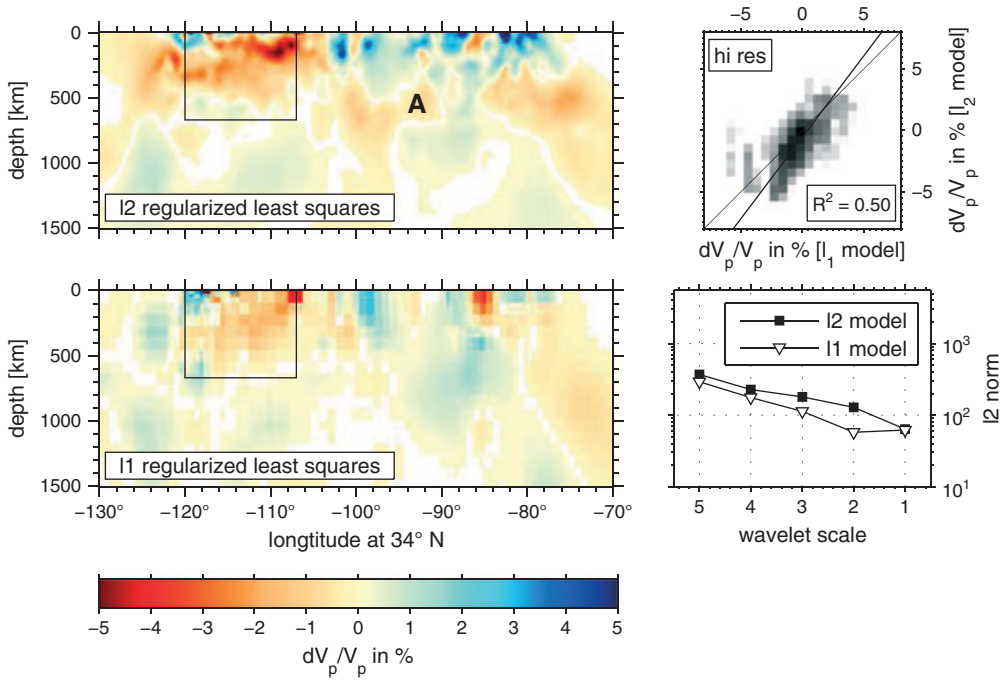

Figure 6. Cross section through the velocity models at 34°N (see Figure 1 for location). The top left panel is the cross section in the *Sigloch* [2008] model. The bottom left one is from this study. In the top right, we represent the values of the velocities as a density scatterplot for the well-resolved (hi-res) part, a total least-squares regression line through the data (thick black line) and the one-to-one line for reference (thin black line). The well-resolved part (rectangles) is located below the dense network in the upper mantle. The remainder of the model domain is less well resolved. In the bottom right, we show energy spectra of the heterogeneity in both cross sections, as the ℓ_2 norms of their wavelet coefficients for different scales. See text for details.

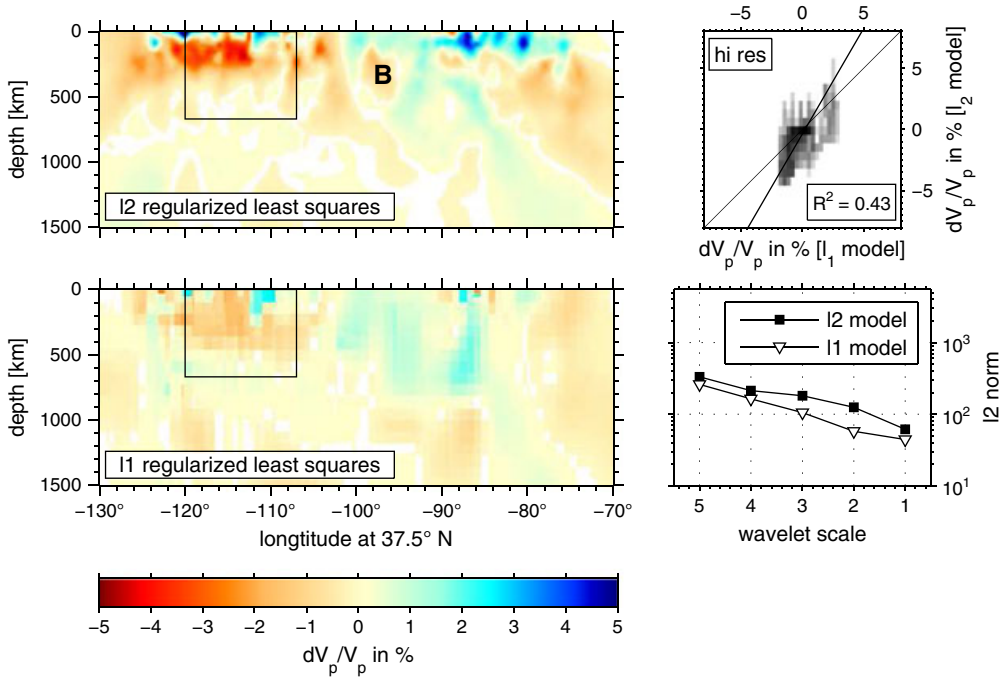

Figure 7. As in Figure 6 but for a cross section at 37.7°N (see Figure 1 for location).

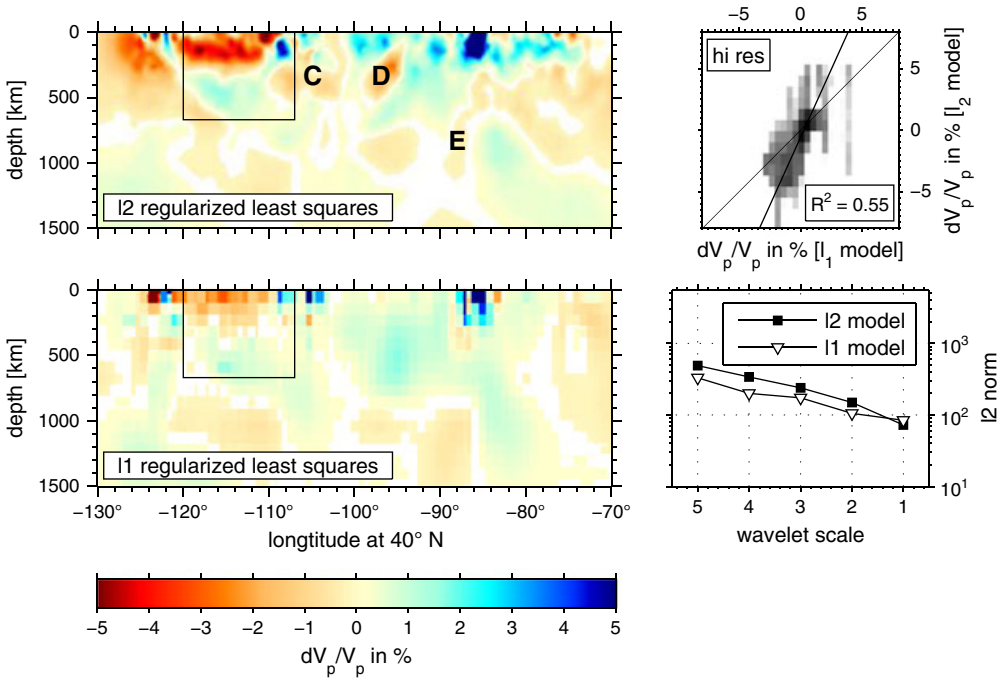

Figure 8. As in Figure 6 but for a cross section at 40°N (see Figure 1 for location).

Sigloch [2008]. The difference between our and their studies lies in the prior information introduced in the inversion: We use the sparsity hypothesis (see section 2.1). A minor difference is that we did not use the regional ISC *P* wave arrival time data set, but this omission should only affect the top 60 km of our model.

[41] The comparison between both models is made in a number of ways. The first is by visual comparison of a few important cross sections. The second is by plotting the values of the voxels in one model versus the other in a given cross section and calculating their correlation coefficients and a total least-squares regression line. The third is by inspection of the scale-dependent decay of the spatial energy of the velocity anomalies for a given section. Guided by the resolution tests presented by *Sigloch* [2008], we distinguish a well-resolved zone that corresponds to the part of the sections below the dense network of USArray stations prior to the year 2008 and down to the bottom of the upper mantle

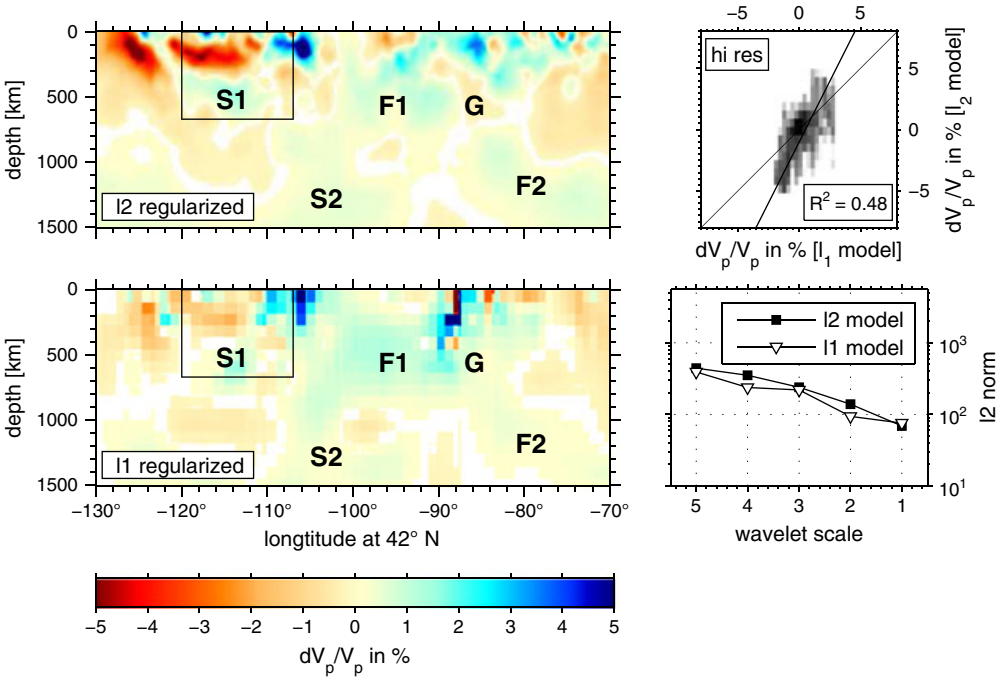

Figure 9. As in Figure 6 but for a cross section at 42°N (see Figure 1 for location).

Table 3. Significance-Tested Correlation Coefficients, R, Between the Model by *Sigloch* [2008] and Our Study for the Four Cross Sections Shown in Figures 6–9, Distinguishing the Well-Resolved and Poorly Resolved Parts

Cross Section Latitude	Resolved Zone	Unresolved Zone
34.0°	0.71	0.37
37.5°	0.65	0.39
40.0°	0.74	0.50
42.0°	0.69	0.43

(660 km depth). The horizontal extension of this zone may evolve slightly with latitude; we outline it by a rectangle of fixed location.

[42] In the well-resolved part of the study area, the cross sections of the models obtained by ℓ_2 smoothing (top left panels of Figures 6–9) and by ℓ_1 norm regularization (bottom left panels of Figures 6–9) are visually very well correlated. Comparing individual pixels gives correlation coefficients that are uniformly around $R \approx 0.7$ at the four latitudes shown. The top right panels of Figures 6–9 contain density estimates of the scatterplots of the voxels from the high-resolution part inside the marked rectangles. A total least-squares regression line, taking into account the uncertainty on both axes, was fit to the data points, and the cross-correlation coefficients R^2 are reported in the labels; R values are further listed in Table 3, both for the well-resolved and poorly resolved model domains. While the ℓ_2 model appears somewhat smoother, and the amplitudes are somewhat smaller in the ℓ_1 model, the close overall agreement is not unexpected: A well-resolved model should not be overly sensitive to the type of regularization used.

[43] The similarity of the models is also largely upheld in the regions where the resolution is poor, namely, in the eastern part of the cross section, and at depth. Thus, the broad appearance of either model is rather similar over the entire model domain. Roughly speaking, we are able to retrieve the same features, but within the resolved zone, the scatter in the pixel-by-pixel comparison between both models is lower than in the poorly resolved zones, and the correlation coefficients are higher (Table 3).

[44] However, not using the ℓ_2-smoothing operator allows some sharp discontinuities in the velocity model to survive the regularization, especially in the first 400 km of the upper mantle. The ℓ_1 approach does not use any smoothing per se. Of course, the very choice of wavelets and scaling functions in the vertical dimension also introduces a degree of spatial coupling of structure in that dimension, but since the wavelets act as a decorrelation filter [*Sweldens*, 1995] and the ℓ_1 scheme is designed to pick out the components that most strongly contribute to the signal, any smoothing introduced by the wavelet transform itself is data adaptive and of a more flexible nature than what is traditionally imposed via voxel-based ℓ_2-regularization tools.

[45] In Figures 7 and 9, some small-scale features seen between –90° and –80° and near the surface are not present in the ℓ_2-regularized model. In our model, fewer features are needed to explain the data. Examples can be seen in Figures 6 and 7, referenced as A and B, and in Figure 8 where the structures referenced as C, D, and E are absent in our model. These examples lie in the part of the model where the resolution is lower. The similarity is greater in the well-resolved part, but exceptions can be found near the western part of that region. In Figure 9, the main individually labeled structures that were discussed by *Sigloch et al.* [2008] are definitely also present in our model.

[46] The sparsity constraint forces the solution to explain the data with structures represented by a minimum number of wavelets. As expected, this tends to simplify the model. This appears clearly under the craton between the longitudes –110°/ –105° and –90° where resolution is low and our model presents a more uniform high-velocity anomaly. Another implication concerns the continuity of features. The connection between F1 and F2 in Figure 9 is a good example. The absence of the G structure connects these two bodies. The wavelet-based minimization outperforms the conventionally obtained results in terms of connecting the positive anomaly in the center right of all cross sections at depths larger than 300 km. We conclude that the effects of regularization are important in the ill-resolved parts of the model, where wavelets are able to reduce the number of disjoint anomalies without smoothing away the smaller scales that can locally be important.

[47] Another quantitative measure of the changes owing to regularization can be obtained by subjecting the cross sections to another wavelet analysis, not necessarily related to the wavelets used in the construction of one of the models, but simply to extract the scale-dependent distribution of energy in the models. For this comparison between the models, we used a D4 wavelet in the horizontal direction and Haar wavelets in the vertical direction, calculated over the four cross sections at latitudes varying from 34°N to 42°N. Since most of the model box belongs to the ill-resolved part, this will dominate in this analysis, and we have made no effort to isolate it from the small volume directly underneath the 2008 USArray data coverage. The bottom right panels in Figures 6–9 show the energy in four wavelet scales, numbered 4 to 1, coarse to fine, with the finest labeled scale corresponding to features about 50 km in vertical extent and 1.25° in the angular dimension. With increasing scale numbers the equivalent length scales of the features double with the support of the wavelets used. The coefficients at the largest quoted scale 5 contain information from the remaining scaling functions. The energy at the shorter length scales is either similar, lower, or only slightly higher (at 40°N and 42°N) in the ℓ_1-regularized model than in the ℓ_2-smoothed model. The decrease of energy with decreasing scale length from 2 to 1 is larger in the ℓ_2 model, and the energy is more consistently decreasing over the other scales in the ℓ_1 model.

[48] Wavelets are a multiscale basis. By applying an ℓ_1 norm, we minimize the number of wavelet coefficients, irrespective of the spatial length scale of the basis function. We expected that this would lead to models with a much simpler structure, but models with a simple structure can also be obtained by smoothing, albeit with an increased tendency to reduce the sharpness of some transitions. In the ℓ_1-regularized model, sharp boundaries have a greater chance at surviving the thresholding process. The data misfit measure χ^2_{red} is similar in both studies (0.407 for Sigloch's model and 0.459 for ours), but fewer features are needed to explain the data for our model.

[49] Wavelets thus are able to perform a dual task: They will smooth out artifacts introduced by data insufficiencies,

as one or a few wavelets are able to bridge unresolved gaps introduced by missing data. But in the case of a well-resolved gradient, there is no need for such smoothing, and the data may sometimes be fitted with smaller amplitudes. While these differences may be slight, even a small difference in the effects of model regularization may have important consequences when interpreted in a geodynamical context, e.g., when exploring the existence of slab tears via the presence or absence of high-velocity structure. No prior length scale is imposed in the inversion through smoothing or ad hoc remeshing of the model. However, the sizes of the larger features at depth are alike in the old and the new solutions, while near the surface, small structures and sharp boundaries are present in the wavelet model. The wavelets choose the size of the structures so as to agree with information present in the data. This point is not to be overlooked, as the choice of gridding can be influential. With 3.6 million voxels we are able to represent the sensitivity kernel rather finely and therefore are able to explore finer-scale structures or increase the resolution.

[50] Since the aim of our paper was primarily to study the effect of the alternative regularization on an actual inversion, we shall not repeat the geodynamical interpretations given by *Sigloch et al.* [2008]. We do point out, however, that the major elements leading to Sigloch's interpretation—such as the slab fragments S1, S2, F1, and F2—are present in both modeling efforts. It is therefore likely that the geological scenarios proposed by *Sigloch* [2011] and *Sigloch and Mihalynuk* [2013] remain valid in the alternative model presented here.

6. Conclusion

[51] We have presented the first application, on actual data, for global seismic tomography, of a new inversion and regularization scheme that employs sparsity constraints on the wavelet representation of the velocity model, via the ℓ_1 norm in wavelet model space. This new methodology is based on the hypothesis of sparsity of the model; that is, we assume that the model is represented by a small number of nonzero coefficients in a known basis. In our case, we chose the wavelet basis and tested new tools that were originally presented by *Simons et al.* [2011]. Finite-frequency kernels are computed on a cubed Earth representation of the model, which allows us to construct a Cartesian coordinate system upon which we can use a large number of wavelet families. For the present study, we chose to use the Cohen-Daubechies-Feauveau CDF 4–2 wavelet family in the angular dimensions and the Haar wavelet in the depth direction.

[52] A comparison of our velocity model with the study by *Sigloch* [2008], who used largely the same data but an ℓ_2 norm regularization with isotropic smoothing and damping, shows for different cross sections that the features within the model are very similar. As expected some discontinuities subsist in the ℓ_1 solutions, as we do not strongly impose model smoothness. For a comparable fit to the data (in terms of its reduced chi-square metric χ^2_{red}), fewer anomalous structures are needed with the ℓ_1 norm wavelet regularization and therefore the latest model is, in some sense, simpler without compromising small scale and details. The advantages of this inversion are (1) that the meshing is regular so that no prior information is needed for its construction and (2) that no prior assumptions are made on the geometry of the features. We only presuppose that these features can be well represented in the wavelet domain. In particular, we conclude that the geometry of slabs or other major structures within the mantle can be revealed by the use of this methodology given an appropriate data set.

[53] **Acknowledgments.** We thank the Editor, Rob Nowack, the anonymous Associate Editor, and two anonymous reviewers for their very thorough comments and pointers to the literature. J.C. and G.N. received support from ERC advanced grant 226837. I.L. is a research associate of the F.R.S.-FNRS (Belgium) and was supported by VUB grant GOA-062 and by the FWO-Vlaanderen grant G.0564.09N. F.J.S. was supported by NSF CAREER grant EAR-1150145. I.C.D was supported by NSF grant DMS-1025418.

References

Abers, G. A., and S. W. Roecker (1991), Deep structure of an arc-continent collision: Earthquake relocation and inversion for upper mantle *P* and *S* wave velocities beneath Papua New Guinea., *J. Geophys. Res.*, *96*(B4), 6379–6401.

Aravkin, A. Y., and T. van Leeuwen (2012), Estimating nuisance parameters in inverse problems, *Inv. Probl.*, *28*, 115016, doi:10.1088/0266–5611/28/11/115016.

Beck, A., and M. Teboulle (2009), A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imag. Sci.*, *2*, 183–202, doi:10.1137/080716542.

Bruckstein, A. M., D. L. Donoho, and M. Elad (2009), From sparse solutions of systems of equations to sparse modeling of signals and images, *SIAM Rev.*, *51*, 34–81.

Candès, E., and M. Wakin (2008), An introduction to compressive sampling, *IEEE Signal Process. Mag.*, *25*, 21–30.

Chevrot, S., and L. Zhao (2007), Multiscale finite-frequency Rayleigh wave tomography of the Kaapvaal Craton, *Geophys. J. Int.*, *169*, 201–215.

Chevrot, S., R. Martin, and D. Komatitsch (2012), Optimized discrete wavelet transforms in the cubed sphere with the lifting scheme—Implications for global finite-frequency tomography, *Geophys. J. Int.*, *191*, 1391–1402, doi:10.1111/j.1365–246X.2012.05686.x.

Chiao, L.-Y., and B.-Y. Kuo (2001), Multiscale seismic tomography, *Geophys. J. Int.*, *145*, 517–527.

Cohen, A., I. C. Daubechies, and J. Feauveau (1992), Biorthogonal bases of compactly supported wavelets, *Comm. Pure Appl. Math.*, *45*, 485–560.

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

Daubechies, I. (1988), Orthonormal bases of compactly supported wavelets, *Comm. Pure Appl. Math.*, *41*, 909–996.

Daubechies, I., M. Defrise, and C. De Mol (2004), An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Comm. Pure Appl. Math., LVII*, *57*, 1413–1457.

Davenport, M. A., M. F. Duarte, Y. C. Eldar, and G. Kutyniok (2012), Introduction to compressed sensing, in *Compressed Sensing: Theory and Applications*, edited by Y. C. Eldar and G. Kutyniok, chap. 1, pp. 1–64, Cambridge Univ. Press, Cambridge, U K.

Donoho, D. L. (1995), Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition, *Appl. Comput. Harmon. Anal.*, *2*, 101–126.

Donoho, D. L. (2006), For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution, *Comm. Pure Appl. Math.*, *59*, 797–829.

Donoho, D. L., and I. Johnstone (1994), Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, *81*, 425–455.

Figueiredo, M. A. T., R. D. Nowak, and S. J. Wright (2007), Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems, *IEEE J. Select. Topics Signal Process.*, *1*, 586–597, doi:10.1109/JSTSP.2007.910281.

Gholami, A., and H. R. Siahkoohi (2010), Regularization of linear and non-linear geophysical ill-posed problems with joint sparsity constraints, *Geophys. J. Int.*, *180*, 871–882, doi:10.1111/j.1365–246X.2009.04453.x.

Hansen, C. (1992), Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Rev.*, *34*, 561–580.

Haupt, J., and R. Nowak (2012), Adaptive sensing for sparse recovery, in *Compressed Sensing: Theory and Applications*, edited by Y. C. Eldar and G. Kutyniok, pp. 269–304, chap. 6, Cambridge Univ. Press, Cambridge, U K.

Hennenfent, G., E. van den Berg, M. P. Friedlander, and F. J. Herrmann (2008), New insights into one-norm solvers from the Pareto curve, *Geophysics*, *73*, A23–A26.

Herrmann, F. J., and Y. Bernabé (2004), Seismic singularities at upper-mantle phase transitions: A site percolation model, *Geophys. J. Int.*, *159*, 949–960.

Herrmann, F. J., and G. Hennenfent (2008), Non-parametric seismic data recovery with curvelet frames, *Geophys. J. Int.*, *173*, 233–248.

Herrmann, F. J., M. P. Friedlander, and Ö. Yilmaz (2012), Fighting the curse of dimensionality: Compressive sensing in exploration seismology, *IEEE Signal Process. Mag.*, *29*, 88–100, doi:10.1109/MSP.2012.2185859.

Jensen, A., and A. la Cour-Harbo (2001), *Ripples in Mathematics*, Springer, Berlin.

Li, X., A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann (2012), Fast randomized full-waveform inversion with compressive sensing, *Geophysics*, *77*, A13–A17.

Li, X.-G., M. Sacchi, and T. J. Ulrych (1996), Wavelet transform inversion with prior scale information, *Geophysics*, *61*, 1379–1385.

Lin, T. T., and F. J. Herrmann (2007), Compressed wavefield extrapolation, *Geophysics*, *72*, SM77–SM93.

Loris, I., G. Nolet, I. Daubechies, and F. A. Dahlen (2007), Tomographic inversion using ℓ_1-norm regularization of wavelet coefficients, *Geophys. J. Int.*, *170*, 359–370.

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

Mallat, S. (2008), *A Wavelet Tour of Signal Processing, The Sparse Way*, 3rd ed., Ann. Phys. (NY), San Diego, Calif.

Montelli, R., G. Nolet, F. A. Dahlen, and G. Masters (2006), A catalogue of deep mantle plumes: New results from finite-frequency tomography, *Geochem. Geophys. Geosys.*, *7*, Q11007, doi:10.1029/2006GC001248.

Paige, C. C., and M. A. Saunders (1982), LSQR: An algorithm for sparse linear equations and sparse least squares, *ACM Trans. Math. Software*, *8*, 43–71.

Piromallo, C., A. P. Vincent, D. A. Yuen, and A. Morelli (2001), Dynamics of the transition zone under Europe inferred from wavelet cross-spectra of seismic tomography, *Phys. Earth Planet. Inter.*, *125*, 125–139.

Ritsema, J., A. Deuss, H. J. van Heijst, and J. H. Woodhouse (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*, 1223–1236.

Ronchi, C., R. Iacono, and P. S. Paolucci (1996), The "cubed sphere": A new method for the solution of partial differential equations in spherical geometry, *J. Comput. Phys.*, *124*, 93–114.

Sigloch, K. (2008), Multiple-frequency body-wave tomography, PhD thesis, Princeton University, Princeton, N J.

Sigloch, K. (2011), Mantle provinces under North America from multifrequency *P* wave tomography, *Geochem. Geophys. Geosys.*, *12*, Q02W08, doi:10.1029/2010GC003421.

Sigloch, K., and M. G. Mihalynuk (2013), Intra-oceanic subduction shaped the assembly of Cordilleran North America, *Nature*, *496*, 50–56, doi:10.1038/nature12019.

Sigloch, K., N. McQuarrie, and G. Nolet (2008), Two-stage subduction history under North America inferred from multiple-frequency tomography, *Nat. Geosci.*, *1*, 458–462.

Simons, F. J., I. Loris, G. Nolet, I. C. Daubechies, S. Voronin, J. S. Judd, P. A. Vetter, J. Charléty, and C. Vonesch (2011), Solving or resolving global tomographic models with spherical wavelets and the scale and sparsity of seismic heterogeneity, *Geophys. J. Int.*, *187*, 969–988.

Strang, G., and T. Nguyen (1997), *Wavelets and Filter Banks*, 2nd ed., Wellesley-Cambridge Press, Wellesley, Mass.

Sweldens, W. (1995), The lifting scheme: A new philosophy in biorthogonal wavelet constructions, Proc. SPIE 2569, Wavelet Applications in Signal and Image Processing III, 68, September 1, 1995, doi:10.1117/12.217619.

Sweldens, W. (1996), The lifting scheme: A custom-design construction of biorthogonal wavelets, *Appl. Comput. Harmon. Anal.*, *3*, 186–200.

Szameit, A., et al. (2012), Sparsity-based single-shot subwavelength coherent diffractive imaging, *Nat. Mater.*, *11*, 455–459, doi:10.1038/nmat3289.

Tian, Y., K. Sigloch, and G. Nolet (2009), Multiple-frequency SH-wave tomography of the western US upper mantle, *Geophys. J. Int.*, *178*, 1384–1402.

Tikhotskii, S., I. Fokin, and D. Schur (2011), Traveltime seismic tomography with adaptive wavelet parameterization, *Izvestiya, Phys. Solid Earth*, *47*, 326–344.

Tikhotsky, S., and U. Achauer (2008), Inversion of controlled-source seismic tomography and gravity data with the self-adaptive wavelet parametrization of velocities and interfaces, *Geophys. J. Int.*, *172*, 619–630, doi:10.1111/j.1365–246X.2007.03648.x.

van den Berg, E., and M. P. Friedlander (2008), Probing the Pareto frontier for basis pursuit solutions, *SIAM J. Sci. Comput.*, *31*, 890–912, doi:10.1137/080714488.

VanDecar, J., and R. Snieder (1994), Obtaining smooth solutions to large, linear, inverse problems, *Geophysics*, *59*, 818–829.

Vonesch, C., and M. Unser (2008), A fast thresholded Landweber algorithm for wavelet-regularized multidimensional deconvolution, *IEEE Trans. Image Proc.*, *17*, 539–549.