Geophysical Journal International

Geophys. J. Int. (2020) **223,** 1899–1918
Advance Access publication 2020 September 08
GJI General Geophysical Methods

doi: 10.1093/gji/ggaa425

A Bayesian 3-D linear gravity inversion for complex density distributions: application to the Puysegur subduction system

Erin Hightower[1], Michael Gurnis[1] and Harm Van Avendonk[2]

[1] *Seismological Laboratory, California Institute of Technology,* 1200 *E California Blvd, Pasadena, CA* 91125, *USA. E-mail: ehightow@caltech.edu*
[2] *Institute for Geophysics, University of Texas,* 10100 *Burnet Road, Austin, TX* 78758, *USA*

Accepted 2020 September 4. Received 2020 September 4; in original form 2020 May 30

SUMMARY

We have developed a linear 3-D gravity inversion method capable of modelling complex geological regions such as subduction margins. Our procedure inverts satellite gravity to determine the best-fitting differential densities of spatially discretized subsurface prisms in a least-squares sense. We use a Bayesian approach to incorporate both data error and prior constraints based on seismic reflection and refraction data. Based on these data, Gaussian priors are applied to the appropriate model parameters as absolute equality constraints. To stabilize the inversion and provide relative equality constraints on the parameters, we utilize a combination of first and second order Tikhonov regularization, which enforces smoothness in the horizontal direction between seismically constrained regions, while allowing for sharper contacts in the vertical. We apply this method to the nascent Puysegur Trench, south of New Zealand, where oceanic lithosphere of the Australian Plate has underthrust Puysegur Ridge and Solander Basin on the Pacific Plate since the Miocene. These models provide insight into the density contrasts, Moho depth, and crustal thickness in the region. The final model has a mean standard deviation on the model parameters of about 17 kg m^{-3}, and a mean absolute error on the predicted gravity of about 3.9 mGal, demonstrating the success of this method for even complex density distributions like those present at subduction zones. The posterior density distribution versus seismic velocity is diagnostic of compositional and structural changes and shows a thin sliver of oceanic crust emplaced between the nascent thrust and the strike slip Puysegur Fault. However, the northern end of the Puysegur Ridge, at the Snares Zone, is predominantly buoyant continental crust, despite its subsidence with respect to the rest of the ridge. These features highlight the mechanical changes unfolding during subduction initiation.

Key words: Gravity anomalies and Earth structure; New Zealand; Inverse theory; Statistical methods; Subduction zone processes.

1 INTRODUCTION

Inverse methods have become increasingly popular for addressing a number of problems in earth science, particularly for subsurface mapping. Gravity inversion, for determining either the densities or depths of bodies of known density in the Earth, has been an established method of mapping the Earth's heterogeneities for some time, though often with emphasis on the non-linear approach. In non-linear gravity inversion, the densities and density contrasts of the subsurface bodies are assumed to be known and one solves for the geometry of the source, usually in terms of depth to a particular interface. These inversions include either methods operating in the spatial domain (Medeiros & Silva 1996; Prutkin & Casten 2009; Camacho *et al.* 2011) or those operating in the wavenumber domain (Parker 1972; Oldenburg 1974; Parker 1995; Chappell & Kusznir 2008; Cowie & Kusznir 2012; Bai *et al.* 2014). However, despite the Fourier method being one of the classical approaches to gravity inversion, wavenumber methods are often less effective in recovering a fully 3-D solution with multiple sources and complex geometry (Bear *et al.* 1995; Geng *et al.* 2019).

With the linear method, the unknowns are the densities of a discretized array of subsurface rectangular prisms and iteration is not required in order to reach model convergence, except in the case of testing variations in model regularization or other constraints. Solving for the 3-D density distribution also indirectly solves for the depth to key interfaces, such as the Moho, because we can interpret such boundaries from sharp transitions in density. While linear gravity inversion is an established method (Bear *et al.* 1995; Li & Oldenburg 1998; Silva *et al.* 2001; Silva Dias *et al.* 2009; Barnoud *et al.* 2016; Welford *et al.* 2018; Geng *et al.* 2019), many of the studies using it only do so for relatively simple geological geometries, such as a single sedimentary basin, mafic intrusion or volcanic

feature (Medeiros & Silva 1996; Silva *et al.* 2001; Barnoud *et al.* 2016). Successful application of this method to crustal scale studies and tectonic margins, with variable approaches to the implementation, also exist (Welford *et al.* 2010, 2018; Geng *et al.* 2019), but few have applied this method to subduction zones. Subduction margins posses a complicated juxtaposition of structure and rock types and significant and sometimes sharp lateral variations in density, as opposed to passive continental margins, which often exhibit a more gradual change in structure and rock type that is more easily handled by smoothed inversions. We construct a 3-D linear gravity inversion for an active subduction zone, demonstrating the successful application of this method to more complex density distributions and bolstering the validity of this method and its use in tectonic applications.

Inversion has the advantage of providing statistical feedback on solution quality. Specifically, within a Bayesian framework, the objective is to determine the posterior distribution of a set of parameters given prior distributions and likelihood functions that describe how the data relate to those unknown parameters (Tarantola 2005; Aster *et al.* 2013; De La Varga & Wellmann 2016; Wellmann *et al.* 2018). The Bayesian approach is particularly useful for geophysical inverse problems, which are in principle ill-posed because they are inherently non-unique. For example, gravity data cannot distinguish between a narrow density anomaly at depth or a wider source near the surface (Li & Oldenburg 1998; Welford *et al.* 2018; Geng *et al.* 2019). Consequently, one must introduce constraints and *a priori* information in order to transform them into well-posed problems. With the Bayesian formulation, we can account for both error in the data and error in our prior information to reduce how that error may be carried over into the final model, and we can quantify the error on our final solution via the covariance and resolution operators. The Bayesian approach we use here offers improvements over traditional gravity inversion and modelling techniques, where one usually removes the effect of the topography and the Moho and analyses the residual. Such an approach requires assuming constant layer densities when in fact those densities are often unknowns, and it requires assuming a known Moho depth that has to manually and iteratively be adjusted by the user. This makes it difficult to fully incorporate lateral changes in density. The Bayesian approach is more flexible and capable of handling complex 3-D geometries because it allows us to constrain where the boundary is *most likely* to be based on seismic data and what the densities are *most likely* to be, while allowing both to vary in accord with the gravity data, the final boundary location being dependent on the differential density. As such, we are able to draw conclusions about the 3-D density distribution in a tectonic setting that would otherwise not be as apparent with traditional forward or inverse gravity methods that require harder constraints or restrictions.

There are a number of common constraints widely used in gravity inversion, including inequality constraints, which specify the lower and upper bounds of parameter estimates; absolute proximity constraints, which specify that model parameters must be close to a specified value, based on geological information at particular points; and relative equality constraints, which specify that the spatial variation of the model parameter values must be smooth (Silva *et al.* 2001). Absolute proximity constraints are rarely used alone because there is often not enough prior information available to constrain all model parameters. An exception would be the minimum Euclidean norm, or similarly zeroth order Tikhonov regularization, which requires all parameter estimates to be as close as possible to null values. This type of regularization is biased towards a solution with minimum density and tends to concentrate mass anomalies toward the surface, which is not entirely physical or useful for our interpretation of the subsurface.

Minimum structure inversion, however, is a commonly used method (Last & Kubik 1983; Li & Oldenburg 1998; Farquharson 2008), utilized by codes such as GRAV3D (Li & Oldenburg 1998). To overcome the inherent insensitivity of gravity to depth and thus the tendency for the inversion to concentrate mass near the surface, these methods often apply a depth weighting (Li & Oldenburg 1998). Applying absolute proximity constraints and inequality constraints to specific regions of the model, however, overcomes the need for a depth weighting (Welford *et al.* 2018; Geng *et al.* 2019). While traditional inverse methods do allow for the adjustment of smoothing parameters, bounds on densities and variable weighting, they usually do so under hard constraints on predefined boundaries where the density is allowed to vary but the geometry of the boundary remains unchanged (Li & Oldenburg 1998; Welford *et al.* 2018). In contrast, the probabilistic approach offers more flexibility. Previous comparisons between such probabilistic methods and approaches such as those used by GRAV3D (Welford *et al.* 2018) highlight these distinctions as well, and we refer the reader to these sources for a more in depth comparison. These comparisons show that while each method has its advantages and disadvantages, a probabilistic approach using sparse seismic Moho constraints may not always lead to better results, particularly when there are significant lateral variations in crustal thickness and composition, as it tends to concentrate more unreasonable densities into different parts of the model to compensate (Welford *et al.* 2018). In contrast to previous applications of this probabilistic method (Barnoud *et al.* 2016; Welford *et al.* 2018; Geng *et al.* 2019), our approach directly incorporates constraints on the interface depths *and* on composition via the mapping of seismic velocities to density, not only at the locations of sparse depth to Moho constraints from seismic lines, but interpolated throughout the model domain and weighted according to the spatial extent of the prior data. We also propagate the error on the seismic velocities into the density prior to ensure that the densities obtained vary within a range that is consistent with the error in the seismic velocities and that the seismic data does not too strongly dominate the final model obtained by the inversion, such that it remains predominantly resolved by the gravity. Moreover, the Bayesian approach allows us to directly evaluate the error and statistical validity of our results in a way that does not assume the seismic data is the full truth.

Due to the non-uniqueness of gravity, however, even with absolute proximity constraints, some sort of smoothing or stabilizing functional is needed to produce a meaningful solution. This can come in the form of relative equality constraints such as either first or second order Tikhonov regularization, which spatially minimizes the first or second derivative of the physical property, respectively. Relative equality constraints by themselves have a tendency to produce a blurred but still valuable model of the density anomalies (Portniaguine & Zhdanov 1999; Silva *et al.* 2001). However, when combined with absolute equality constraints, this inversion technique is often able to produce accurate representations of the source geometry and density (Medeiros & Silva 1996; Silva *et al.* 2001). As such, our method uses a combination of absolute and relative equality constraints in the form of Gaussian priors based on existing geophysical data and a combination of first and second order Tikhonov regularization.

There is distinction in the literature between traditional regularization methods and proper Bayesian approaches to inverse problems. Traditionally, regularization modifies the function relating the data to the source of its signal, in an effort to eliminate the

unstable problem by replacing it with a similar stable one. This often involves a penalty on the inversion that guarantees a unique solution (Calvetti & Somersalo 2018). The Bayesian approach, on the other hand, by modelling the solution as a random variable, allows one to use the exact function relating the data to its source and offers the flexibility of obtaining multiple reasonable solutions, as the final posterior model is in fact a probability distribution. However, the non-uniqueness of gravity inversion in particular requires some form of regularization. The regularization method that best bridges the classical deterministic theory and the Bayesian approach is Tikhonov regularization because instead of modifying the model function, it solves a minimization problem (Calvetti & Somersalo 2018). In that sense, Tikhonov regularization is essentially a smoothness prior and can be implemented within a probabilistic framework, allowing the inversion problem to remain Bayesian even though it involves regularization.

2 METHODS

2.1 Calculation of forward gravity

We model the subsurface density and structure of a defined region and its associated effect on the gravity by discretizing the subsurface into a finite number of rectangular blocks. The gravitational attraction of each rectangular prism is calculated and then summed to compute the gravity field. The gravitational attraction of a homogeneous right rectangular prism relative to an observation point on the surface is given as in Turcotte & Schubert (2014) as

$$\Delta g = \Gamma \Delta\rho \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \mu_{ijk}[\Delta z_k arctan\left(\frac{\Delta x_i \Delta y_i}{\Delta z_k R_{ijk}}\right) - \Delta x_i ln(R_{ijk} + \Delta y_j) - \Delta y_j ln(R_{ijk} + \Delta x_i)], \quad (1)$$

where $\Delta x_i = (x_i - x_p)$, $\Delta y_j = (y_j - y_p)$, $\Delta z_k = (z_k - z_p)$ and $\mu_{ijk} = (-1)^i(-1)^j(-1)^k$. $\Delta\rho$ is the density contrast of the prism, and Γ is the universal gravitational constant. x_p, y_p and z_p are the coordinates of the measurement point, and x_i, y_j and z_k are the coordinates of the corners of the prism, where $(i, j, k) = (1, 2)$. R_{ijk} is the distance from the measurement point to a corner at x_i, y_j, z_k and is given by $R_{ijk} = (\Delta x_i^2 + \Delta y_j^2 + \Delta z_k^2)^{1/2}$.

The sum defines the geometry of the prism relative to the observation point and can be extended to the case of multiple prisms, such that each prism in the domain has a single geometry coefficient for each gravity observation point. We invert gravity data at N observation points to obtain the best-fitting estimate of the densities of M subsurface prisms, or M model parameters. Eq. (1) then results in an $N \times M$ matrix $\mathbf{G}$ that describes the geometry of each prism relative to each observation point times Γ. The gravity anomaly at any observation point due to the combined attraction of all the prisms is the product of this matrix and $\boldsymbol{\Delta\rho}$, which is an $M \times 1$ vector containing the differential density of each prism, expressed as

$$\Delta\mathbf{g} = \boldsymbol{G\Delta\rho}. \quad (2)$$

2.2 Linear least-squares inversion

We adopt the method for linear least-squares inversion as given in Aster *et al.* (2013) and Tarantola (2005). For N data points and M model parameters, where $g_i(\mathbf{m})$ is the model prediction of the ith datum (the Δ has been omitted for clarity), the least-squares misfit is:

$$F(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{N}(d_i - g_i(\mathbf{m}))^2. \quad (3)$$

For a linear model such as that given in eq. (2), the model derivative is independent of the model parameters, and our prediction can be written directly as $\mathbf{Gm}$. The Gauss–Newton solution of the model parameters that minimizes the least-squares misfit in eq. (3) is thus:

$$\mathbf{m} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T(\mathbf{d}). \quad (4)$$

The data $\mathbf{d}$ are the observed gravity anomaly values, and the model parameters to be estimated are the differential densities of each discretized block in the subsurface.

We accommodate data errors and prior constraints on the model parameters in the inversion via a Bayesian approach. Bayes theorem states that the probability of the model parameters, given the data, is proportional to the product of (1) the probability of producing those data with the model and (2) the probability of the model itself.

$$P(\mathbf{m}|\mathbf{d}) \propto P(\mathbf{d}|\mathbf{m})P(\mathbf{m}). \quad (5)$$

$P(\mathbf{m})$ is a prior that we use to restrict the model parameters to certain values given our existing geological knowledge.

In including the data error in the least-squares solution, we make the key simplifying assumption that the data are independent. In the case of gravity, we are incorporating the relative attraction of both adjacent and distal blocks of mass, and if the data are gridded with some form of interpolation, then they are arguably not truly independent. However, given the complexity of the problem and its physical geometry, the interdependence of the data is difficult to quantify and the simplifying assumption that the data are independent is sufficient to perform the inversion. We assume each data point can be represented by a Gaussian distribution with known error such that we can define a new least-squares misfit:

$$F(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{N}\left(\frac{d_i - g_i(\mathbf{m})}{\sigma_{d_i}}\right)^2, \quad (6)$$

where we are now minimizing the difference between the known and predicted gravity, given the error in the gravity data. From Bayes Theorem, minimizing this new misfit $F(\mathbf{m})$ is equivalent to maximizing $P(\mathbf{m}|\mathbf{d})$. To incorporate the data error into the model parameter solution, we define a diagonal and symmetric weight matrix $\mathbf{C_d}$ with the data variance on the diagonal. The solution becomes:

$$\mathbf{m} = (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{d}. \quad (7)$$

2.3 Tikhonov regularization

Linear least squares, even when using the generalized inverse or the truncated generalized inverse to handle small singular values, is often insufficient for many inverse problems due to non-uniqueness and instability, especially for high-dimensional problems. Thus, a form of regularization must be applied. We use a combination of first and second order Tikhonov regularization, which stabilizes the inversion and acts as a relative equality constraint on the values of the model parameters. First order Tikhonov regularization minimizes the square of the first spatial derivative of the model parameters (i.e. the gradient), thus serving to flatten the solution. Second order Tikhonov minimizes the square of the second spatial derivative of the model parameters (i.e. the curvature) and hence smooths the solution. Zeroth order Tikhonov, on the other hand, favors models

that are small and is identical to applying a Gaussian prior with a mean of zero and minimizing the square of the model parameter values themselves.

As Tikhonov regularization is equivalent to applying a prior that enforces either small values, flatness, or smoothness, we can derive the regularized solution by adjusting the misfit equation to reflect the additional minimization of the model parameters or their first or second derivatives.

$$F(\mathbf{m}) = \frac{1}{2}(\mathbf{d}-\mathbf{Gm})^T\mathbf{C_d}^{-1}(\mathbf{d}-\mathbf{Gm}) + \lambda^2(\mathbf{Lm})^T(\mathbf{Lm}), \quad (8)$$

L is either the identity matrix, a first derivative finite difference operator, or a second derivative finite difference operator for zeroth, first, or second order Tikhonov regularization, respectively. λ is a constant controlling the strength of the regularization. As the misfit remains exactly quadratic with the addition of the Tikhonov regularization term, the inverse problem remains linear, and the weighted and regularized linear least-squares solution becomes

$$\mathbf{m} = (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G} + \lambda^2\mathbf{L}^T\mathbf{L})^{-1}\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{d}. \quad (9)$$

For 3-D models, first and second order Tikhonov are implemented using the sums of the finite-difference approximations to the first or second derivatives in each direction, respectively. Because the discretization of the grid can be different in the x, y and z directions, we apply three different regularizations, with associated constants α for the x-, β for the y- and ζ for the z-direction. The derivation of the Tikhonov regularization matrices is given in Appendix A. For three-dimensions, the weighted Tikhonov regularized solution is

$$\mathbf{m} = (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G} + \alpha^2{\mathbf{L_x}}^T\mathbf{L_x} + \beta^2{\mathbf{L_y}}^T\mathbf{L_y} + \zeta^2{\mathbf{L_z}}^T\mathbf{L_z})^{-1} (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{d}). \quad (10)$$

Without a flatness constraint in the far-field, abrupt density changes at the edges of the model domain result in a classical gravity edge effect. Consequently, to ensure mathematical stability, we impose an infinite edge boundary condition, which allows the gravity to smoothly continue off the edges of the model area. We accomplish this condition by padding the domain with edge prisms that are sufficiently long that they extend far beyond the edge of the gravity grid (on the order of 1000 km for the regional problem with which we test the method). We also enforce this condition during the inversion by using first order Tikhonov regularization with a strong regularization coefficient to minimize the difference between the edge parameters and the adjacent values so that their predicted densities are the same. Thus, we apply different orders and strengths of Tikhonov regularization to the edges and the interior of the model simultaneously. The interior of the model has second order Tikhonov imposed in the horizontal directions to allow for smooth continuity of density bodies in the subsurface, and first order Tikhonov is applied in the vertical direction, as it is better equipped to allow for sharp contacts between layers of rock, while strong first order is applied on the boundary.

As before, this variable order Tikhonov regularization can be achieved by redefining the misfit equation, where separate **L** matrices apply different weights to different sets of model parameters and different directions. The full Tikhonov regularized solution, with boundary conditions applied, is

$$\mathbf{m} = (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G} + \mathbf{L})^{-1}(\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{d}), \quad (11)$$

where

$$\mathbf{L} = \alpha^2{\mathbf{L_x}}^T\mathbf{L_x} + \beta^2{\mathbf{L_y}}^T\mathbf{L_y} + \zeta^2{\mathbf{L_z}}^T\mathbf{L_z} + b^2{\mathbf{B_x}}^T\mathbf{B_x} + b^2{\mathbf{B_y}}^T\mathbf{B_y}, \quad (12)$$

b is the weight of the first order Tikhonov regularization applied to the boundary condition. $\mathbf{B_x}$ and $\mathbf{B_y}$ are the regularization matrices that apply the boundary conditions in the x and y directions, respectively.

2.4 Priors

Meaningful solutions consistent with existing geological knowledge are obtained by applying absolute equality constraints as Gaussian priors. In this approach, each parameter is forced to be close to a mean value but is allowed to vary within a specified range. Different regions of the model domain can have different priors depending on (1) what we suspect the densities of the rocks in those areas are and (2) how confident we are in those values based on their location relative to the other data we have. The prior on each parameter is given by the Gaussian probability density function

$$P(m_k) = \frac{1}{\sigma_p\sqrt{2\pi}}exp(-\frac{1}{2\sigma_p^2}(m_k-\mu_p)^2), \quad (13)$$

where m_k is the estimated model parameter value, μ_p is the expected value of that model parameter based on our prior information and σ_p is the standard deviation of the prior for that parameter.

As with the data error and Tikhonov regularization, we define a new misfit by adding the exponential component of the Gaussian prior to the existing misfit:

$$F(\mathbf{m}) = \frac{1}{2}(\mathbf{d}-\mathbf{Gm})^T\mathbf{C_d}^{-1}(\mathbf{d}-\mathbf{Gm}) + \alpha^2(\mathbf{Lm})^T(\mathbf{Lm}) + \frac{1}{2}(\boldsymbol{\mu_p}-\mathbf{m})^T\mathbf{C_p}^{-1}(\boldsymbol{\mu_p}-\mathbf{m}). \quad (14)$$

Defining the prior covariance operator $\mathbf{C_p}$ as an $M \times M$ diagonal matrix with the variance of the prior on the diagonal, we arrive at the final data weighted, Tikhonov regularized solution with prior constraints

$$\mathbf{m} = (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G} + \mathbf{L} + \mathbf{C_p}^{-1})^{-1}(\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{d} + \mathbf{C_p}^{-1}\boldsymbol{\mu_p}), \quad (15)$$

where **L** is defined as in eq. (12). This is the final solution vector used in our inversion. The row or column number of the elements along the diagonal of $\mathbf{C_p}$ correspond to the index number of that model parameter. Likewise, $\boldsymbol{\mu_p}$ is an $M \times 1$ vector for which each element corresponds to the density of single prism. To apply different priors to different model parameters, one need only use the coordinates of the model parameter centroids within the desired region to find the appropriate model parameter index and apply a value to that element. If an element on the diagonal of $\mathbf{C_p}^{-1}$ is zero, then no prior is applied to that model parameter.

2.5 Quantifying error

A key advantage of a Bayesian approach is that it allows us to statistically evaluate the solution, via the posterior covariance matrix **C** of the model parameters and the resolution matrix **R**. The covariance matrix is defined as the inverse of the Hessian:

$$\mathbf{C} = (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G} + \mathbf{L} + \mathbf{C_p}^{-1})^{-1}. \quad (16)$$

Here the values of **m** estimated by the inversion are the centre-points of the posterior Gaussian, and the diagonal values of the covariance matrix **C** are their associated variances.

The resolution matrix is determined from the covariance matrix (Tarantola 2005):

$$\mathbf{R} = \mathbf{I} - \mathbf{C}\mathbf{C_p}^{-1}, \tag{17}$$

where $\mathbf{I}$ is the identity matrix. If the resolution matrix equals the identity matrix, the model is fully resolved by the data. This particular formulation of the resolution operator primarily allows us to distinguish between those parameters that are resolved by inversion of the gravity data and those that are resolved by the prior. Mathematically, this can be written as:

$$tr(\mathbf{I}) = tr(\mathbf{R}) + tr(\mathbf{C}\mathbf{C_p}^{-1}) \tag{18}$$

meaning the total number of model parameters is the sum of the number of parameters resolved by the data and the number of parameters resolved by the prior information (Tarantola 2005). Higher resolution (values closer to 1) means those parameter values have mostly been determined by the inversion—in other words, we have learned something from the gravity that we did not know *a priori*. On the other hand, low resolution (values closer to 0) means the values of those parameters are almost entirely attributed to the prior. This is the case for regions of the model where the prior is very strong, that is a very small prior variance.

Ultimately, solution quality is based on the mean absolute error of the gravity and the mean standard deviation of the model parameters as determined from the diagonal of the covariance matrix, as well as visual inspection of the model to determine its geological reasonability. Even with relative and absolute equality constraints, gravity inversion remains non-unique and there are a number of model solutions that could fit the data. It is possible to obtain a solution that minimizes the misfit as required but that still appears geologically unreasonable and must be disregarded as the most likely posterior distribution of densities. However, the regularization and priors ensure enough stability in the model that with the appropriate regularization parameters α, β and ζ, the model obtained is geologically sound and in line with our standing geophysical knowledge.

3 SYNTHETIC TESTS

Estimating optimal regularization parameters is difficult for gravity inversion. We use an iterative technique on a series of synthetic tests to determine α and ζ values that produce (1) the best fit between the predicted and observed gravity and (2) the most geologically reasonable solution, which for the synthetic models, is a nearly complete recovery of the known density distribution. We conduct these synthetic tests on a simplified lower resolution model of a subduction system. In all synthetic tests, we construct a density model, compute the forward gravity as given by eq. (1) and add Gaussian noise to the gravity using a similar standard deviation to that of the data set we will later use (about 1.7 mGal). We invert this gravity for a range of Tikhonov regularization parameters and orders, with or without priors on specific sets of model parameters, while attempting to recover the known density distribution and judging the stability of the inversion.

The performance of the inversion when used with first and second order Tikhonov is tested using a simplified synthetic 3-D model of a subduction zone (depicted in representative cross-sections in the bottom row of Figs 1 and 2). We test various combinations of the horizontal regularization coefficient α and the vertical regularization coefficient ζ for the cases of only first order Tikhonov, only second order Tikhonov, and a combination of second order in the horizontal and first order in the vertical. For each of these cases, we test four additional classes of constraints: no priors, priors enforced only on parameters within the water column, priors enforced only on parameters within the water and crustal layers, and priors on all parameters, including the mantle. The prism size is about 17.5 km in the x-direction, 22.5 km in the y-direction, and increases from about 206 to 2060 m from shallow to deeper depths in the z-direction. The α and ζ values tested range from 10^{-3} to 10^{8}. There are a total of 10, 648 model parameters and 22, 500 data points, yielding an overdetermined system. The synthetic density model is constructed with a seawater density of 1027 kg m^{-3}, oceanic crustal density of 2900 kg m^{-3}, sediment density of 2300 kg m^{-3}, continental crustal density of 2700 kg m^{-3} and mantle density of 3300 kg m^{-3}. We define differential density, $\Delta\rho$, by subtracting the lateral average of each layer from the true density of each prism in that layer. The prior densities, when applied, match those differential densities. The standard deviation of the priors, when applied, are 5 kg m^{-3} for seawater, 80 kg m^{-3} for the sedimentary and crustal rocks and 100 kg m^{-3} for the mantle.

The results for these synthetic tests are summarized in Figs S1–S4, which show gridded results for each combination of α and ζ in panels corresponding to the order(s) of Tikhonov regularization used (panel rows) and the set of priors used (panel columns). Grey regions demarcate α and ζ combinations where the regularization strength is too low to produce stable results. The minimum of each test for both the mean absolute error (MAE) on the gravity and the MAE on the model parameters is plotted in each of these figures as well. Fig. S1 depicts the MAE between the true gravity field of the synthetic model and the gravity predicted by the recovered density distribution. Changes in the gravity misfit are much more dependent on the order of regularization than they are on the presence of a prior. For first order Tikhonov alone, the misfit increases dramatically above α values of 10^4 because the model becomes too flat to correctly reproduce the shorter wavelength variations in the gravity field. For second order Tikhonov, stability is achieved at ζ values of 10^2 in cases with limited priors, above which the gravity error remains reasonably low until α values of about 10^7. For the combination of first and second order Tikhonov, the error remains reasonably low until an α value of 10^7 and between ζ values of 10^{-1} and 10^3. The lowest error on the gravity amongst all the tests is about 1.29 mGal, which is less than the noise level of 1.7 mGal, and occurs for the case of first order Tikhonov with no priors for $\alpha = 10^0$ and $\zeta = 10^{-1}$. The lowest gravity error occurs for the case of no priors because without priors the model is allowed to take whatever shape it must, subject to the smoothness constraint, to fit the data, again highlighting the inherent non-uniqueness of the gravity.

However, to achieve a geologically reasonable model, priors must be applied. For the case of enforcing a prior on all parameters, the minimum gravity error is still only 1.36 mGal, so the fit to the gravity data is not compromised by adding priors, while the fit to the true density model is dramatically improved. Fig. S2 illustrates the MAE between the predicted model parameter values and the true model parameter values of the known density model. For most combinations of different regularization orders and priors, too small of an α or ζ value and the regularization is not strong enough to provide a smooth and continuous density distribution, yielding non-physical fluctuations in the density values (Figs 1 and 2, columns 1 and 2). Alternative cross sections with results using different regularization strengths are shown in supplementary Figs S5 and S6. For α values that are too large, the solution smooths over any density variations almost entirely. For cases with no priors or limited priors, the misfit decreases with increasing ζ, but for cases with more priors, the misfit begins to increase again with larger ζ values after

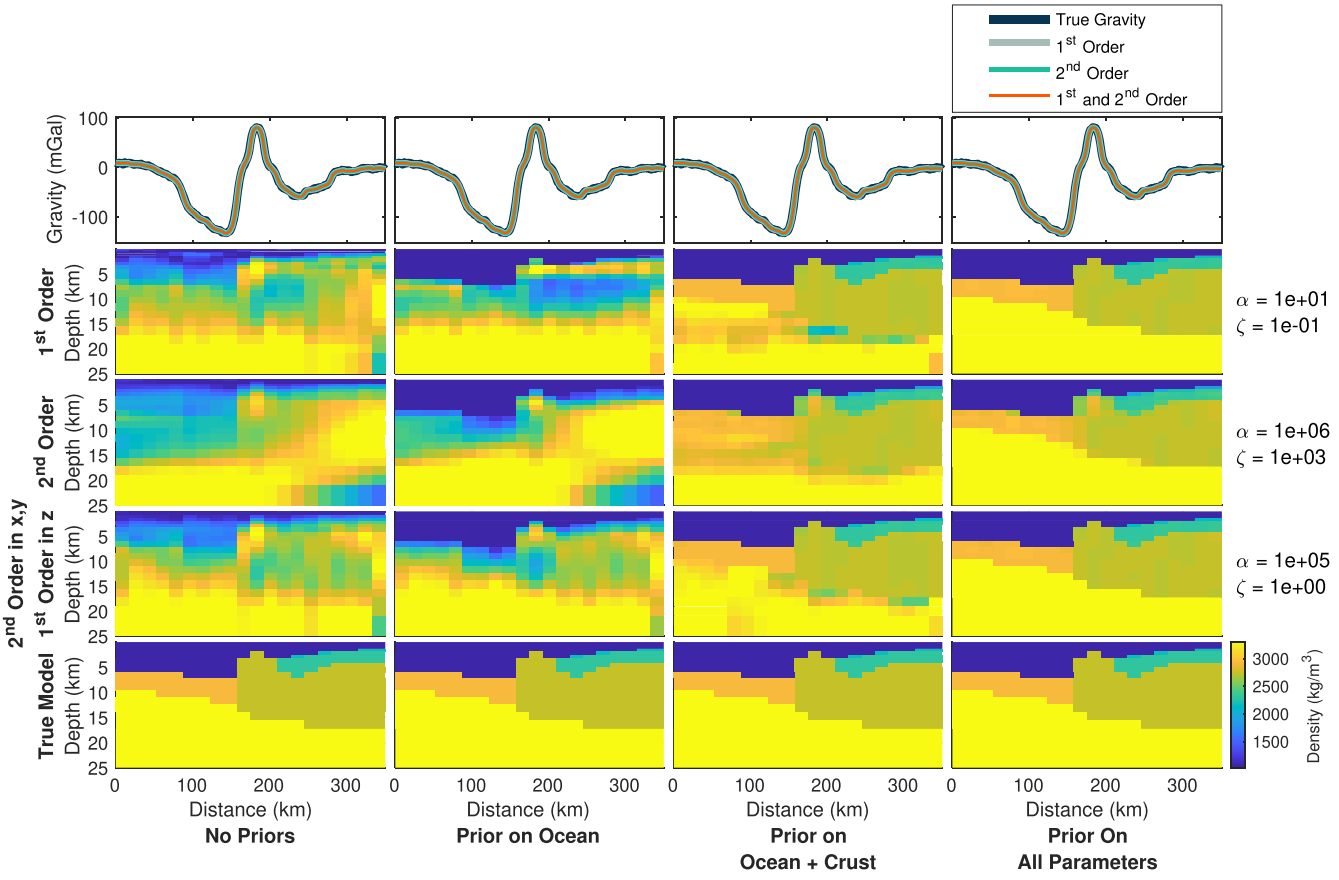

Figure 1. Representative cross-section in the *x*-direction of the synthetic inversion results for α, ζ combinations that produce some of the lowest errors for their respective order of Tikhonov regularization, as determined by comparing the test results depicted in Figs S1–S4. Row 1: gravity profiles for each of the three cases depicted in the panels below. Dark blue line: true gravity produced by the synthetic model, with noise; gray line: gravity from inversion using only first order Tikhonov; light blue line: gravity from inversion using only second order Tikhonov; orange line: gravity from inversion using second order Tikhonov in the horizontal and first order in the vertical. Row 2: cross-sections of the density model recovered from using only first order Tikhonov with $\alpha = 10^1$ and $\zeta = 10^{-1}$ for the cases of no priors, priors only on the ocean water parameters, priors on the ocean and crustal parameters, and priors on all parameters. Row 3: cross-sections of the density model recovered from using only second order Tikhonov with $\alpha = 10^6$ and $\zeta = 10^3$ for each of the different prior cases. Row 4: cross-sections of the density model recovered from using a combination of first and second order Tikhonov with $\alpha = 10^5$ and $\zeta = 10^0$ for each of the different prior cases. Row 5: cross-section of true synthetic density model for comparison.

achieving its minimum. However, the misfit decreases overall as we apply more priors throughout the model domain, starting with the ocean. Though the results from applying a prior only to the ocean do not look dramatically improved over the case of no priors, in practice, the prior on the ocean is one of the most important constraints because it is the most certain. It eliminates any need for the model to determine where the seafloor is and forces the inversion to put higher densities in the crust and mantle where they belong. This is evident in the cross-sections in Figs 1 and 2. The minimum absolute error on the model parameters amongst all tests is approximately 10.1 kg m^{-3} and occurs when using priors on all parameters and $\alpha = 10^1$, $\alpha = 10^4$ and $\alpha = 10^4$, and $\zeta = 10^{-3}$, $\zeta = 10^1$ and $\zeta = 10^{-1}$ for first, second, and combination-first-and-second Tikhonov, respectively.

Comparing the MAE of the model parameters, given the known density distribution, to the standard deviation of the model parameters as determined from the diagonal of the covariance matrix (Fig. S3) allows us to determine how the covariance matrix reflects uncertainty in the presence of *a priori* model constraints. For all cases except that of second order Tikhonov with no priors, the posterior standard deviation on the model parameters decreases with increasing α and ζ and is consistently lowest for the case where priors are enforced on all model parameters. For the α and ζ values where the MAE on the model parameters was lowest (red square in Fig. S2, lower right-hand panel), the mean posterior standard deviation on the model parameters is comparatively 56.3 kg m^{-3}, a value that, while higher, is still reasonably within a range necessary to distinguish one rock layer from another. The standard deviation from the covariance matrix continuously decreases with increasing regularization, while the MAE starts to increase after some minimum when priors are applied, because unlike with the MAE, the minimization in the gradient or curvature between the model parameters enforced by the Tikhonov regularization tends to dominate the definition of the covariance matrix. As the weight of regularization increases, the Tikhonov component of the misfit equation (eq. 14) is reduced, and as such, the posterior covariance is reduced as well. Too large of an α or ζ value can cause oversmoothing of the model parameters, and as such the standard deviation of the posterior solution is not always as accurate an estimator of the error on the model parameters away from the 'correct' density distribution as the MAE is.

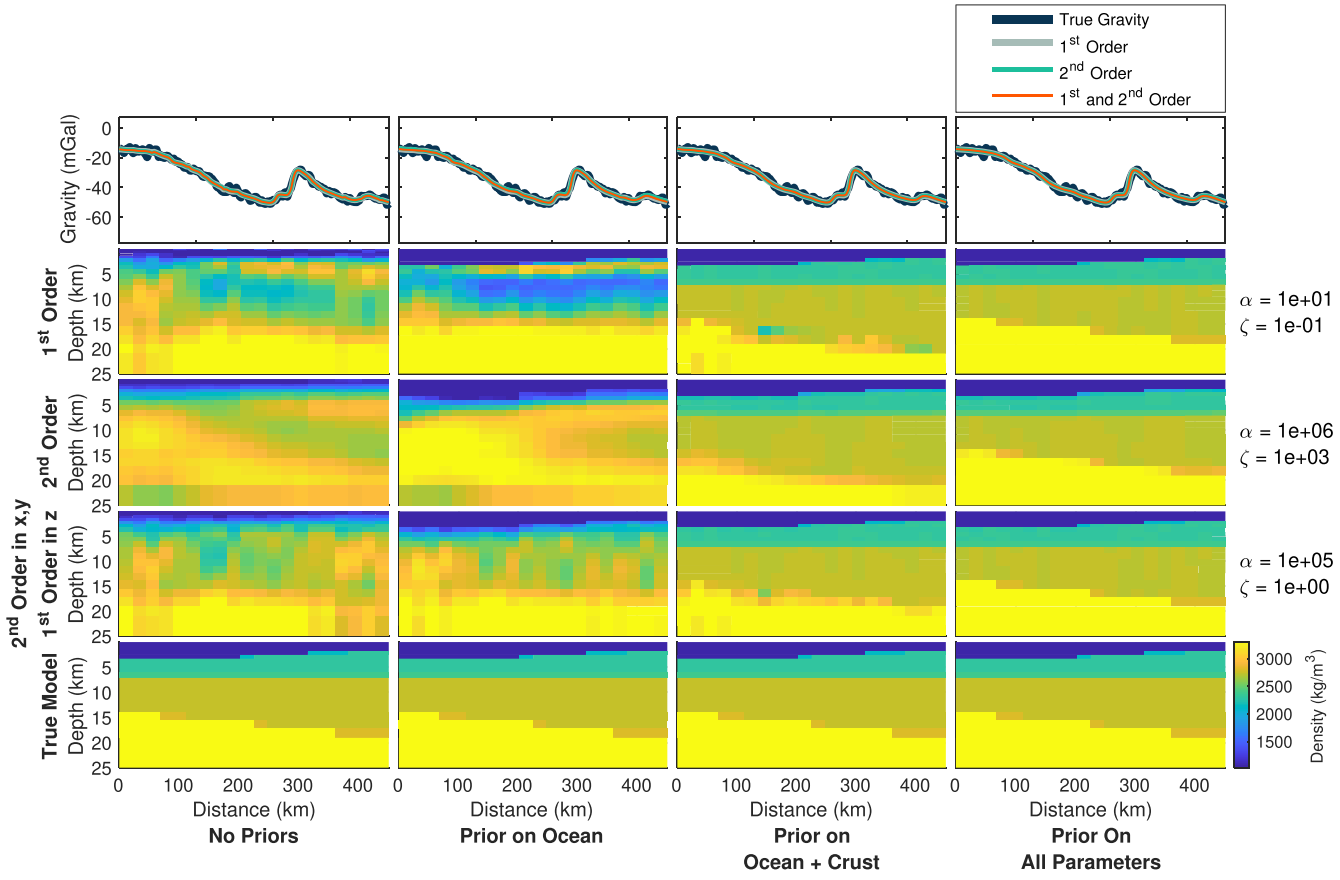

Figure 2. Representative cross-section in the y-direction of the synthetic inversion results for α and ζ combinations that produced some of the lowest errors for their respective order of Tikhonov regularization, as determined by comparing the test results depicted in Figs S1–S4. Row 1: gravity profiles for each of the three cases depicted in the panels below. Dark blue line: true gravity produced by the synthetic model, with noise; gray line: gravity from inversion using only first order Tikhonov; light blue line: gravity from inversion using only second order Tikhonov; orange line: gravity from inversion using second order Tikhonov in the horizontal and first order in the vertical. Row 2: cross-sections of the density model recovered from using only first order Tikhonov with $\alpha = 10^1$ and $\zeta = 10^{-1}$ for the cases of no priors, priors only on the ocean water parameters, priors on the ocean and crustal parameters, and priors on all parameters. Row 3: cross-sections of the density model recovered from using only second order Tikhonov with $\alpha = 10^6$ and $\zeta = 10^3$ for each of the different prior cases. Row 4: cross-sections of the density model recovered from using a combination of first and second order Tikhonov with $\alpha = 10^5$ and $\zeta = 10^0$ for each of the different prior cases. Row 5: cross-section of true synthetic density model for comparison.

However, because we do not know the correct density distribution in a study with real data, as we do in the synthetic tests, we can only use the α and ζ combination of the minimum MAE from the synthetic models as a proxy for what the ideal regularization coefficients must be in order to produce the best geological model. The covariance matrix still provides information on how well the model parameters are estimated, but we should expect errors as high as around $50-60\,\mathrm{kg\,m^{-3}}$ to be indicative of a good model because we do not want to fully minimize the Tikhonov component.

We can also use the resolution matrix **R** to quantify how much we have actually learned about the subsurface density structure from inverting the gravity data, as opposed to what we already knew from our prior. The mean resolution of all the model parameters for each of the tests is depicted for each combination of α and ζ in Fig. S4. The resolution should be interpreted as the fraction of that model parameter estimate that can be attributed to the inversion of the gravity data itself, as opposed to the prior. Resolution values close to 1 mean the model is well resolved by the gravity, not the prior. Hence, the tests for the case of no priors have a resolution of 1 because those models are resolved entirely by the gravity. Resolution values close to 0 mean the model is mostly resolved by the prior information alone and not the gravity: that is the gravity inversion did not tell us anything we did not already know from the prior. In this way, a resolution of 0 does not necessarily mean the values of the model parameters are wrong in a geological sense, just that the inversion was not useful. In some regions of the model, such as the ocean layer, where we know the density, it is desirable to have low resolution values because we want these regions to be entirely constrained by the prior and not affected by the inversion. As such there is a clear relationship between the MAE on the model parameters and the resolution: lower σ_p (i.e. a stronger prior) correlates with lower resolution and hence lower MAE on the model parameters. Going from the case of no priors to that of priors on all parameters, we can clearly see that the dependence on the prior increases for a greater number of α, ζ combinations as expected (Fig. S4). We can also see that the more or the stronger priors we apply, the less regularization is needed to produce a stable and reasonable model (Fig. S5).

Ideally, we are trying to obtain a model that both best fits the gravity and matches the prior data and so we neither want a model that is entirely determined by the gravity nor entirely determined by the prior. Thus, a very high resolution is not necessarily ideal.

Rather, we would expect resolution to increase with distance from the locations where we have prior constraints, exhibiting a spatial dependence. Therefore, a mean resolution somewhere in the middle may be considered reasonable, which is consistent with α and ζ values in the range of $10^4 - 5 \times 10^6$ and $10^{-1} - 10^1$, respectively (Fig. S4), as well as the best α and ζ values as determined by the MAE on the model parameters (Fig. S2). As an example, representative cross sections of the 3-D model results, using $\alpha = 10^1$ and $\zeta = 10^{-1}$ for first order, $\alpha = 10^6$ and $\zeta = 10^3$ for second order and $\alpha = 10^5$ and $\zeta = 10^0$ for combination first and second order, are shown in Figs 1 and 2. Similarly, representative cross-sections of the 3-D model domain using α and ζ values that produced the lowest MAE on the model parameters for each of the different combinations of regularization order and priors are shown in Figs S5 and S6. The top panel shows the predicted gravity profile produced by each of the models against the true gravity calculated from the known density model. The subsequent three rows depict the resulting models for first, second and combination first and second order Tikhonov, respectively. The bottom panel illustrates the corresponding cross-section of the true synthetic density model that we are trying to recover in each of these inversions, for comparison. Different α, ζ pairs are ideal for different orders of Tikhonov regularization. Higher regularization constants are needed for second order Tikhonov; those same coefficients would, on the other hand, oversmooth the first order models.

For these combinations of α and ζ, the accuracy of the resulting density models changes drastically across the different applied priors. However, the gravity signal for each model is essentially the same and matches the true gravity well, with an MAE of only about 1.3–1.4 mGal for each case, demonstrating the effective non-uniqueness of the gravity (Figs 1 and 2, Row 1). Thus, priors are necessary to improve the model. When priors are applied to all parameters, all three regularization options recover the known density model, though at higher values of α, combination first and second order Tikhonov is better at recovering the density distribution. Ultimately, the recovered model is more sensitive to changes in ζ than in α, and for low values of α, first order regularization appears sufficient. However, across all tests, the combination of first and second order Tikhonov consistently produced the most stable results and was the most successful at recovering the known density distribution. Moreover, increasing the resolution of the subsurface model (i.e. adding more model parameters) tends to require increasing the regularization strength, so the combination first and second order Tikhonov is more stable for larger models, as first order becomes too strong, flattening out the model completely, at large α values.

Even without any priors, some semblance of the structure is recovered for the example in Fig. S5 when using first order regularization, though with such high α values, structure is better recovered when using the first and second order combination. For first and second order, there is often an unrealistic degree of fluctuation in the density values for the case of no priors and priors only on the ocean (Figs 1 and 2). Ultimately, different combinations of the α and ζ values can yield similarly satisfactory models, but based on the above results for the MAE on the gravity and model parameters and the covariance and resolution matrices, and considering the increase in model resolution for the regional study, using a combination of first and second order Tikhonov with $\alpha = 10^5$ or $\alpha = 10^6$ and $\zeta = 10^0$ with priors of varying certainty on all parameters produces the best results. These are the values that will be applied to the subsequent regional case study of the Puysegur region offshore southern New Zealand.

4 APPLICATION TO THE PUYSEGUR REGION

Gravity inversion of an active tectonic margin is challenging because of the complicated structures and source geometries and the sharp lateral changes in density across the boundary. Those very compositional contrasts across and along such an active margin play a large role in governing the tectonic processes taking place. Because dynamic processes often dominate the gravity field and influence local topography, gravity modelling at these locations can shed light on important aspects of subduction (Toth & Gurnis 1998; Krien & Fleitout 2008). The Puysegur subduction zone is an attractive test case for subduction initiation in particular because of its young age and the transition from developed subduction in the north to incipient underthrusting in the south (Gurnis *et al.* 2004, 2019). As such, the margin provides a progressive snapshot of the subduction initiation process along strike. Puysegur also exhibits unusual gravity anomalies, the origin of which can inform us about the regional dynamics and motivates detailed study of Puysegur with a gravity inversion.

4.1 Regional setting

The Puysegur-Fiordland subduction zone lies at the northern end of the Macquarie Ridge Complex (MRC) and the southern tip of South Island, New Zealand. Present day plate motion is predominantly dextral strike-slip, with highly oblique subduction of the Australian Plate (AUS) northeastwards beneath the Pacific Plate (PAC) at the Puysegur Ridge and Fiordland (Fig. 3a; Sutherland 1995; DeMets *et al.* 2010). The Puysegur margin has evolved from a spreading ridge into a subduction zone. Spreading along the PAC-AUS margin began approximately 45 Ma, then became increasingly oblique as the AUS-PAC Euler pole migrated to the southeast during the Miocene, eventually rotating into a strike-slip plate boundary (Sutherland 1995; Lebrun *et al.* 2003). This evolution is evident in the curvilinear fracture zones that merge along the MRC and are prominent in the gravity field and bathymetry. Oblique convergence led to subduction beneath the Fiordland boundary starting around 16–10 Ma, beneath the northern extent of the Puysegur segment about 11–8 Ma, and beneath the southernmost extent of the Puysegur Ridge within the last several million years (Lebrun *et al.* 2003; Sutherland *et al.* 2006).

The crustal structure and tectonics related to the above kinematic history were investigated in detail with seismic reflection, seismic refraction, and bathymetric mapping during the recent South Island Subduction Initiation Experiment (SISIE, Gurnis *et al.* 2019). Puysegur has the advantage of being a small subduction zone with a well known plate kinematic history before and during subduction initiation, making it accessible for studying the process of subduction initiation and for constructing a regional gravity inverse model at a relatively high resolution.

The margin possesses distinctive, high amplitude gravity anomalies, which as of yet have poorly constrained structural and compositional interpretations and which have implications for the dynamic processes taking place in the region. The MRC is characterized by long and narrow bathymetric and gravitational highs and lows along strike (Fig. 3b). The southern part of Puysegur Ridge is characterized by a 100 to 150 mGal gravity high adjacent to the –100 to –150 mGal gravity low of the trench. In contrast, a significant approximately –150 mGal gravity low exists over the northern Puysegur Ridge—a region known as the Snares Zone (Fig. 3; Gurnis *et al.* 2019). This region is of particular interest in our gravity inversion

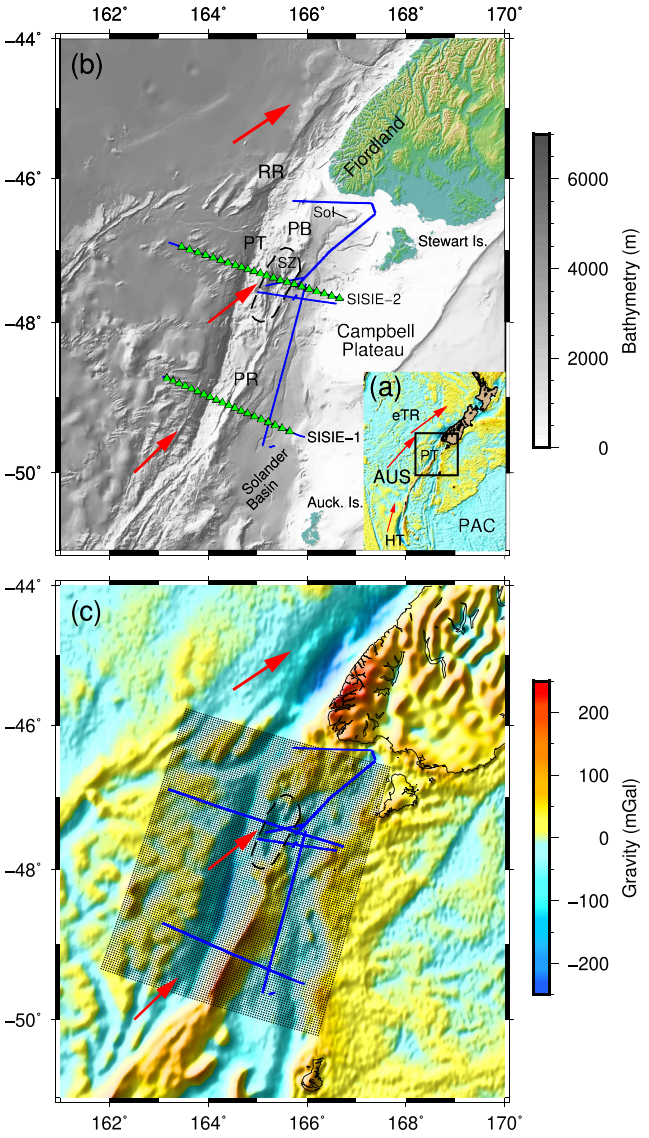

Figure 3. (a) Puysegur survey area, outlined by the black rectangle. The Macquarie Ridge Complex is the long, narrow gravity high/low feature running between the Australian (AUS) and Pacific (PAC) plates from the Hjort Trench (HT) in the south to the Puysegur Trench (PT) in the north. eTR is the extinct Tasman Ridge. Base map is free-air gravity (Sandwell *et al.* 2019). (b) Bathymetry of Puysegur region from the NIWA grid (Mitchell *et al.* 2012). Solid blue lines are MCS lines. Triangles represent the locations of OBS. SISIE-1 and SISIE-2 are combined OBS and MCS lines. Black dashed line outlines the Snares Zone (SZ) bathymetric low. RR: Resolution Ridge; Sol: Solander Island; PB: Puysegur Bank; PR: Puysegur Ridge; and PT: Puysegur Trench. Red arrows are the modern relative plate motion (DeMets *et al.* 2010) of the AUS plate with respect to the PAC. (c) Satellite free-air gravity for the Puysegur study region from the Sandwell *et al.* (2019) global marine gravity grid, v. 29.1. Labels, seismic lines, and plate motion vectors are the same as in (a) and (b). The grid of black dots are the locations of the gravity data points used in the inversion.

because it has subsided with respect to the rest of Puysegur Ridge by nearly 2 km (Collot *et al.* 1995). If composed of buoyant crust, this subsidence has implications for the subduction initiation process and the force balance on the system. In addition to addressing questions about these anomalies, gravity modelling can help stitch together the information obtained seismically to provide a more complete 3-D picture of the structures and rock types in the region.

4.2 Prior geophysical constraints

Prior constraints on a gravity inversion can come from a number of geophysical data, including seismic, bathymetric, borehole data and more. For the investigation of the Puysegur subduction system, we utilize bathymetric and seismic data collected from the SISIE marine geophysical expedition (Gurnis *et al.* 2019; Shuck *et al.* submitted), as well as sediment thickness estimates from the NOAA sediment thickness database (Straume *et al.* 2019), to constrain the gravity inversion. These data include the regional NIWA bathymetry grid (Mitchell *et al.* 2012), horizons picked from seismic reflection profiles, and seismic velocity models constructed from ocean bottom seismometer (OBS) seismic refraction analysis (Gurnis *et al.* 2019; Shuck *et al.* submitted). The NIWA grid is based only on shiptrack multibeam data and not calculated from the gravity like the global bathymetry data sets. This ensures the prior remains independent of the gravity data.

The seismic velocity models were constructed using a tomographic inversion of marine seismic refraction data gathered during the SISIE cruise. A total of 43 short-period OBSs were deployed on two east–west transects across the Puysegur Trench (Fig. 3a). The wide-angle seismic data records show reflected and refracted arrivals that help constrain the seismic velocities, depth to basement and Moho of both the Australian and Pacific Plates. We correlated arrival times between neighboring stations to identify refracted and reflected phases and checked the reciprocity on opposite source–receiver pairs. The average maximum source–receiver offset at which we observed seismic refractions was 80 km. We assigned traveltime uncertainties between 40 and 150 ms to account for noise on wide-angle data. We applied a regularized tomographic inversion of the wide-angle traveltimes to image the seismic velocities of the crust and uppermost mantle along the two transects (Van Avendonk *et al.* 2004). The resulting seismic velocity models for SISIE-1 and SISIE-2 have an rms data misfit of 90 and 80 ms, respectively, which is comparable to the assigned data errors.

The 2-D seismic velocity images along SISIE-1 and SISIE-2 show the nature of the oceanic crust of the incoming AUS Plate and the crustal structure of the overriding Puysegur Ridge and Solander Basin (Fig. 4). In the deeper parts of the basin, the top of basement was constrained by wide-angle seismic refractions. However, we determined the depth to basement from multichannel seismic reflection images (Shuck *et al.* submitted) where the sediment cover was thin. We were able to determine the Moho depth outboard of the trench on the AUS Plate and beneath the Solander Basin. However, the thickness of Puysegur Ridge is not well resolved from the OBS refraction data alone because the observed wide-angle refractions did not turn to such depths near the plate boundary. The gravity model can thus help constrain the thickness of the crust at the ridge. Nevertheless, Moho depths as determined from the seismic velocity models (deeper bold gray line in Fig. 4) are included in our prior. Like the other horizons, the Moho is not a 'hard' constraint but rather a probabilistic constraint on where the top of mantle is most likely to be. This flexibility is a reflection of the fact that there is uncertainty in the seismic data, and the Bayesian method means we do not have to take it completely at face value. The seismic velocities along SISIE-1 and SISIE-2 (Fig. 4) confirm that relatively thin oceanic crust of the AUS Plate has higher seismic velocities than the rifted continental crust of the PAC Plate (Gurnis *et al.* 2019). Consequently, there should be a substantial density contrast across the margin that should be evident in the gravity. The composition of Puysegur Ridge appears predominantly continental as well, though this is questionable and a point of interest for the gravity inversion.

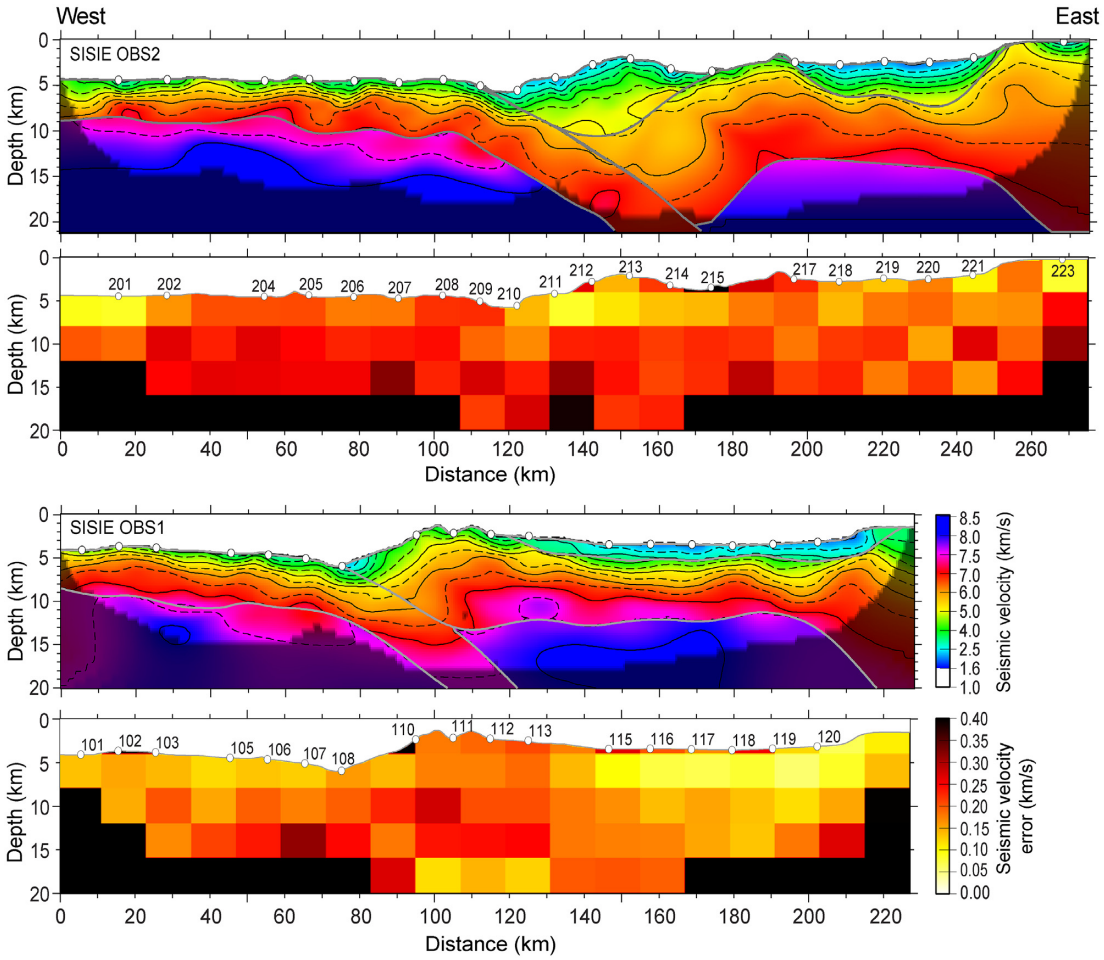

Figure 4. Seismic velocity models used in the prior. (a, c) Seismic velocity profiles for OBS lines SISIE-2 and SISIE-1, respectively. Grey lines are the sediment-basement contact and Moho interpretations. White dots are the locations of OBS. Dark shaded area is where the model has unreliable resolution. (b, d) Standard deviation of the seismic velocities based on seismic ray tracing for OBS lines SISIE-2 and SISIE-1, respectively. Black regions indicate areas with unreliable velocities. White dots are the locations of OBS and their corresponding numbers.

Based on the seismic velocities, we can constrain the thickness of the incoming AUS Plate to be about 7 km, with isolated pockets of sediment, usually less than 500 m in thickness. Due to the spatial resolution of the gravity model, sediment on the AUS Plate usually does not appear in the model except in places where it is relatively thick. The seismic reflection profiles reveal that sediment thickness in the Solander Basin is as thick as 6 km in places, averaging about 2–3 km for the majority of the basin (Shuck *et al.* submitted). Between our seismic lines, we also constrain the top of basement using the NOAA global sediment thickness database (Straume *et al.* 2019). The Snares Zone on the northern end of Puysegur Ridge is filled with up to 1 km of sediment, and both the layering observed in the seismic reflection profiles and the low seismic velocities on the western half of Puysegur Ridge suggest it is composed of deformed sediments, more than 10 km in width and 3 km in depth. However, below about 5.5 km depth, the seismic reflection data are inconclusive as to whether the accretionary wedge consists of sedimentary rock or crystalline basement (Gurnis *et al.* 2019); the gravity inversion can shed light on the compositions of these rocks.

The decollement between the overriding and subducting plates is visible on the seismic reflection images from SISIE-1 and SISIE-2 (Gurnis *et al.* 2019; Shuck *et al.* submitted). This horizon is used to constrain the top of the slab in the prior. The vertical, strike-slip Puysegur fault that cuts through the middle of the Snares Zone also appears to be present in the seismic reflection profiles (Shuck *et al.* submitted). While this fault is not included in the prior information directly, its presence could explain potential density differences observed in the final model.

We invert the Sandwell *et al.* (2019) global 1 min marine gravity grid, v. 29.1, for the region within the black grid in Fig. 3(b), which for the Puysegur region has a standard error of about 1.7–2 mGal (Sandwell *et al.* 2013, 2019). We include horizons that represent the seafloor, the sediment-basement contact, and the interpreted Moho from the velocity models (Fig. 4). Seismic velocities along the profile lines were converted to density using the empirical Nafe-Drake equation (Ludwig *et al.* 1970; Brocher 2005). Those densities were then extrapolated from the 2-D SISIE transects to each model parameter using a 3-D interpolation scheme. Likewise, surfaces for the horizons are interpolated from the scattered data points of the 2-D seismic data and the basement as determined from sediment thickness. For certain regions of the model, the prisms that fall between certain surfaces are assigned a specific prior. For example, prisms that fall between the AUS basement and Moho are given a prior oceanic crustal density of 2900 kg m^{-3} and prisms below the interpreted seismic Moho on the PAC Plate are assigned a prior density of 3300 kg m^{-3}, as the estimated seismic velocity in the

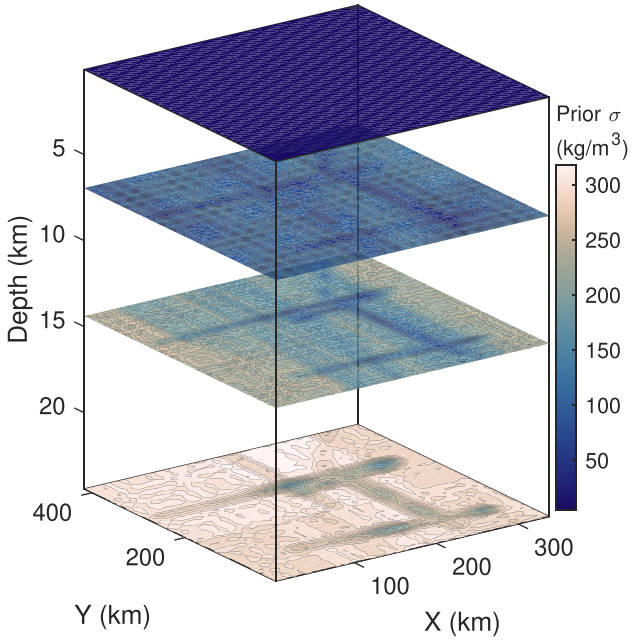

Figure 5. Variation of the standard deviation of the prior in 3-D space. Because the prior has a higher certainty along the trajectory of the seismic lines, the standard deviation is lowest along the survey lines, and increases exponentially away from the lines and their respective horizons and away from the seafloor. The ocean layer is essentially fixed by the bathymetry and thus has the lowest prior error.

models falls below an acceptable resolution below about 15 km depth. Otherwise, the prior densities used are those obtained directly from the velocity conversion.

For the prior, we have the highest degree of certainty on the densities of the prisms that lie along our seismic lines. We estimated the local standard deviation in the seismic velocity model with a forward ray tracing test. The uncertainty assigned to the model was the range in seismic velocity perturbations that would not raise the traveltime misfit more than 5 ms. These errors are for blocks of 10 km by 4 km. The lowest error is approximately 0.05 km s^{-1} and the highest is approximately 0.35 km s^{-1}. These standard deviations of the velocities are mapped into density using standard error propagation methods and the Nafe-Drake relationship. We 3-D interpolate these density errors to the locations of the prism centroids, which then serve as the starting values for the standard deviations on the prior. Certainty on the parameter values decreases from the initial value as we move away from the seismic lines, which we implement in the model by using a higher standard deviation farther from the lines, allowing the gravity to dominate the resulting density values in areas where we do not have seismic data. This is accomplished with a 3-D nearest neighbor algorithm that calculates the distance each prism centroid is from its closest data point. The standard deviation determined from propagation of error is then weighted via a smoothly varying functional—exponential decay of the increasing form—of nearest neighbor distance from the seismic and bathymetric data points. In this way, our prior includes both the error on the initial velocity model and the uncertainty due to spatial separation from our prior information. Horizontal slices of the spatially variable prior uncertainty mapped into 3-D space are shown in (Fig. 5). The Tikhonov regularization then ensures the model retains a smooth solution laterally, so values everywhere are to some degree constrained by those along the seismic lines. The degree to which the solution values are the result of the prior data or the gravity inversion itself can be visualized in the resolution matrix. The majority of the model domain is determined by gravity data and thus not overly biased by the prior information, except in places where we want it to be, such as in the ocean layer.

Ultimately, the final model has 64 000 model parameters with a horizontal resolution of about 9 km and a vertically increasing resolution of about 110–1130 m. We invert 92 953 gravity data points from the marine gravity grid, using a horizontal second order Tikhonov regularization coefficient of 5×10^6 and a vertical first order Tikhonov regularization coefficient of 10^0. We include priors on all parameters, though with spatially varying standard deviation as described above.

4.3 Results

We predict the gravity field from the final density model and compare it to the observed gravity, as well as the residual between the two (Fig. 6). The mean absolute error on the gravity produced from the final model is about 3.9 mGals, which is less than 2 per cent of the maximum anomaly in the study area (220 mGal). All the prominent features of the satellite gravity are well-recovered, including the prominent lows in the Snares Zone and the trench and the gravity high over the southern portion of Puysegur Ridge. Some of the finer features in the gravity are not fully recovered due to model resolution. The highest errors on the gravity, as shown by the residual, are mostly concentrated over areas with the largest gravity anomalies and where there is a sharp change in bathymetry, such as over the Puysegur Ridge and the edge of the Campbell Plateau. This is likely due to the trade-off between the regularization trying to smooth features laterally and the inversion trying to match these sharp changes in the gravity and bathymetry. Nevertheless, the highest error on the gravity is only 33.5 mGal, which on the scale of the major anomalies in the study area is still minor.

The model results for the 3-D density distribution are presented in representative cross-sections in Figs 7, 8 and 9, with the prior density distribution and posterior standard deviation plotted for comparison. The resolution and covariance matrices also illustrate the 3-D distribution of error in the posterior model (Fig. 10). The full resolution matrix exhibits a sharp diagonal with elements close to one, demonstrating that the model parameters are well resolved by the inversion. Looking at only the diagonal components, on the other hand, where each element of the diagonal represents the resolution of a particular model parameter as determined by the gravity, gives us a better sense of how the resolution varies throughout the model domain. As each element is associated with a particular parameter, we can map the diagonal of **R** into 3-D space (Fig. 10a). This 3-D resolution illustrates which parameters are resolved mostly by the gravity and which are not. The resolution is almost zero in the ocean layer because those parameters are determined entirely by the prior and thus are not resolved by the gravity. The resolution of parameters along the seismic lines is also lower because these parameters are weighted more by the prior. The resolution matrix shows an increase in the degree to which parameters are resolved by the gravity with depth.

However, barring the degree to which the parameters are determined by the prior, there is a fall off in the certainty of the solution with depth, as evident from the posterior covariance matrix, the square root of the diagonal of which is also mapped into 3-D space and visualized in Fig. 10(b). This shows the spatial distribution of the standard deviation of the posterior estimate of **m**. The mean

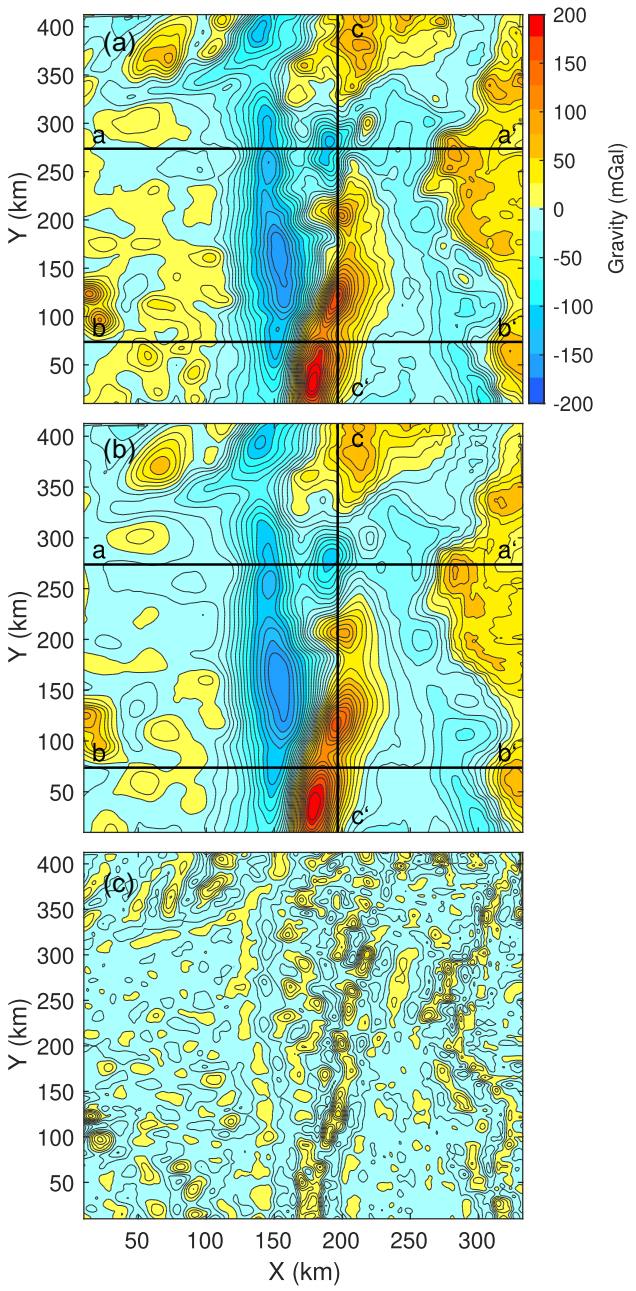

Figure 6. Gravity results for the final model. (a) Observed gravity field as extracted from the Sandwell *et al.* (2019) gravity grid, v 29.1. Black lines are the locations of cross-sections shown in Figs 7, 8 and 9. (b) Gravity predicted from the final density model determined by the inversion. (c) Residual gravity between the observed and predicted gravity grids, calculated as the absolute difference between each point on the grid.

standard deviation on the model parameters as determined from the diagonal of the covariance matrix for the entire model is 17 kg m^{-3}. There is a fall-off in accuracy with depth, ranging from about 10 to 15 kg m^{-3} in the shallow crust along the seismic lines to about 30 kg m^{-3} on average in the deepest layer. The maximum model parameter standard error is 68 kg m^{-3}, concentrated at the bottom and at the edges of the model, where there is less coverage by the gravity data and less constraint by the prior.

The most notable features of the final density model are the densities and structures of the Snares Zone and along Puysegur Ridge. The inversion requires a low density body beneath the central and eastern portion of the Snares Zone, extending to about 18–20 km depth (Figs 7 and 9) and is mostly consistent with the prior velocity models. However, the western half of Puysegur Ridge, below about 5.5 km depth, is consistently higher density than predicted from the velocity models. The southern cross section (Fig. 8), on the other hand, shows an elevated mantle beneath the Puysegur Ridge, more-so than suggested by the velocity prior. In all cases, we are mostly unable to resolve a slab structure, despite its presence in the prior.

To get a broad sense of the density and crustal variations within the final model, as well as how they compare to the prior and the seismic velocities, we look at the posterior densities of each prism versus their respective V_p values used to determine the prior (Fig. 11). The points are colored by the block of the model in which they reside, as determined by interpolating surfaces between the horizons on the seismic reflection lines from the SISIE survey. Based on these surfaces, prisms are either in the sediments (gray points), the AUS Plate crust (blue points), or the PAC Plate crust (burnt orange points); prisms within the mantle are not shown for clarity. There is scatter even in the prior data points because only the prisms lying along the seismic lines were converted directly with the Nafe-Drake equation; the other prism densities are then 3-D interpolated. The scatter is greatest within sedimentary units where rocks can vary over a relatively large range of densities and where there is substantial shallow structural complexity from the velocity models for the interpolation to accommodate. To more clearly illustrate the variation in structure along the ridge, we also determine the Moho depth from the density model, interpreted at the points where the density first exceeds 3200 kg m^{-3} (Fig. 12b). We also compute the crustal thickness (Fig. 12c) by subtracting the bathymetry (Fig. 12a) from the Moho. The crust is notably thicker beneath the Snares Zone, about 18 km thick, than it is beneath the southern part of Puysegur Ridge, where it is as thin as 7–8 km. The Moho shallows to around 10–12 km depth under the southern part of the Solander Basin and deepens to about 18 km in the northern part of the Basin, and even further to 23 km or greater beneath the Campbell Plateau.

5 DISCUSSION

The method of linear 3-D gravity inversion can be applied not only to simple, local scale structural geometries, but also complex density distributions across active plate margins. The Bayesian method allows for direct inclusion of existing geophysical data as priors and statistical feedback on the quality of the final model. Due to the non-uniqueness of gravity, which is clearly demonstrated by the relative insensitivity of the predicted gravity to changes in the prior (Figs 1 and 2), the final model is ultimately dependent on the prior and the strength of the regularization.

The synthetic tests demonstrate how the resulting models are often more sensitive to changes in regularization than they are the geophysical prior. The 3-D resolution matrix likewise shows how different parameters are determined more by the prior than by the gravity or vice versa. The Tikhonov regularization is a smoothness prior and goes into the definition of the resolution matrix (eq. 17), so when a parameter has a low resolution, the inversion is more strongly constrained by the existing geophysical information *and* the smoothness requirement than by the gravity. Differentiating the degree to which that parameter is determined by the geophysical prior versus the regularization is more difficult. Nevertheless, the majority of the model domain is resolved predominantly by the

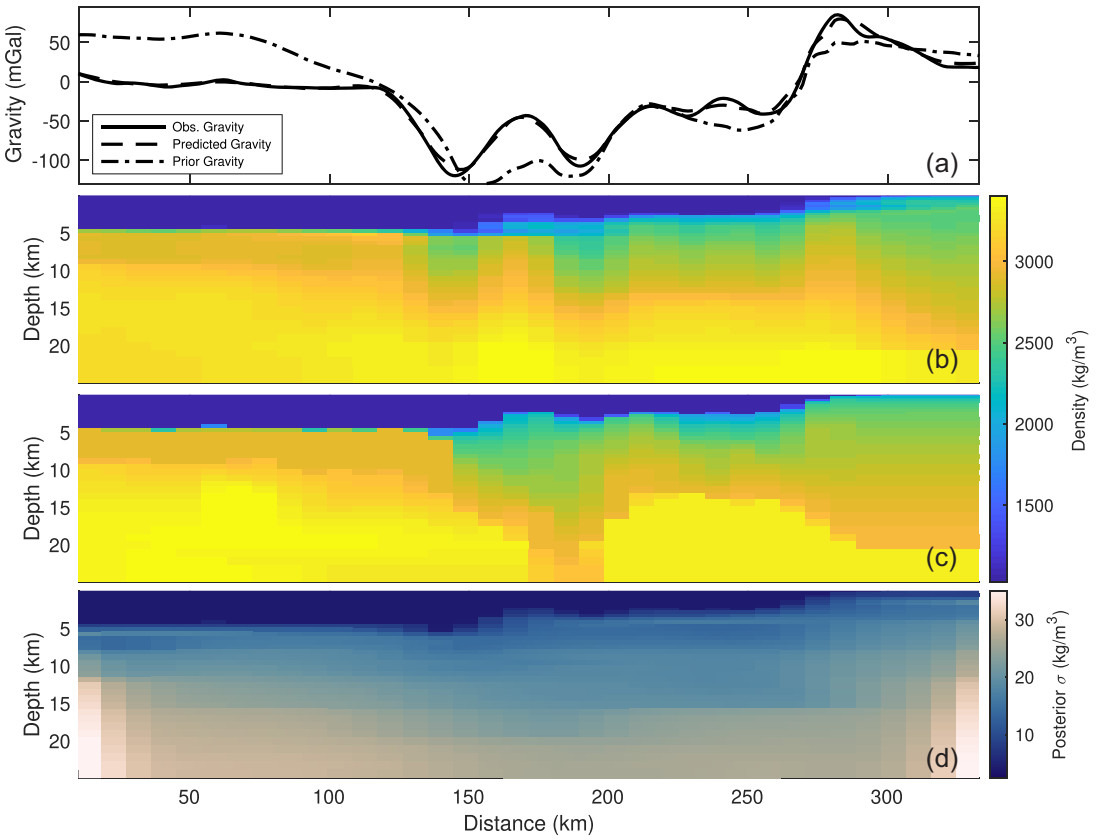

Figure 7. East–west cross section of the final 3-D density model for line a-a' in Fig. 6, roughly parallel to seismic line SISIE-2. (a) Gravity profiles for the density cross-section. Solid line is the observed gravity; dashed line is the predicted gravity from the final model slice shown in panel B; dashed–dotted line is the gravity from only the prior density model (panel c). (b) Predicted density distribution from the gravity inversion. (c) Prior density distribution used to constrain the gravity inversion. (d) Posterior standard deviation of the density of each prism shown in the cross-section, as determined from the posterior covariance matrix. Colourbar is saturated at 35 kg m^{-3}. Panels (b) and (d) together represent the posterior distribution for the model parameters shown in this cross-section.

gravity data. Only within the ocean layers and along the shallow portion of the seismic lines does the prior dominate the posterior solution, demonstrating that in the regions where we do not have seismic coverage, and to some extent in the regions where we do, we have learned something from the gravity.

Ultimately, the goal of obtaining a realistic density model from the inversion is to place constraints on the composition of key features and structures that control subduction and subduction initiation regionally. As the composition of Puysegur Ridge and the origin of the Snares Zone are key motivators for the gravity inversion and for understanding subduction initiation, these regions are highlighted in the comparison of the posterior densities to seismic velocities in Fig. 11. Prisms corresponding to the western and eastern halves of Puysegur Ridge at the Snares Zone are shown by pink and maroon points, respectively. The western half of the ridge plots in two distinct regions, a cluster lying predominantly between 2700 and 2900 kg m^{-3} and a cluster lying below 2100 kg m^{-3}, the latter of which corresponds to the sedimentary units within the shallow portion of the Snares Zone bathymetric depression and the accreted sedimentary portion of the western half of the ridge (Fig. 11, red-orange points), which is also clearly visible on the seismic reflection images from SISIE-2 (Shuck *et al.* submitted).

The difference between the western and eastern halves of Puysegur Ridge at the Snares Zone is notable, with the western half averaging around 2803 kg m^{-3} and the eastern half averaging around 2750 kg m^{-3}—the difference of which is more than three times as much as the mean standard deviation of the prisms within the Snares Zone, about 15.1 kg m^{-3} (Fig. 10). This is especially significant in light of the difference between the final density model and the prior. The prior densities for the Puysegur Ridge at the Snares Zone, particularly for the western half, average around 2500−2700 kg m^{-3} and are consistent with a continental crustal interpretation (Figs 7c and 11a). However, the gravity consistently requires the presence of a higher density body of around 2700−3100 kg m^{-3} on the western half of Puysegur Ridge in order to fit the observed gravity signal (Fig. 7). These densities, however, are not inconsistent with the velocity models because highly fractured or deformed rock can have a much lower seismic velocity while still maintaining a high density, so what the seismic velocity models seem to indicate is deformed sediment or continental crust, could in fact be fractured oceanic crustal rock (Barton 1986). We postulate that basement rock of the western half of Puysegur Ridge is compositionally distinct from that of the east and is most likely a sliver of oceanic crust that has been emplaced laterally against the continental crust of the eastern half via the strike slip motion of the Puysegur Fault, which runs through the Snares Zone. This inference is also consistent with the seismic interpretations in Shuck *et al.* (submitted).

The under-prediction of the densities on the western half of Puysegur Ridge by the seismic velocities via the Nafe-Drake curve and the large amount of scatter in the posterior densities relative to

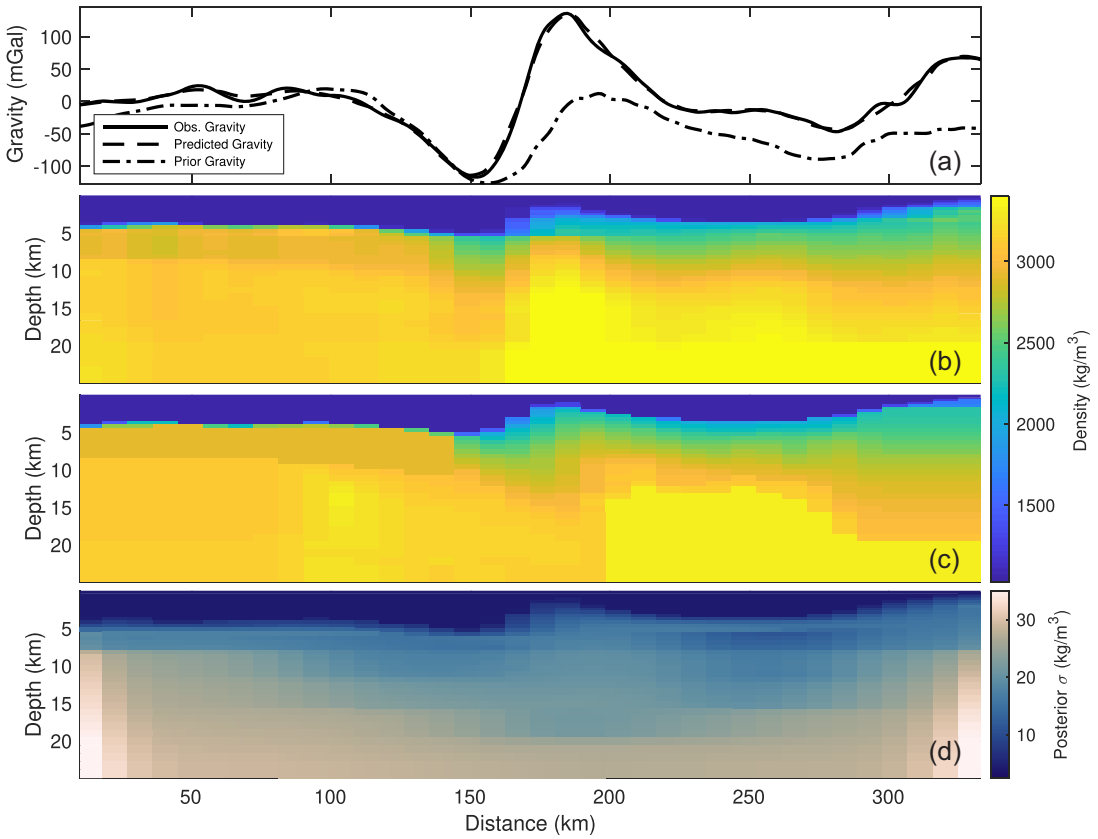

Figure 8. East-west cross section of the final 3-D density model for line b–b' in Fig. 6, roughly parallel to seismic line SISIE-1. (a) Gravity profiles for the density cross-section. Solid line is the observed gravity; dashed line is the predicted gravity from the final model slice shown in panel (b); dashed–dotted line is the gravity from only the prior density model (panel c). (b) Predicted density distribution from the gravity inversion. (c) Prior density distribution used to constrain the gravity inversion. (d) Posterior standard deviation of the density of each prism shown in the cross-section, as determined from the posterior covariance matrix. Colourbar is saturated at 35 kg m^{-3}. Panels (b) and (d) together represent the posterior distribution for the model parameters shown in this cross-section.

that curve put limitations on the degree to which the Nafe-Drake relationship can be used to predict densities without further information, as has been noted by previous authors (Barton 1986). The Nafe-Drake equation, though valid for velocities between 1.5 and 8.5 km s^{-1}, was based empirically on continental crustal data from California (Ludwig *et al.* 1970; Brocher 2005) and as such may not be accurate for oceanic crust. However, a comparison between the Nafe-Drake predictions of Brocher (2005) and theoretical seismic velocity and density predictions from mineral physics calculations using the MinVel Subduction Factory Toolbox (Abers & Hacker 2016; Sowers & Boyd 2019) reveal that differences between the two predictions are less than 1 per cent on average, though can be as high as 37 per cent for specific rock types (Sowers & Boyd 2019). There is also the question of whether thermal effects may impact the accuracy of the Nafe-Drake prediction and the model density estimates. However, an analysis of the possible perturbations to the velocity and density estimates of the Brocher (2005) relationship under a hot geotherm calculated using the MinVel toolbox, using the half-space cooling model with a plate age of 25 Ma for rocks in oceanic regimes and a typical continental geotherm with a conservatively high surface heat flux of 120 mW m^{-2} for continental regimes, demonstrate that elevated temperature has a negligible impact on the Nafe-Drake predictions relative to the range of densities in our model domain (Figs 11b and c). The rock compositions used in this analysis include basalt (Hacker *et al.* 2003), harzburgitic mantle (Hacker *et al.* 2003), Fiordland orthogneiss (Bradshaw 1990) and a combination of pelagic clays and biogenic ooze (Li & Schoonmaker 2003; Patel *et al.* 2020). The absolute densities estimated for each of these rock types differ insubstantially between low (surface) and warm (25 km depth) temperatures, and the velocity and density both change in accord with one another with that change in temperature, such that the predictive relationship between them remains the same (Fig. 11c, Sowers & Boyd 2019). Puysegur itself is also not a notably hot subduction zone. Despite the young age of the subduction front, the crust that is being subducted is not particularly young, spreading in the Tasman Sea having ceased around 53 Ma, though spreading in Emerald Basin south of the study area continued until around 10–20 Ma (Lebrun *et al.* 2003). Thus, we find it is not necessary to incorporate any thermal effect into our model and that the Nafe-Drake relationship is a reasonable one in light of any possible thermal perturbations and its performance relative to mineral physics estimates.

Some of the differences in density between the prior and posterior also likely arise from error in the 3-D interpolation scheme, but the difference in densities between the two even across the Snares Zone, where we have direct seismic data, suggests a significance in the under-prediction of many of the posterior densities by the Nafe-Drake equation. However, this does not invalidate its use as a prior, but rather highlights the advantage of using it in the context of a Bayesian approach. Rather than using seismic velocity as the

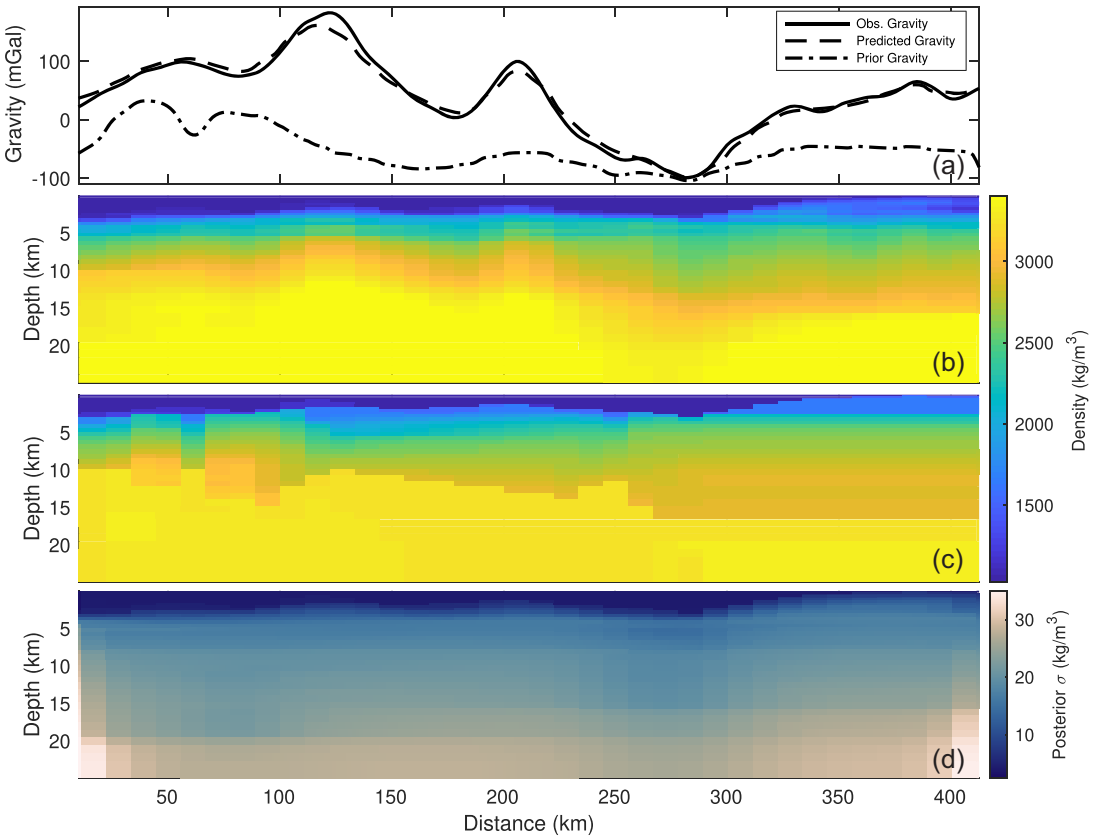

Figure 9. North–south cross-section of the final 3-D density model for line c–c' in Fig. 6, roughly parallel to the Puysegur Ridge. (a) Gravity profiles for the density cross-section. Solid line is the observed gravity; dashed line is the predicted gravity from the final model slice shown in panel (b); dashed–dotted line is the gravity from only the prior density model (panel c). (b) Predicted density distribution from the gravity inversion. (c) Prior density distribution used to constrain the gravity inversion. (d) Posterior standard deviation of the density of each prism shown in the cross-section, as determined from the posterior covariance matrix. Colourbar is saturated at 35 kg m^{-3}. Panels (b) and (d) together represent the posterior distribution for the model parameters shown in this cross-section.

only indication of a rock's density, we use it as a guide for the rock's possible density and weight that estimate of density accordingly. As such, the Bayesian approach allows for a more reasonable and flexible use of a common velocity–density relationship that otherwise, by itself, may be erroneous in its estimation of rock type.

For this reason, the gravity inversion is an invaluable supplement to our seismic study in estimating rock compositions and structure and in particular to spatially filling the gaps between where we have seismic information. Gravity at short wavelength strongly reflects topography (or bathymetry, Sandwell *et al.* 2014; Turcotte & Schubert 2014); however, if the bathymetry is fully constrained in the inversion and cannot by itself reproduce the gravity signal, then perturbations to the gravity must be coming from other sources—namely lateral density variations that may be governed by Moho geometry. As such, the shape of the interpreted Moho (Fig. 12b) strongly mirrors the gravity. Traditional gravity modelling techniques avoid this by removing the signal from the Moho/the isostatic effect and looking at the residual (Oldenburg 1974; Bai *et al.* 2014). However, this assumes constant densities in the respective layers and sometimes a fixed interface. Because we do not explicitly impose such assumptions with the Bayesian inverse approach, but rather constrain the 3-D densities and hence the structure probabilistically, the resulting Moho, though it does mirror the gravity, is likely a good approximation to the true Moho. Taking the southern line, SISIE-1, as an example, ultimately to match the gravity high over the ridge, there can be either (1) an elevated Moho or (2) anomalously high densities in the crust. In the absence of fixing either of these, the algorithm has no knowledge about which is the correct choice to fit the gravity, and the easiest way to fit the gravity is to create a density distribution increasing in depth with a shape mirroring that of the gravity. This is why inclusion of the Bayesian priors is so important. We can see the effect of the prior versus that of the gravity beneath the Campbell Plateau in Fig. 8(b), where there is smearing at the base of the crust relative to the prior in panel (c). The gravity in combination with the regularization wants to put the Moho higher to smoothly mirror the gravity signal. The prior, on the other hand, pulls the Moho down, but not so much so that the predicted gravity is depressed. As we can see in Fig. 8(a), the gravity from only the prior is too low to match the observations. This means that, given the inclusion of the prior, the combination of density and structure returned by the gravity inversion is probably the most reasonable estimate of the true structure. In other words, it is the most likely combination of (1) adjusting the Moho depth and (2) adjusting the density that can be obtained in light of our existing knowledge. It is the Bayesian approach that allows us to do this so effectively. It also means, given we have applied a strong prior along this transect, the fact that the gravity still pulls the Moho up under the Ridge despite the constraint is all the more significant and suggests this is not just an artefact of reflecting the shape of the gravity.

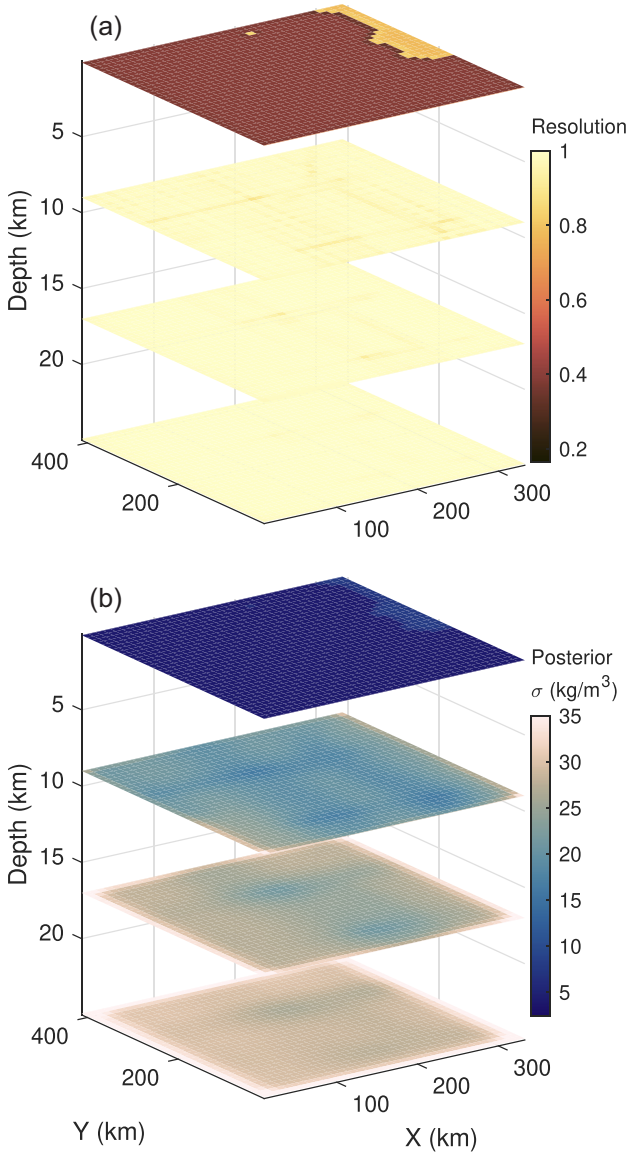

Figure 10. (a) Diagonal of the resolution matrix mapped into 3-D space. Slices are shown at depths of 1, 9, 17 and 25 km. The resolution represents the fraction of each model parameter value that is resolved by the gravity as opposed to the prior information. (b) Posterior standard deviation of the model parameters (square root of the diagonal of the covariance matrix) mapped into 3-D space. Slice depths are the same as in A. Colourbar is saturated at 35 kg m^{-3}.

This large gravity high over the southern portion of Puysegur Ridge cannot be explained solely by the bathymetry and requires a mass excess (Fig. 8). Similarly, the large gravity low over the Snares Zone also cannot be reproduced by the bathymetry alone, and hence requires a mass deficit to produce the observed gravity (Fig. 7). In other words, the density profiles and Moho and crustal thickness maps demonstrate there is relatively shallow mantle beneath the southern Puysegur Ridge and unusually thick crust beneath the Snares Zone; unusual in that the region is bathymetrically low, yet predominantly composed of buoyant continental crust, except for the very western side as previously discussed. The Solander Basin, which is composed of rifted continental margin crust, evidenced by both the seismic data (Gurnis *et al.* 2019; Shuck *et al.* submitted) and the densities, progressively thins to the south, where the basin

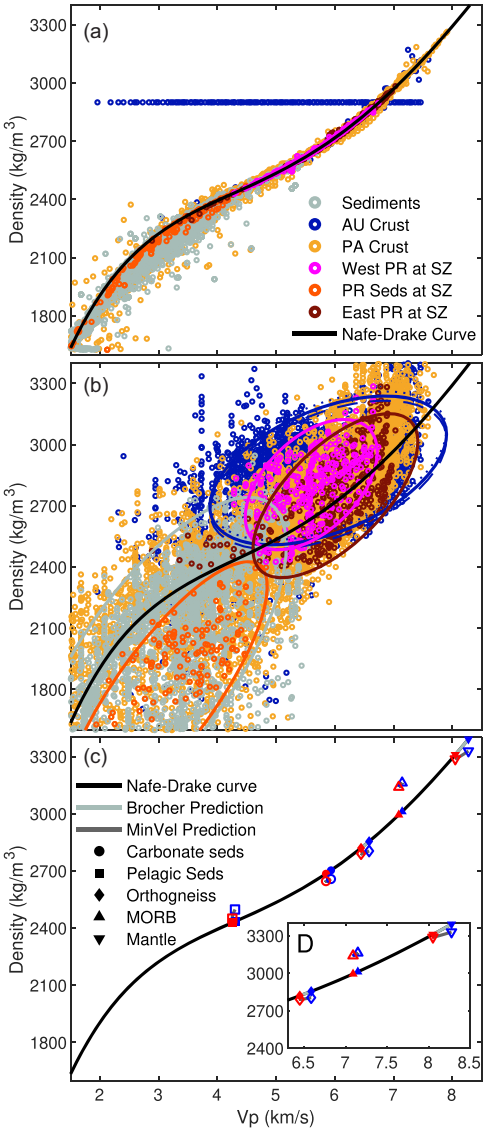

Figure 11. Density versus V_p relative to the Nafe-Drake equation (black line). (a) Density versus V_p for the prisms in the prior. Densities were calculated from V_p using the Nafe-Drake equation (black-line) along the seismic lines; the remaining prism centroid densities were 3-D interpolated from those points, producing the observed scatter. The prior for the oceanic crust was set to 2900 kg m^{-3} instead of using values directly from the equation (blue circles). (b) Posterior density from the inversion model versus V_p for each prism centroid. Prisms in the mantle have been omitted for clarity. Colours and their corresponding 2σ error ellipses represent different regions of the model as defined by the structural horizons in the prior. Dotted ellipse represents the shift in the density prediction resulting from low temperature conditions and dashed ellipse represents the shift due to high temperature conditions, as calculated from the MinVel predictions in panel (c). Similar ellipses can be computed for the other crustal blocks, but in all cases, the effect is negligible, so they have been omitted for clarity. Colours are as in panel (a). (c) Comparison of Brocher (2005) density predictions (filled symbols) to MinVel density predictions (open symbols) for low (surface) temperature conditions (blue symbols) and hotter (25 km depth) temperature conditions (red symbols) for characteristic rock types present in the model domain. Carbonate and pelagic sediment compositions are estimated from values in Li & Schoonmaker (2003) and Patel *et al.* (2020). Composition of Fiordland Orthogneiss, taken to represent regional continental crustal rock, is from Bradshaw (1990). Composition of MORB and harzburgitic mantle is from Hacker *et al.* (2003).

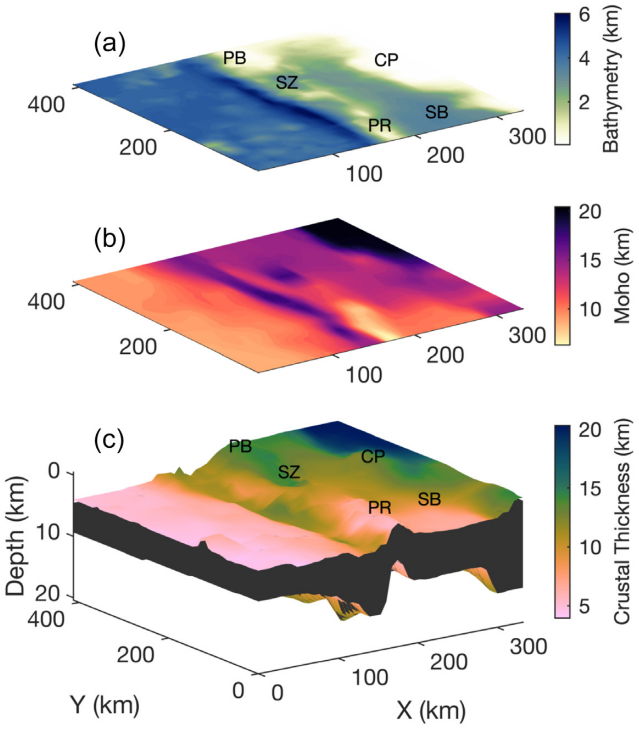

Figure 12. (a) Bathymetry for the Puysegur study area used in the computation of crustal thickness. PB, Puysegur Bank; SZ, Snares Zone; CP, Campbell Plateau; PR, Puysegur Ridge and SB, Solander Basin. (b) Moho depth interpreted from the 3-D density model at the points where the density first exceeds 3200 kg m^{-3}. (c) Crustal thickness for the Puysegur study area calculated by subtracting the bathymetry from the Moho depth and overlain on the bathymetric surface. The crustal volume is filled to the base of the crust using the Moho surface in panel (b). Text labels are as in panel (a).

experienced more extension during the rifting phase in the Eocene to Oligocene prior to the development of the strike-slip and subduction margin (Lebrun *et al.* 2003). Based on the crustal thickness results as estimated from the gravity, we estimate the continent–ocean transition in the southern Solander Basin to be around 50°S or even further south of the model domain, which is roughly consistent with Shuck *et al.* (submitted).

Another notable feature of the inversion results is the inability to resolve a slab structure, despite its presence in the prior and the seismically observable décollement between the two plates on seismic reflection data. The absence of descending crust in the final density model is likely due to the obliquity of subduction. A seismic Benioff zone extending to 150 km depth puts the slab northwards of the gravity study area, beneath Fiordland (Sutherland *et al.* 2006; Eberhart-Phillips & Reyners 2001). It is also possible that while the slab is present, it is not required to recover the local scale gravity signal, which is dominated by the bathymetry and shallow crustal structure.

6 CONCLUSIONS

The inversion technique presented inverts gravity data for 3-D density distributions within a Bayesian framework without the need for iteration and with the direct incorporation of prior geophysical constraints. Previous applications of linear gravity inversion, as opposed to the commonly used non-linear and wavenumber domain methods, have predominantly been for geometrically and structurally simpler density anomalies, though have also successfully been applied to crustal scale and tectonic studies. We have demonstrated this method can also be successfully applied to more geologically complex regions with significant lateral variations in density and structure by applying it to an active subduction zone.

The resulting density models provide a more complete picture of the subsurface, filling in the gaps between where there is seismic data and allowing us to estimate the Moho depth and crustal thickness. The crustal thickness and density models reveal the presence of buoyant, yet subsided, continental crust beneath the central and eastern portions of the Puysegur Ridge at the Snares Zone, whereas the western half of the ridge is most likely a sliver of oceanic crust. In contrast, an elevated mantle underlies the southern portion of Puysegur Ridge. The features observed in the Snares Zone and along the Ridge have implications for the structures and rock compositions that control subduction initiation and the changing state of stress during the initiation process, and they support the idea that the margin is transitioning to a state of self-sustaining subduction in the north. These results will allow us to make further calculations of the regional stress and effective topography that can be used to constrain geodynamic models that are the target of future research.

ACKNOWLEDGEMENTS

Supported by the National Science Foundation through awards OCE-1654766 (to Caltech) and OCE-1654689 (to UT Austin). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1745301. We thank IHS-Markit for a university educational license for Kingdom Software, provided to Caltech, for seismic data visualization and interpretation. We thank Kim Welford, Scott King and Sean Gulick for their helpful and insightful comments on the manuscript and Rupert Sutherland, Brandon Shuck and Joann Stock for helpful discussion. **Data Availability:** https://ngdc.noaa.gov/mgg/sedthick/Gravity data used in the inversion, Sandwell et al. (2019) global 1 min marine gravity grid, are publicly available from the Scripps Institute of Oceanography at https://topex.ucsd.edu/marine_grav/mar_grav.html. Bathymetric data used in the inversion (Mitchell et al. 2012) is publicly available from the New Zealand Institute for Water and Atmospheric Research (NIWA) at https://niwa.co.nz/our-science/oceans/bathymetry. Seismic data from the SISIE cruise, MGL1803, is available at the Lamont-Doherty Earth Observatory, Marine Geoscience Data System (MGDS), Academic Seismic Portal: http://www.marine-geo.org/collections/#!/collection/Seismic#summary. Sediment thickness data use to constrain the inversion (Straume et al. 2019) are publicly available from NOAA at https://ngdc.noaa.gov/mgg/sedthick/. The gravity inversion code will be available on CaltechDATA (https://data.caltech.edu/) and GitHub, and per communication with the corresponding author.

REFERENCES

Abers, G.A. &Hacker, 2016. A MATLAB toolbox and excel workbook for calculating the densities, seismic wave speeds, and major element composition of minerals and rocks at pressure and temperature, *Geochem. Geophys. Geosyst.,* **17**(11), 616–624.

Aster, R.C., Borchers, B. & Thurber, C.H., 2013. *Parameter Estimation and Inverse Problems,* 2nd edn, Elsevier Inc.

Bai, Y., Williams, S.E., Dietmar Müller, R., Liu, Z. & Hosseinpour, M., 2014. Mapping crustal thickness using marine gravity data: methods and uncertainties, *Geophysics,* **79**(2), G27–G36.

Barnoud, A., Coutant, O., Bouligand, C., Gunawan, H. & Deroussi, S., 2016. 3-D linear inversion of gravity data: method and application to Basse-Terre volcanic island, Guadeloupe, Lesser Antilles, *Geophys. J. Int.,* **205**(1), 562–574.

Barton, P.J., 1986. The relationship between seismic velocity and density in the continental crust—a useful constraint?, *Geophys. J. R. astr. Soc.,* **87**(1), 195–208.

Bear, G.W., Al-shukri, H.J. & Rudman, A.J., 1995. Linear inversion of gravity data for 3-D density distributions, *Geophysics,* **60**(5), 1354–1364.

Bradshaw, J.Y., 1990. Geology of crystalline rocks of Northern Fiordland: details of the granulite facies western fiordland orthogneiss and associated rock units, *New Zeal. J. Geol. Geophys.,* **33**(3), 465–484.

Brocher, T.M., 2005. Empirical relations between elastic wavespeeds and density in the Earth's crust, *Bull. seism. Soc. Am.,* **95**(6), 2081–2092.

Calvetti, D. & Somersalo, E., 2018. Inverse problems: from regularization to Bayesian inference, *Wiley Interdiscip. Rev. Comput. Stat.,* **10**(3), doi:10.1002/wics.1427.

Camacho, A.G., Fernández, J. & Gottsmann, J., 2011. A new gravity inversion method for multiple subhorizontal discontinuity interfaces and shallow basins, *J. geophys. Res.,* **116**(B02413), 1–13.

Chappell, A.R. & Kusznir, N.J., 2008. Three-dimensional gravity inversion for Moho depth at rifted continental margins incorporating a lithosphere thermal gravity anomaly correction, *Geophys. J. Int.,* **174,** 1–13.

Collot, J.-Y., Lamarche, G., Wood, R.A., Delteil, J., Sosson, M., Lebrun, J.-F. & Coffin, M.F., 1995. Morphostructure of an incipient subduction zone along a transform plate boundary: Puysegur ridge and trench, *Geology,* **23**(6), 519–522.

Cowie, L. & Kusznir, N., 2012. Mapping crustal thickness and oceanic lithosphere distribution in the Eastern Mediterranean using gravity inversion, *Petrol. Geosci.,* **18,** 373–380.

De La Varga, M. & Wellmann, J.F., 2016. Structural geologic modeling as an inference problem: a Bayesian perspective, *Interpretation,* **4**(3), SM1–SM16.

DeMets, C., Gordon, R.G. & Argus, D.F., 2010. Geologically current plate motions, *Geophys. J. Int.,* **181**(1), 1–80.

Eberhart-Phillips, D. & Reyners, M., 2001. A complex, young subduction zone imaged by three-dimensional seismic velocity, Fiordland , New Zealand, *Geophys. J. Int.,* **146,** 731–746.

Farquharson, C.G., 2008. Constructing piecewise-constant models in multi-dimensional minimum-structure inversions, *Geophysics,* **73**(1), K1–K9.

Geng, M., Welford, J.K., Farquharson, C.G. & Hu, X., 2019. Gravity modeling for crustal-scale models of rifted continental margins using a constrained 3D inversion method, *Geophysics,* **84**(4), G25–G39.

Gurnis, M., Hall, C. & Lavier, L., 2004. Evolving force balance during incipient subduction, *Geochem. Geophys. Geosyst.,* **5**(7), 1–31.

Gurnis, M., *et al.*, 2019. Incipient subduction at the contact with stretched continental crust: the Puysegur Trench, *Earth Planet. Sci. Lett.,* **520**(1), 212–219.

Hacker, B.R., Abers, G.A. & Peacock, S.M., 2003. Subduction factory 1. Theoretical mineralogy, densities, seismic wave speeds, and H_2O contents, *J. geophys. Res.,* **108**(B1), 1–26.

Krien, Y. & Fleitout, L., 2008. Gravity above subduction zones and forces controlling plate motions, *J. geophys. Res.,* **113**(B09407), 1–20.

Last, B.J. & Kubik, K., 1983. Compact gravity inversion., *Geophysics,* **48**(6), 713–721.

Lebrun, J.-F., Lamarche, G. & Collot, J.-Y., 2003. Subduction initiation at a strike-slip plate boundary: the Cenozoic Pacific-Australian plate boundary , south of New Zealand, *J. geophys. Res.,* **108**(B9 2453), 1–18.

Li, Y. & Oldenburg, D.W., 1998. 3-D inversion of gravity data, *Geophysics,* **63**(1), 109–119.

Li, Y.H. & Schoonmaker, J.E., 2003. Chemical composition and mineralogy of marine sediments, *Treat. Geochem.,* **7–9,** 1–35.

Ludwig, W., Nafe, J. & Drake, C., 1970. Seismic refraction, in *Sea,* Vol. 4, pp. 53–84, Wiley-Interscience.

Medeiros, W.E. & Silva, J. B.C., 1996. Geophysical inversion using approximate equality constraints, *Geophysics,* **61**(6), 1678–1688.

Mitchell, J., Mackay, K., Neil, H., Mackay, E., Pallentin, A. & Notman, P., 2012. Undersea New Zealand, 1:5,000,000.

Oldenburg, D.W., 1974. The inversion and interpretation of gravity anomalies, *Geophysics,* **39**(4), 526–536.

Parker, R.L., 1972. The rapid calculation of potential anomalies, *Geophys. J. R. astr. Soc.,* **31,** 447–455.

Parker, R.L., 1995. Improved Fourier terrain correction, Part I, *Geophysics,* **60**(4), 1007–1017.

Patel, J., Sutherland, R., Gurnis, M., Van Avendonk, H., Gulick, S.P., Shuck, B., Stock, J. & Hightower, E., 2020. Stratigraphic architecture of Solander Basin records Southern Ocean currents and subduction initiation beneath southwest New Zealand, *Basin Res.,* 1–24, doi:10.1111/bre.12473.

Portniaguine, O. & Zhdanov, M.S., 1999. Focusing geophysical inversion images, *Geophysics,* **64**(3), 874–887.

Prutkin, I. & Casten, U., 2009. Efficient gravity data inversion for 3D topography of a contact surface with application to the Hellenic subduction zone, *Comput. Geosci.,* **35**(2), 225–233.

Sandwell, D., Garcia, E., Soofi, K., Wessel, P., Chandler, M. & Smith, W.H., 2013. Toward 1-mGal accuracy in global marine gravity from CryoSat-2, Envisat, and Jason-1, *Leading Edge,* **32**(8), 892–899.

Sandwell, D.T., Müller, R.D., Smith, W.H., Garcia, E. & Francis, R., 2014. New global marine gravity model from CryoSat-2 and Jason-1 reveals buried tectonic structure, *Science,* **346**(6205), 65–67.

Sandwell, D.T., Harper, H., Tozer, B. & Smith, W.H., 2019. Gravity field recovery from geodetic altimeter missions, *Adv. Space Res.,* doi:10.1016/j.asr.2019.09.011, in press

Shuck, B., *et al.*, submitted. Strike-slip enables subduction initiation beneath a failed rift: new seismic constraints from Puysegur margin, New Zealand, *Tectonics,* doi:10.1002/essoar.10503735.1

Silva, J., Medeiros, W.E. & Barbosa, V., 2001. Potential-field inversion: choosing the appropriate technique to solve a geologic problem, *Geophysics,* **66**(2), 511–520.

Silva Dias, F.J.S., Barbosa, V.C.F. & Silva, J.B.C., 2009. 3D gravity inversion through an adaptive-learning procedure, *Geophysics,* **74**(3), I9–I21.

Sowers, T. & Boyd, O., 2019. Petrologic and mineral physics database for use with the U.S. Geological Survey National Crustal Model, USGS Open-File Rep. 2019-1035.

Straume, E.O. *et al.*, 2019. GlobSed: updated total sediment thickness in the World's Oceans, *Geochem. Geophys. Geosyst.,* **20**(4), 1756–1772.

Sutherland, R., 1995. The Australia-Pacific boundary and Cenozoic plate motions in the SW Pacific: some constraints from Geosat data, *Tectonics,* **14**(4), 819–831.

Sutherland, R., Barnes, P. & Uruski, C., 2006. Miocene-recent deformation, surface elevation, and volcanic intrusion of the overriding plate during subduction initiation, offshore southern Fiordland, Puysegur margin, southwest New Zealand, *New Zeal. J. Geol. Geophys.,* **49**(1), 131–149.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* Society for Industrial and Applied Mathematics.

Toth, J. & Gurnis, M., 1998. Dynamics of subduction initiation at preexisting fault zones, *J. geophys. Res.,* **103**(B8), 18 053–18 067.

Turcotte, D. & Schubert, G., 2014. *Geodynamics,* 3rd edn, Cambridge Univ. Press.

Van Avendonk, H.J., Shillington, D.J., Holbrook, W.S. & Hornbach, M.J., 2004. Inferring crustal structure in the Aleutian island arc from a sparse wide-angle seismic data set, *Geochem. Geophys. Geosyst.,* **5**(8), doi:10.1029/2003GC000664.

Welford, J.K., Shannon, P.M., O'Reilly, B.M. & Hall, J., 2010. Lithospheric density variations and Moho structure of the Irish Atlantic continental margin from constrained 3-D gravity inversion, *Geophys. J. Int.,* **183**(1), 79–95.

Welford, J.K., Peace, A.L., Geng, M., Dehler, S.A. & Dickie, K., 2018. Crustal structure of Baffin Bay from constrained three-dimensional gravity inversion and deformable plate tectonic models, *Geophys. J. Int.,* **214**(2), 1281–1300.

Wellmann, J.F., De La Varga, M., Murdie, R.E., Gessner, K. & Jessell, M., 2018. Uncertainty estimation for a geological model of the Sandstone greenstone belt, Western Australia—sights from integrated geological and geophysical inversion in a Bayesian inference framework, *Geol. Soc., Lond., Spec. Publ.,* **453,** 41–56.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Figure S1 Mean absolute error between the gravity from the true density model and that predicted by the inversion for each combination of α and ζ which are labeled for every other value. Panel rows represent either first or second order Tikhonov regularization or a combination of the two. Panel columns represent, from left to right, inversion with no priors, inversion with priors only on prisms that fall within the ocean, inversion with priors on prisms in the ocean and crustal rocks, and inversion with priors on all prisms, including the mantle. Red circles mark the α, ζ combination corresponding to the minimum MAE on the gravity; red squares mark the α, ζ combination corresponding to the minimum MAE on the model parameters relative to the true model. Colorbar is saturated at 25 mGal. Gray regions correspond to α, ζ combinations that yield unstable or unreasonable results.

Figure S2 Mean absolute error between the predicted model parameter values and the known model parameter values from the synthetic model for each combination of α and ζ which are labeled for every other value. Panel rows represent either first or second order Tikhonov regularization or a combination of the two. Panel columns represent, from left to right, inversion with no priors, inversion with priors only on prisms that fall within the ocean, inversion with priors on prisms in the ocean and crustal rocks, and inversion with priors on all prisms, including the mantle. Red circles mark the α, ζ combination corresponding to the minimum MAE on the gravity; red squares mark the α, ζ combination corresponding to the minimum MAE on the model parameters relative to the true model. Colorbar is saturated at 800 kg m^{-3}. Grey regions correspond to α, ζ combinations that yield unstable or unreasonable results.

Figure S3 Mean standard deviation on the model parameters as determined from the diagonal of the covariance matrix **C** for each combination of α and ζ which are labeled for every other value. Panel rows represent either first or second order Tikhonov regularization or a combination of the two. Panel columns represent, from left to right, inversion with no priors, inversion with priors only on prisms that fall within the ocean, inversion with priors on prisms in the ocean and crustal rocks, and inversion with priors on all prisms, including the mantle. Red circles mark the α, ζ combination corresponding to the minimum MAE on the gravity; red squares mark the α, ζ combination corresponding to the minimum MAE on the model parameters relative to the true model. Colorbar is saturated at 800 kg m^{-3}. Gray regions correspond to α, ζ combinations that yield unstable or unreasonable results.

Figure S4 Mean resolution of the model parameters as determined from the diagonal of the resolution matrix **R** for each combination of α and ζ which are labeled for every other value. Panel rows represent either first or second order Tikhonov regularization or a combination of the two. Panel columns represent, from left to right, inversion with no priors, inversion with priors only on prisms that fall within the ocean, inversion with priors on prisms in the ocean and crustal rocks, and inversion with priors on all prisms, including the mantle. Red circles mark the α, ζ combination corresponding to the minimum MAE on the gravity; red squares mark the α, ζ combination corresponding to the minimum MAE on the model parameters relative to the true model. Grey regions correspond to α, ζ combinations that yield unstable or unreasonable results. Lower resolution means that model parameters are determined more by the prior than they are the gravity data itself. Resolution values of 1 or near 1 mean model parameter values are resolved more by the gravity data than the prior.

Figure S5 Representative cross section in the x-direction of the 3-D inversion results for the α and α combinations that produced the minimum MAE on the model parameters for each of the regularization order and prior combinations, as determined from the test results depicted in Figs S1–S4. Row 1: gravity profiles for each of the three cases depicted in the panels below. Dark blue line: true gravity produced by the synthetic model, with noise; gray line: gravity from inversion using only first order Tikhonov; light blue line: gravity from inversion using only second order Tikhonov; orange line: gravity from inversion using second order Tikhonov in the horizontal and first order in the vertical. Row 2: cross-sections of the density model recovered from using only first order Tikhonov for the cases of no priors, priors only on the ocean water parameters, priors on the ocean and crustal parameters, and priors on all parameters, each with their respective minimum model parameter MAE α, ζ combinations. Row 3: cross-sections of the density model recovered from using only second order Tikhonov for each of the different prior cases. Row 4: cross-sections of the density model recovered from using a combination of first and second order Tikhonov for each of the different prior cases. Row 5: cross-section of true synthetic density model for comparison.

Figure S6 Representative cross-section in the y-direction of the inversion results for α and α combinations that produced the minimum MAE on the model parameters for each of the regularization order and prior combinations, as determined by comparing the test results depicted in Figs S1–S4. Row 1: gravity profiles for each of the three cases depicted in the panels below. Dark blue line: true gravity produced by the synthetic model, with noise; gray line: gravity from inversion using only first order Tikhonov; light blue line: gravity from inversion using only second order Tikhonov; orange line: gravity from inversion using second order Tikhonov in the horizontal and first order in the vertical. Row 2: cross-sections of the density model recovered from using only first order Tikhonov for the cases of no priors, priors only on the ocean water parameters, priors on the ocean and crustal parameters, and priors on all parameters, each with their respective minimum model parameter MAE α, ζ combinations. Row 3: cross-sections of the density model recovered from using only second order Tikhonov for each of the different prior cases. Row 4: cross-sections of the density model recovered from using a combination of first and second order Tikhonov for each of the different prior cases. Row 5: cross-section of true synthetic density model for comparison.

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

APPENDIX A: TIKHONOV REGULARIZATION

Tikhonov regularization is implemented using different regularization matrices for each of the x, y and z directions. For first order Tikhonov regularization and the 1-D case, the finite difference approximation to the first derivative is

$$\frac{\partial m_k}{\partial x} = \frac{1}{\Delta x}(-m_k + m_{k+1}), \tag{A1}$$

which can be represented in the form of an upper bidiagonal matrix operator **L1** acting on a vector of the spatially discretized model

parameters. The result is an $M-1 \times M$ matrix.

$$\frac{\partial m_k}{\partial x} = \frac{1}{\Delta x}\begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{bmatrix} \tag{A2}$$

Because the discretization can vary within the x, y and z directions, the **L** matrices are unique for each of those directions and Δx, Δy or Δz may vary for each adjacent pair of model parameters being regularized. When this is the case, the $1/\Delta x$ term is brought inside **L**.

For second order Tikhonov regularization and for the 1-D case, the finite difference approximation to the second derivative is

$$\frac{\partial^2 m_k}{\partial x^2} = \frac{1}{\Delta x^2}(m_{k-1} - 2m_k + m_{k+1}) \tag{A3}$$

which can likewise be represented in the form of an upper tridiagonal matrix operator **L2** acting on a vector of the model parameters, where **L2** is an $M-2 \times M$ matrix.

$$\frac{\partial^2 m_k}{\partial x^2} = \frac{1}{\Delta x^2}\begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{bmatrix}\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{bmatrix} \tag{A4}$$

The second derivative finite difference operator can be written in terms of the first derivative finite difference approximation as

$$\begin{aligned} \frac{\partial^2 m_k}{\partial x^2} &= \frac{1}{\delta x_j}\left(\frac{\partial m_{k+1}}{\partial x_{i+1}} - \frac{\partial m_k}{\partial x_i}\right) \\ &= \frac{1}{\delta x_j}\left(\frac{-m_k + m_{k+1}}{\Delta x_{i+1}} - \left(\frac{-m_{k-1} + m_k}{\Delta x_i}\right)\right). \end{aligned} \tag{A5}$$

The **L2** matrix is thus calculated from the **L1** matrix for the x and y directions. For the z-direction, we use only first order Tikhonov regularization.