submitted to *Geophys. J. Int.*

Robust Bayesian moment tensor inversion with optimal transport misfits: layered medium approximations to the 3D SEG-EAGE overthrust velocity model

Andrea Scarinci[1,2], Umair bin Waheed[3], Chen Gu[2,4], Xiang Ren[4], Ben Mansour Dia[5] Sanlinn Kaka[3], Michael Fehler[2], Youssef Marzouk[1,2]

[1]*Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA* 02139*, USA. Email: scarinci@mit.edu*

[2]*Earth Resources Laboratory, Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA* 02139*, USA.*

[3]*Department of Geosciences, College of Petroleum Engineering and Geosciences, King Fahd University of Petroleum and Minerals, Dhahran,* 31261*, Kingdom of Saudi Arabia.*

[4]*Department of Civil Engineering, Tsinghua University, Beijing,* 100084*, China.*

[5]*Center for Integrative Petroleum Research, College of Petroleum Engineering and Geosciences, King Fahd University of Petroleum and Minerals, Dhahran,* 31261*, Kingdom of Saudi Arabia.*

SUMMARY

A velocity model is generally an imperfect representation of the subsurface, which cannot precisely account for the three-dimensional inhomogeneities of Earth structure. We present a Bayesian moment tensor inversion framework for applications where reliable, tomography-based, velocity model reconstructions are not available. In particular, synthetic data generated using a three-dimensional model (SEG-EAGE Overthrust) are inverted using a layered medium model. We use a likelihood function derived from an

optimal transport distance—specifically, the transport-Lagrangian distance introduced by Thorpe et al. (2017)—and show that this formulation yields inferences that are robust to misspecification of the velocity model. We establish several quantitative metrics to evaluate the performance of the proposed Bayesian framework, comparing it to Bayesian inversion with a standard Gaussian likelihood. We also show that the non-double-couple component of the recovered mechanisms drastically diminishes when the impact of velocity model misspecification is mitigated.

Key words: Waveform inversion – Statistical methods – Inverse theory – Probability distributions – Earthquake source observations – Induced seismicity.

1 INTRODUCTION

Velocity model error or misspecification is a determining factor in the quality of many seismic inverse problem solutions. Due to the difficulties of characterizing the subsurface medium (e.g., different rock types, three-dimensional spatial heterogeneities), velocity models used in practice are generally approximate and inaccurate. Mischaracterization of the velocity, however, can impact one's ability to infer other quantities of interest, such as the hypocenter and the moment tensor (focal mechanism) of a seismic event. In deterministic full waveform inversion (FWI), misspecification can exacerbate the well known phenomenon of cycle-skipping, which traps optimizers in local minima (Gauthier et al. 1986). In the Bayesian setting, model misspecification can lead to overconfidence in the posterior distribution, i.e., under-reporting of uncertainty (Gu et al. 2018; Kleijn et al. 2012). The most direct approach to mitigating the impact of model misspecification is to introduce better physical models (when feasible) or improved statistical discrepancy models (Kennedy & O'Hagan 2001). These approaches, however, typically increase computational cost and may compromise parameter identifiability. As an example, in moment tensor inversion, using a simple layered-medium model for the propagation velocity can be orders of magnitude less computationally expensive than using a fully three-dimensional elastic wave propagation model. Moreover, such sophisticated models are typically not available: on the one hand, data for learning the velocity jointly with the focal mechanism in such a three-dimensional setting may be difficult to collect; on the other hand, the duality between source location and velocity estimation can become a confounding factor when trying to estimate the focal mechanism itself (Gu et al. 2018).

Data-model misfit functions used in deterministic full waveform inversion are often based on L_p norms or variations thereof: the signal value observed at time index i is compared to that of the simu-

lated signal at the same time. L_2-norm based misfit functions are particularly common, as they match the statistical notion of additive Gaussian noise. Due to misspecification of the velocity models just described, however, it may often occur that portions of the observed signal are anticipated or delayed with respect to the model predictions. When this occurs, the use of L_p norms can have unintended consequences: two waveforms that look very similar in "shape" to the human observer may appear far apart under an L_p norm due to misalignment and warping of the time coordinates. These norms, in fact, ignore the inter-relationships between signal values at different times, and simply treat the observed and modeled signals as vectors of amplitude values. Many corrections and modifications to address this kind of problem have been proposed, including the pioneering work of Luo & Schuster (1991); Gee & Jordan (1992), who achieve better waveform arrival-time matching through sequential perturbations of the velocity model or the fitting of dedicated waveform functionals. In this paper, we investigate the benefits of using an alternative, optimal transport (OT)-based, misfit function to measure discrepancies between observed and model-predicted data. Recent literature has demonstrated the applicability of optimal transport distances to seismic imaging problems in a deterministic setting (Engquist & Froese 2014; Métivier et al. 2018, 2019; Yang et al. 2018). In this context, OT has been shown to produce drastic reductions in the non-convexity of the objective function, especially when compared to misfits based solely on L_p distances. Rigorous mathematical treatment (Engquist & Yang 2020) has in fact shown that the 1-D quadratic Wasserstein distance between probability density functions is convex with respect to dilation and translation. When applying this distance to signals, however, it is necessary to normalize and positivize the signals accordingly. These requirements introduce data transformations that are not typically justifiable within the physics of our problem. We, therefore, focus instead on a particular Wasserstein distance that does *not* require signal positivation and normalization, making it more suitable for seismic waves. This distance is referred to as the transport–Lagrangian (TL) distance (Thorpe et al. 2017; Thorpe & Slepcev 2017; Kolouri et al. 2016) and can be interpreted as the result of solving an optimal transport problem between the graphs of two functions, thus performing a signal matching that allows translation in both amplitude and time.

While the benefits of using this kind of distance have recently been explored in a number of deterministic inverse problems and applications (Thorpe et al. 2017), including seismology (Métivier et al. 2018, 2019; Pladys et al. 2021), in this paper we formulate and explore its integration within a fully *Bayesian statistical framework*, focusing on moment tensor inversion. Our emphasis is on the interaction of the TL distance (and posterior distributions derived therefrom) with *model misspecification*, rather than on issues of cycle skipping and convexification, which are less relevant here. In this setting, we will evaluate the extent to which the TL distance can operate as a natural "feature extractor" that disregards information in the data not relevant to inference of the quantity of interest

(i.e., the earthquake focal mechanism), thereby reducing the impact of model misspecification. We will demonstrate this behavior empirically, under realistic scenarios of velocity model misspecification and for a variety of earthquake focal mechanisms. To assess the quality of our Bayesian inversion results—i.e., of a given posterior distribution over the moment tensor—we will employ quantitative statistical criteria (e.g., proper scoring rules), additional physically-motivated statistics, and further multivariate assessments of the posterior distribution.

2 BACKGROUND

2.1 Velocity model uncertainty

Spatial inhomogeneities, and the difficulty of directly observing the Earth's structure, have induced seismologists to find alternative strategies to velocity model building. How to construct reliable velocity models is a longstanding issue in seismology, to which a definitive answer has yet to be provided (Yang & Ma 2019; Socco et al. 2010). A common approach that we will consider throughout this paper is to generate model waveforms using a layered medium model (e.g., Dreger & Woods (2002)). This model is often derived from well logs or from some other analysis of the subsurface, such as one derived from arrival-time tomography (Guilhem et al. 2014) or kinematic source representation (Sánchez-Reyes et al. 2018). This modeling process adds uncertainty to the results of any associated inverse problem. In general, looking at the *effects* of layered medium approximations to 3D velocity models is also at present an under-developed area of research.

Because the propagation velocity of seismic waves impacts the time at which the wave features reach a station, velocity modeling errors can translate into the type of misspecification outlined in the previous section. As an example, Figure 1 shows a pair of waveforms—specifically, displacements $u_i(\mathbf{x}, t)$ for some direction of displacement i and a fixed surface location $\mathbf{x}$—resulting from two different velocity models but the same seismic event. The orange waveform results from a three-dimensional velocity model, while the blue waveform results from a two-dimensional layered medium model designed to approximate the 3D model. (Details on how the layered model is constructed will be given in Section 4.1 below.) It is evident that some kind of warping occurs between the two traces, which are otherwise similar in "shape."

2.2 The forward model

Our forward model for moment tensor inversion begins with Green's functions $G_{i,j}(\mathbf{x}, t; \mathbf{V}, \mathbf{x}_{\text{src}})$ of the 3D elastic wave equation (Shearer 2009), where $\mathbf{x}_{\text{src}} \in \mathbb{R}^3$ is the source position, $\mathbf{V}$ represents a velocity model, and $\mathbf{x} \in \mathbb{R}^3$ and $t \in \mathbb{R}$ are the points in space and time, respectively, where the Green's

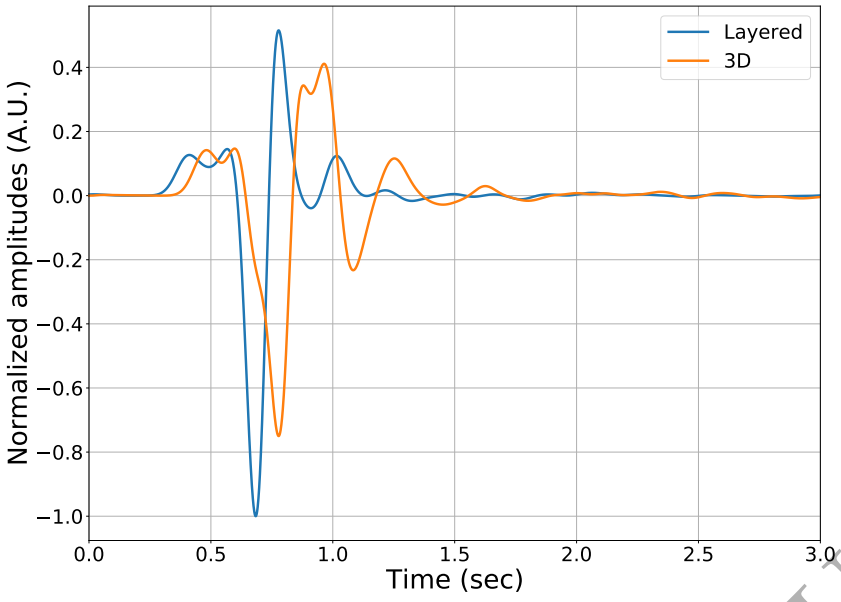

Figure 1. Sample waveforms coming from two different velocity models: the 3D SEG/EAGE overthrust model, and a 2D layered-medium velocity model based on the 3D model.

function is evaluated. Here the index $i \in \{1, 2, 3\}$ denotes the orientation of displacement associated with the Green's function, and $j \in \{1, 2, \ldots, 6\}$ corresponds to one of the six independent elements of the moment tensor $\mathbf{m} \in \mathbb{R}^6$ describing an impulse event. The waveforms predicted at a station location $\mathbf{x}_s$, for a source event at $\mathbf{x}_{\text{src}}$ with moment tensor $\mathbf{m} = \{m_j\}_{j=1}^6$, are therefore expressed as:

$$u_i(\mathbf{x}_s, t) = \sum_{j=1}^{6} G_{i,j}(\mathbf{x}_s, t; \mathbf{V}, \mathbf{x}_{\text{src}})\, m_j \tag{1}$$

In practice, we will discretize the waveforms u_i and the associated Green's functions $G_{i,j}$ at n time points $\mathbf{t} = \{t_k\}_{k=1}^n$; thus we will write, more compactly and for a given station location $\mathbf{x}_s$,

$$\mathbf{u}_i(\mathbf{m}; \mathbf{x}_s, \mathbf{x}_{\text{src}}) = \mathbf{G}_i(\mathbf{x}_s, \mathbf{V}, \mathbf{x}_{\text{src}}) \cdot \mathbf{m}^\top, \tag{2}$$

where $\mathbf{u}_i(\mathbf{m}; \mathbf{x}_s, \mathbf{x}_{\text{src}}) \in \mathbb{R}^n$ and $\mathbf{G}_i(\mathbf{x}_s, \mathbf{V}, \mathbf{x}_{\text{src}})$ is an $n \times 6$ matrix, for each orientation of displacement i. Below we may suppress the explicit dependence of $\mathbf{u}_i$ and $\mathbf{G}_i$ on $\mathbf{x}_s$, $\mathbf{x}_{\text{src}}$, and $\mathbf{V}$ unless it is specifically needed.

A typical statistical model then assumes that observed waveforms are perturbed by additive i.i.d. Gaussian noise at each time point:

$$\mathbf{y}_i = \mathbf{u}_i(\mathbf{m}) + \mathbf{e}_i, \quad \text{with } \mathbf{e}_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I}). \tag{3}$$

where $\mathbf{y} \in \mathbb{R}^n$ is again a discretized version of the underlying *noisy* continuous-time waveform. These assumptions lead to a Gaussian likelihood function, which, when paired with a conjugate Gaussian prior, yield a closed-form expression for the Gaussian posterior distribution on $\mathbf{m}$. In our setting, however, both the presence of misspecification and the choice of an alternative transport-based misfit measure call for a different formulation of the Bayesian update. We will discuss this methodological question in Section 3.

2.3 Methods for solving the seismic inverse problem

Moment tensor inversion with waveform data possesses many similarities to full waveform inversion. We therefore give a brief review of the relevant literature. The full waveform inversion problem has been traditionally and primarily addressed in a deterministic sense, and while most of the methods we discuss do not specifically aim to tackle velocity model error, the variety and nature of the proposed approaches convey the complexity of the task and are in a sense a symptom of model misspecification.

2.3.1 Optimization approaches

The most traditional approach to seismic inversion is non-probabilistic, and consists in recovering model parameters by optimizing a misfit function defined over the observed $\mathbf{y}$ and model-predicted $\mathbf{u}$ waveforms (Virieux & Operto 2009). Typically, misfit functions are chosen to be L_p norms, with the squared L_2 norm, as in a least squares problem, being the most popular choice (Odile et al. 1986). A number of techniques (Newton, truncated Newton (Métivier et al. 2013), Gauss-Newton, or gradient descent) are used to solve, under various assumptions, the optimization problem.

Alternative measures of misfit have also been proposed ((Luo & Schuster 1991; Gee & Jordan 1992)), such as the L_1 norm, secant and mixed L_1–L_2 norms (Huber penalty) (Crase et al. 1990), as well as misfit functions based on optimal transport (Métivier et al. 2016b; Chen & Peter 2018; Métivier et al. 2016a; Michael et al. 2013). Of a similar flavor are cross-correlation approaches (Van Leeuwen & Mulder 2010), together with deconvolution approaches (Luo & Sava 2011; Warner & Guasch 2016; Guasch et al. 2019). The aim of these alternative measures is to minimize the impact of phase/travel time differences, or relative phase shifts, while generally mitigating the well-known phenomenon of cycle skipping (Warner & Guasch 2014). With the same objective, it is also worth mentioning methods based on instantaneous phase differences and envelope ratios between observed and synthetic seismograms (Bozdağ et al. 2011; Rickers et al. 2012; Luo et al. 2018). No matter the choice of misfit function, building a reliable initial model for optimization remains a difficult task, especially so when this includes estimating the velocity $\mathbf{V}$. On the other hand, starting with a highly misspecified model inevitably leads to poor parameter estimates. In addition, deterministic inversion typically yields only

single-point estimates, which do not convey any information about the uncertainty that surrounds the final solution.

2.3.2 Bayesian formulations

An extensive and growing body of work has pursued full waveform inversion in a Bayesian framework (Bosch et al. 2010; Duijndam 1988; Gouveia & Scales 1998; Sen & Stoffa 1996; Ray et al. 2017; Gunning & Glinsky 2007; Zhu et al. 2016; Singh et al. 2018; Izzatullah et al. 2019; Gebraad et al. 2020). A main motivation behind these efforts is that a Bayesian framework offers, in principle, a more complete representation of one's state of knowledge about the inverse solution. This goal is particularly relevant in a misspecified setting, though as noted in previous literature, misspecification can cause Bayesian methods to become poorly "calibrated"—e.g., to under-report uncertainty Miller & Dunson (2018).

A central challenge of Bayesian seismic inversion is the computational cost of characterizing the posterior distribution and its marginals. For this reason, a large part of the literature on the topic has focused on strategies to reduce the computational burden. In Sen & Stoffa (1996); De Figueiredo et al. (2017) a number of sampling approaches have been considered, from Metropolis-Hastings Markov chain Monte Carlo (MCMC) to Gibbs sampling. An approach to mitigating the high dimensionality in certain seismic Bayesian inverse problems is the trans-dimensional Markov chain Monte Carlo sampler implemented in Ray et al. (2017). Simplifications of the forward problem to increase computational speed have also been proposed. In Arild & Henning (2003) the authors define a Bayesian framework that relies on a forward model that is linearized in the velocity $\mathbf{V}$. The error and prior distributions are chosen to be Gaussian, which allows the posterior distribution to be characterized analytically. Other Bayesian inversion attempts with linearized models can be found in Gunning & Glinsky (2007); Grana & Della Rossa (2010).

Gu et al. (2018) attempted a full Bayesian moment tensor inversion without any simplification of the forward model or linearization around specific solutions. To overcome the computational challenges and increase the robustness of the solution, a number of sampling strategies were implemented to exploit conditional linearities, and the associated conditional Gaussianities, in the problem. These include marginal-then-conditional sampling, pre-computing a library of velocity models and source locations, as well as *coarsening* as described in Miller & Dunson (2018). The results achieved through this implementation were satisfactory when the velocity model was known and set to a specific value. As soon as uncertainty was introduced in $\mathbf{V}$, the solutions of the problem exhibited a high degree of instability, indicating model misspecification issues.

2.4 Optimal transport distances

A resurgence of interest in formulations and numerical methodologies for *optimal transportation* has led to, in the past few years, many applications in the field of signal analysis. Optimal transport is attractive in the present seismological inverse problem, as it enables the richer kind of comparison of waveforms that we seek. Optimal transport (OT) is in general (cf. the Kantorovich problem) a way of finding a *coupling* of two probability measures that minimizes a certain total transportation cost (Villani 2009; Peyré et al. 2019). In the very specific case of discrete/empirical probability measures with equal numbers of equally weighted points in their support, the OT problem reduces to an *assignment* problem (Peyré et al. 2019) between the points in support of each distribution. The transportation cost is often taken to be the distance or the squared distance between these points; the associated minimum total cost, over all possible assignments, is then the 1-Wasserstein distance or the 2-Wasserstein distance, respectively.

A distinctive feature of Wasserstein distances versus dynamic time warping (Müller 2007) is that optimal transport does not ensure causality. This may seem a limitation in its application to waveform comparison because of the inherent sequential nature of time signals. When dealing with velocity model misspecification, however, this freedom can actually be beneficial—in that inconsistencies in velocity modeling can produce either anticipation *or* delay in the reproduction of some parts of the observed signal.

One way of relating the OT problem to the comparison of time-dependent signals is to treat the signals as univariate probability density functions. For the resulting OT problem to be well posed, however, it is necessary for these input signals to be normalized (i.e., integrating to one) and positive, as these conditions are necessarily satisfied by probability density functions. Yet waveforms are not densities—i.e., they do not in general sum/integrate to one and are not in general non-negative. A common workaround to this problem is to shift the signal along the ordinate axis to make it positive and then to divide it by its L_1 norm, i.e., the sum of all of its points (Yang et al. 2018; Métivier et al. 2016b; Thorpe et al. 2017). This adjustment also allows for a fast, analytical, computation of the Wasserstein distance via formulas that apply only in one dimension. Attempts at using the Wasserstein distance in this fashion have been made in the field of waveform inversion (Yang et al. 2018; Métivier et al. 2016b). Promising results were achieved in these works for velocity inversion. OT-based misfit functions have proven to be beneficial in terms of reducing cycle-skipping effects (Brossier et al. 2015; Warner et al. 2013). While computationally convenient, the transformation of the signals that is required is somewhat artificial and is not justified by the physics of the problem. In addition, the transformation can distort the signal, smoothing out amplitude versus frequency differences (Thorpe et al. 2017). In a general sense, any *a priori* transformation of the data introduces the possibility of

artifacts in the results of the inversion that can be hard to predict and quantify. For this reason, when applied to field data, these techniques may prove to be less reliable.

A different approach that avoids these pitfalls is to use the so-called transport–Lagrangian (TL) distance (Thorpe & Slepcev 2017), which is a specific instantiation of the Wasserstein distance adapted to signals. Consider two real-valued signals $y(t), u(t) : \mathbb{R} \to \mathbb{R}$ as the observed data and modeled waveforms. For simplicity, here we focus on the case where both signals have been discretized on n points $\mathbf{t} = (t_i)_{i=1}^n$. Let $\mathbf{y} = (y(t_i))_{i=1}^n$ and $\mathbf{u} = (u(t_j))_{i=1}^n$. Then the TL distance can be written as the solution of the following minimization problem:

$$\mathrm{TL}_p^\lambda(\mathbf{y}, \mathbf{u}) = \min_{\boldsymbol{\sigma} \in \mathrm{Perm(N)}} ||\mathbf{y}_{\boldsymbol{\sigma}} - \mathbf{u}||_p^p + \lambda ||\mathbf{t}_{\boldsymbol{\sigma}} - \mathbf{t}||_p^p \tag{4}$$

where $\boldsymbol{\sigma}$ indicates any permutation of the elements of the vector $\mathbf{y}$, i.e., $\mathbf{y}_{\boldsymbol{\sigma}} = [y_{\sigma(1)}, \ldots, y_{\sigma(i)}, \ldots, y_{\sigma(n)}]$. The first term of the above expression is simply an L_p norm of amplitude differences, but between $\mathbf{u}$ and a *transported* version of the signal $\mathbf{y}$. The term is meant to control the amount of across-time transport induced by the optimization over $\boldsymbol{\sigma}$. The parameter λ is a degree of freedom chosen to control the relative weights of the two terms of the objective function. We focus on $p = 2$, i.e., the TL_2^λ distance, which will allow for a direct comparison to the case where a classic L_2 misfit is used. The above formulation can also be interpreted as optimal transport between uniform measures on the *graphs* of $y(t)$ and $u(t)$; in discrete form, it is optimal transport between two equally-weighted empirical measures supported on $\mathbb{R}^2$, i.e., at points $\{(t_1, y(t_1)), (t_2, y(t_2)), \ldots, (t_n, y(t_n))\}$ and $\{(t_1, u(t_1)), (t_2, u(t_2)), \ldots, (t_n, u(t_n))\}$.

The TL distance operates directly on the (discretized) signals, and thus avoids unnatural data transformations while still allowing an OT formulation. Also, while computing the Wasserstein distance in general discrete settings amounts to solving a linear programming problem (with $O(\max(n, m)^3)$ complexity, n and m being the dimensions of the discretized signals), for the special case of $n = m$, one can adopt more specialized algorithms that solve an assignment problem. Our algorithm of choice for such problems is the auction algorithm (Bertsekas 1981), which exhibits a nearly quadratic average complexity of $O(n^2 \log n)$ for problems with $n < 1000$ (Schwartz 1994; Métivier et al. 2019). Finally, the choice of the parameter λ is of crucial importance for a successful use of this distance. In general, setting $\lambda \to \infty$ reverts the TL_2 distance to the L_2 norm, while sending $\lambda \to 0$ allows for rather large amounts of horizontal transport, almost neglecting amplitude matching—which is, for most applications, the most informative feature of the data. Empirically, we have found that a good choice for λ is to ensure that the scale of the time vector values ($\mathcal{A}$) and that of the amplitude values ($\mathcal{T}$) are somewhat comparable, i.e., $\lambda = \frac{\mathcal{A}}{\mathcal{T}}$. This is in accordance with related literature (Métivier et al. 2019).

A detailed analysis of the TL distance as an objective function in deterministic seismic inversion (including differentiability and efficient computation of its gradients) has been given by Métivier et al. (2019). Improvements in the convexity of the misfit have emerged as the primary effect of the choice of such a distance measure (Pladys et al. 2021).

3 A CONSISTENT BAYESIAN FRAMEWORK FOR OPTIMAL TRANSPORT DISTANCES

In this section, we answer the following question: how can we build a coherent Bayesian framework around the TL distance as a misfit statistic? One approach could be to seek a tractable expression for the likelihood function based on this statistic—i.e., for the probability density $p\left(\mathrm{TL}_p^{\lambda}(\mathbf{y},\mathbf{u})|\mathbf{m}\right)$— under the standard assumption of additive Gaussian noise on $\mathbf{y}$ (3). (In this setting, one could also set $p = 2$ to directly compare with the L_2 norm misfit.) There are two main impediments to this strategy. First, calculating the TL distance involves solving an optimal transport problem, which is itself a minimization problem. There is no closed-form expression for the solution of this problem. Evaluating the likelihood function—which reflects the *distribution* of the TL distance $\mathrm{TL}_2^{\lambda}(\mathbf{y},\mathbf{u})$ under random perturbations of $\mathbf{y}$— therefore becomes challenging, if not impossible. Second, as we have described earlier, our goal is to perform inference in a misspecified setting. Even if the likelihood expression above were tractable, it presumes that the data follow exactly from the statistical model (3). This assumption is violated under misspecification of the velocity, and therefore direct use of a likelihood based on the distribution of the TL statistic would introduce inconsistency.

3.1 Gibbs posteriors

We seek an alternative inference approach that allows both for more general misfit functions and for model misspecification. To this end, we adopt the framework introduced by Bissiri et al. (2016), which is "a coherent procedure for general Bayesian inference which is based on the updating of a prior belief distribution to a posterior when the parameter of interest is connected to observations via a loss function." The resulting posterior belief distributions are known as Gibbs posteriors. This framework relaxes a key requirement of parametric Bayesian inference: that a parametric family containing the true data-generating distribution be known. In the misspecified setting, this assumption is not satisfied; for example, without the correct velocity model $\mathbf{V}$, we cannot generate precisely the observed waveforms, or even the observed waveform up to an additive Gaussian noise perturbation.

A full motivation and derivation of the Gibbs posterior are given in Bissiri et al. (2016), but we recall here the essential features. Let $\boldsymbol{\theta} \in \Theta$ generically represent the parameters of interest. (For

moment tensor inversion, we have $\boldsymbol{\theta} \equiv \mathbf{m}$.) The central idea is to construct and *minimize* an objective function $\mathcal{L}$ over probability measures on Θ; this objective is defined by the observations $\mathbf{y}$, a given *loss function* $\ell(\boldsymbol{\theta}, \mathbf{y})$, and a prior probability measure p_0 on Θ, reflecting our prior beliefs. An appropriate objective turns out to be the sum of two terms, one reflecting an *expected loss* to the data $\mathbf{y}$ and the other reflecting Kullback–Leibler divergence (Cover (1999)) from the prior:

$$\mathcal{L}(p; \mathbf{y}, \ell, p_0) = \int \ell(\boldsymbol{\theta}, \mathbf{y}) p(d\boldsymbol{\theta}) + \mathcal{D}_{\mathrm{KL}}(p \| p_0). \tag{5}$$

Here p should be understood as a candidate probability measure over Θ. A coherent belief update is given by the minimizer of (5), which can be written in closed form:

$$p^* = \arg\min_p \mathcal{L}(p; \mathbf{y}, \ell, p_0) = \frac{\exp\left(-\ell(\boldsymbol{\theta}, \mathbf{y})\right)\, p_0(\boldsymbol{\theta})}{\int \exp\left(-\ell(\boldsymbol{\theta}, \mathbf{y})\right)\, p_0(d\boldsymbol{\theta})}. \tag{6}$$

This minimizer, by construction, balances prior information with fidelity to the observed data, where the latter is quantified through the loss function ℓ. It also has the form of a Bayesian update, where the denominator is a normalizing constant and the first term in the numerator (the exponentiated negative loss) takes the place of a likelihood function. If it is known that the data are generated from a given parametric family of distributions (e.g., $p(\mathbf{y}|\boldsymbol{\theta})$), then choosing the loss to be the negative log-likelihood, $\ell(\boldsymbol{\theta}, \mathbf{y}) = -\log p(\mathbf{y}|\boldsymbol{\theta})$, reverts (6) precisely to Bayes' rule. The classical Bayesian update is therefore a special case of this more general framework for updating belief distributions.

In our experiments, to use the Gibbs posterior framework with the TL misfit, we will choose ℓ to be the TL distance between the observed and modeled waveforms, for each orientation of displacement and each observation station. The total loss is then the sum of these TL distances, over all stations and all three orientations of displacement. More specifically, let $\mathbf{y}_i$ be the (discretized) observed waveform and $\mathbf{u}_i(\mathbf{m})$ be the (discretized) modeled waveform given a moment tensor $\mathbf{m}$, for a given station and displacement direction (jointly indexed by i). For this pair, we put $\ell(\mathbf{m}, \mathbf{y}_i) = \mathrm{TL}_2^{\lambda}(\mathbf{u}_i(\mathbf{m}), \mathbf{y}_i) - \log s$; summing the loss over all pairs yields

$$\exp\left(-\ell(\mathbf{m}, \mathbf{y})\right) = s^N \exp\left(-s \sum_{i=1}^{N} \mathrm{TL}_2^{\lambda}(\mathbf{u}_i(\mathbf{m}), \mathbf{y}_i)\right), \tag{7}$$

which is analogous to the expression suggested in Motamed & Appelo (2019), but using TL distances. Here, the parameter s acts as a scaling factor, and N is the number of waveform pairs (one pair for each station and orientation of displacement). The parameter s is analogous to an unknown variance hyperparameter in a traditional Bayesian setup, but it plays no role in the data-generating process; rather, it is necessary to ensure that the values taken by the loss function are of a scale that can produce meaningful posterior distributions after being exponentiated. In other words, it is a necessary adjustment to integrate any given loss function with a prior-to-posterior update that is not derived from

an explicit statistical model of the data-generating process. We treat s as a hyperparameter and infer it from the data, along with with $\mathbf{m}$, in a hierarchical Bayesian framework.

Specifically, in our numerical experiments below, we endow s with a Gamma prior distribution as in Motamed & Appelo (2019). This allows us to use a Metropolis-within-Gibbs scheme to explore the posterior distribution: at each step of Markov chain Monte Carlo (MCMC), the value of s is updated via a conjugate Gibbs proposal (i.e., a sample from the full conditional distribution of s, which is again Gamma), and the value of $\mathbf{m}$ is adjusted using an adaptive Gaussian proposal (Haario et al. 2001) screened by a Metropolis-Hastings accept/reject step. The choice of values for the shape a and rate b parameters of the Gamma prior are particularly critical to obtaining a meaningful posterior. In our experiments, we set $a = 100$ and $b = 1$, such that the average value of s under the prior is 10^2. This compensates for the scale of the sum of TL distances, which is $\mathcal{O}(10^{-2})$ in our problem setting, and leads to a combined factor of $s \sum_{i=1}^{N} \mathrm{TL}_2^{\lambda}(\mathbf{u}_i(\mathbf{m}), \mathbf{y}_i)) = \mathcal{O}(1)$.

To summarize: we write our Gibbs posterior as $p^*(\mathbf{m}, s|\mathbf{y}) \propto \exp\left(-\ell(\mathbf{m}, \mathbf{y})\right) p_0(\mathbf{m}) p_0(s)$. Here $p_0(s)$ is the Gamma prior density described above, and $p_0(\mathbf{m})$ is chosen to be a uniform distribution on the 6-dimensional L_∞ unit ball (i.e., $\mathbf{m} \sim \mathcal{U}([-1, 1]^6)$). The MCMC algorithm run for J steps produces a sequence of realizations $\{(\mathbf{m}^{(j)}, s^{(j)})\}_{j=1}^{J}$. Marginalization is then trivial: the sequence $\{\mathbf{m}^{(j)}\}_{j=1}^{J}$ is (asymptotically) distributed according to $p^*(\mathbf{m}|\mathbf{y}) = \int p^*(\mathbf{m}, s|\mathbf{y}) ds$, i.e., the posterior distribution on the moment tensor components.

3.2 Quantitative metrics for posterior evaluation

While Bayesian inference is now widely used in applications, often it is not obvious what constitutes a "good" posterior. This question is particularly relevant and fundamental in misspecified settings, i.e., where the inference machinery does not contain a model of the true data-generating process. In such context, how much uncertainty is the right amount? Should the true value of the parameter always be expected to lie in the highest probability regions of the posterior (e.g., at the center of a Gaussian or unimodal posterior)? Partial answers exist in literature, but largely depend on the more fundamental question of "what does one want to use the posterior for?"

3.2.1 Continuous rank probability scores

We seek a methodology that quantifies how well our inference method performs relative to a true known answer (e.g., synthetic data or data-generating parameter values). To achieve this quantification, we need a *scoring rule*, which is a well known concept in Bayesian inference (Gneiting & Raftery 2007). Let H be the true data distribution and therefore a perfect forecaster. If G is instead an inference-based forecaster, a scoring rule $S(G, H)$ assesses the predictive accuracy of G with re-

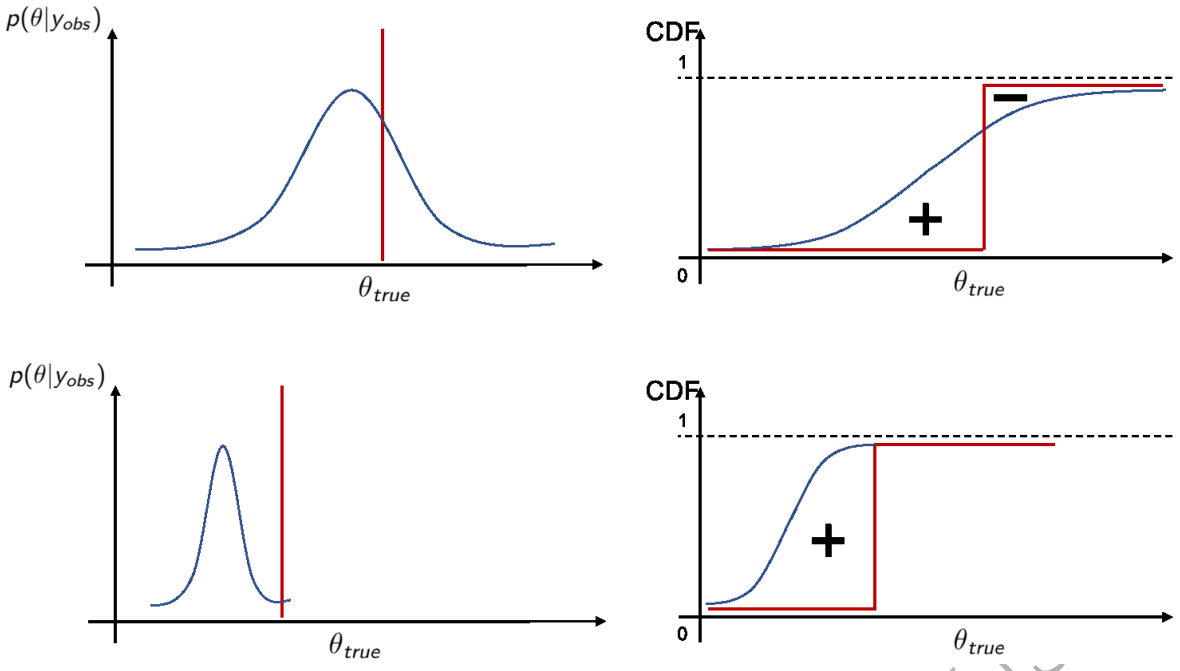

Figure 2. Variability (top) and bias (bottom) quantification in CRPS scores; here, red lines represent the perfect forecaster and blue lines represent a candidate inference-based forecaster.

spect to H. A scoring rule is said to be *proper* if $S(H, H) = \min_G S(G, H)$, i.e., if the lowest score corresponds to $G = H$. While this idea is often used to compare the true data distribution to a posterior predictive distribution, the same idea can be applied to parameters of interest, rather than data. In this case, a perfect forecaster would be a Dirac distribution centered on the true parameter value, i.e., $H(\theta) = \delta_{\theta=\theta_{\text{true}}}(\theta)$, while G could be any distribution with probability density $p_\theta(\theta)$ (e.g., a posterior distribution). For simplicity and computational tractability, we consider univariate scoring rules (i.e., rules for scalar θ). Some popular rules include the quadratic Brier score, $S(G, H) = \int_{-\infty}^{+\infty} \left(\delta_{\theta=\theta_{\text{true}}}(\theta) - p_\theta(\theta)\right)^2 d\theta$, and the continuous ranked probability score (CRPS),

$$S(G, H) = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{\theta} p_\theta(z)dz - \mathbb{1}_{\theta \leq \theta_{\text{true}}}(\theta) \right)^2 d\theta. \tag{8}$$

For both of these scores, a value of zero is the minimum, achieved if and only if p_θ is a Dirac distribution centered at θ_{true}. In this paper, we focus on the CRPS score (8). This score jointly captures the bias and variance of the forecaster G relative to θ_{true}, as illustrated in Figure 2. This ability follows from the fact that the CRPS compares the cumulative distribution functions of H and G, rather than their probability densities (which are not monotone increasing functions).

Note that if we had instead used scoring rules based on the data y and its corresponding posterior predictive distribution, the true data distribution is unknown outside of a synthetic setting, and thus the perfect forecaster is in practice represented by the empirical distribution of some held-out obser-

vational data (typically separate from the sample used to build the inference-based forecaster). In our numerical experiments below, however, we will know the true value of the parameters θ_{true} and thus can apply the CRPS score directly to the posterior distribution $p(\theta|\mathbf{y})$. In particular, we will compute a separate CRPS score for each scalar element of the moment tensor, $\theta = m_i$, $i = 1, \ldots, 6$; we do so because multi-dimensional generalizations of the CRPS score are extremely expensive to compute. These scalar scores will reflect both the bias and variance of each posterior marginal. In Sections 4.3 and 4.4, we will introduce additional *multivariate* schemes for assessing posterior quality, based on distances and angles between samples $\mathbf{m} \in \mathbb{R}^6$ and on the physical interpretation of the moment tensor.

4 EXPERIMENT: THE SEG/EAGE OVERTHRUST MODEL

4.1 Velocity models and inverse problem setup

Our earthquakes are simulated with the three-dimensional velocity model derived from the SEG/EAGE Overthrust model (House et al. 1996; Aminzadeh et al. 1994, 1995, 1996). We choose this model because it contains structural complexity that may not be easily represented using layered-medium models. We use a 15×15 km region located in the southwestern portion of the Overthrust model. The model extends to a depth of 4.7 km. Since only the P-wave velocity (V_p) model is given, we derive the S-wave velocity (V_s) using a variable V_p/V_s ratio in the range [2, 1.7], where V_p/V_s near the surface is close to 2 and it approaches 1.7 at the bottom of the model. The density model is obtained using Gardner's relation ($\rho = 310 V_p^{1/4}$).

Figure 3 shows a horizontal cross-section of the velocity at the source depth (1.1 km), the horizontal positions of the receivers (in blue), which are located at the surface, and the position of the source (in yellow). Figure 4 shows East-West cross sections of the model taken at the source location, which is at the position of the yellow star. The source position was taken to be near the fault that cuts the anticline. We used a total of six stations located at the surface and surrounding the source. We report in Table 1 the locations of the receivers and source. We simulated three-component waveforms for a single earthquake (with {strike, dip, rake} = $\{40^\circ, 50^\circ, 280^\circ\}$, respectively) in the elastic 3D model using SPECFEM (Komatitsch & Vilotte 1998). The source time history was taken to be a pulse that is nearly white between frequencies of approximately 1 and 13 Hz. These waveforms are taken as our earthquake waveforms, i.e., our observed data $\mathbf{y}$.

We then derived layered-medium models to be representations of the 3D structure obtained from well logs. We took vertical profiles of the velocity and density models. We averaged the P-wave velocity over 500 m depth intervals to approximate how one might obtain a layered medium model from

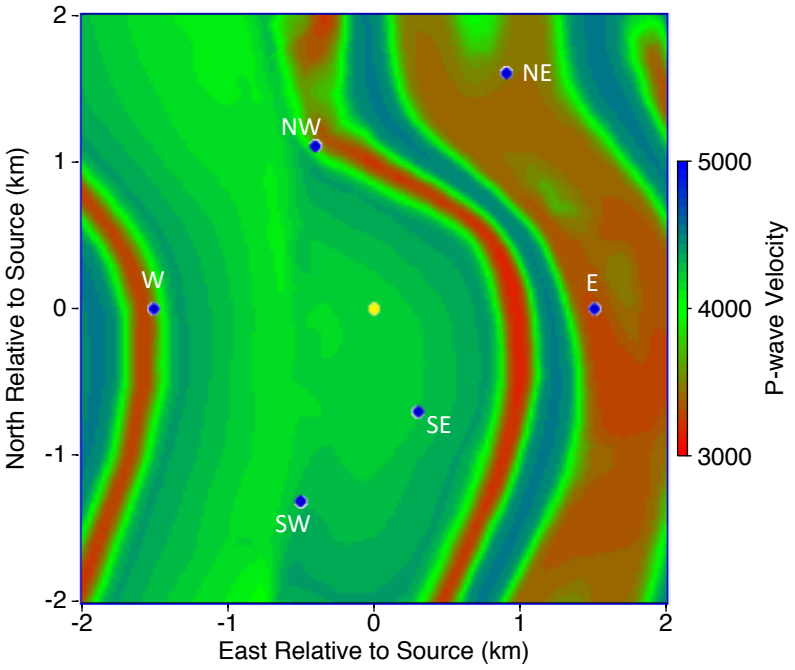

Figure 3. Horizontal cross section of the P-velocity model at the source depth (yellow dot). Locations of stations at the surface of the model are shown in blue.

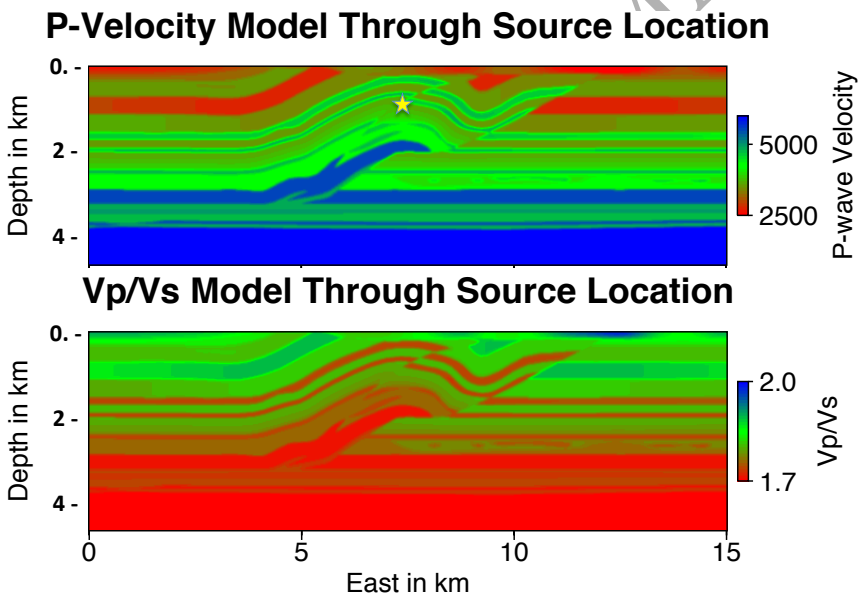

Figure 4. East-West vertical cross sections through SEG/EAGE Overthrust model at the position of the source (yellow star). Upper plot shows P-velocity model and lower plot shows ratio of P to S-wave velocities.

Stations	North	East	Depth
NW	1100	-400	10
NE	1600	900	10
SW	-1300	-500	10
SE	-700	300	10
W	10	-1500	10
E	10	1500	10
SOURCE	0	0	1100

Table 1. Location of the source and receivers. All distances expressed in meters and relative to the source.

a well log. To this averaged (smoothed) model, we added some noise equal to 2% at the top of the model and 10% at the base of the model to mimic increasing uncertainty in well logs with increasing depth. Further, we used a constant ratio of $V_p/V_s = 1.73$ to obtain the S-wave velocity. The density was taken to be constant at 2000 km/m^3. We used vertical profiles at each station and the source location to yield a total of seven layered velocity models. Figure 5 shows the velocity profile derived from the well log at the source location (on the right), as well as the layered-medium models obtained by averaging model properties over depth (and adding some noise) at each of the other well-log locations (on the left).

For each of the seven layered medium models, we simulate three-component waveforms at each surface station using Axitra, a discrete-wavenumber reflectivity modeling approach (Coutant 1990). We initially validated the ouputs of Axitra by also simulating an earthquake for a layered medium model in SPECFEM, and verifying that the outputs of both codes were visually identical. We then used Axitra to construct the Green's functions, and hence the forward model, used in inversion. Specifically, for each layered medium velocity model and for each of the six moment tensor components, we

simulated three-component waveforms at each station, using the same source time history as was used for the 3D earthquake simulation in SPECFEM. For each velocity model, we thus obtained a set of Green's functions used to simulate the waveforms $\mathbf{u}(\mathbf{m})$ at each station for a candidate $\mathbf{m}$ according to (2).

We summarize our inversion workflow in the box below.

Algorithm 1 Workflow for testing moment tensor inversion under model misspecification

1: Select $\mathbf{m}_{\text{true}}$ and $\mathbf{x}_{\text{true}}$;

2: Generate waveform data $\mathbf{y}_{i,k}$, for each orientation of displacement $i \in \{1,2,3\}$ and station $k \in \{\text{NW, NE, SW, SE, W, E, Source}\}$, using SPECFEM:

$$\mathbf{y}_{i,k} = \mathbf{u}_{\text{SPECFEM},i}(\mathbf{m}_{\text{true}}; \mathbf{x}_{s,k}, \mathbf{x}_{\text{true}}, \mathbf{V}_{\text{3D}}) + \mathbf{e}_{i,k}, \text{ where } \mathbf{e}_{i,k} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$$

with $\sigma = 10^{-3}$ (approx 1.5 orders of magnitude lower than the signal amplitude).

3: Sample the posterior distribution of $\mathbf{m}$ (via the MCMC scheme described in Section 3.1), using the following misspecified forward model in the generalized likelihoood (7): $\mathbf{u}_{i,k}(\mathbf{m}) = \mathbf{u}_{\text{Axitra},i}(\mathbf{m}; \mathbf{x}_{s,k}, \mathbf{x}_{\text{src}}, \mathbf{V}_{\text{layered}})$ where $\mathbf{x}_{\text{src}}$ is not necessarily equal to $\mathbf{x}_{\text{true}}$ (details in experiments below).

In the experiments below, we run our MCMC chains for at least $J = 10^6$ steps each; this number is probably rather higher than needed, but ensures thorough posterior exploration. The average wall-clock time per MCMC step was 0.18 ms when using an L_2 misfit and 0.65 ms when using the TL misfit. The hardware specifications are as follows: RAM 31.3 GiB, processor: Intel Core i7-8700K CPU at 3.7GHz, 12 cores.

4.2 Inversion results

Our inversion setup, as detailed in Section 4.1, includes a realistic form of model misspecification by construction: the observed data $\mathbf{y}$ are simulated using a complex 3D velocity model via SPECFEM, but the moment tensor $\mathbf{m}$ is inferred using one of seven layered approximations to this model. We now proceed to recover the moment tensor using each layered velocity model, applying a Bayesian formulation that employs either the TL or L_2 distance as a loss function (the L_2 loss is the standard approach, corresponding to the additive Gaussian likelihood model, and follows simply by replacing $\text{TL}_2^{\lambda}(\mathbf{u}_i(\mathbf{m}), \mathbf{y}_i)$ in (7) with $\|\mathbf{u}_i(\mathbf{m}) - \mathbf{y}_i\|_2^2$). We illustrate the one and two-parameter marginal posteriors of $\mathbf{m}$ for the North-West velocity model in Figure 6. The TL approach provides significantly better recovery: it exhibits smaller variance and closer alignment with $\mathbf{m}_{\text{true}}$. Similar results (not shown here) are obtained with the other velocity models. For a more quantitative comparison of

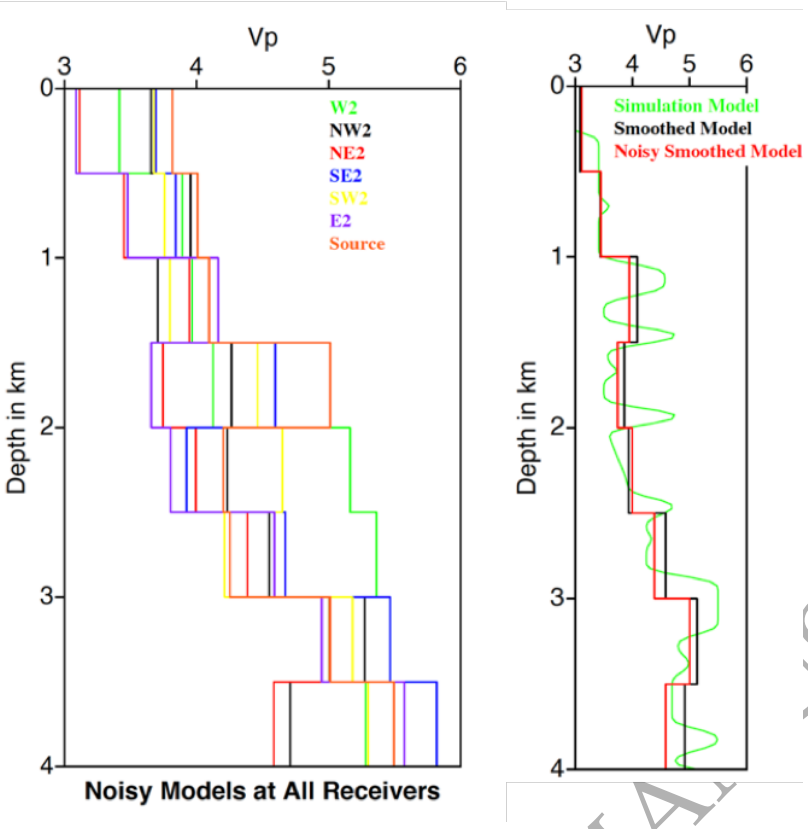

Figure 5. On the right: vertical velocity profile ("well log") of the 3D model at the source location (green), with smoothed (black) and noisy (red) profiles used to build the layered medium models. On the left: velocity profile for the layered medium model built at each station location.

the posteriors, we report CRPS scores obtained with each velocity model in Table 2. The CRPS values in the first seven lines of the table are averaged across each of the six components of $\mathbf{m}$. The bottom line then shows the mean score—and mean *difference* in CRPS scores obtained with the TL distance versus the L_2—across all seven velocity models.

For all of the layered velocity models, the TL misfit provides better inference and uncertainty quantification for the moment tensor. Lower CRPS scores indicate that the TL-based posterior distributions are on average less biased, and exhibit less variance, than those obtained with the standard L_2 distance. This translates into more reliable moment tensor estimates even in the present misspecified setting—i.e., when a realistic 3D velocity model is represented (incorrectly) by a layered medium model constructed from well logs.

Looking at the variability of the CRPS scores across the models, it also appears that the posteriors obtained with the L_2 distance are more sensitive to the velocity model used for inversion than the TL posteriors. (Consider a CRPS standard deviation of 0.05 in the L_2 case, versus 0.01 for TL.) The interpretation of this behavior in geophysical terms requires further investigation, but indicates

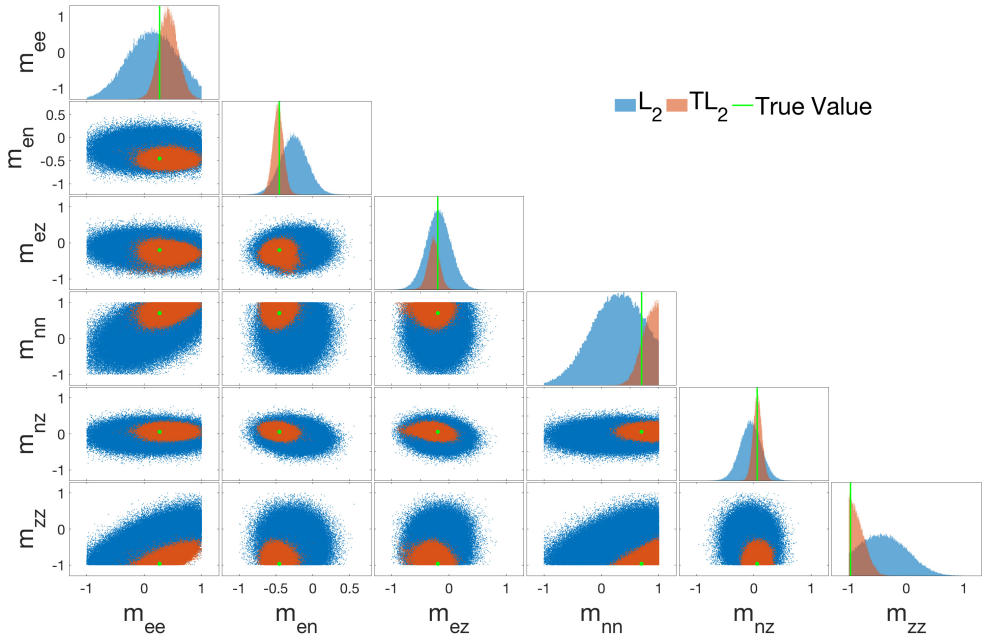

Figure 6. Matrix plot of one- and two-dimensional marginal posterior distributions obtained with the NW velocity model, for each moment tensor component and misfit measure. Green lines/dots indicate the true (data-generating) value of each moment tensor component.

Well log	$\mathbf{TL}_2$	$\boldsymbol{L_2}$	$\boldsymbol{\Delta_{L_2-\mathrm{TL}}}$
NW	0.0528	0.1685	0.1157
NE	0.0637	0.1785	0.1317
SW	0.0710	0.2005	0.1295
SE	0.0637	0.1785	0.1148
W	0.0635	0.1173	0.0539
E	0.0837	0.1665	0.0828
SOURCE	0.0799	0.3008	0.2209
Mean (Std)	**0.0694** 0.0105	**0.1907** 0.0562	**0.1213** 0.0198

Table 2. Average CRPS scores for univariate marginal posteriors, resulting from different layered velocity models.

$\mathbf{m}$	$\mathbf{TL}_2$	$\boldsymbol{L_2}$	$\boldsymbol{\Delta_{L_2-\mathrm{TL}}}$
m_{ee}	0.1410	0.1294	-0.0116
m_{en}	0.0399	0.1809	0.1410
m_{ez}	0.0455	0.0625	0.0169
m_{nn}	0.0951	0.2607	0.1656
m_{nz}	0.0202	0.0724	0.0522
m_{zz}	0.0744	0.4381	0.3637

Table 3. CRPS score for each moment tensor component, averaged across velocity models.

that relative TL distance between waveforms is less sensitive than L_2 to variations of the velocity model. This, in turn, suggests that OT misfit measures may exhibit some robustness to variations in experimental design (i.e., choice of station and well log location).

It is also interesting to calculate a CRPS score for each moment tensor element, averaging across all seven velocity models. We report these results in Table 3. These scores are lower in the TL case for all components of $\mathbf{m}$ except for m_{ee}. Visual inspection of the m_{ee} posterior marginal distribution suggestions that the L_2 posteriors, while in general more dispersive (i.e., higher variance), exhibit proportionally less bias, which in turn produces a lower score. At the moment, we do not have an explanation in geophysical terms for this behavior. Since it does not seem to be linked to any specific velocity model, however, possible causes may be related to the chosen network configuration.

4.3 Inner products and stereonets

As an additional way of assessing the quality of our inversion results, we calculate the posterior mean of the normalized inner product between the true moment tensor $\mathbf{m}_{\text{true}}$ and its inferred value. We estimate this quantity using posterior samples $\{\mathbf{m}^{(j)}\}$ as $\frac{1}{J}\sum_{j=1}^{J}\frac{\mathbf{m}_{\text{true}}\cdot\mathbf{m}^{(j)}}{\|\mathbf{m}_{\text{true}}\|_2\|\mathbf{m}^{(j)}\|_2}$. Values of this posterior statistic necessarily fall in the range $[-1,1]$. A value of one means that the posterior samples are perfectly aligned with the true moment tensor. This statistic is of particular interest since it looks at the six components of the moment tensor *jointly*, rather than individually/marginally as does the CRPS. In fact, one might argue that a moment tensor sampled from the posterior has a geophysical meaning only when analyzed in its entirety. We report in Table 4 the results for each velocity model, and for

Well log	$\mathbf{TL}_2$	$\boldsymbol{L_2}$	$\boldsymbol{\Delta_{L_2-\mathrm{TL}}}$
NW	0.9613	0.5952	0.3661
NE	0.9484	0.5811	0.3673
SW	0.9475	0.4841	0.4634
SE	0.959	0.5709	0.3881
W	0.9551	0.7718	0.1833
E	0.9452	0.6845	0.2607
SOURCE	0.9588	0.0302	0.9286
Mean (Std)	**0.9528** 0.0065	**0.6146** 0.2390	**0.4225** 0.2410

Table 4. Average (normalized) inner product between $\mathbf{m}_{\text{true}}$ and samples from the posterior distribution on $\mathbf{m}$, for each velocity model and loss formulation. A value of one indicates perfect alignment.

the TL and L_2 loss functions. Through this multivariate measure of posterior quality, the TL approach manifests itself as the clear winner, with an average score of 0.953 versus 0.615 for the L_2 approach.

A related quality measure is the Euclidean distance (in $\mathbb{R}^6$) of each posterior sample from the true moment tensor, i.e., $\|\mathbf{m}^{(j)} - \mathbf{m}_{\text{true}}\|_2$. We plot histograms of this quantity in Figure 7. It is apparent that samples from the L_2-based posteriors tend to be much further from the correct value than the TL-based posterior distributions. Additionally, the variance of the distance appears to be higher in the L_2 case than the TL, confirming a trend already observed in CRPS scores.

We conclude this section with a more physical interpretation of the moment tensor inversion results: stereonet plots of the normals to fault planes. In Figure 8 we report the results for station NW. The red dots represent the correct answer. Compared to the poles obtained from the L_2 analysis, the TL poles are more centered around the correct answer and at the same time much more tightly clustered.

4.4 Impact of model misspecification on the recovery of double couple earthquakes

The moment tensor as used throughout this paper is more general than the purely double-couple (DC) mechanism and can be in fact decomposed into DC, isotropic (ISO), and compensated linear vector dipole (CLVD) components, with the latter two representing the non-DC part of the focal mechanism. We refer to (Vavryčuk 2015) for the details of how to perform this decomposition. In this section,

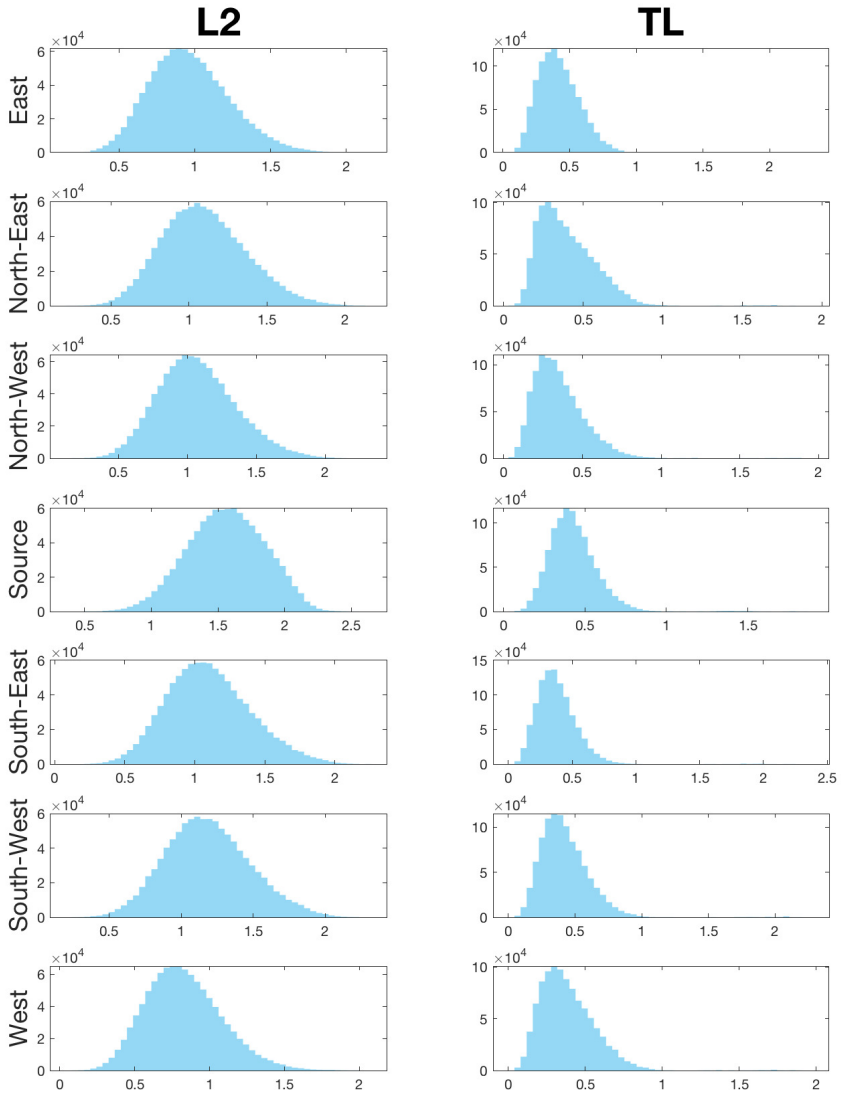

Figure 7. Histogram of Euclidean distances between posterior samples and the true moment tensor value, for each velocity model (row) and contrasting the L_2 and TL formulations (columns).

we assess the amount of non-DC component recovered through the two inversion methods, TL and standard L_2.

We are interested in assessing whether the characterization of isotropic or CLVD components in moment tensor inversion is a byproduct of model misspecification, rather than information actually coming from the data. Recall that the moment tensor used to generate the data, before noise perturbations, is a pure double couple. As a simple test, we analyze whether there is any change in non-DC

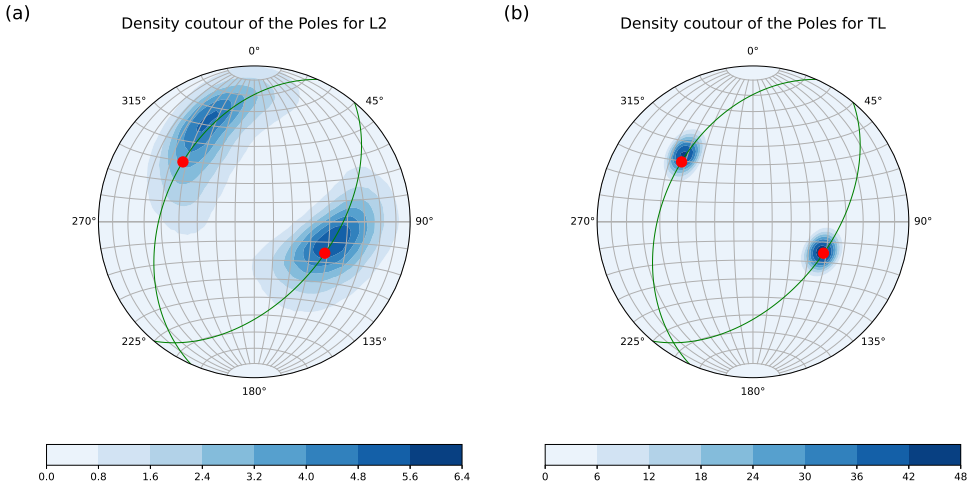

Figure 8. Stereonet contour plot of posterior samples resulting from inversion with the NW layered velocity model, for (a) the L_2-based posteriors and (b) the TL-based posteriors.

percentage for events recovered through the TL-based posterior versus the events recovered through the L_2-based posterior. We report the results in Figure 9. We decompose each posterior sample $\mathbf{m}^{(j)}$ into the three components, and aggregate the percentages over all posterior samples to produce each bar plot. We see a clear reduction in non-DC components (CLVD and ISO) in each of the TL posteriors, versus the L_2 posteriors. These results seem to confirm the hypothesis, at least in this case, that the recovery of non-DC mechanisms is mostly linked to the presence of model misspecification, rather than intrinsic to the data.

For additional perspective, Figure 10 presents bar plots where we classify an event (i.e., a posterior sample of $\mathbf{m}$) as DC if the percentage of DC component in the event is above 60%. We see that even though the true event is pure DC, the majority of events obtained with the L_2 analysis fail to satisfy this criterion. The TL analysis, on the other hand, yields a dramatically higher percentage of events with a primarily DC component.

4.5 Experiment extension: robustness under different focal mechanisms

So far, we have tested the performance of inversion with the TL distance under different velocity models. We have, however, used the same data set, i.e., waveform data $\mathbf{y}$ generated through the 3D Overthrust model from a single DC event. In this section, we instead use a single layered velocity model for inversion (obtained through a well log at the NW station location), while generating data

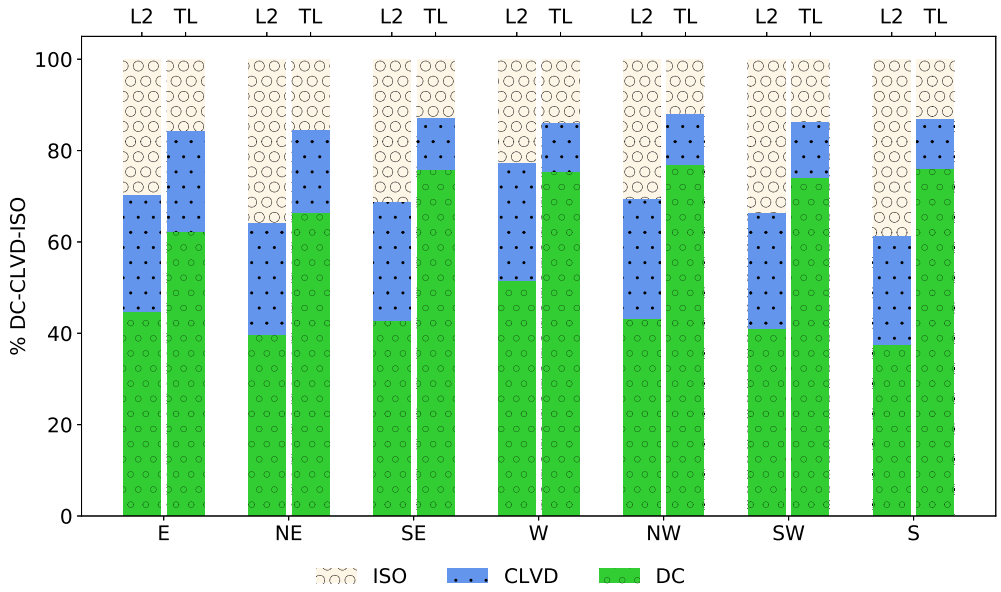

Figure 9. DC-ISO-CLVD decomposition of samples from the posterior distribution resulting from each velocity model.

from the Overthrust model using different values of the moment tensor. Our objective is to verify that the results described so far are not simply dependent on a specific event.

We generate eight alternative data sets by choosing focal mechanisms that reflect real earthquakes. All of the events are taken from the Harvard CMT catalogue (Huang et al. 1997; Dziewonski et al. 1981; Ekström & Nettles 1997; Chen et al. 2001; Ekström et al. 2012). Characteristics of the chosen events are given in Tables 5 and 6.

These events represent a variety of earthquakes with different percentages of DC and CLVD components. In Table 7 we report univariate CRPS scores, averaged across moment tensor components, for each event. According to these scores, it can be seen that for all events except 070886A, the TL misfit yields improved inversion performance compared to L_2. For event 070886A, the CRPS score for the posterior obtained with TL loss is significantly larger than the CRPS score for other events. We cannot provide a particularly intuitive explanation of why this specific event yields such behavior, but a multivariate perspective on the posterior adds important nuance to these results. In Figures 11, 12, and 13, we show posterior distributions for the first three events (070886A, 12487G, and 062992L), with the associated stereonets and contour plots given in Figure 14. Generally speaking, the TL contour plots show *less biased* and *less dispersed* fault plane recovery. Even for the first event (070886A), although the amount of bias obtained with the TL misfit is higher than that obtained with the L_2, the variance reduction is striking. On this note, as already stated in Section 4.3, the univariate CRPS score might be particularly penalizing when quantifying bias, since it is not calculated for $\mathbf{m}$ as a six-dimensional

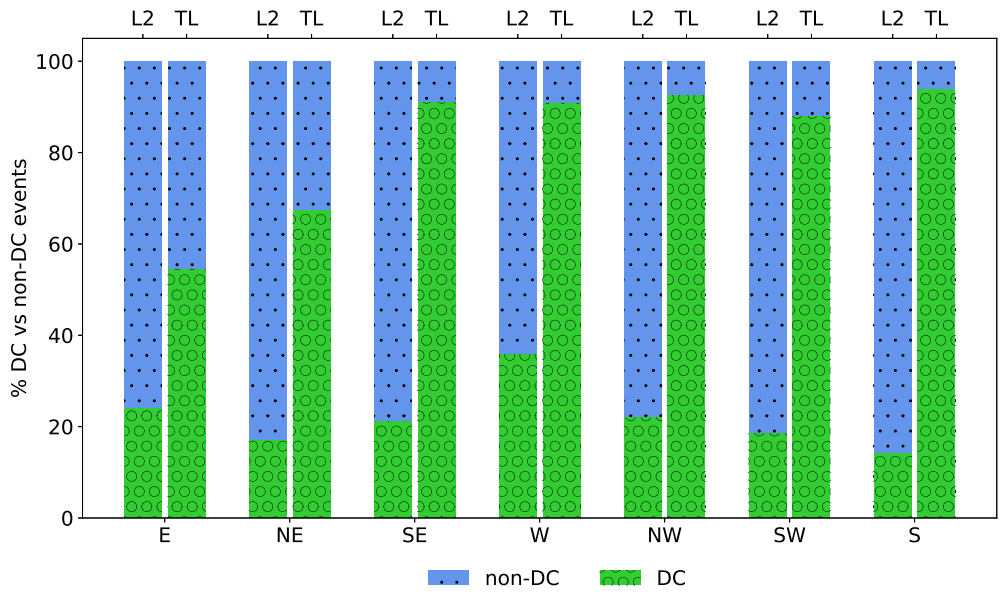

Figure 10. Share of events with a higher than 60% DC component, for each velocity model and for the L_2 and TL-based posteriors.

Event Name	$\mathbf{m_{nn}}$	$\mathbf{m_{ne}}$	$\mathbf{m_{nz}}$	$\mathbf{m_{ee}}$	$\mathbf{m_{ez}}$	$\mathbf{m_{zz}}$
070886A	-9.933	3	5.247	4.644	-8.325	5.29
12487G	-7.28	0.384	-0.945	6.744	1.105	0.536
062992L	-1.438	-1.178	0.296	1.413	-0.415	0.025
092904C	-4.75	-1.11	0.002	5.1	0.011	-0.35
201804051929A	-0.547	-1.25	-0.125	0.523	0.061	0.025
201507271812A	0.987	-2	-0.059	-0.676	0.004	-0.311
201511190742A	2.07	-1.11	-0.486	-1.63	-0.076	-0.436
201511300949A	3.23	-0.651	-0.438	-2.22	-0.325	-1.01

Table 5. List of selected events from the Harvard CMT catalogue (Dziewonski et al. 1981; Ekström et al. 2012). In computations, moment tensors have been normalized so their scalar moments are one.

Event Name	Strike -1	Dip - 1	Rake -1	Strike - 2	Dip -2	Rake -2	DC%	CLVD%
070886A	294	37	156	44	76	55	99.4	0.5
12487G	133	78	178	224	88	12	87.2	12.7
062992L	334	77	173	66	83	13	90.9	9.1
092904C	231	90	0	141	90	180	86.6	13.4
201804051929A	168	84	178	258	88	6	97.4	2.5
201507271812A	191	89	180	281	90	1	73.2	26.8
201511190742A	209	78	179	299	89	12	60.8	39.2
201511300949A	219	75	-173	127	83	-15	43.5	56.5

Table 6. List of selected events from the Harvard CMT catalogue (Dziewonski et al. 1981; Ekström et al. 2012): strike, dip, rake, and DC versus non-DC percentage.

vector, but rather for each component independently. The stereonets provide, instead, a more complete and physically interpretable visualization of the differences between the posteriors obtained with the two misfits. In this view, the inversion depicted in Figure 14(b) (obtained with the TL distance) might be more desirable than that of Figure 14(a). The other seven cases are unambiguous, in that the TL results are superior in both the univariate metrics and fault plane visualizations.

4.6 Experiment extension: misspecification of the source location

So far, we have only considered cases in which the source location is assumed to be known and well specified. In many contexts, however, source location information is not available and some estimation technique is needed to place the source within the model. In the next case study, we adopt a technique based on ray tracing from first-time arrivals. Rather than picking arrival times on the traces simulated in the 3D model, we used ray tracing through the 3D model to calculate the traveltimes from the correct source location to each receiver. To mimic what might be done with field data, we used these traveltimes in a grid search scheme to estimate the location of the event in the layered-medium model. We calculated Green's functions in the NW layered medium model (Figure 5) from this source location to each station. The resulting traces showed good alignments of both P and S arrivals between the traces calculated at the correct source location in the 3D model and those calculated using the shifted source location in the layered medium model. We took this alignment as an indication of a reasonable estimation of the location from the traveltimes. The newly derived Green's function traces were used both for the inversion with L_2 and TL misfit, since in both cases we would be operating

Event Name	$\mathbf{TL}_2$	$\boldsymbol{L_2}$	$\boldsymbol{\Delta_{L_2-\mathrm{TL}}}$
070886A	0.7000	0.2027	-0.4973
12487G	0.0282	0.1804	0.1522
062992L	0.0546	0.1702	0.1156
092904C	0.0268	0.1842	0.1574
201804051929A	0.0657	0.1497	0.0840
201507271812A	0.0757	0.1754	0.0997
201511190742A	0.0540	0.1990	0.1450
201511300949A	0.0608	0.1735	0.1127
Mean (Std)	**0.1332** 0.2148	**0.1794** 0.0157	**0.0462** 0.2068

Table 7. Average CRPS scores for univariate marginal posteriors, for different seismic events.

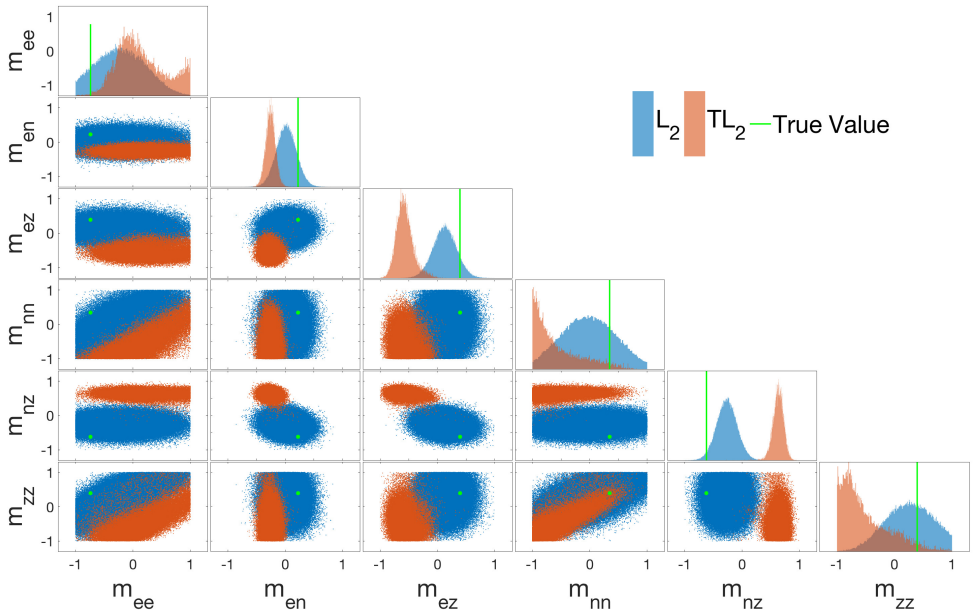

Figure 11. Event 070886A: matrix plot of one- and two-dimensional marginal posterior distributions obtained with a layered velocity model and either the TL or L_2 misfit.

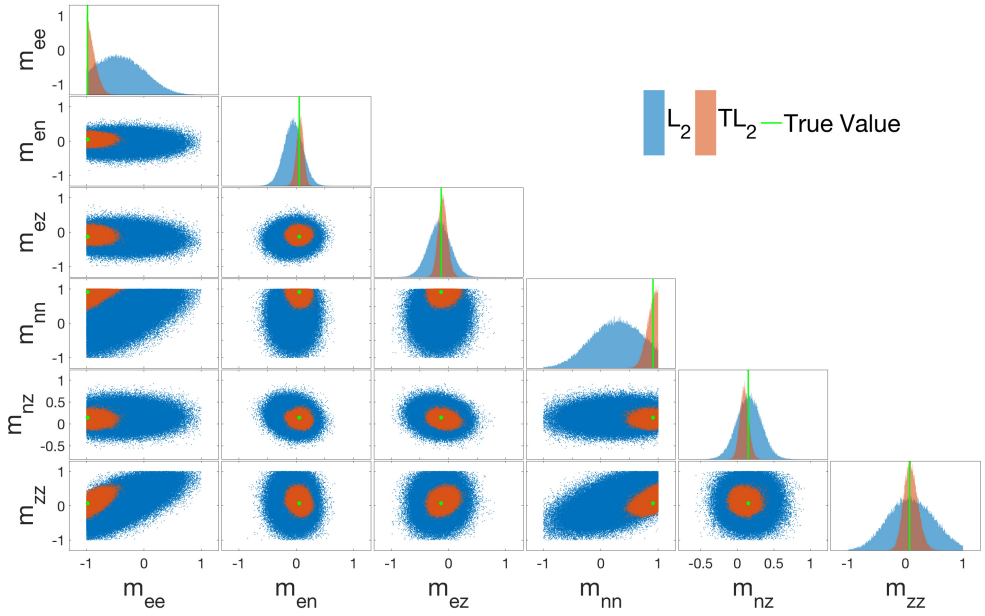

Figure 12. Event 12487G: matrix plot of one- and two-dimensional marginal posterior distributions obtained with a layered velocity model and either the TL or L_2 misfit.

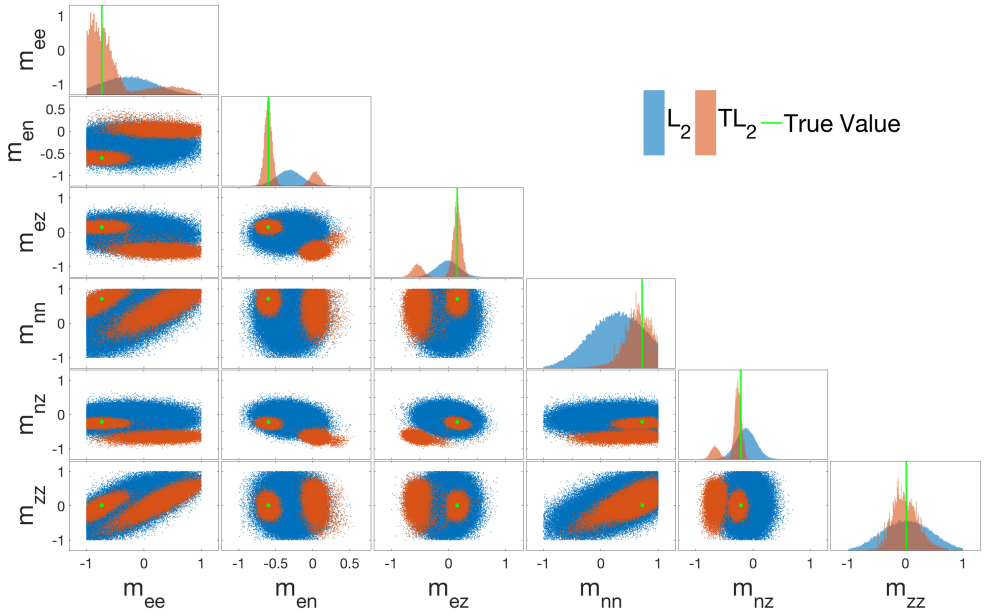

Figure 13. Event 062992L: matrix plot of one- and two-dimensional marginal posterior distributions obtained with a layered velocity model and either the TL or L_2 misfit.

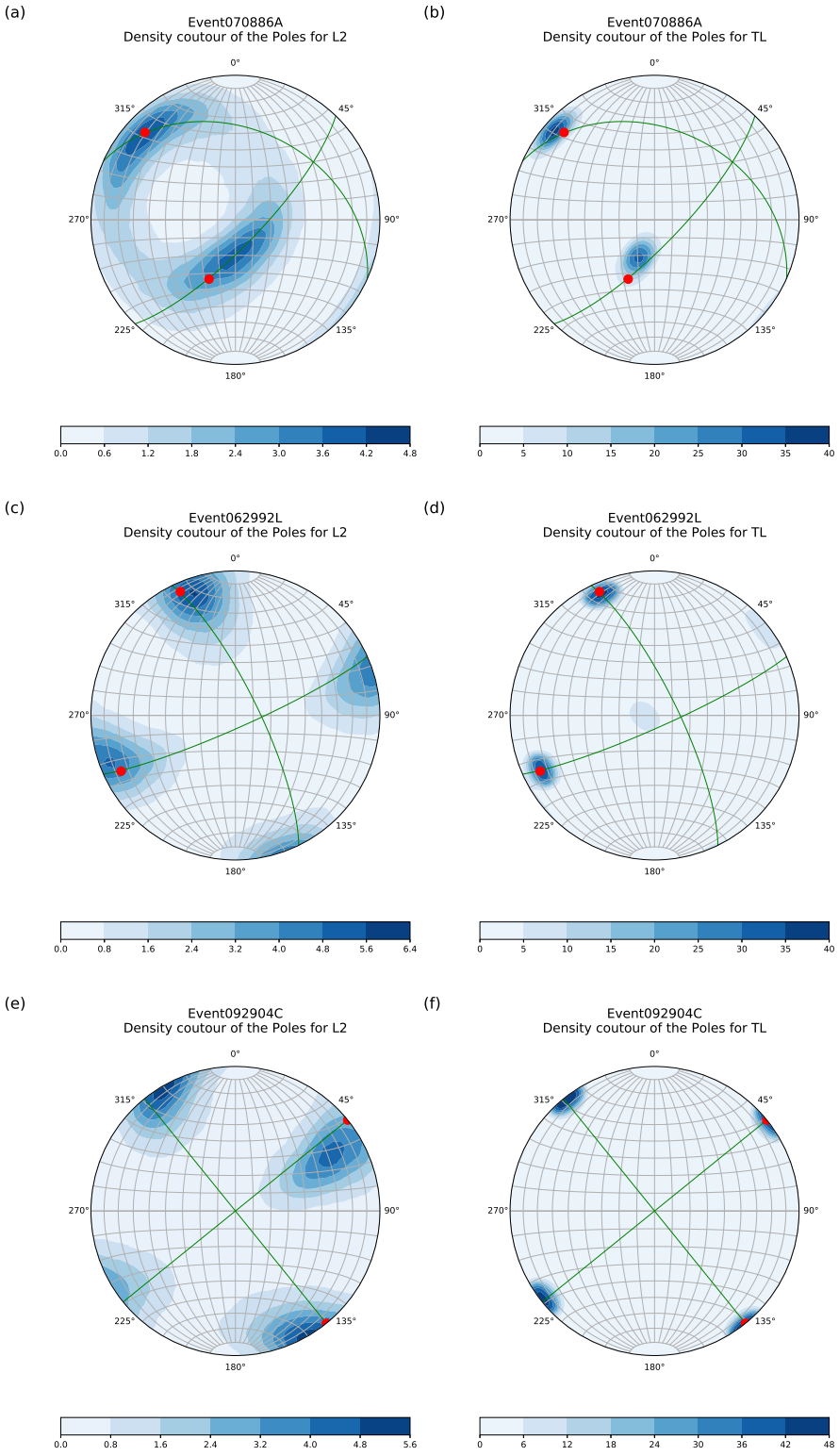

Figure 14. (a)-(f): Stereonet plot with contour lines for Event 070886A, Event 062992L, and Event 092904C with L_2 and TL misfits. The green lines and red dots denote the true fault planes and poles.

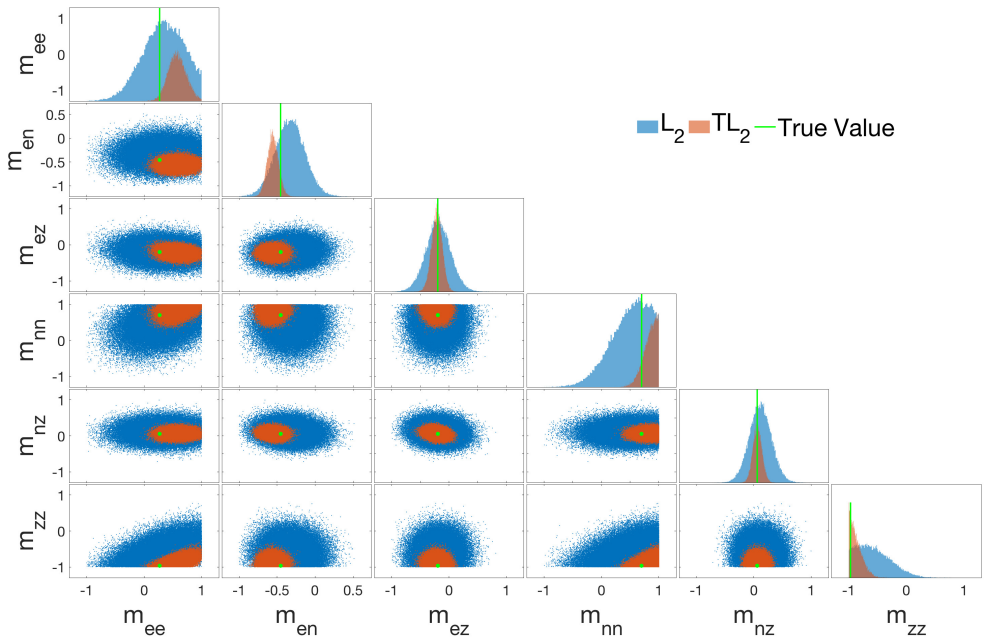

Figure 15. Matrix plot of one- and two-dimensional marginal posterior distributions obtained with the North-West velocity model, ray-traced source location and either the TL or L_2 misfit.

under the same instance of source location misspecification. For this test (and the ones that follow in this section) we used a single moment tensor value ({strike, dip, rake} = $\{40°, 50°, 280°\}$) and a single velocity model (NW), as we did not expect significantly different behaviors across velocity models and focal mechanisms. Figure 15 shows the posterior distributions obtained in this case: from a qualitative viewpoint, the TL still outperforms the L_2 bringing in a major reduction in variance. The L_2, however, exhibits less bias than it did in the corresponding case without source location (Figure 6). We explain this result with the fact that the specific technique used for locating the source acts as a partial compensator for time delays introduced by velocity model errors. In other words, it introduces a certain amount of time shifting in the Green's function traces that facilitates the recovery by the L_2. The TL does not benefit as much from this adjustment since it is within its own nature to compensate for this kind of discrepancy. We also report the now familiar quantitative measures for posterior and posterior samples evaluation (Table 8). These results confirm what emerged from a simple inspection of the posterior distributions.

4.7 Experiment extension: attenuation modeling

As mentioned in the introduction, TL does not simply act as a surrogate for a static time-shift aimed at compensating velocity model errors. It can also be useful under circumstances in which the dis-

Criterion	L_2	TL
Euclidean Distance	0.7801	0.4252
Inner Product	0.8000	0.9575
CRPS	0.1044	0.0780

Table 8. Quantitative measures of posterior quality for the North-West velocity model, ray-traced source location and either the TL or L_2 misfit. Recall that for Euclidean distance and CRPS, lower values are better; for the inner product, higher values (closer to one) are better.

crepancy between the traces arises from amplitude differences. To further investigate this aspect and the performance of the TL vs. L_2 with increased realism of the SPECFEM-generated waveforms, we present a number of tests for which the data-generating model includes the effects produced by attenuation. We repeated the experiment just discussed (Section 4.6), the only difference being that instead of using the elastic model for data generation (SPECFEM), an anelastic model with a quality factor Q_p of 50 and Q_s of 100 was used.

SPECFEM has the capability to model waveforms that include the effects of attenuation, which is specified as independent values of Q for P and S waves. Included in the simulation is the value of a reference frequency, which we choose as 20 Hz because that leads to causal waveforms at frequencies below 20 Hz. The source location used in SPECFEM was the correct source location of the 3D model. The inversion was conducted using the Green's functions calculated for the layered-medium model (NW) with the source at the location determined using ray tracing and earthquake location as described above.

The results (Figure 16 and Table 9) show once more the benefits brought by the use of the TL misfit. The introduction of attenuation modeling, while slightly producing an improvement in the performance of the TL, actually introduces more bias in the L_2-based posteriors compared to the equivalent posteriors obtained in the previous test case with attenuation-free data (Figure 15). This test case (as well as the ones that follow) indicates that the TL does not simply act as an alternative time-shifting mechanism, but is able to better quantify the discrepancy between traces that also differ in amplitude. The introduction of the attenuation modeling in fact further highlights the benefits of using the TL vs. the L_2.

4.8 Experiment extension: L_2 misfit and cross-correlation alignment

With the newly derived waveform data that includes the effects of attenuation, it is worth testing the performance of the TL vs. L_2 when the L_2 also benefits from the use of maximum cross-correlation to time-adjust the AXITRA modeled Green's functions. We cross-correlated waveforms calculated in

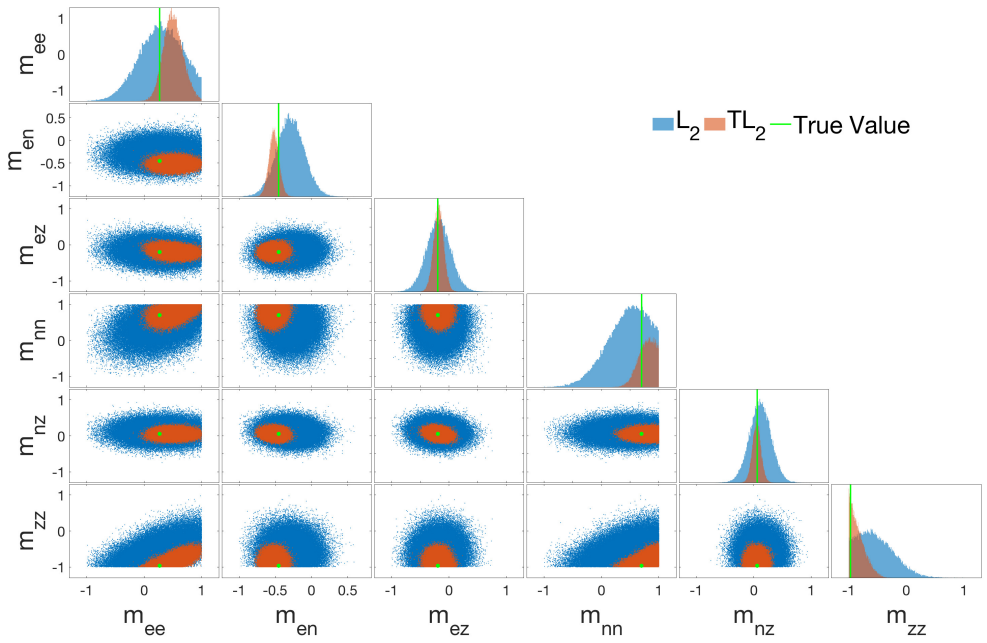

Figure 16. Matrix plot of one- and two-dimensional marginal posterior distributions obtained with the North-West velocity model, ray-traced source location, attenuation factor $Q_p = 50$ and $Q_s = 100$ and either the TL or L_2 misfit

the 3D medium for the earthquake by SPECFEM and waveforms calculated for an earthquake with the same moment tensor in layered model NW. Using the maximum time shift determined by cross-correlation for all three components of waveforms for each station, we shifted the Green's functions. This meant that the overall trace of each Green's function trace had optimal alignment with the earthquake waveform.

While in this setting the source location is not ray-traced for the TL nor the L_2, the L_2 is expected to benefit (at least partially) from the time alignment of the waveforms. The posteriors reported in Figure 17 once again confirm the increased variance reduction brought by the TL as well as the reduction in bias for the L_2, which largely benefits from the adjustment introduced by cross-correlation.

Criterion	L_2	TL
Euclidean Distance	0.8278	0.3764
Inner Product	0.7718	0.9587
CRPS	0.1153	0.0590

Table 9. Quantitative measures of posterior quality for the North-West velocity model, ray-traced source location, attenuation factor $Q_p = 50$ and $Q_s = 100$ and either the TL or L_2 misfit.

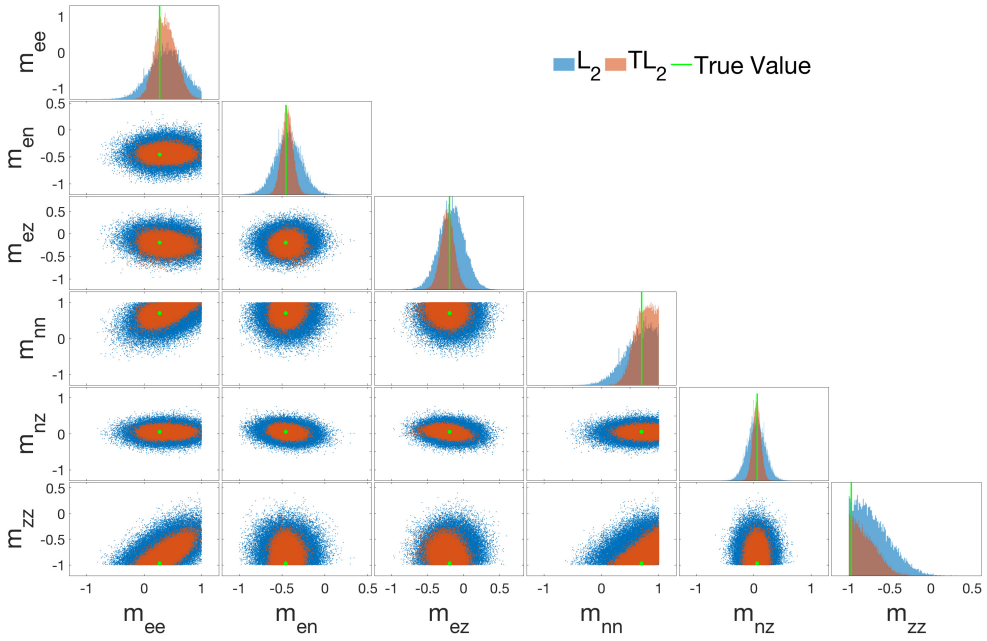

Figure 17. Matrix plot of one- and two-dimensional marginal posterior distributions obtained with the North-West velocity model, attenuation factor $Q_p = 50$ and $Q_s = 100$ and either the TL or L_2 with cross-correlation alignment misfit

On this note, it is important to remember that while computationally less expensive than the TL, cross-correlation alignment requires some decisions to make about the portion of trace to use for each waveform and whether to perform separate cross-correlation analyses for each component of waveforms.

4.9 Experiment extension: increased model error

We conclude this series of experiments with a test that, although rather extreme, exemplifies in a more direct way the fact the TL brings the most benefit when the amount of model misspecification is rather large. In summary, we repeat the last experiment (Section 4.8 – attenuation and cross-correlation for

Criterion	L_2	TL
Euclidean Distance	0.5659	0.3450
Inner Product	0.8999	0.9610
CRPS	0.0704	0.0443

Table 10. Quantitative measures of posterior quality for the North-West velocity model, attenuation factor $Q_p = 50$ and $Q_s = 100$ and either the TL or L_2 with cross-correlation alignment misfit.

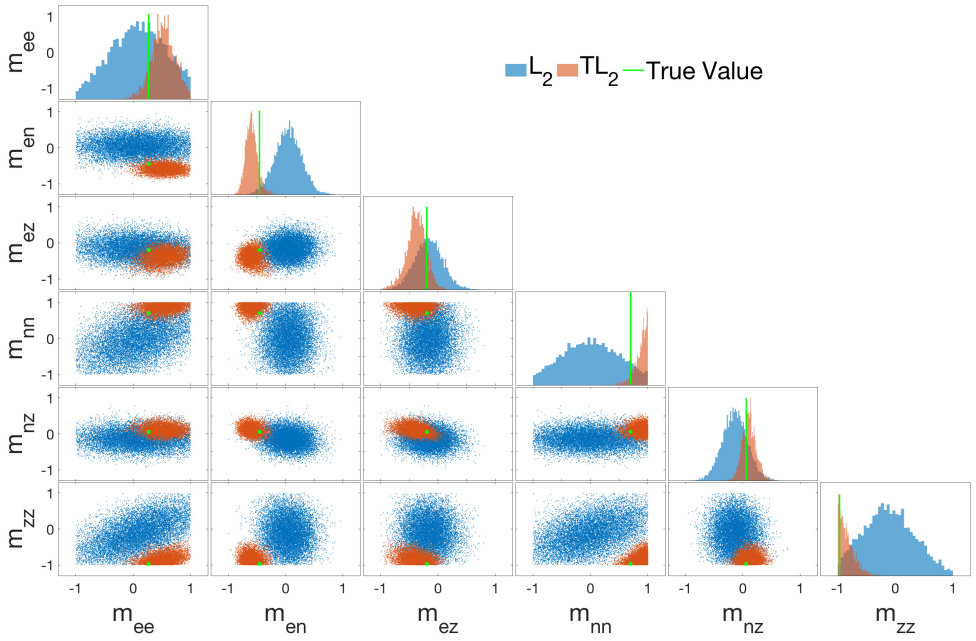

Figure 18. Matrix plot of one- and two-dimensional marginal posterior distributions obtained with a new North-West velocity model (P-wave velocity to S-wave velocity of 1.6), attenuation factor $Q_p = 50$ and $Q_s = 100$ and either the TL or L_2 with cross-correlation alignment misfit.

the L_2) with a newly misspecified layered model NW. Rather than defining the S-wave velocities for the layered model using "well logs" from the 3D model, we choose a constant ratio of P-wave velocity to S-wave velocity to illustrate the case where S-wave velocity information is not available and a poor choice is made in choosing the velocity ratio. We choose a ratio of 1.6. The effect of varying the ratio is to essentially compress or expand the time between P and S phases that comprise the waveforms depending on the spatial variation in this ratio in the true model.

We report the posterior distributions in Figure 18. In this case, the amount of improvement brought by the TL is significantly higher than that of the cross-correlated L_2 both in terms of bias and variance. This is also confirmed by the quantitative performance measures (Table 11). In summary, the TL is particularly beneficial when the amount of model misspecification is high enough for the optimal transport to effectively re-arrange the amplitude-time point pairs. By contrast, when the discrepancy between the signals is high, the L_2 approach is particularly penalized. When the dissimilarities between the signals are instead smaller, then the TL and L_2 approaches tend to behave similarly to the point of being identical the when the TL optimal assignment is the identity.

Criterion	L_2	TL
Euclidean Distance	1.4876	0.5144
Inner Product	0.1247	0.9417
CRPS	0.2901	0.0968

Table 11. Quantitative measures of posterior quality for a new North-West velocity model (P-wave velocity to S-wave velocity of 1.6), attenuation factor $Q_p = 50$ and $Q_s = 100$ and either the TL or L_2 with cross-correlation alignment misfit.

5 CONCLUSIONS AND DISCUSSION

In this paper, we proposed and evaluated an optimal transport-based Bayesian inference framework for a realistic scenario of moment tensor inversion with misspecified velocity models. Our framework uses the transport-Lagrangian distance as a loss function to define a Gibbs posterior. Data were generated using the SEG-EAGE Overthrust 3D velocity model. Velocity models used for inversion were layered-medium approximations of this 3D model. We demonstrated the reliability of the methodology in recovering the correct moment tensors under various scenarios of model misspecification, and for a variety of source mechanisms; in particular, we saw significant improvements over Bayesian moment tensor inversion with standard quadratic misfit functions (e.g., Gaussian likelihood models). We quantitatively assessed the validity of these results through a number of statistical and geophysical criteria. Finally, we showed how reducing the impact of model misspecification via our optimal transport approach led to a significant decrease in the non-double-couple component of the recovered focal mechanisms.

There are of course a number of important open questions following from the present study, and many avenues for future development and improvement. First, it is important to note that the type of model misspecification at hand (e.g., misspecified P-wave and S-wave velocities, attenuation) could certainly affect the performance of Bayesian inversion with the TL distance, and its performance relative to other misfits. We have explored many of these, and in almost all cases observed the superiority of TL to L_2, but one cannot make strict conclusions about misspecifications that are outside of the present scope; hence these may merit future investigation. Another important path towards more general statements and guarantees is theoretical analysis. Empirical studies cannot guarantee that the proposed approach will always be superior. Therefore, it would be desirable to develop a theoretical understanding of specific forms or scenarios of model misspecification (from the sources of misspecification to their impact on waveforms) and their interaction with the TL misfit. As explained in Section 1, the intuition behind TL distance is better "shape matching" of waveforms—e.g., the ability to detect shape similarity despite time-inhomogeneous phase shifts, amplitude mismatch, and other

forms of signal warping. Making this intuition precise could involve parameterizing specific aspects of "shape" and contrasting the dependence of the TL and L_2-derived likelihood models on these parameters; in turn, one would want to elucidate the extent to which such shape variations depend on the moment tensor $\mathbf{m}$ or instead on the various physical factors that influence the Green's function $\mathbf{G}$. Separating these dependencies seems crucial to understanding the extent to which moment tensor inversion can be made robust to misspecification. That said, it is reasonable to expect that for model misspecifications producing discrepancies of phase or amplitude within the regime covered here, the use of a TL distance in Bayesian inversion should provide good results. For particularly safety critical applications, a possibility is always to solve the inverse problem using both TL and standard methods, and to compare the plausibility of the answers in case of significant discrepancies.

Thus far, our moment tensor inversion method has targeted local-scale seismic data, e.g., microseismic data in oil/gas fields. The expansion of this method to regional-scale moment tensor inversion where dispersed surface waves are present needs further study. Another important caveat is that the computational cost of computing the TL distance may be non-negligible (see Section 4.1); in our implementation, it is roughly a factor of 3 or 4 larger than that of computing a simple L_2 distance. This can be relevant when MCMC chains need to be run for a large number of steps to achieve sufficient mixing.

Future work should also involve applying this methodology to field microseismic data, rather than synthetic data, and evaluate the scientific value of TL-based Bayesian inversion in such settings. A parallel line of theoretical work could seek more general suggestions for misfit metrics in moment tensor inversion, to mitigate different types of velocity misspecification, and embed these metrics within the Gibbs posterior construction presented here. Finally, we believe that the benefits of using optimal transport distances in misspecified Bayesian inverse problems involving time series data could extend to other problems within and outside seismology.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the College of Petroleum Engineering & Geosciences at King Fahd University of Petroleum and Minerals, Global Partnership Program with MIT. We also thank Yanhua Yuan for her significant help on 3D waveform modeling with attenuation.

DATA AVAILABLITY STATEMENT

The data underlying this article will be shared on request to the corresponding author.

REFERENCES

Aminzadeh, F., Burkhard, N., Nicoletis, L., Rocca, F., & Wyatt, K., 1994. SEG/EAEG 3-D modeling project: 2nd update, *The Leading Edge*, **13**(9), 949–952.

Aminzadeh, F., Burkhard, N., Kunz, T., Nicoletis, L., & Rocca, F., 1995. 3-D modeling project: 3rd report, *The Leading Edge*, **14**(2), 125–128.

Aminzadeh, F., Burkhard, N., Long, J., Kunz, T., & Duclos, P., 1996. Three dimensional SEG/EAEG models: an update, *The Leading Edge*, **15**(2), 131–134.

Arild, B. & Henning, O., 2003. Bayesian linearized AVO inversion, *Geophysics*, **68**(1), 185–198, doi: 10.1190/1.1543206.

Bertsekas, D. P., 1981. A new algorithm for the assignment problem, *Mathematical Programming*, **21**(1), 152–171.

Bissiri, P. G., Holmes, C. C., & Walker, S. G., 2016. A general framework for updating belief distributions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**(5), 1103–1130.

Bosch, M., Mukerji, T., & Gonzalez, E., 2010. Seismic inversion for reservoir properties combining statistical rock physics and geostatistics: a review, *Geophysics*, **75**(5), 75a165–75a176.

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

Brossier, R., Operto, S., & Virieux, J., 2015. Velocity model building from seismic reflection data by full-waveform inversion, *Geophysical Prospecting*, **63**(2), 354–367.

Chen, F. & Peter, D., 2018. Constructing misfit function for full waveform inversion based on sliced Wasserstein distance, in *80th EAGE Conference and Exhibition 2018*, Ifema, Madrid, Spain, doi: 10.3997/2214-4609.201801030.

Chen, P. F., Nettles, M., Okal, E. A., & Ekström, G., 2001. Centroid moment tensor solutions for intermediate-depth earthquakes of the wwssn–hglp era (1962–1975), *Physics of the Earth and Planetary Interiors*, **124**(1-2), 1–7.

Coutant, O., 1990. Programme de simulation numerique AXITRA, rapport LGIT, *Univ. Joseph Fourier, Grenoble*.

Cover, T. M., 1999. Elements of information theory, p. 19.

Crase, E., Picat, A., Noble, M., McDonald, J., & Tarantola, A., 1990. Robust elastic nonlinear waveform inversion: Application to real data, *Geophysics*, **55**(5), 527–538.

De Figueiredo, L. P., Grana, D., Santos, M., Figueiredo, W., Roisenberg, M., & Neto, G. S., 2017. Bayesian seismic inversion based on rock-physics prior modeling for the joint estimation of acoustic impedance, porosity and lithofacies, *Journal of Computational Physics*, **336**, 128–142, doi: 10.1016/j.jcp.2017.02.013.

Dreger, D. & Woods, B., 2002. Regional distance seismic moment tensors of nuclear explosions, *Tectonophysics*, **356**(1-3), 139–156.

Duijndam, A., 1988. Bayesian estimation in seismic inversion. Part I: principles, *Geophysical Prospecting*, **36**(8), 878–898.

Dziewonski, A. M., Chou, T., & Woodhouse, J. H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *Journal of Geophysical Research: Solid Earth*, **86**(B4), 2825–2852.

Ekström, G. & Nettles, M., 1997. Calibration of the HGLP seismograph network and centroid-moment tensor analysis of significant earthquakes of 1976, *Physics of the Earth and Planetary Interiors*, **101**(3-4), 219–243.

Ekström, G., Nettles, M., & Dziewoński, A., 2012. The global CMT project 2004–2010: Centroid-moment tensors for 13017 earthquakes, *Physics of the Earth and Planetary Interiors*, **200**, 1–9.

Engquist, B. & Froese, B. D., 2014. Application of the Wasserstein metric to seismic signals, *Communications in Mathematical Sciences*, **12**(1), 979–988.

Engquist, B. & Yang, Y., 2020. Optimal transport based seismic inversion: Beyond cycle skipping, *ArXiv*, arXiv:2002.00031.

Gauthier, O., Virieux, J., & Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: Numerical results, *Geophysics*, **51**(7), 1387–1403.

Gebraad, L., Boehm, C., & Fichtner, A., 2020. Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo, *Journal of Geophysical Research: Solid Earth*, **125**(3), e2019JB018428.

Gee, L. S. & Jordan, T. H., 1992. Generalized seismological data functionals, *Geophysical Journal International*, **111**(2), 363–390.

Gneiting, T. & Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation, *Journal of the American statistical Association*, **102**(477), 359–378.

Gouveia, W. P. & Scales, J. A., 1998. Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis, *Journal of Geophysical Research: Solid Earth*, **103**(B2), 2759–2779.

Grana, D. & Della Rossa, E., 2010. Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion, *Geophysics*, **75**(3), O21–o37, doi: 10.1190/1.3386676.

Gu, C., Marzouk, Y. M., & Toksöz, M. N., 2018. Waveform-based Bayesian full moment tensor inversion and uncertainty determination for the induced seismicity in an oil/gas field, *Geophysical Journal International*, **212**(3), 1963–1985.

Guasch, L., Warner, M., & Ravaut, C., 2019. Adaptive waveform inversion: Practice, *Geophysics*, **84**(3), R447–r461.

Guilhem, A., Hutchings, L., Dreger, D. S., & Johnson, L., 2014. Moment tensor inversions of m˜ 3 earthquakes in the geysers geothermal fields, california, *Journal of Geophysical Research: Solid Earth*, **119**(3), 2121–2137.

Gunning, J. & Glinsky, M. E., 2007. Detection of reservoir quality using Bayesian seismic inversion, *Geophysics*, **72**(3), R37–r49, doi: 10.1190/1.2713043.

Haario, H., Saksman, E., Tamminen, J., et al., 2001. An adaptive Metropolis algorithm, *Bernoulli*, **7**(2), 223–242.

House, L. S., Fehler, M., Barhen, J., Aminzadeh, F., & Larsen, S., 1996. A national laboratory industry collaboration to use SEG/EAEG model data sets, *The Leading Edge*, **15**(2), 135–136.

Huang, W. C., Okal, E. A., Ekström, G., & Salganik, M. P., 1997. Centroid moment tensor solutions for deep

earthquakes predating the digital era: the world-wide standardized seismograph network dataset (1962–1976), *Physics of the earth and planetary interiors*, **99**(1-2), 121–129.

Izzatullah, M., van Leeuwen, T., & Peter, D., 2019. Bayesian uncertainty estimation for full waveform inversion: A numerical study, in *SEG International Exposition and Annual Meeting*, OnePetro.

Kennedy, M. C. & O'Hagan, A., 2001. Bayesian calibration of computer models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(3), 425–464, doi: 10.1214/12-EJS675.

Kleijn, B., Van der Vaart, A., et al., 2012. The Bernstein-von Mises theorem under misspecification, *Electronic Journal of Statistics*, **6**, 354–381, doi: 10.1214/12-EJS675.

Kolouri, S., Park, S., Thorpe, M., Slepčev, D., & Rohde, G. K., 2016. Transport-based analysis, modeling, and learning from signal and data distributions, *arXiv preprint arXiv:1609.04767*.

Komatitsch, D. & Vilotte, J.-P., 1998. The spectral element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures, *Bulletin of the seismological society of America*, **88**(2), 368–392.

Luo, J., Wu, R. S., & Gao, F., 2018. Time-domain full waveform inversion using instantaneous phase information with damping, *Journal of Geophysics and Engineering*, **15**(3), 1032–1041.

Luo, S. & Sava, P., 2011. A deconvolution-based objective function for wave-equation inversion, in *SEG Technical Program Expanded Abstracts 2011*, pp. 2788–2792, Society of Exploration Geophysicists.

Luo, Y. & Schuster, G. T., 1991. Wave equation inversion of skeletalized geophysical data, *Geophysical Journal International*, **105**(2), 289–294.

Métivier, L., Brossier, R., Virieux, J., & Operto, S., 2013. Full waveform inversion and the truncated Newton method, *SIAM Journal of Scientific Computing*, **35**(2), B401–b437, doi: 10.1137/120877854.

Métivier, L., Brossier, R., Merigot, Q., Oudet, E., & Virieux, J., 2016a. An optimal transport approach for seismic tomography: Application to 3D full waveform inversion, *IOPscience: Inverse Problems*, **32**(11), R59–r80, doi: 10.1190/GEO2012-0338.1.

Métivier, L., Brossier, R., Merigot, Q., Oudet, E., & Virieux, J., 2016b. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion, *Geophysics Journal International*, **205**(6), 345–377, doi: 10.1093/gji/ggw014.

Métivier, L., Allain, A., Brossier, R., Mérigot, Q., Oudet, E., & Virieux, J., 2018. A graph-space approach to optimal transport for full waveform inversion, in *SEG Technical Program Expanded Abstracts 2018*, pp. 1158–1162, Society of Exploration Geophysicists.

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

Michael, W., Andrew, R., Tenice, N., Joanna, M., Adrian, U., Nikhil, S., Vetle, V., Ivan, S., Lluis, G., Caroline, W., Graham, C., & Bertrand, A., 2013. Anisotropic 3D full-waveform inversion, *Geophysics*, **78**(2), R59–R80, doi: 10.1190/GEO2012-0338.1.

Miller, J. W. & Dunson, D. B., 2018. Robust Bayesian inference via coarsening, *Journal of the American Statistical Association*, (just-accepted), 1–31, doi: 10.1080/01621459.2018.1469995.

Motamed, M. & Appelo, D., 2019. Wasserstein metric-driven Bayesian inversion with applications to signal processing, *International Journal for Uncertainty Quantification*, **9**(4).

Müller, M., 2007. Information retrieval for music and motion, chap. Dynamic time warping, pp. 69–84, Springer.

Odile, G., Virieux, J., & Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: Numerical results, *Geophysics*, **51**(7), 1387–1403.

Peyré, G., Cuturi, M., et al., 2019. Computational optimal transport: With applications to data science, *Foundations and Trends® in Machine Learning*, **11**(5-6), 355–607.

Pladys, A., Brossier, R., Li, Y., & Métivier, L., 2021. On cycle-skipping and misfit functions modification for full-wave inversion: comparison of five recent approaches, *Geophysics*, **86**(4), 1–85.

Ray, A., Kaplan, S., Washbourne, J., & Albertin, U., 2017. Low frequency full waveform seismic inversion within a tree based Bayesian framework, *Geophysical Journal International*, **212**(1), 522–542.

Rickers, F., Fichtner, A., & Trampert, J., 2012. Imaging mantle plumes with instantaneous phase measurements of diffracted waves, *Geophysical Journal International*, **190**(1), 650–664.

Sánchez-Reyes, H. S., Tago, J., Métivier, L., Cruz-Atienza, V., & Virieux, J., 2018. An evolutive linear kinematic source inversion, *Journal of Geophysical Research: Solid Earth*, **123**(6), 4859–4885.

Schwartz, B., 1994. A computational analysis of the auction algorithm, *European journal of operational research*, **74**(1), 161–169.

Sen, M. K. & Stoffa, P. L., 1996. Bayesian inference, Gibbs sampler, and uncertainty estimation in geophysical inversion, *Geophysical Prospecting*, **44**(2), 313–350.

Shearer, P. M., 2009. *Introduction to seismology*, Cambridge University Press.

Singh, S., Tsvankin, I., & Naeini, E. Z., 2018. Bayesian framework for elastic full-waveform inversion with facies information, *The Leading Edge*, **37**(12), 924–931.

Socco, L. V., Foti, S., & Boiero, D., 2010. Surface-wave analysis for building near-surface velocity models—established approaches and new perspectives, *Geophysics*, **75**(5), 75a83–75a102.

Thorpe, M. & Slepcev, D., 2017. Transportation l_p distances: Properties and extensions.

Thorpe, M., Park, S., Kolouri, S., Rohde, G. K., & Slepcev, D., 2017. A transportation l_p distance for signal analysis, *Journal of Mathematical Imaging and Vision*, **59**(2), 187–210.

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

Vavryčuk, V., 2015. Moment tensor decompositions revisited, *Journal of Seismology*, **19**(1), 231–252.

Villani, C., 2009. *Optimal transport: old and new*, vol. 338, Springer.

Virieux, J. & Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics, *Geophysics*, **74**(6), Wcc1–wcc26.

Warner, M. & Guasch, L., 2014. Adaptive waveform inversion-FWI without cycle skipping: theory, in *76th EAGE Conference and Exhibition 2014*, vol. 2014, pp. 1–5, European Association of Geoscientists & Engineers.

Warner, M. & Guasch, L., 2016. Adaptive waveform inversion: Theory, *Geophysics*, **81**(6), R429–r445.

Warner, M., Nangoo, T., Shah, N., Umpleby, A., & Morgan, J., 2013. Full-waveform inversion of cycle-skipped seismic data by frequency down-shifting, in *SEG Technical Program Expanded Abstracts 2013*, pp. 903–907, Society of Exploration Geophysicists.

Yang, F. & Ma, J., 2019. Deep-learning inversion: A next-generation seismic velocity model building method, *Geophysics*, **84**(4), R583–r599.

Yang, Y., Engquist, B., Sun, J., & Hamfeldt, B. F., 2018. Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion, *Geophysics*, **83**(1), R43–r62.

Zhu, H., Li, S., Fomel, S., Stadler, G., & Ghattas, O., 2016. A Bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration, *Geophysics*, **81**(5), R307–r323.

APPENDIX

Throughout the paper we have maintained the assumption of having to deal with data affected by uncorrelated Gaussian noise. To test the TL under a more realistic setting, we repeated the experiment of Section 4.8, but we added noise to simulated SPECFEM traces using actual noise recorded by the Groningen (Netherlands) network. We choose a magnitude 0.5 earthquake and computed the signal-to-noise for the event as a function of frequency using a portion of the trace prior to the earthquake as noise. We choose six stations located within a few km of the event so that the stations were at roughly the same distance as the stations in our test data. Prior to adding noise to our test data, the Groningen noise traces were scaled in the frequency domain so that the signal-to-noise of our test data would be similar to that in the Groningen data. The posterior distributions resulting from this test are reported in Figure 19. It can be seen that the addition of correlated noise did not particularly affect the performance of the TL versus L_2 (cf. Figure 17).

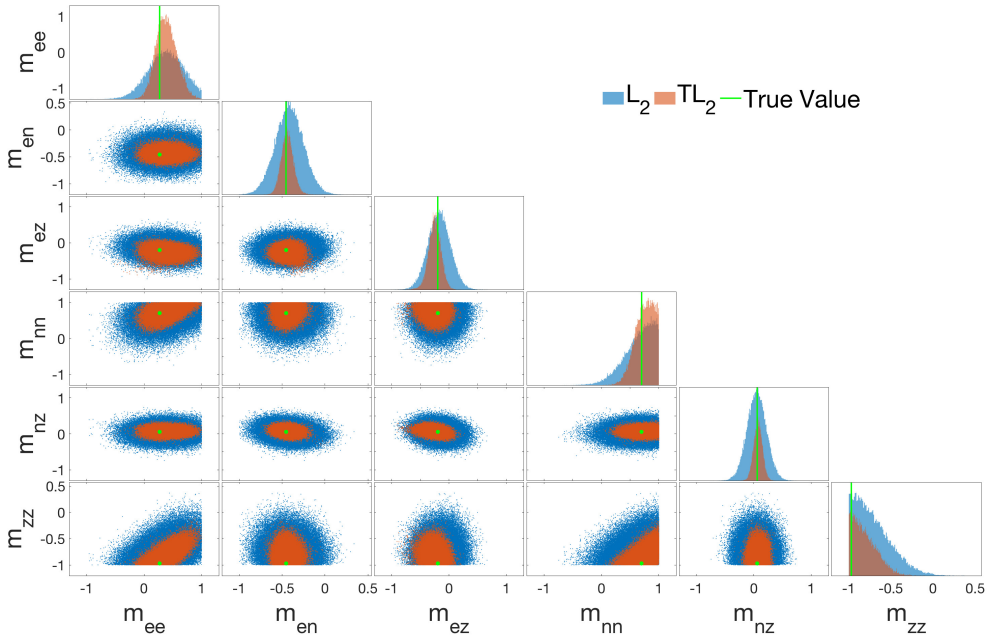

Figure 19. Matrix plot of one- and two-dimensional marginal posterior distributions obtained with the North-West velocity model, attenuation factor $Q_p = 50$ and $Q_s = 100$, Groningen noise and either the TL or L_2 with cross-correlation alignment misfit