submitted to *Geophys. J. Int.*

Uncertainty quantification for regularized inversion of electromagnetic geophysical data. Part I: Motivation and Theory.

Daniel Blatter[a], Matthias Morzfeld[a], Kerry Key[b], Steven Constable[a]

[a] *Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92037, USA*

[b] *Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA.*

E-mail: dblatter@ucsd.edu

SUMMARY

We present a method for computing a meaningful uncertainty quantification (UQ) for regularized inversion of electromagnetic (EM) geophysical data that combines the machineries of regularized inversion and Bayesian sampling with a "randomize-then-optimize" (RTO) approach. The RTO procedure is to perturb the canonical objective function in such a way that the minimizers of the perturbations closely follow a Bayesian posterior distribution. In practice, this means that we can compute UQ for a regularized inversion by running standard inversion/optimization algorithms in a parallel for-loop with only minor modification of existing codes. Our work is split into two parts. In Part I we review RTO and extend the methodology to estimate the regularization penalty weight on the fly, not unlike in the Occam inversion. We call the resulting algorithm the RTO-TKO and explain that it samples from a biased distribution which we numerically demonstrate to be nearby the Bayesian posterior distribution. In return for accepting this small bias, the advan-

tage of RTO-TKO over asymptotically unbiased samplers is that it significantly accelerates convergence and leverages computational parallelism, which makes it highly scalable to 2D and 3D EM problems. In Part II, we showcase the versatility and computational efficiency of RTO-TKO and apply it to a variety of EM inversions in 1D and 2D, carefully comparing the RTO-TKO results to established UQ estimates using other methods. We further investigate scalability to 3D, and discuss the influence of prior assumptions and model parameterizations on the UQ.

1 INTRODUCTION

Regularized inversion is well-established and remains to this day the "workhorse" algorithm in geophysics. The result of a regularized inversion is a single model that minimizes data misfit while simultaneously satisfying regularization constraints (see, e.g., Constable et al. 1987; Parker 1994; Newman & Alumbaugh 2000; Aster et al. 2011; Fournier & Oldenburg 2019). Uncertainty quantification (UQ) is an important aspect of geophysical inversion and is essential to constraining Earth properties like temperature, melt fraction, or pore fluid content, from the physical properties directly sensed by geophysical methods, such as electrical resistivity and seismic velocity.

Presently, there is no satisfactory avenue to computing a UQ for a regularized inversion, especially for inversions of electromagnetic (EM) data in 2D or 3D. Linearization (see, e.g., Tarantola 2005), for example, is computationally feasible, but it is problematic because it is only valid in the immediate vicinity of a reference model and can greatly under- or over-estimate uncertainty (Dettmer & Dosso 2013). Bayesian sampling goes beyond linearization and seeks to place probabilistic bounds on the acceptable model space by finding the range of models that fit the data and prior assumptions (or, equivalently, satisfy regularization constraints). However, Bayesian sampling is typically implemented via a Markov chain Monte Carlo (MCMC) approach which, due to extremely slow convergence, limits applicability to 1D EM and some 2D problems (Blatter et al. 2021, see below for more detail on the limitations of MCMC).

In this work we describe the mathematical and computational framework for how to compute and interpret uncertainty in regularized inversion, starting from a well-defined, non-trivial model class—similar in spirit to the Occam inversion (Constable et al. 1987), but with a Bayesian twist. To achieve our goal, we indeed combine regularized inversion with Bayesian

sampling, resulting in an algorithm that we call the "RTO-TKO." The RTO-TKO is designed to be computationally efficient, grid invariant (see below for explanations), and scalable to 2D (shown here), and probably 3D (left for future work). Moreover, RTO-TKO is easy to use because it repurposes regularized inversion codes and does not require much further tuning.

The computational advantages of the RTO-TKO come about because it does not sample the Bayesian posterior distribution, but a "biased" distribution that is often nearby. We demonstrate numerically that this bias is small in EM problems compared to other sources of error/bias, such as assumptions about model errors, model error covariances, or the errors introduced by an asymptotically unbiased sampler at finite chain length. The (small) bias of the RTO-TKO is also the inspiration for its abbreviation: TKO stands for "technical knock-out," where a true "knock out" would be reserved for an unbiased method that can still be scaled to very large problems.

Our work is split into two parts. Part I focuses on the mathematical foundations of the RTO-TKO method. Specifically, we explain the sampling algorithm, its advantages and shortcomings, and we present how to automatically adjust the regularization strength—again, similar in spirit to the Occam inversion (Constable et al. 1987). We illustrate the use of RTO-TKO on simplified toy problems and on a 1D DC resistivity problem (field-data). Part II showcases the RTO-TKO applied to a large number of EM problems, including the magnetotelluric (MT) method, controlled source EM, and joint inversions. We demonstrate its computational efficiency by using the RTO-TKO to invert 2D MT field-data and by comparing the computational costs of RTO-TKO and a recently developed trans-dimensional MCMC sampler (Blatter et al. 2021). Part II also details the (large) effect of prior assumptions (regularization and model parameterization) on the UQ and discusses the practical consequences of these mathematical facts.

2 MOTIVATION, OVERVIEW AND CONTEXT

Before getting into the details of the RTO-TKO, we motivate our thinking and provide an informal overview of how the algorithm works and how it achieves computational efficiency. We also explain how our work fits into the current literature on Bayesian sampling in geophysics.

2.1 RTO basics

"Randomize-then-optimize" (RTO) has been studied in applied mathematics (Bardsley et al. 2014, 2015; Wang et al. 2017; Bardsley et al. 2020; Bardsley & Cui 2021), and variations

of the idea are used in applications ranging from numerical weather prediction (where it is called "ensemble of data assimilation," Bonavita et al. (2012)) to reservoir modeling/history matching, (where variations of RTO are called "randomized-maximum likelihood," Gu & Oliver (2007); Chen & Oliver (2012); Oliver (2017); Stordal & Nævdal (2018)). The key insight behind RTO is that the canonical objective function of regularized inversion can be recast in terms of a Bayesian posterior distribution where the data misfit term corresponds to the likelihood and the model regularization term corresponds to the prior (see also Calvetti & Somersalo 2018; Vignoli et al. 2021).

The basic ideas behind RTO are as follows. The deterministic solution to the regularized inverse problem has a high posterior probability because it fits the data (high likelihood) and satisfies the regularization constraints (high prior probability). In fact, the regularized inverse solution maximizes the posterior probability (Stewart 2010). To explore the full range of models that satisfy the data and regularization constraints (rather than sampling only the maximum probability model), RTO perturbs the objective function so that the minimizers of the perturbations closely follow an associated Bayesian posterior distribution. In practice, this means that RTO allows us to compute a nonlinear UQ by repurposing the machinery of regularized inversion within a Bayesian framework. The samples that are generated by solving (perturbed) optimization problems can be viewed as proposals for a Markov chain Monte Carlo (MCMC) sampler and are accepted/rejected using the typical Metropolis-Hastings machinery (Metropolis et al. 1953; Hastings 1970).

The main goal of this paper series is to demonstrate that the accept/reject step (i) slows down convergence and prevents effective use of parallel computing; and (ii) is unnecessary because the bias caused by omitting the accept/reject step is small. The latter is a delicate issue, in particular because a mathematical theory for when the bias is small is missing. We take a pragmatic approach and demonstrate, numerically, that RTO (and our extension called the RTO-TKO) yields useful UQ for regularized inversion in EM geophysics. We cannot guarantee, however, that our method would achieve a similarly small bias if applied to other geophysical methods. On the other hand, RTO ideas have been used widely and with great success in Earth science (Bonavita et al. 2012; Gu & Oliver 2007; Chen & Oliver 2012; Oliver 2017; Stordal & Nævdal 2018) suggesting the ideas may be broadly applicable. It remains for the applied mathematics community to develop the theory to robustly quantify the bias resulting from neglecting the Metropolis-Hastings machinery.

2.2 Limitations of MCMC

There exists a large number of MCMC methods that can sample a Bayesian posterior distribution, e.g., Random Walk Metropolis (RWM), Metropolis adjusted Langevin algorithm (MALA), Hamiltonian Monte Carlo (HMC, Neal 2011; Duane et al. 1987), or ensemble samplers (Goodman & Weare 2010; Christen & Fox 2010). Unfortunately, none of these samplers scale well with the dimension of the problem, limiting the use of MCMC in EM geophysics to 1D problems and some 2D problems (e.g., Chen et al. 2012; Rosas-Carbajal et al. 2014), and more broadly to problems with only a handful of unknowns.

The reasons for this unfortunate fact are well known. An MCMC sampler must be "tuned" to keep the step size small enough for optimal acceptance probability, yet large enough to explore the space in a reasonable amount of time. For RWM, this means the step size is inversely proportional to the dimension of the problem, n (Beskos et al. 2009; Roberts et al. 1997; Roberts & Rosenthal 1998). Moreover, the "optimal" acceptance rate is 0.234 (Roberts et al. 1997), meaning that even an optimally-tuned RWM algorithm is quite inefficient, with about three quarters of all proposed models ultimately wasted. Similarly, the optimal step size for MALA is $O(n^{-1/3})$ (Robert & Rosenthal 2001) and for HMC it is $O(n^{-1/4})$ (Beskos et al. 2013). We note that for other samplers (e.g., the ensemble samplers of Goodman & Weare 2010; Christen & Fox 2010), such scalings are unknown but are likely no better (see also Morzfeld et al. 2019).

The inverse scaling of step size with dimension implies that an MCMC sampler needs draw a very large number of samples in order to give reliable results in high dimensional problems—even when setting issues of reaching stationarity (the infamous 'burn-in' period) to the side. Drawing this many samples can quickly become impractical if each step in the Markov chain requires running a computationally demanding numerical model, as in EM geophysics. To put things simply, MCMC explores the model space sequentially and slowly, so that convergence is usually too slow for problems of significant size. Moreover, whether or not an MCMC sampler has converged is difficult to assess because it is challenging to say whether, after n iterations, the algorithm has drawn samples from all the high probability regions of model space. Even more generally, a direct consequence of the Metropolis-Hastings accept/reject step is that MCMC is fundamentally serial: model $\boldsymbol{m}_{i+1}$ in the chain depends upon model $\boldsymbol{m}_i$. For this reason, MCMC cannot leverage HPC resources efficiently (besides running chains independently and in parallel, which does not address issues of reaching stationarity, see below for more detail). For all these reasons, the practical applications of MCMC have been limited to 1D and some lightweight 2D problems in EM geophysics.

While the focus of this paper series is on regularized models, many recent geophysical works on Bayesian sampling use trans-dimensional (trans-D) Bayesian sampling (Green 1995). Trans-D sampling, introduced to geophysics by Malinverno (2002), makes use of Bayesian parsimony as an implicit form of regularizaton, but trans-D samplers typically do not include additional (explicit) regularization. The implicit regularization induced by the parsimony is such that trans-D samplers prefer simpler models with fewer parameters to more complex models with more parameters (MacKay 2003; Schoniger et al. 2015). Trans-D sampling methods have been successfully used in geophysics in 1D (e.g. Minsley 2011; Ray et al. 2014; Dettmer et al. 2015; Blatter et al. 2018) and 2D (e.g. Bodin & Sambridge 2009; Hawkins & Sambridge 2015; Galetti & Curtis 2018; Blatter et al. 2021), though most 2D examples are in seismology rather than EM geophysics. And while there has been some recent success with applying trans-D sampling to 3D geophysics problems (Brett et al. 2021), the high cost of 3D forward modeling, the large number of forward evaluations needed for MCMC algorithms to converge (Agostinetti & Bodin 2018), and the fundamentally serial nature of MCMC make application of these methods to high-dimensional models difficult.

2.3 Trading computational efficiency for a small bias

Our main motivation is to overcome the computational bottleneck associated with MCMC and we do so by using RTO to sample a "biased" distribution, which is not equal to the (desired) Bayesian posterior distribution, but which is often a good approximation of it. We do this by simply omitting the Metropolis-Hasting accept/reject step – all RTO proposals are accepted. We propose biased sampling for the following reasons:

(i) The bias is often small and we numerically demonstrate that this is so in a large number of EM problems (see also Part II of this series). This means in particular that one obtains essentially the same UQ under biased sampling and under asymptotically exact sampling.

(ii) Accepting a small bias brings about significant computational advantages and allows us to use RTO at scale on 2D (and perhaps even 3D) EM geophysics problems.

(iii) RTO, with a small bias, has proven successful in other applications, see, e.g., Emerick & Reynolds (2013); Gao et al. (2006), but also the original RTO paper of Bardsley et al. (2014), as well as the literature on randomized-maximum likelihood, e.g., Wang et al. (2018); Oliver (2017). More generally, biased sampling, e.g., via an ensemble Kalman filter, has proven very effective in (global) numerical weather prediction and in physical oceanography.

Indeed, RTO drastically changes how the sampling is done: Rather than making small

local changes, RTO generates global samples via optimization of a perturbed objective function and, since the accept/reject step is omitted, RTO exhibits rapid convergence because it avoids largely unnecessary rejections and being forced to take small, local steps. The computational efficiency of RTO – both in terms of the total flops required and the total run time of the algorithm – is due to two key facts. First, each sample has, by construction, a high posterior probability. While the computational resources that go into the construction of a single sample are high (compared to, say RWM), this investment pays off because no samples are wasted so that one only needs a small number of samples for an approximate UQ (see below and Part II for more details). Second, the samples are independent of each other and, due their independence, RTO can leverage high performance computing (HPC) more effectively than MCMC samplers.

We emphasize that we do not carelessly trade bias for efficiency. Indeed, we verify numerically that the bias is small in a large number of test problems (toy and field data) by comparing to established UQ algorithms and, whenever possible, to unbiased MCMC. Finally, we note that even an asymptotically exact MCMC sampler is "biased" in practical application, since the number of MCMC steps is by necessity finite. Indeed, we show that the bias introduced by running an asymptotically unbiased MCMC sampler with finite chain length can exceed the bias introduced by the RTO. Additionally, statistical assumptions about errors in the data, e.g., Gaussian errors with known variances/covariances, are often hard to justify from first principles, calling into question why the precise Bayesian posterior distribution is the gold standard.

In summary, our main motivation for omitting the accept/reject step in RTO is that we cannot have everything – computational efficiency and asymptotic (infinite number of samples) convergence. In view of the bias being negligibly small in many practical situations (such as inversion of EM geophysical data, as demonstrated here), trading bias for efficiency seems to be a pragmatic way forward for pushing UQ beyond simplified or 1D inverse problems.

2.4 The RTO-TKO

We extend the well-known RTO method to sample the regularization penalty weight (see Section 3.3), letting the data and the prior determine the appropriate range of values for this important parameter – much in the spirit of an Occam inversion. We call this extension the RTO-TKO, because it essentially amounts to using an RTO step twice within a Gibbs sampler (see also Bardsley & Cui 2021). We design the RTO-TKO to be invariant under grid refinement (Stewart 2010; Chen et al. 2018; Dunlop et al. 2020). This means the UQ remains

unchanged if the underlying parameter grid is made finer, i.e., the choice of the number of model parameters does not influence the UQ (so long as the grid is sufficiently fine to capture model structure resolvable by the data). Grid invariance is an important property, because without grid invariance, the results are "biased" by the choice of grid. Within RTO-TKO, grid invariance is surprisingly easy to implement via a clever change of variables (see below for more detail).

2.5 Which UQ do you want?

Finally, we note that geophysical data rarely can constrain all aspects of the model. For this reason, the prior has a large effect on the posterior distribution and, therefore, on the resulting UQ. Throughout this two-part article, we carefully investigate the practical consequences of these important issues (for the first time, as far as we know). For example, the mathematical and computational framework we describe relies heavily on well-established regularized inversion and computes a UQ for such models. We contrast this 'regularized' UQ with results obtained by state-of-the art trans-D Bayesian samplers, which are not regularized, besides the implicit regularization induced by Bayesian parsimony. As we will demonstrate, the two UQs are *not* the same, but this fact should not be used to say that one UQ is "better" than another. Instead, we need to be aware that UQ is always "local" not "global,"—prior assumptions, including regularization (implicit or explicit) and model parameterization, will always have a profound influence on posterior uncertainty (see Part II for more details).

3 METHODS

3.1 Randomize-then-optimize: Transforming regularized inversion codes into parallel samplers

Regularized inversion finds an optimal model by minimizing an objective function that has the following canonical form:

$$\min_{\boldsymbol{m}} \quad f(\boldsymbol{m}) = \frac{1}{2}\left\| C_d^{-1/2}\big(\boldsymbol{F}(\boldsymbol{m}) - \boldsymbol{d}\big) \right\|^2 + \frac{\mu}{2}|| \, L\boldsymbol{m} \, ||^2 \tag{1}$$

In the above, f is the objective function, $\boldsymbol{m}$ is a vector of model parameters, $\boldsymbol{d}$ is a vector of data, $\boldsymbol{F}$ is a forward modeling operator, $\boldsymbol{F}(\boldsymbol{m})$ is a vector of modeled data, C_d is the measurement noise covariance matrix, L is a regularization operator (often a first or second derivative), and μ is the regularization penalty weight. Throughout this paper, we use vertical

bars to denote the 2-norm, i.e., $||\boldsymbol{x}|| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$, where $\boldsymbol{x}$ is a vector and superscript T denotes the transpose.

Nonlinear uncertainty estimates can be obtained by realizing that the objective function has a probabilistic interpretation. Taking the negative exponential of the objective function defines a probability density function (pdf):

$$p(\boldsymbol{m}|\boldsymbol{d}) \propto \exp\left(-f(\boldsymbol{m})\right) \propto \exp\left(-\frac{1}{2}\left|\left| C_d^{-1/2}\Big(\boldsymbol{F}(\boldsymbol{m}) - \boldsymbol{d}\Big)\right|\right|^2 - \frac{\mu}{2}\left|\left| L\boldsymbol{m}\right|\right|^2\right). \tag{2}$$

This pdf can be identified as a Bayesian posterior distribution, $p(\boldsymbol{m}|\boldsymbol{d}) \propto p(\boldsymbol{d}|\boldsymbol{m})p(\boldsymbol{m})$, where the data misfit term defines the likelihood, and the regularization term defines the prior:

$$p(\boldsymbol{d}|\boldsymbol{m}) = \exp\left(-\frac{1}{2}\left|\left| C_d^{-1/2}\Big(\boldsymbol{F}(\boldsymbol{m}) - \boldsymbol{d}\Big)\right|\right|^2\right), \quad p(\boldsymbol{m}) = \exp\left(-\frac{\mu}{2}\left|\left| L\boldsymbol{m}\right|\right|^2\right). \tag{3}$$

In fact, the minimizer of the optimization problem in Eq. 1 maximizes the posterior probability. More generally, models that have a high posterior probability result from the product of a low data misfit and/or low regularization term, with the regularization strength ensuring that highly regularized models that don't fit the data are not permitted. To quantify uncertainty in the solution of an inverse problem, one can thus compute all models that have a large posterior probability.

Efficient sampling of the space of all high posterior probability models is possible by transforming the deterministic optimization problem in Eq. 1 into a stochastic optimization problem. This requires making two simple adjustments. The first change is to perturb the data, so that the uncertainty in the data is reflected in the solutions to Eq. 1. The second change is to regularize against a random model that satisfies the regularization constraint (but which does not necessarily have a small data misfit). This second change allows uncertainty in the regularization (the prior model covariance) to be reflected in the range of models that solve Eq. 1. With these two changes, Eq. 1 becomes

$$\min_{\boldsymbol{m}} \quad f(\boldsymbol{m}) \;=\; \frac{1}{2}\left|\left| C_d^{-1/2}\Big(\boldsymbol{F}(\boldsymbol{m}) - \tilde{\boldsymbol{d}}\Big)\right|\right|^2 + \frac{\mu}{2}\left|\left| L(\boldsymbol{m} - \tilde{\boldsymbol{m}})\right|\right|^2 \tag{4}$$

where

$$\tilde{\boldsymbol{d}} \sim \mathcal{N}(\boldsymbol{d}, C_d), \quad \tilde{\boldsymbol{m}} \sim \mathcal{N}(0, \frac{1}{\mu}(L^TL)^{-1}) \tag{5}$$

are the perturbed data and prior model; we use the common notation that $\mathcal{N} \sim (\boldsymbol{a}, B)$ is the Gaussian distribution with mean $\boldsymbol{a}$ and covariance matrix B. Note that the perturbations are in line with the prior and assumptions about measurement noise. The perturbed data $\tilde{\boldsymbol{d}}$ are Gaussian with mean equal to the data $\boldsymbol{d}$ and covariance C_d. The perturbed model $\tilde{\boldsymbol{m}}$, which

we call the "prior model," is Gaussian with mean zero and covariance matrix $\frac{1}{\mu}(L^TL)^{-1}$, which is the prior covariance matrix implied by Eq. 3. Throughout Part I, we assume that the prior covariance is symmetric positive definite and, hence invertible. This is often not true in practice, where L is a (discretized) derivative, leading to non-invertible prior covariances. We discuss these practical issues in Part II of this paper series.

Because the optimization problem in Eq. 4 is stochastic, its solution, which we call the RTO solution, is also stochastic. It can be shown that if the modeling function $\boldsymbol{F}(\boldsymbol{m})$ is linear, then the distribution of the RTO solutions is equal to the posterior distribution, see Bardsley et al. (2014), or the Appendix A, where we also show that both perturbations, to the data and the prior, are indeed required (which may be surprising, but also familiar from ensemble Kalman filtering). If the model is nonlinear, the distribution of the RTO solutions is not equal to the posterior distribution and we refer to the discrepancy between the RTO sampling distribution and the targeted posterior distribution (Eq. 2) as a "bias."

This bias can be removed by introducing an accept/reject step. There are strong computational reasons not to do this, however. First, even an optimally tuned accept/reject criterion inevitably rejects a significant portion of proposed models. But the most important issue here is parallelism. The accept/reject process turns RTO into a serial algorithm, because $\boldsymbol{m}_{i+1}$ is accepted or rejected by comparing its posterior probability relative to that of $\boldsymbol{m}_i$. This makes each step dependent on the previous step. If instead the accept/reject step is omitted and every RTO model is accepted into the ensemble, each sample is independent of the others, and the algorithm becomes truly parallel – any number of CPUs can be utilized to draw RTO samples independently of one another. Because the algorithm is scalable in this way, the run time required to obtain a number of samples sufficient to adequately estimate the model parameter uncertainty can be reduced to the degree desired or to the limit of the HPC resources available. This means that, in theory, UQ can be obtained for problems of arbitrary size so long as the solution to (4) can be obtained. In this case, the limiting factor is the amount of HPC resources available, not a prohibitively long run time.

While there is no theory for the conditions under which the bias should be small, a small bias has been observed in very different applications, ranging from numerical weather prediction to reservoir management (Bonavita et al. 2012; Oliver 2017). Here, we follow this lead and verify a small bias in toy problems and in 1D field data inversions, where we can compare to an unbiased solution produced using traditional MCMC. In Part II, we demonstrate the usefulness of the RTO solution in 2D by solving a 2D magnetotelluric (MT) problem. In these cases, the computational limitations of traditional MCMC are such that the RTO solutions

for $\underline{i = 1 : N_{\text{samples}}}$ **do**
  Draw perturbed data set: $\tilde{\boldsymbol{d}} \sim \mathcal{N}(\boldsymbol{d}, C_d)$
  Draw prior model: $\tilde{\boldsymbol{m}} \sim \mathcal{N}(0, \frac{1}{\mu}(L^T L)^{-1})$
  Solve Eq. 4 to get the model $\boldsymbol{m}_i$
end

Algorithm 1: The RTO algorithm is remarkably simple: repeatedly minimize a perturbed objective function. Since all perturbed data and prior models are independent, computing the models $\boldsymbol{m}_i$ is embarrassingly parallel (executing the for loop in parallel, using as many cores as are available)

cannot be compared to the unbiased, target Bayesian posterior. We can, however, compare the RTO solution to regularized inversions and (computationally challenging, but feasible) trans-D MCMC inversions.

The RTO algorithm described above (and summarized in pseudo-code in Algorithm 1) is delightfully simple. The simplicity of RTO is such that only minor modifications are needed to turn an existing regularized inversion code into a RTO sampler. This means that nonlinear uncertainty estimation should be easily accessible wherever robust and efficient regularized inversions are available, such as in EM inversion where tools such as MARE2DEM (Key 2016) and ModEM (Kelbert et al. 2014) have found wide use. We also emphasize that RTO does not require any tuning beyond what is needed to perform a single, deterministic inversion.

Generating perturbed data and prior models within the RTO is straightforward. With respect to the perturbed data, we note that C_d is often diagonal so that generating perturbed data, $\tilde{\boldsymbol{d}}$, is trivial. Generating a prior model, $\tilde{\boldsymbol{m}}$, is only slightly more complicated and amounts to a linear solve:

$$\sqrt{\mu} L \tilde{\boldsymbol{m}} = \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(0, I) \tag{6}$$

where I is the identity matrix with the same size as L. If desired, one can also incorporate bounds on the model parameters and deal with singular L by solving Eq. 6 with a constrained least-squares solver (see Part II for more details).

3.2 Numerical illustrations of RTO on toy problems

To illustrate RTO and to give insight into how RTO produces an ensemble of high probability models, we consider several simplified, one- or two-parameter toy problems. The toy problems are nonlinear and most are characterized by more than one mode. This demonstrates, among other things, that RTO is not a linear (or nearly linear) method and that it is capable of

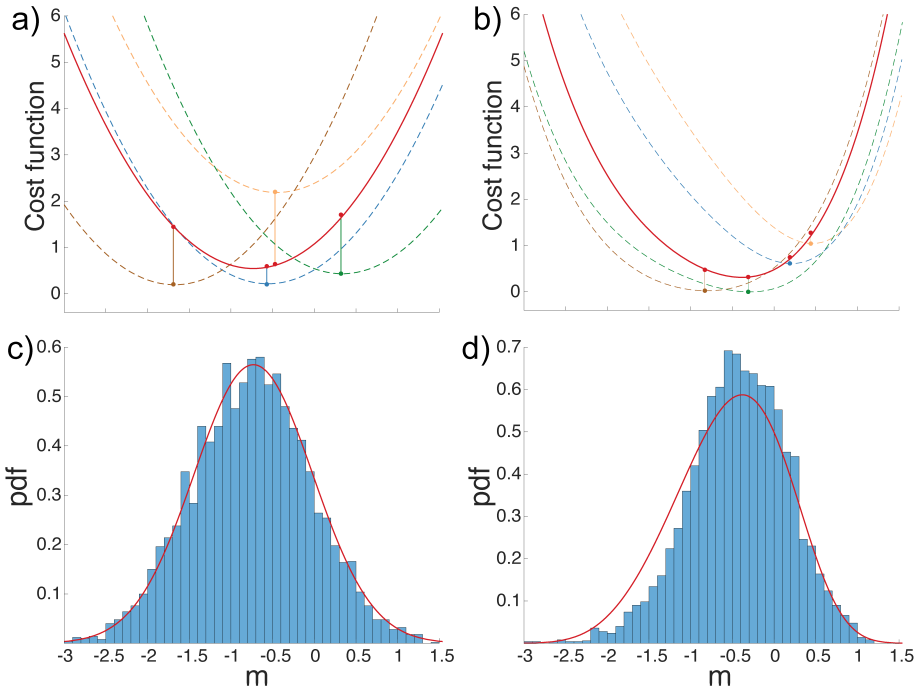

Figure 1. RTO generates samples of high posterior probability by repeatedly perturbing and optimizing an objective function. (a-b) The minimizers of perturbed objective functions (dashed lines) in the case of linear and nonlinear modeling functions, respectively, are clustered around the minimizer of the unperturbed objective function (solid line). (c-d) The ensemble of RTO samples is distributed according to the Bayesian posterior (solid line) in the linear case, while in the nonlinear case there is a small bias.

accurately sampling from complex posterior probability landscapes. Readers familiar with RTO or the RTO literature can safely skip this section as it largely serves as an RTO tutorial for geophysics.

3.2.1 Linear and mildly nonlinear problems

We first consider a linear and a nonlinear model

$$F_{\rm l}(m) = m + \frac{1}{2}, \quad F_{\rm nl}(m) = \frac{1}{2}m^2 + m. \tag{7}$$

For both models, we chose $\tilde{d} \sim \mathcal{N}(-1, 1)$ and $\tilde{m} \sim \mathcal{N}(0, 1)$. Consider first the linear model. Fig. 1a shows the unperturbed objective function (Eq. 1, solid red line) and several examples of randomly perturbed objective functions (Eq. 4, dashed lines). The solutions m of the perturbed objective functions are shown as colored dots. Their relative cost in terms of the unperturbed objective function is also shown. RTO makes use of deterministic inversion to

estimate model parameter uncertainty by sampling models away from the minimum of the canonical, unperturbed objective function, but that are still compatible with the data and prior model assumptions, expressed through regularization. Fig. 1c shows a histogram of 5,000 RTO samples along with the Bayesian posterior distribution (in red). The distribution of models generated by RTO matches the Bayesian posterior in the linear case, as it should.

Fig. 1b shows perturbed objective functions corresponding to the nonlinear model. The objective functions are not parabolas (as in the linear case), but the same principles apply. In Fig. 1d we show a histogram of 5,000 RTO samples along with the true Bayesian posterior distribution (in red). We see that RTO is sampling from a distribution very similar, but not exactly equal, to the Bayesian posterior, which means that the bias introduced by neglecting the accept/reject step is small in this example.

3.2.2 Bimodal and multi-modal posterior distributions

The next example is adapted from Wang et al. (2018) and demonstrates that RTO (without an accept/reject step) can handle bimodal distributions. Specifically, we consider the posterior distribution defined by

$$p(m|d) = \exp\left(-\frac{1}{2}\left(\frac{d-m^2}{\sigma}\right)^2 - \frac{1}{2}(m-M)^2\right), \tag{8}$$

where the datum is $d = 1$, the data variance is $\sigma^2 = 0.4^2$ and the prior mean is $M = -1$ (the prior variance is set to one). We apply RTO, initializing each optimization by drawing from the prior ($\mathcal{N}(-1, 1)$), and compare the results to those from two different MCMC samplers—an RWM sampler (tuned) and the "emcee" ensemble sampler of Goodman & Weare (2010), an affine invariant sampler.

We run all samplers at "constant cost." We estimate the cost of one RTO sample by the number of iterations needed during the associated optimization. The cost of an RWM step is taken to be that of a single function evaluation (thus not accounting for the cost of initial tuning, which is irrelevant in this toy problem, but which is substantial in applications). The cost of the ensemble sampler is estimated as a single function evaluation per "walker" (ensemble member) and we set the number of walkers to four. (The number of walkers for emcee must be larger than the number of parameters, which means that this algorithm does not scale to high-dimensional problems, but emcee is an effective method for low-dimensional problems).

Figure 2 summarizes our numerical experiments and shows histograms of the samples the three algorithms produced (all samplers are run at an equal computational cost). Panels (a)-

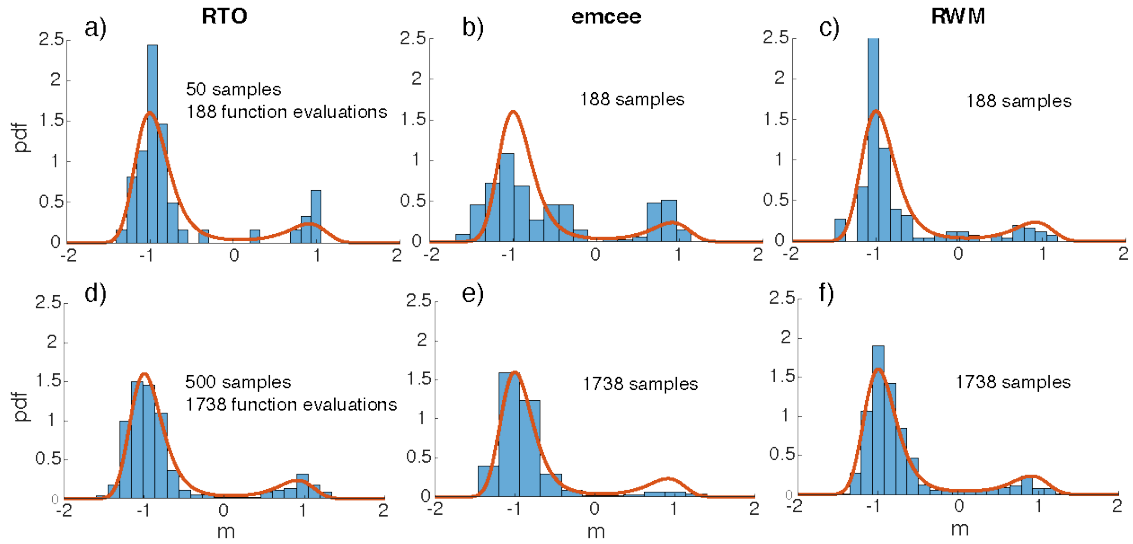

Figure 2. RTO (a,d), emcee (b,e) and RWM (c,f) applied to a bimodal posterior distribution (red). All samplers are run at constant cost. (a)-(c) simulate a resource-constrained scenario, while (d)-(f) simulate the case where samples can be easily drawn. The bias introduced by RTO is comparable to the bias introduced by running asymptotically unbiased samplers (RWM and emcee) at finite chain length.

(c) simulate a computationally resource-constrained situation, where due to the size of the problem the number of samples drawn was insufficient to achieve asymptotic convergence of MCMC. Panels (d)-(f) simulate a scenario where far more samples can be drawn.

Examining panels (a)-(c), we note three important points. First, RTO succeeds in sampling a bimodal distribution, demonstrating that the algorithm can handle strongly nonlinear problems. Second, comparing RTO to MCMC (emcee and RWM) we note that RTO achieves a comparable solution with significantly fewer samples. In particular, the bias between the sampling distribution (blue histograms) and the target distribution (red lines) is no worse for RTO than for MCMC. Third, the bias between the MCMC sampling distribution (which is guaranteed to converge to the correct solution as the number of samples becomes infinite) is significant for finite length chains. We point out here the obvious constraint that, in practice, MCMC samplers will *always* be run at finite chain lengths, and that this unavoidable bias in the MCMC posterior is difficult to estimate, especially for large problems where the practical limitations in chain length are most severe (panel (b) shows this clearly).

In panels (d)-(f), where computational constraints are less severe, the MCMC bias is reduced. But so is the RTO bias, which despite drawing fewer samples is no worse than that

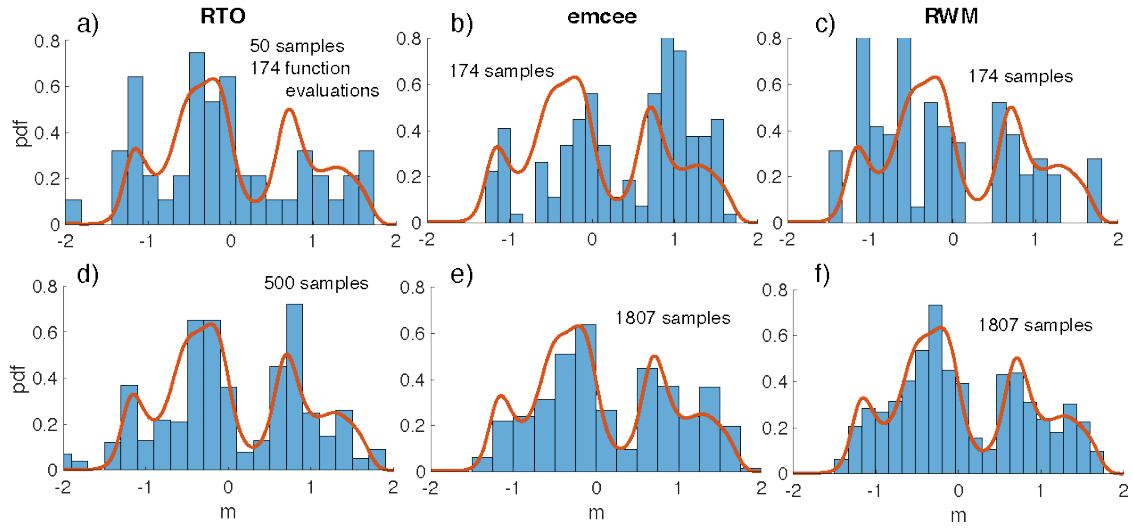

Figure 3. RTO (a,d), emcee (b,e) and RWM (c,f) applied to a multi-modal posterior distribution (red). All samplers are run at constant cost. (a)-(c) simulate a resource-constrained scenario, while (d)-(f) simulate the case where samples can be easily drawn. The bias introduced by RTO is comparable to the bias introduced by running asymptotically unbiased samplers (RWM and emcee) at finite chain length. As sample size grows, the MCMC bias narrows while a small RTO bias remains

of finite length MCMC chains. Panel (d) in particular demonstrates that RTO can converge to a distribution that is very similar to the target posterior, even for strongly nonlinear problems.

We further consider a problem with an even more complex and multi-modal posterior distribution defined by

$$p(m|d) = \exp\left(-\frac{1}{2}\left(d - F(m)\right)^2 - \frac{1}{2}m^2\right), \tag{9}$$

where the data are $d = 1$ and the model is $f(m) = m^2 - \sin(5m)$. We apply the same numerical experiments as above, simulating a computationally-constrained scenario and a scenario where computational constraints are less severe (Figure 3).

The first point to note from panels (a) and (d) of Figure 3 is that RTO succeeds in sampling from a strongly nonlinear target distribution. In the resource-constrained scenario—panels (a)-(c)—we again see that the bias is significant in the sampling distributions obtained via MCMC methods as well as in the RTO sampling distribution, but that the RTO bias is no worse than that obtained via MCMC.

In the resource-unconstrained scenario—panels (d)-(f)—we see the MCMC bias diminishing as the number of samples grows. This is expected, since MCMC enjoys guarantees of asymptotic convergence, while some finite RTO bias always remains. These toy examples illustrate that if a problem is small enough for MCMC to handle, it remains the preferred

method. If, however, the number of parameters to be sampled is large and the cost of the forward problem is high, as is frequently the case in typical EM geophysics problems, RTO may be the only feasible option. Importantly, this conclusion remains true even for strongly nonlinear problems. The RTO distribution shown in panel (d), while biased, is without a doubt a large improvement over a single point estimate, which is standard procedure in EM geophysics today.

3.2.3 Two-parameter toy problem

The computational advantages of RTO become more apparent in high dimensional problems (which is what this paper is about), and can be illustrated even by a two-parameter problem, adapted from Bardsley et al. (2014). Here, the model is

$$F(\mathbf{m}) = m_1 \left(\mathbf{1} - \exp(-m_2 \mathbf{x})\right). \tag{10}$$

where m_1, m_2 are the components of the model parameter vector $\mathbf{m}$, and where $\mathbf{x} = [1, 3, 5, 7, 9]^T$ are "locations." The symbol $\mathbf{1}$ denotes a 5×1 vector whose elements are all equal to one; the exponential function is applied element-wise to a vector. The data are $\mathbf{d} = [0.076, 0.258, 0.369, 0.492, 0.559]^T$ and the data variance is $\sigma^2 = 0.014^2$ (Bardsley et al. 2014). The posterior distribution is

$$p(\boldsymbol{m}|\boldsymbol{d}) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{F}(\boldsymbol{m}) - \boldsymbol{d}\right\|^2 - \frac{1}{2}\left\|\boldsymbol{m}\right\|^2\right). \tag{11}$$

We again run the same numerical experiments as before on this two-parameter problem and compare RTO to emcee and RWM. As a reference solution (which is not as easy to compute as in the one-parameter problems), we show the result of the emcee sampler with 10^6 steps (for each of the four walkers, resulting in $4 \cdot 10^6$ overall samples – we carefully verified that this sampler is close to convergence with this large number of samples). The results of the numerical experiments are summarized in Figure 4, where we show the approximations of one-dimensional marginal distributions ($p(m_1|\mathbf{d})$ and $p(m_2|\mathbf{d})$).

The same conclusions as before are again apparent. In the resource-constrained scenario, RTO provides a reliable—if biased—estimate of uncertainty with relatively few samples while the MCMC sampling distributions exhibit strong bias when estimated with finite length chains. In the resource-unconstrained scenario, the MCMC samplers slowly converge to a low-bias solution while the RTO uncertainty estimate remains biased (although not beyond usefulness). Again, we conclude that if the size of the problem is amenable to MCMC sampling, this methods is preferable. RTO, however, has an edge on unbiased samplers if the number of

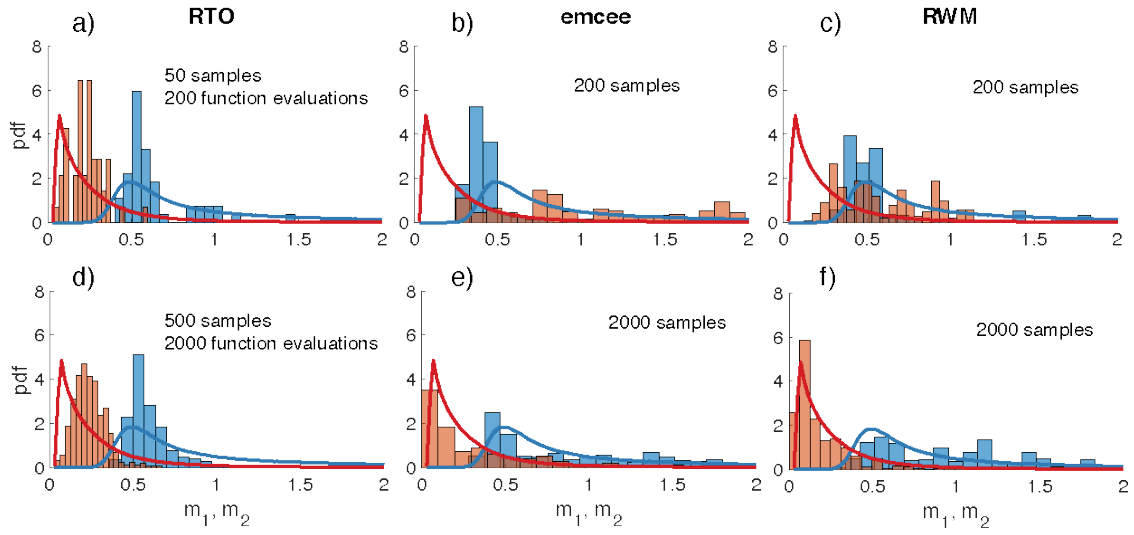

Figure 4. RTO (a,d), emcee (b,e) and RWM (c,f) applied to a two-parameter posterior distribution. The reference solutions (red and blue lines) are obtained by running the emcee sampler for 10^6 steps per walker. The blue histograms are approximations of the marginal $p(m_1|\mathbf{d})$, and the red histograms are approximations of $p(m_2|\mathbf{d})$. All samplers are run at constant cost. In the resource-constrained scenario,(a)-(c), the bias introduced by RTO is smaller that the bias introduced by running asymptotically unbiased samplers (RWM and emcee) at finite chain length. With far more samples, the bias in the MCMC sampling distributions diminishes, as expected, while the RTO bias remains.

samples is severely limited—which is nearly always the case in the high-dimensional problems that are relevant to EM geophysics and many other fields of earth science.

3.3 RTO-TKO: Hierarchically sampling the regularization penalty weight

So far, we have assumed that an appropriate regularization penalty weight μ is known *a priori* (see equation 1) or can be reliably determined. In practice, however, this is challenging. One solution is to use the well known trade-off between data fit and model regularization to find the maximum value of μ consistent with a specified/desired level of data misfit (e.g., RMS 1.0). This, indeed, is how the Occam inversion determines the regularization strength. Using data misfit to determine a minimum optimal value for μ, however, is not possible since there is always room to improve the data fit by further reducing the regularization strength. Eventually, however, this will cause the optimization to start fitting the data noise, resulting in spurious models. In addition, as the regularization strength decreases the objective function becomes very complex with many shallow local minima that make finding a solution challenging. As a result, the maximum μ consistent with a specified RMS is a common choice for

the regularization strength. This amounts to a prior bias toward highly regularized (smooth) models, however, which has a corresponding effect on the posterior (it tends to lower the posterior variance).

An alternative strategy is to use a hierarchical Bayesian framework where μ is treated as an unknown and a posterior distribution is defined over the model $\boldsymbol{m}$ and the regularization penalty μ:

$$p(\boldsymbol{m}, \mu|\boldsymbol{d}) \propto p(\boldsymbol{d}|\boldsymbol{m}, \mu)p(\boldsymbol{m}|\mu)p(\mu). \tag{12}$$

Here, $p(\mu)$ is a prior on μ and the remaining terms are as before, but now conditional on μ. To avoid guiding the inversion towards a (somehow) pre-determined μ, the prior for μ should be chosen to be "wide," e.g., a uniform distribution over a large interval, or other distributions with a large support (see also Part II of this paper series).

Writing the "hierarchical" posterior in terms of the model and regularization, we get

$$p(\boldsymbol{m}, \mu|\boldsymbol{d}) \propto \mu^{n/2} \exp\left(-\frac{1}{2}\left\| C_d^{-1/2}\Big(\boldsymbol{F}(\boldsymbol{m}) - \boldsymbol{d}\Big)\right\|^2 - \frac{\mu}{2}\left\| L\boldsymbol{m}\right\|^2\right) p(\mu) \tag{13}$$

where n is the number of model parameters (e.g. the number of layers in a 1D parametrization). We note that the discretization of the forward model appears explicitly in the posterior distribution via the factor $\mu^{n/2}$. This means that the discretization of the model has a direct influence on the posterior uncertainty. This dependence of the solution on the number of model parameters is particularly strong in 2D and 3D geometries where n is very large.

To avoid this issue, we formulate the problem to be grid invariant, so that the discretization becomes irrelevant to the solution. This is achieved by a nonlinear change of variables

$$\boldsymbol{\xi} = \sqrt{\mu} L\boldsymbol{m}, \tag{14}$$

so that the posterior distribution becomes

$$p(\boldsymbol{\xi}, \mu|\boldsymbol{d}) \propto \exp\left(-\frac{1}{2}\left\| C_d^{-1/2}\left(\tilde{\boldsymbol{F}}(\boldsymbol{\xi}, \mu) - \boldsymbol{d}\right)\right\|^2 - \frac{1}{2}\left\|\boldsymbol{\xi}\right\|^2\right) p(\mu), \tag{15}$$

where

$$\tilde{\boldsymbol{F}}(\boldsymbol{\xi}, \mu) = \boldsymbol{F}\left(\frac{1}{\sqrt{\mu}} L^{-1}\boldsymbol{\xi}\right). \tag{16}$$

The change of variables, first described in (Stewart 2010; Chen et al. 2018; Dunlop et al. 2020), thus "normalizes" the model prior to be the standard Gaussian distribution, independent of μ. This comes at the expense of making the modified forward model, $\tilde{\boldsymbol{F}}$, a function of the regularization parameter. Most importantly, however, the dependence on the number of model parameters n disappears after changing variables.

The numerical solution of the hierarchical problem consists of drawing samples $\boldsymbol{m}^i$, μ^i,

$i = 1, \ldots, N_{\text{samples}}$, from the hierarchical posterior in Eq. 15. We do this via Gibbs sampling, i.e., we sample, in sequence, the conditionals

$$p(\boldsymbol{\xi}^{i+1}|\boldsymbol{d}, \mu^i) \propto \exp\left(-\frac{1}{2}\left\| C_d^{-1/2}\left(\tilde{\boldsymbol{F}}(\boldsymbol{\xi}^{i+1}, \mu^i) - \boldsymbol{d}\right)\right\|^2 - \frac{1}{2}\left\|\boldsymbol{\xi}^{i+1}\right\|^2\right), \tag{17}$$

$$p(\mu^{i+1}|\boldsymbol{d}, \boldsymbol{\xi}^{i+1}) \propto \exp\left(-\frac{1}{2}\left\| C_d^{-1/2}\left(\tilde{\boldsymbol{F}}(\boldsymbol{\xi}^{i+1}, \mu^{i+1}) - \boldsymbol{d}\right)\right\|^2\right) p(\mu^{i+1}), \tag{18}$$

where i is the iteration number in the Markov chain. It can be shown (and it is well-known), that a Gibbs sampler accepts every sample by design so that there is no need to evaluate an accept/reject probability. One iteration of a Gibbs sampler, which generates one pair of samples ($\boldsymbol{m}^{i+1}$, μ^{i+1}), amounts to:

(i) compute $\boldsymbol{\xi}^{i+1}$, for a fixed regularization weight μ^i, by sampling from $p(\boldsymbol{\xi}^{i+1}|\boldsymbol{d}, \mu^i)$;

(ii) compute μ^{i+1}, using $\boldsymbol{\xi}^{i+1}$ from step (i), by sampling $p(\mu^{i+1}|\boldsymbol{d}, \boldsymbol{\xi}^{i+1})$

We apply the RTO idea to both of these steps. For step (i), we note that because μ is constant in Eq. 17, the first step of the Gibbs sampler, sampling from $p(\boldsymbol{\xi}^{i+1}|\boldsymbol{d}, \mu^i)$ can be done by optimizing Eq. 4. Step (i) thus amounts to performing one RTO step as in the previous section.

For step (ii), we recognize that, following the core idea behind RTO, sampling from Eq. 18 can be done by optimizing the objective function defined by the negative logarithm of $p(\mu|\boldsymbol{d}, \boldsymbol{\xi})$:

$$\min_{\mu} \quad \frac{1}{2}\left\| C_d^{-1/2}\left(\tilde{\boldsymbol{F}}(\boldsymbol{\xi}^{i+1}, \mu) - \tilde{\boldsymbol{d}}\right)\right\|^2 - \log\left(p(\mu)\right), \tag{19}$$

where the data perturbation $\tilde{\boldsymbol{d}}$ is drawn from the same distribution as before (note that for steps (i) and (ii) separate instances of $\tilde{\boldsymbol{d}}$ must be drawn), and where the perturbation applied to the prior for μ depends on the choice of prior. To keep things simple, we now choose a uniform prior for μ, which implies that the optimization problem becomes

$$\begin{aligned} \min_{\mu} \quad & \tfrac{1}{2}\left\| C_d^{-1/2}\left(\tilde{\boldsymbol{F}}(\boldsymbol{\xi}^{i+1}, \mu) - \tilde{\boldsymbol{d}}\right)\right\|^2 \\ \text{s.t.} \quad & \mu_{\text{l}} \leq \mu \leq \mu_{\text{u}} \end{aligned} \tag{20}$$

where μ_{l} and μ_{u} are lower and upper (prior) bounds for μ.

In order to optimize Eq. 20, we need to express $\tilde{\boldsymbol{F}}(\boldsymbol{\xi}^{i+1}, \mu)$ in terms of $\boldsymbol{F}(\boldsymbol{m}^{i+1}, \mu)$. This can be done via the change of variables in Eq. 16 and by recognizing that, in step (ii), $\boldsymbol{\xi}^{i+1}$ is

for $\underline{i = 0 : N_{\text{samples}}}$ **do**

1. Solve stochastic optimization problem for $\boldsymbol{m}^{i+1}$ *at constant* μ^i

Draw perturbed data set: $\tilde{\boldsymbol{d}} \sim \mathcal{N}(\boldsymbol{d}, C_d)$

Draw prior model: $\tilde{\boldsymbol{m}} \sim \mathcal{N}(0, \frac{1}{\mu}(L^T L)^{-1})$

Solve Eq. 4 to get $\boldsymbol{m}^{i+1}$

2. Solve stochastic optimization problem for μ^{i+1} *at constant* $\boldsymbol{m}^{i+1}$

Draw another perturbed data set: $\tilde{\boldsymbol{d}} \sim \mathcal{N}(\boldsymbol{d}, C_d)$

Draw a prior μ (depends on p(μ))

Solve Eq. 19 to get μ^{i+1}

end

Algorithm 2: The RTO-TKO algorithm consists of two RTO steps, one for sampling the model parameters at fixed μ, the other for sampling μ for a fixed model.

held constant. Thus, we have

$$\tilde{\boldsymbol{F}}\left(\boldsymbol{\xi}^{i+1}, \mu^{i+1}\right) = \boldsymbol{F}\left(\frac{1}{\sqrt{\mu^{i+1}}} L^{-1} \xi^{i+1}\right) \tag{21}$$

$$= \boldsymbol{F}\left(\sqrt{\frac{\mu^i}{\mu^{i+1}}} \boldsymbol{m}^{i+1}\right) \tag{22}$$

where we have used $L^{-1}\boldsymbol{\xi} = \sqrt{\mu}\boldsymbol{m}$ and the fact that, in step (i), $\boldsymbol{m}^{i+1}$ was determined for fixed μ^i. As a result, step (ii), the 'TKO' step of RTO-TKO, amounts to solving (for a uniform prior on μ)

$$\begin{array}{ll} \min\limits_{\mu} & \frac{1}{2}\left\| C_d^{-1/2}\left(\boldsymbol{F}\left(\sqrt{\frac{\mu_{ref}}{\mu}}\boldsymbol{m}\right) - \tilde{\boldsymbol{d}}\right)\right\|^2 \\ \text{s.t.} & \mu_{\text{l}} \le \mu \le \mu_{\text{u}} \end{array} \tag{23}$$

where μ_{ref} is the value of the regularization penalty used in step (i), the 'RTO' step. In Eq. 23, μ plays the role of a 'stretch factor', stretching (for small values of μ) or compressing (for large values of μ) a constant model $\boldsymbol{m}$. The optimization problem in the 'TKO' step (Eq. 23) seeks the value of the regularization penalty weight (stretch factor) that minimizes the misfit to the perturbed data set, $\tilde{\boldsymbol{d}}$. The Gibbs iterations can repeat until the desired number of samples ($\boldsymbol{m}^i$ and μ^i) is generated. The sampling algorithm, which we call RTO-TKO, is summarized in pseudo-code in Algorithm 2.

In summary, we use RTO twice within a hierarchical Gibbs sampling framework. This "one-two punch" generates samples $\boldsymbol{m}^i$ and then μ^i at each step of the algorithm. Since we use RTO without an accept/reject criterion, we introduce bias during both steps. We have already argued that the bias in step (i), the fixed-μ RTO step, can be expected to be small.

[H]

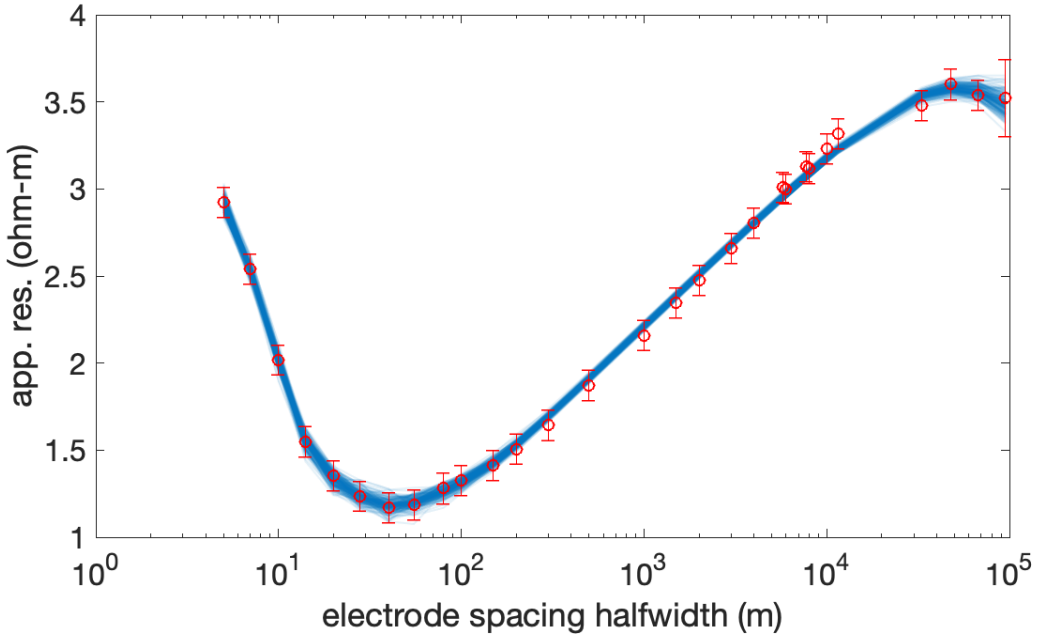

Figure 5. DC apparent resistivity as a function of electrode spacing halfwidth (Constable et al. 1984). The model responses of 200 randomly-selected models from an RTO-generated model ensemble with a fixed regularization penalty weight are shown in blue.

We show that this is indeed the case for a 1D DC resistivity data problem, and that the bias introduced by step (ii) (the one for sampling the regularization penalty μ) is also small. As before, we obtain large computational gains because we accept a small bias—the same reasoning as above applies, with a caveat for parallelism, which we discuss in Section 5. The small bias is why we call the sampler RTO-TKO, 'TKO' standing for "technical knock-out," rather than a "knock-out," which we reserve for an algorithm (as yet undiscovered) that can achieve large computational gains *without* bias.

4 RESULTS

In what follows, we demonstrate several desirable attributes of RTO-TKO by inverting DC resistivity data (Constable et al. 1984) for 1D models of subsurface electrical resistivity. Each subsurface model consists of a fixed grid of layers, each of which is assigned a resistivity value that is constant across the layer. The DC resistivity data and data uncertainty are shown in Fig. 5. The data covariance is assumed to be diagonal.

First, we demonstrate that the bias in RTO and RTO-TKO is small. We then demonstrate

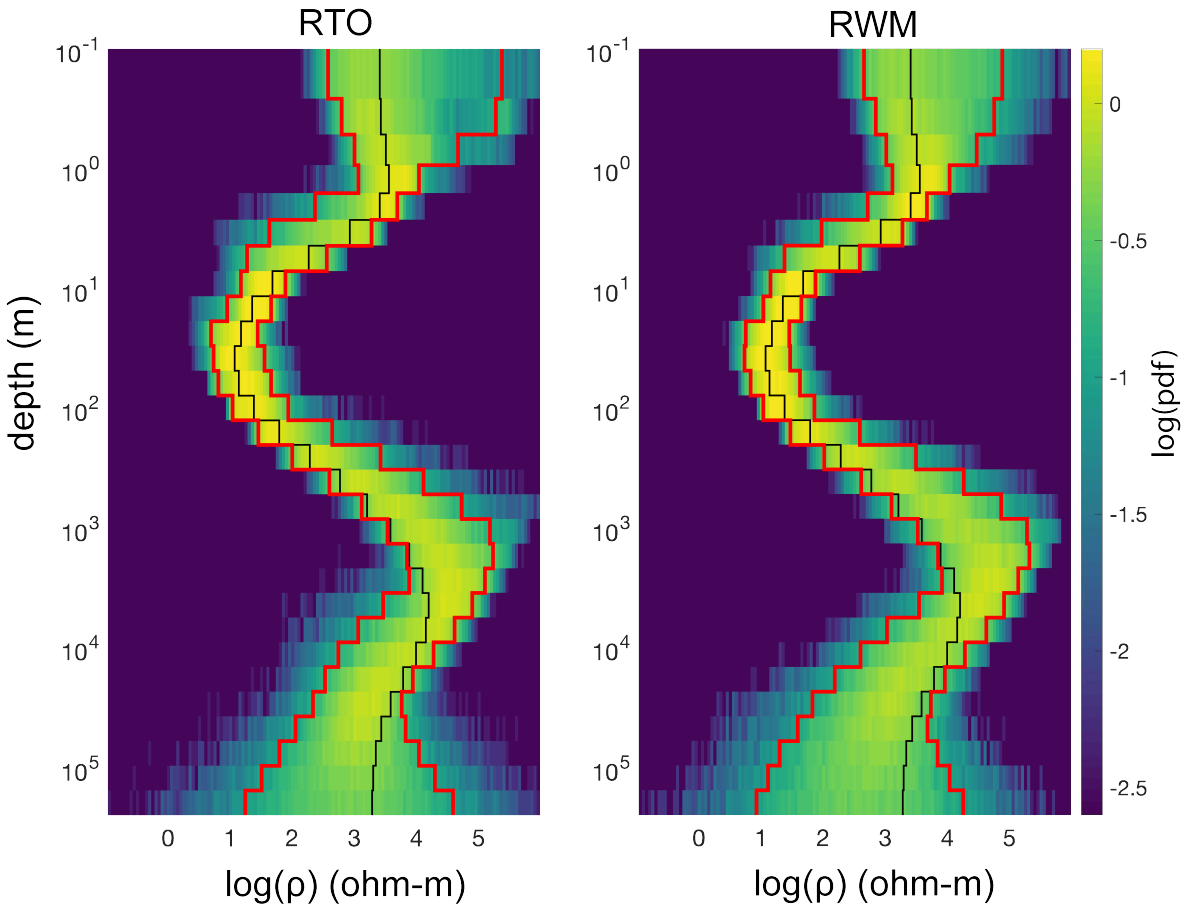

Figure 6. Distribution of model parameter uncertainty as a function of depth estimated using the RTO ensemble (left) and the RWM ensemble (right) at fixed $\mu = 3.2$. Warmer colors indicated regions of higher probability. The left and right red lines are the 5th and 95th percentiles of the distribution at each depth, respectively, and the black line is the Occam inversion result (Constable et al. 1987). While only the RWM distribution is guaranteed to converge to the target Bayesian posterior (for infinitely many samples), in practice the distribution of uncertainty produced from RTO models is often very similar.

the impact of μ on parameter uncertainty and how RTO-TKO makes an a priori choice of μ unnecessary (much in the spirit of an Occam inversion). In the final portion of this section, we show that the uncertainty estimated by RTO-TKO is unaffected as the parameter grid is successively refined.

4.1 Small RTO and RTO-TKO bias

We argue above that the distribution sampled by RTO is not the same as the Bayesian posterior (see toy problems in Section 3.2), but that the difference between the RTO solution

and the target posterior is often negligibly small. To demonstrate this on field data, we invert the DC resistivity data in Fig. 5 using RTO and we compare the result to the reference, target Bayesian posterior distribution produced by inverting the field data using traditional MCMC.

We first consider the case where the regularization parameter μ is fixed at 3.2 and approximate the Bayesian posterior distribution by applying RWM. RWM is asymptotically unbiased (as $N_{\text{samples}} \to \infty$) and we produce a large number of samples, since computational constraints are not an issue in this 1D example.

Fig. 6 shows the marginal distributions of posterior probability for electrical resistivity as a function of depth for both RTO and RWM (left and right plots, respectively). The RTO posterior uncertainty was estimated using 10,000 RTO samples, while the RWM posterior uncertainty results from five million RWM samples, of which every 500th sample was used to estimate the posterior. The differences between the posterior uncertainties estimated via RTO and RWM are small. In both panels of Fig. 6 the black line can be used as a reference since it represents the same inversion result—obtained using the Occam inversion method (Constable et al. 1987), which automatically selected a final regularization penalty weight of 19.8. We note that the Occam solution is well within the confidence intervals of the (approximately) Bayesian solution, but that it is not "centered" within the uncertainty of the Bayesian solution. We discuss this observation, and what it means for UQ via RTO and RTO-TKO in Part II of this paper series.

Fig. 6 represents a qualitative comparision of RTO with an unbiased technique. To quantitatively assess how similar the two posterior distributions of RTO and RWM are, we compute the Kullback-Leibler (KL) divergence. The KL divergence measures the dissimilarity between two probability distributions and is defined by (see, e.g., MacKay 2003)

$$D_{KL}(P||Q) = \sum_i P(x_i) \log\left(\frac{P(x_i)}{Q(x_i)}\right) \tag{24}$$

where D_{KL} is the KL divergence, P and Q are the distributions being compared, and i denotes the discrete bins in a numerically estimated distribution. If P represents an estimate of the posterior distribution made using RTO and Q the estimate made using RWM, then $D_{KL}(P||Q)$ is a quantitative measure of how dissimilar the two distributions in Fig. 6 are. For more on the use of the KL divergence in a Bayesian setting see, *inter alia*, Pinski et al. (2015); Blatter et al. (2018).

For each subsurface layer, we computed the KL divergence between the marginal RTO and RWM distributions. The median result was 0.63, while the sum over all 30 layers was 27.3. Because the KL divergence is unitless, for comparison we also computed the KL divergence

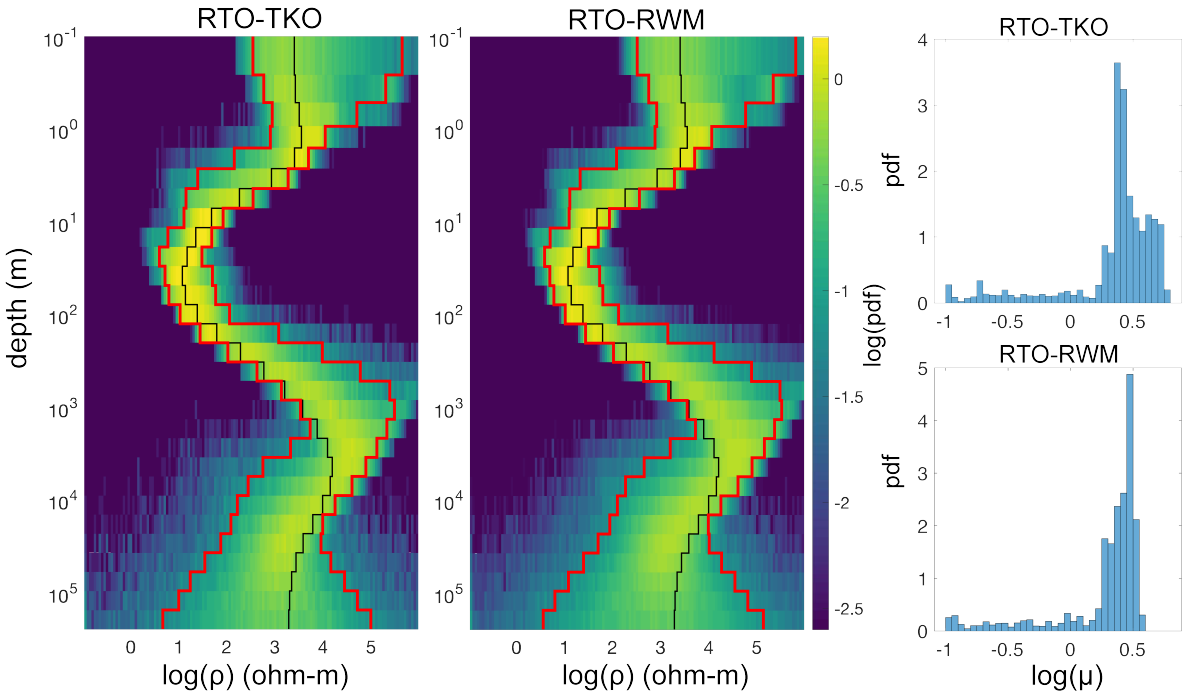

Figure 7. The posterior model parameter uncertainty estimated from the field DC resistivity data using RTO-TKO (left) is highly similar to that estimated using RTO-RWM (center). The regularization penalty weight distributions for RTO-TKO and RTO-RWM (right) are likewise very similar. Once again, the red lines delineate the 90% credible interval and the black line is the Occam inversion result (identical in both panels).

between the RWM distribution and a uniform distribution (using the same histogram binning). The median and sum over all the layers were 87.3 and 2,631, respectively. The two orders of magnitude lower KL divergence between the RTO and RWM results show these distributions are quite similar.

We now consider a hierarchical setup in which the regularization penalty μ is not fixed and show that RTO-TKO introduces only a small bias in the DC resistivity field problem. As before, we need to compute a reference solution to which the RTO-TKO solution can be compared. We found it difficult to compute a truly unbiased Bayesian solution. Using a RWM algorithm to sample the hiearchical posterior distribution over both the model and regularization weight is extremely slow to converge and convergence is very difficult to assess, even in this relatively simple problem. Likewise, affine invariant MCMC (the emcee algorithm) proved impractical due to the need for a very large number of walkers. We thus decided to use a "proxy" as a reference solution, which mimics the setup of the RTO-TKO. Specifically, we use a Gibbs sampling framework and use RTO for step (i) to sample the model (which

we just showed has small bias). For step (ii) of the Gibbs sampler, we us an (asymptotically unbiased) RWM method to sample the regularization penalty μ. The proxy, which we refer to as RTO-RWM, thus exposes additional bias due to the second RTO step when sampling the regularization penalty μ.

The posterior probability distributions obtained via RTO-TKO and RTO-RWM are shown in Fig. 7. By visual inspection, the differences between them are difficult to detect. The quantitative KL divergence tests likewise reveal that the distributions obtained via RTO-TKO and RTO-RWM are very similar. The median KL divergence between the two distributions over all 30 layers was 0.2 while the sum over all layers was 6.6. The median KL divergence between the RTO-RWM distribution and a uniform distribution, meanwhile, was 83.5, while its sum over all layers was 2,521.7.

We have shown that the RTO-TKO 'bias' is small (both qualitatively and quantitatively) for highly nonlinear, multimodal one- and two-parameter problems, as well as for the 1D DC resistivity method. We caution, however, that this does not necessarily guarantee that the RTO-TKO bias will be acceptably small for all other methods, since RTO still lacks a general mathematical theory describing the size of the bias. As such, we assert that RTO-TKO works well for EM geophysics problems and suggest that further work needs to be done to verify this for other methods.

Finally, we demonstrate the need for a hierarchical formulation by showing the large impact an *a priori* choice of μ has on uncertainty estimates. To this effect, we invert the DC resistivity data using the RTO algorithm (fixed μ) for three different choices of the regularization penalty weight: $\mu = 1$, $\mu = 3.2$, and $\mu = 10$. Fig. 8 shows the posterior uncertainty estimates for these three inversions. All other inversion parameters were held constant. The largest value of the regularization penalty weight yields the tightest posterior uncertainty, as expected, while the smaller value permits much more model variability and hence larger variance. Put simply: nearly any level of posterior uncertainty can be achieved by chosing μ accordingly. The hierarchical setup, in which the regularization penalty μ is treated as an unknown, and which can be efficiently implemented via RTO-TKO, does away with *a priori* choice of regularization and, much in the spirit of an Occam inversion, determines the appropriate level of regularization during the inversion. It is of course also possible to determine the regularization parameter μ in some other way (e.g., via Occam) and then use the RTO (without TKO). The RTO-TKO is an alternative method that also reveals uncertainty in the regularization parameter which could be useful in practice.

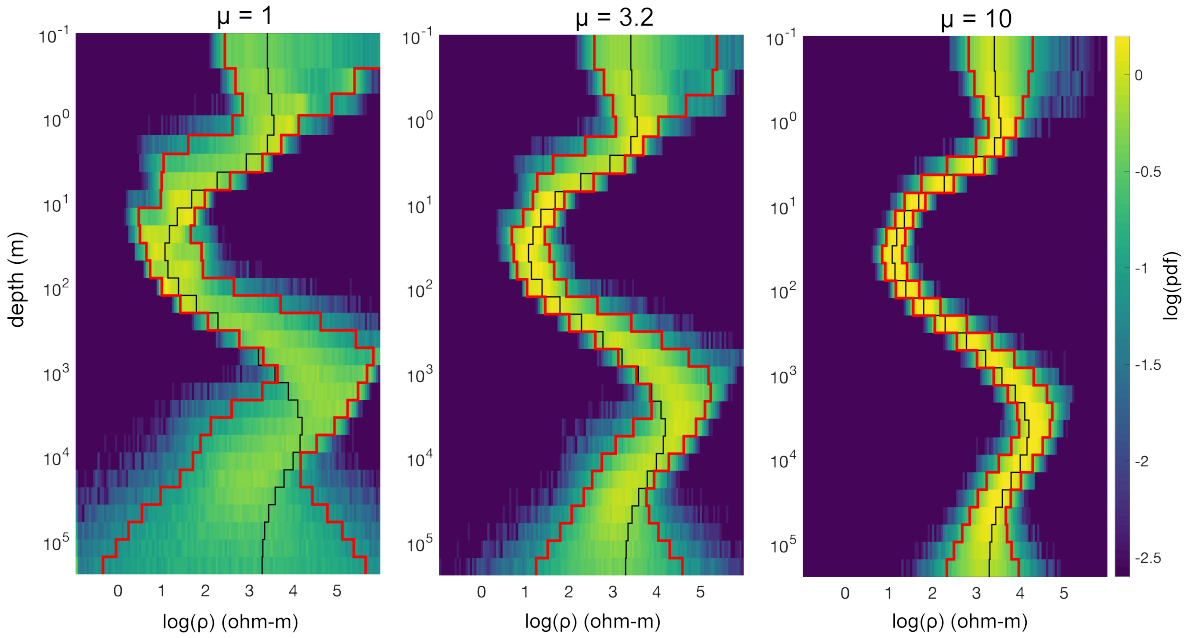

Figure 8. Model parameter uncertainty for the DC resistivity data estimated using RTO for three different choices of regularization penalty weight: 1, 3.2, and 10 (left, center, right, respectively). Similar to Fig 6, warmer colors indicate regions of higher probability, the left and right red lines are the 5th and 95th percentiles of the distribution at each depth, respectively, and the black line is the Occam inversion result. Smaller μ leads to larger posterior variance.

4.2 Grid invariance of RTO-TKO

RTO-TKO is designed to be grid invariant. This means that, so long as the number of model parameters is large enough to accommodate the structural features required to fit the data, increasing the number of parameters will not affect the overall solution, sampling efficiency, or posterior uncertainty. To demonstrate grid invariance, we inverted the DC resistivity data using RTO-TKO with 30, 60, and 120 layers. Fig. 9 shows the estimated posterior uncertainties, which do not exhibit significant differences in variance.

In fact, grid invariance has connections with trans-D MCMC algorithms that do not make use of any explicit model regularization. Instead, they rely on Bayesian parsimony (Malinverno 2002; MacKay 2003) as an implicit form of regularization: all else being equal, trans-D inversion prefers models with fewer parameters. Guided by this principle, trans-D MCMC algorithms allow the data (and model prior) to select the appropriate degree of model complexity (regularization). While RTO-TKO requires that a grid be specified and that a regularization term be present, it hierarchically samples the strength of the regularization. Guided by the principle of model smoothness (a form of parsimony), RTO-TKO similarly

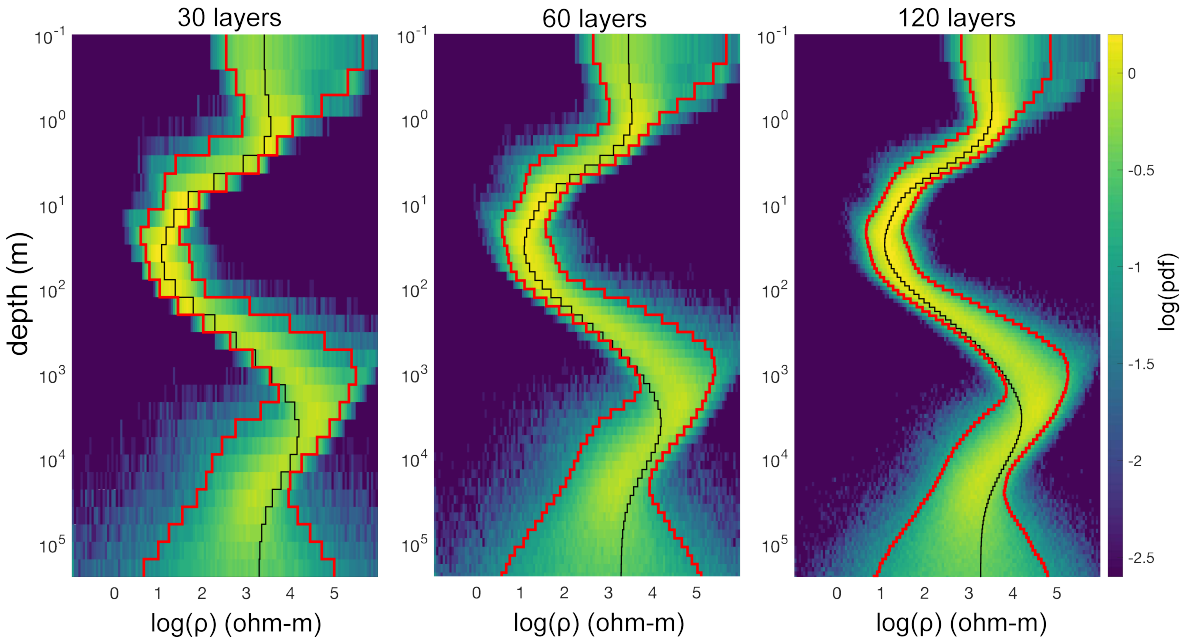

Figure 9. The posterior model parameter uncertainty estimated from the DC resistivity data using RTO-TKO is invariant under grid refinement. Inversions using 30, 60, and 120 layers do not exhibit significant differences in posterior variance. As usual, the black lines represent the respective Occam inversions.

lets the data and and model prior determine the appropriate degree of model complexity. And since adding additional model parameters through a refining of the parameter grid does not affect the RTO-TKO solution, the sampling distribution of μ does not affect the RTO-TKO estimate of the posterior uncertainty (provided the grid is fine enough to resolve all relevant scales needed to fit the data). RTO-TKO is computationally more efficient than trans-D MCMC, however, which suffers from long burn-in times and very long correlation times within the chain. We disuss these issue further in Part II of this paper series, where we also show a comparison of trans-D MCMC and RTO-TKO on 2D field data inversions.

4.3 Applicability of RTO-TKO

We have demonstrated that RTO-TKO produces meaningful UQ for the DC resistivity problem. In particular, we have shown that the RTO-TKO 'bias' is negligibly small for this problem by direct comparison to MCMC methods that were able to sample long enough (five million samples) that we could be reasonably certain they had asymptotically converged. In Part II of this series, we demonstrate a meaningful UQ for individual and joint inversion of MT and CSEM data, including a 2D MT problem. This represents a significant advancement in ca-

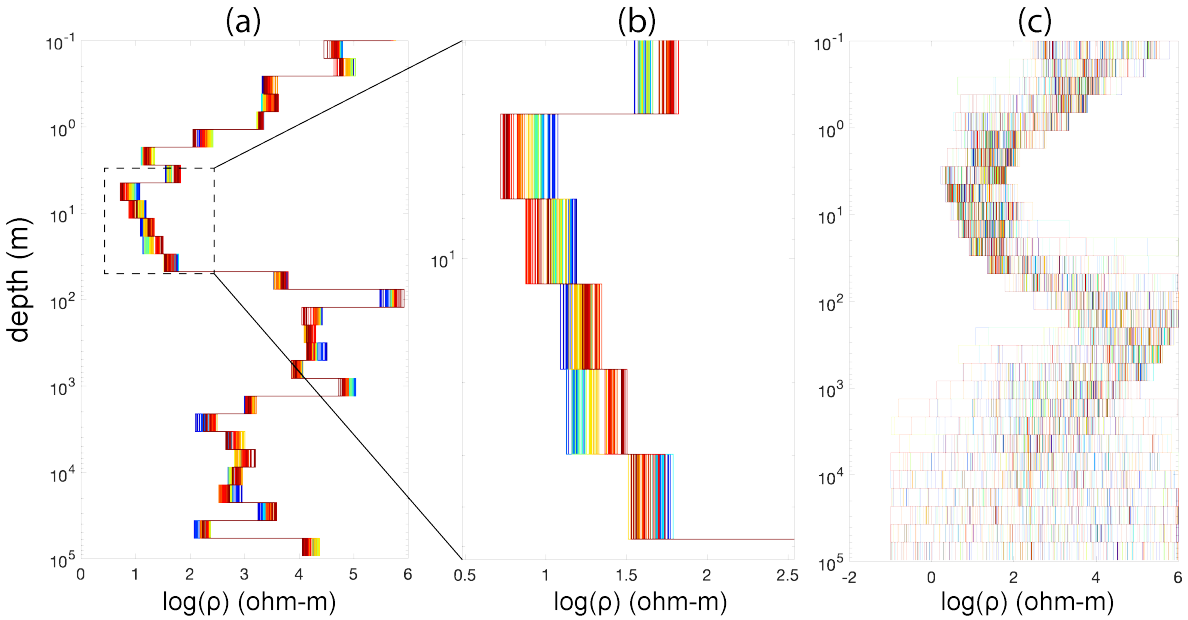

Figure 10. A randomly chosen sequence of 200 consecutive models drawn from hierarchical RWM (a) and RTO-TKO (c) model ensembles. Color indicates position in the sequence, with earlier models plotted in cooler colors while later models are shown in warmer colors. (b) The sequence from (a) is shown in greater detail. Note the slow, gradual shift in color in (a-b), indicating strong correlation, contrasted with the lack of correlation in (c).

pability for the EM geophysics community where to date the standard practice is to perform regularized inversions without UQ analysis.

We stress, however, that because no rigorous mathematical theory exists to stipulate the conditions under which the RTO-TKO bias is acceptably small, we cannot guarantee similar results for methods other than the ones we consider here—namely, those that belong to diffusive EM geophysics. Further careful analysis is needed to determine the breadth of geophysical methods to which RTO-TKO can usefully be applied.

5 COMPUTATIONAL ADVANTAGES OF RTO-TKO

One crucial advantage of RTO-TKO over serial MCMC samplers is that the RTO-TKO samples are (nearly) independent of one another while those drawn by MCMC samplers are not. The near-independence (rather than complete independence) is due to the hierarchical sampling over the scalar regularization parameter, μ. In fact, RWM (see above), has a low acceptance rate even after fine-tuning the method (about 0.234 for optimal tuning, see Roberts et al. 1997). For this reason, the models in the RWM Markov chain are highly correlated with

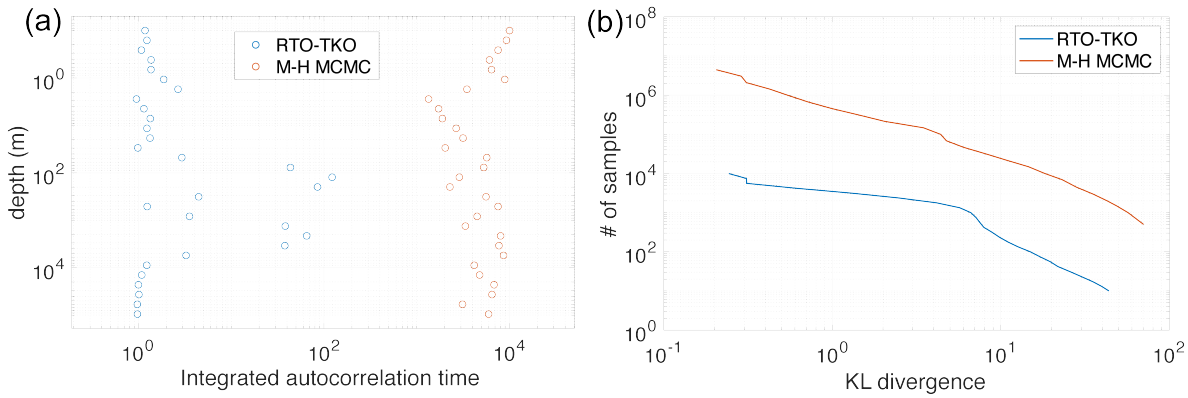

Figure 11. (a) Integrated autocorrelation time estimated for 10,000 RTO-TKO samples (blue) and 4 million RWM samples (orange) using the Gamma method of Wolff (2004). While the RTO-TKO samples are more or less independent (median IACT of 1.3), the RWM samples are highly correlated (median of 3,200). (b) Normalized KL divergence between partial and converged estimates of marginal posterior uncertainty as a function of the number of samples used to make the estimate, for both RTO-TKO (blue) and RWM (orange) model ensembles. The plotted values are averages over all model parameters. To achieve a similar level of divergence from the converged posterior estimate, RWM requires roughly 100x more samples.

one another. Fig. 10a-b shows a randomly chosen model sequence taken from a hierarchical RWM inversion of the DC resistivity data. The cooler colors indicate models taken at the beginning of the sequence while models at the end of the sequence are plotted in warmer colors. The color patterns slowly shift from blue to red, clearly showing a high degree of correlation between successive models. Over this 200 iteration sequence, the range of models sampled by the RWM is limited to a small neighborhood of the first model in the sequence. In addition, strong trade-offs are noticeable, evidenced by some layers smoothly transitioning from high to low resistivity while others do the opposite. The overall posterior uncertainty is not clearly visible in the range of models plotted in Fig. 10a-b.

By contrast, the overall posterior uncertainty is noticeable in Fig. 10c (compare with Fig. 7), which shows a 200 model sequence from an RTO-TKO inversion of the DC resistivity data. As before, color here also indicates relative position in that sequence, with cooler colors coming earlier and warmer colors later. The lack of any coherent color transition at any depth indicates that the RTO-TKO samples are (nearly) independent of one another (and that the hierarchical sampling over μ does little to disturb this independence). As such, RTO-TKO requires only a few samples to obtain a reasonable estimate of the posterior uncertainty (we investigate this in greater detail in Part II of this paper series).

To quantify the degree of correlation between models in both of these methods, we computed the integrated autocorrelation time (IACT) for a 10,000 model RTO-TKO ensemble and a four million model hierarchical RWM ensemble, both obtained by inverting the DC resistivity data. The IACT is a measure of the time (in algorithm steps) it takes for an algorithm to 'forget' where it started. In our case, it is a measure of the distance between independent samples in the ensemble. We computed the IACT using the Gamma method described in Wolff (2004). (Recall that IACT measures correlation between successive samples in the Markov chain, but not correlations within the model parameters, see, e.g., Sokal (1997)).

Fig. 11a shows the estimated IACT for RTO-TKO (blue) and RWM (orange) ensembles. It is not obvious how to interpret IACT in a multivariate setting, but while there is some fluctuation the RTO-TKO ensemble IACT hovers around 1, with a median of 1.3, meaning that for the most part the RTO-TKO samples are independent of one another. We emphasize that any degree of correlation between the models is introduced by the hierarchical problem formulation, in which we estimate the regularization penalty μ simultaneously with the model parameters (for a fixed regularization, RTO samples are provably independent). Notably, the model parameters for which the RTO-TKO IACT is greatest are located where the model is changing most rapidly, but even here the estimated RTO-TKO IACT (~ 100) is much less than the RWM IACT ($\sim 5,000$). In summary, the RWM samples exhibit an IACT roughly three orders of magnitude larger than RTO-TKO, with a median value around 3,200. While IACT is an imperfect measure of each sample's unique information content, it is clear that the RWM samples are more strongly correlated than the RTO-TKO samples, and, therefore, a large fraction of them are redundant. In the foregoing analysis, we are, of course, discussing the correlation across consecutive models generated using each algorithm, not the correlation within individual models produced by regularization.

Another measure of this redundancy is how fast (in terms of number of samples) it takes for a Bayesian sampling algorithm to converge to its target distribution. This can be measured using the KL divergence, where P now represents an estimate of the posterior distribution made using n consecutive samples of a model ensemble and Q represents another estimate of the same distribution made using $N > n$ samples sufficient to have converged. In this way, $D_{KL}(P||Q)$ is a measure of the remaining divergence between the partial estimate and the converged estimate.

To measure the number of samples required by RTO-TKO and hierarchical RWM to achieve an equivalent degree of convergence to their respective target distributions, we generated a second model ensemble of the same size as before (10,000 RTO-TKO samples and 4

million RWM samples) using each of these algorithms. We estimated the marginal distribution of probability density for each subsurface layer resistivity using the first n samples from these new model ensembles (P). We then computed the KL divergence between these distributions and the posterior estimated using the entirety of the respective first model ensembles (Q). All posterior distributions were estimated using the same binning. For each value of n, we then averaged the KL divergence over all the model parameters.

Fig. 11b shows the results. For a given level of divergence from the converged posterior estimate, RWM requires roughly two orders of magnitude more model samples. By this measure, each RTO-TKO sample contributes roughly 100x more unique information about model parameter uncertainty than each RWM sample.

Computationally speaking, the independence of its samples represents a major advantage of RTO-TKO over RWM. At first glance, RTO-TKO would seem to be considerably more expensive, since each sample requires solving an expensive optimization problem while RWM only requires a single forward computation per sample drawn. Yet Figs. 10 and 11 indicate that the information about model parameter uncertainty contributed by each additional RWM sample is largely redundant, while the independence of the RTO-TKO samples means each of its samples contributes new, independent information. As such, far fewer RTO-TKO samples need to be drawn, leading to a significant reduction in computational cost. Fig. 11 loosely suggests that so long as solving each stochastic optimization problem takes fewer than 100 forward computations, RTO-TKO should be faster to converge than MCMC in terms of total flops. This is explored in greater detail using a 2D MT field data set in Part II.

Yet efficiency in terms of total flops is not RTO-TKO's only advantage. The (near) independence of RTO-TKO samples means that we can leverage parallelism by running several RTO-TKO chains on different sets of CPUs. Further, there is no need for intercommunication between processors executing RTO sampling, so the RTO approach could run in parallel without any special parallel programming constructs or a fast network for CPU interconnections; the only necessary criterion is that the random number generators used for the data and model perturbations need to start with different seeds on each parallel process. Only when samples are (nearly) independent, can the result of several short chains run in parallel be comparable to the result of a single, long chain. Moreover, RTO-TKO does not suffer from a burn-in period as MCMC does. This means that the total run time required to draw a given number of samples can be reduced linearly with the number of CPUs simultaneously drawing RTO samples. Being able to efficiently leverage parallelism is increasingly important as HPC

resources are becoming less expensive but the speed of individual CPUs is no longer rapidly increasing.

Suppose a given problem requires 1 s for a forward evaluation and about 60 s for an RTO-TKO iteration (e.g., Occam + TKO) for a single compute node on a cluster. If 10,000 samples must be drawn to accurately approximate the parameter uncertainty, then a single compute node would take roughly 7 days to obtain the desired uncertainty estimate. If 100 compute nodes are available, however, the model parameter uncertainty could be obtained in less than two hours. On the other hand, if one RWM sample can be drawn in 1 s (the cost of a forward evaluation), but one million samples are required to ensure convergence, then the time required to invert the data using RWM would be approximately 12 days. We present these ideas in more detail in Part II where we show that 2D Bayesian sampling is indeed feasible to do in a single day with RTO-TKO and a (modest) computer cluster.

In making this computational comparison, it is not our intent to suggest that RTO-TKO should replace MCMC where it is computationally practical to use the latter. Rather, we are suggesting that RTO-TKO can be used to provide meaningful UQ where MCMC cannot, for practical reasons, be used at all. We have shown (and will further demonstrate in Part II) that RTO-TKO works well in efficiently obtaining meaningful UQ for models obtained from regularized inversion of EM geophysical data, and that it does so at reasonable cost, even for problems of significant size.

6 CONCLUSIONS

This work describes a sampling algorithm, the RTO-TKO, that can compute a meaningful UQ for regularized models at a reasonable computational cost, thus enabling UQ in 2D (and possibly 3D) EM inverse problems. The mathematics behind RTO-TKO make it remarkably simple to turn regularized inversion algorithms into UQ algortihms.

In this paper, Part I of a two part series, we introduce the basic ideas and algorithms, and describe important mathematical aspects of RTO-TKO. Specifically, we argue that immense computational advantages are obtained when accepting a small 'bias' between an RTO-TKO sampling distribution and a targeted Bayesian posterior distribution. From a practical perspective, our work implies that one can obtain uncertainty estimates by running existing inversion machinery in a parallel for-loop. We illustrated these ideas in simplified toy problems and in a 1D DC resistivity problem with field-data. Our work is of great relevance to the geophysics community because obtaining quantitative uncertainty for inverted model parameters is increasingly important to advancing our understanding of the Earth—but the

currently available algorithms for doing so either neglect nonlinearity or are prohibitively expensive. We caution, however, that while we demonstrate in this paper (and in Part II) that RTO provides meaningful UQ for various diffusive EM geophysical methods, we cannot guarantee it will prove as useful for other methods. Further work is required to test the breadth of RTO-TKO's applicability beyond EM geophysics.

In Part II of this two-part series, we focus on practical aspects of RTO-TKO in geophysical inversions, in particular in EM geophysical methods. Most importantly, we verify our speculations about the computational efficiency of RTO-TKO by solving a 2D MT problem via RTO-TKO. Part II also addresses issues of UQ related to prior assumptions and model parameterizations. We compare and contrast uncertainty estimates obtained by a variety of methods, including the RTO-TKO and trans-D MCMC, and discuss the practical implications.

DATA AVAILABILITY

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

7 ACKNOWLEDGEMENTS

DB was supported by the Green Foundation's John W Miles postdoctoral fellowship in theoretical and computational geophysics. MM was supported by the US Office of Naval Research (ONR) grant N00014-21-1-2309. KK was supported by the Electromagnetic Methods Research Consortium at Columbia University.

REFERENCES

Agostinetti, N. P. & Bodin, T., 2018. Flexible Coupling in Joint Inversions: A Bayesian Structure Decoupling Algorithm, *Journal of Geophysical Research: Solid Earth*, **123**(10), 8798–8826.

Aster, R., Borchers, B., & Thurber, C., 2011. *Parameter Estimation and Inverse Problems*, Academic Press, 2nd edn.

Bardsley, J. & Cui, T., 2021. Optimization-based markov chain monte carlo methods for nonlinear hierarchical statistical inverse problems, *SIAM/ASA Journal on Uncertainty Quantification*, **9**(1), 29–64.

Bardsley, J. M., Solonen, A., Haario, H., & Laine, M., 2014. Randomize-Then-Optimize: A Method for Sampling from Posterior Distributions in Nonlinear Inverse Problems, *SIAM Journal on Scientific Computing*.

Bardsley, J. M., Seppänen, A., Solonen, A., Haario, H., & Kaipio, J., 2015. Randomize-Then-Optimize for Sampling and Uncertainty Quantification in Electrical Impedance Tomography, *SIAM/ASA Journal on Uncertainty Quantification*.

Bardsley, J. M., Cui, T., Marzouk, Y. M., & Wang, Z., 2020. Scalable Optimization-Based Sampling on Function Space, *SIAM Journal on Scientific Computing*.

Beskos, A., Roberts, G., & Stuart, A., 2009. Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions, *Annals of Applied Probability*, **19**(3), 863–898.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., & Stuart, A., 2013. Optimal tuning of the hybrid Monte Carlo algorithm, *Bernoulli*, **19**(5), 1501–1534.

Blatter, D., Key, K., Ray, A., Foley, N., Tulaczyk, S., & Auken, E., 2018. Trans-dimensional Bayesian inversion of airborne transient EM data from Taylor Glacier, Antarctica, *Geophysical Journal International*, **214**(3), 1919–1936.

Blatter, D., Key, K., & Ray, A., 2021. Two-dimensional Bayesian inversion of magnetotelluric data using trans-dimensional Gaussian processes, *Geophysical Journal International*, **226**(1), 548–563.

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *Geophysical Journal International*, **178**(3), 1411–1436.

Bonavita, M., Isaksen, L., & Hólm, E., 2012. On the use of EDA background error variances in the ECMWF 4D-Var, *Quarterly Journal of the Royal Meteorological Society*, **138**(667), 1540–1559.

Brett, H., Hawkins, R., Lythgoe, K., Waszek, L., & Deuss, A., 2021. 3D Transdimensional Seismic Tomography of the Inner Core, in EGU General Assembly, Copernicus Meetings.

Calvetti, D. & Somersalo, E., 2018. Inverse problems: From regularization to Bayesian inference, *Wiley Interdisciplinary Reviews: Computational Statistics*, **10**(3), e1427.

Chen, J., Hoversten, G. M., Key, K., Nordquist, G., & Cumming, W., 2012. Stochastic inversion of magnetotelluric data using a sharp boundary parameterization and application to a geothermal site, *Geophysics*, **77**(4), E265–E279.

Chen, V., Dunlop, M. M., Papaspiliopoulos, O., & Stuart, A., 2018. Robust MCMC Sampling with Non-Gaussian and Hierarchical Priors, *arXiv:1803.03344*.

Chen, Y. & Oliver, D. S., 2012. Ensemble Randomized Maximum Likelihood Method as an Iterative Ensemble Smoother, *Mathematical Geosciences*, **44**(1), 1–26.

Christen, J. & Fox, C., 2010. A general purpose sampling algorithm for continuous distributions (the t-walk), *Bayesian Analysis*, **5**(2), 263–281.

Constable, S. C., McElhinny, M. W., & McFadden, P. L., 1984. Deep Schlumberger sounding and the crustal resistivity structure of central Australia, *Geophysical Journal International*, **79**, 893–910.

Constable, S. C., Parker, R. L., & Constable, C. G., 1987. Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics*, **52**(3), 289–300.

Dettmer, J. & Dosso, S. E., 2013. Probabilistic two-dimensional water-column and seabed inversion with self-adapting parameterizations, *The Journal of the ...*, **133**(5), 2612–2623.

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

Duane, S., Kennedy, A., Pendleton, B., & Roweth, D., 1987. Hybrid Monte Carlo, *Physics Letters B*, **195**, 216–222.

Dunlop, M. M., Helin, T., & Stuart, A. M., 2020. Hyperparameter Estimation in Bayesian MAP Estimation: Parameterizations and Consistency, *SIAM Journal of Computational Mathematics*, **6**, 69–100.

Emerick, A. A. & Reynolds, A. C., 2013. Investigation of the sampling performance of ensemble-based methods with a simple reservoir model, *Computational Geosciences*, **17**, 325–350.

Fournier, D. & Oldenburg, D. W., 2019. Inversion using spatially variable mixed ℓp norms, *Geophysical Journal International*, **218**(1), 268–282.

Galetti, E. & Curtis, A., 2018. Transdimensional Electrical Resistivity Tomography, *Journal of Geophysical Research: Solid Earth*, **123**(8), 6347–6377.

Gao, G., Zafari, M., & Reynolds, A. C., 2006. Quantifying uncertainty for the PUNQS3 problem in a Bayesian setting with RML and EnKF, *SPE Journal*, **11**.

Goodman, J. & Weare, J., 2010. Ensemble samplers with affine invariance, *Communications in Applied Mathematics and Computational Science*, **5**(1), 65–80.

Gu, Y. & Oliver, D. S., 2007. An iterative ensemble Kalman filter for multiphase fluid flow data assimilation, *SPE Journal*, **4**, 438–446.

Hastings, W. K., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, **57**(1), 97.

Hawkins, R. & Sambridge, M., 2015. Geophysical imaging using trans-dimensional trees, *Geophysical Journal International*, **203**(2), 972–1000.

Kelbert, A., Meqbel, N., Egbert, G. D., & Tandon, K., 2014. ModEM: A modular system for inversion of electromagnetic geophysical data, *Computers & Geosciences*, **66**, 40–53.

Key, K., 2016. MARE2DEM: a 2-D inversion code for controlled-source electromagnetic and magnetotelluric data, *Geophysical Journal International*, **207**(1), 571–588.

MacKay, D., 2003. *Information Theory, Inference and Learning Algorithms*, Cambridge University Press.

Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophysical Journal International*, **151**(3), 675–688.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E., 1953. Equation of State Calculations by Fast Computing Machines, *The Journal of Chemical Physics*, **21**(6), 1087–1092.

Minsley, B. J., 2011. A trans-dimensional Bayesian Markov chain Monte Carlo algorithm for model assessment using frequency-domain electromagnetic data, *Geophysical Journal International*, **187**(1),

252–272.

Morzfeld, M., Tong, X., & Marzouk, Y., 2019. Localization for MCMC: sampling high-dimensional posterior distributions with local structure, *Journal of Computational Physics*, **380**, 1–28.

Neal, R. M., 2011. *MCMC Using Hamiltonian Dynamics*, chap. Chapter 5, CRC Press.

Newman, G. A. & Alumbaugh, D. L., 2000. Three-dimensional magnetotelluric inversion using non-linear conjugate gradients, *Geophysical Journal International*, **140**(2), 410–424.

Oliver, D. S., 2017. Metropolized Randomized Maximum Likelihood for Improved Sampling from Multimodal Distributions, *SIAM/ASA Journal on Uncertainty Quantification*.

Parker, R. L., 1994. *Geophysical Inverse Theory*, Princeton University Press.

Pinski, F. J., Simpson, G., Stuart, A. M., & Weber, H., 2015. Kullback–Leibler Approximation for Probability Measures on Infinite Dimensional Spaces, *SIAM Journal on Mathematical Analysis*.

Ray, A., Key, K., Bodin, T., Myer, D., & Constable, S., 2014. Bayesian inversion of marine CSEM data from the Scarborough gas field using a transdimensional 2-D parametrization, *Geophysical Journal International*, **199**(3), 1847–1860.

Robert, G. & Rosenthal, J., 2001. Optimal scaling for various Metropolis-Hastings algorithms, *Statistical science*, **16**(4), 351–367–64.

Roberts, G. & Rosenthal, J., 1998. Optimal scaling of discrete approximations to Langevin diffusions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**, 255–268.

Roberts, G. O., Gelman, A., & Gilks, W. R., 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms, *Annals of Applied Probability*, **7**(1), 110–120.

Rosas-Carbajal, M., Linde, N., Kalscheuer, T., & Vrugt, J. A., 2014. Two-dimensional probabilistic inversion of plane-wave electromagnetic data: methodology, model constraints and joint inversion with electrical resistivity data, *Geophysical Journal International*, **196**(3), 1508–1524.

Schoniger, A., Illman, W. A., Wöhling, T., & Nowak, W., 2015. Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection, *Journal of Hydrology*, **531**(Part 1), 96–110.

Sokal, A., 1997. *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms*, vol. 361 of **Functional Integration. NATO ASI Series (Series B: Physics)**, Springer, Boston, MA.

Stewart, A. M., 2010. Inverse problems: a Bayesian perspective, *Acta Numerica*, **19**, 451–559.

Stordal, A. S. & Nævdal, G., 2018. A modified randomized maximum likelihood for improved Bayesian history matching, *Computational Geosciences*, **22**(1), 29–41.

Tarantola, A., 2005. *Inverse problem theory and methods for model parameter estimation*, Society for Industrial and Applied Mathematics.

Vignoli, G., Guillemoteau, J., Barreto, J., & Rossi, M., 2021. Reconstruction, with tunable sparsity levels, of shear wave velocity profiles from surface wave data, *Geophysical Journal International*, **225**(3), 1935–1951.

Wang, K., Bui-Thanh, T., & Ghattas, O., 2018. A Randomized Maximum A Posteriori Method for Posterior Sampling of High Dimensional Nonlinear Bayesian Inverse Problems, *SIAM Journal on Scientific Computing*, **40**(1).

Wang, Z., Bardsley, J. M., Solonen, A., Cui, T., & Marzouk, Y. M., 2017. Bayesian Inverse Problems with l_1 Priors: A Randomize-Then-Optimize Approach, *SIAM Journal on Scientific Computing*.

Wolff, U., 2004. Monte Carlo errors with less errors, *Computer Physics Communications*, **156**, 143–153.

APPENDIX A: POSTERIOR COVARIANCE FOR LINEAR INVERSIONS AND THE RTO SOLUTION

If the likelihood and prior are given by

$$p(\boldsymbol{d}|\boldsymbol{m}) \propto \exp\left(-\frac{1}{2}\left\| C_d^{-1/2}\Big(\boldsymbol{F}(\boldsymbol{m}) - \boldsymbol{d}\Big)\right\|^2\right). \tag{A.1}$$

and

$$p(\boldsymbol{m}) \propto \exp\left(-\frac{1}{2}\left|\left| C_m^{-1/2}(\boldsymbol{m} - \overline{\boldsymbol{m}})\right|\right|^2\right) \tag{A.2}$$

respectively, and $\boldsymbol{F}(\boldsymbol{m}) = G\boldsymbol{m}$ is linear, it can be shown that the posterior mean and covariance are given by

$$\hat{\boldsymbol{m}} = \overline{\boldsymbol{m}} + K(\boldsymbol{d} - G\overline{\boldsymbol{m}}) \tag{A.3}$$

$$\hat{C_m} = (I - KG)C_m \tag{A.4}$$

where C_m is the prior model covariance (e.g. $C_m = \frac{1}{\mu}(L^T L)^{-1}$) and $K = C_m G^T(GC_m G^T + C_d)^{-1}$. In this appendix, the various variables are the same as defined in the main text of the paper.

We wish to show that choosing $\tilde{\boldsymbol{d}} \sim \mathcal{N}(\boldsymbol{d}, C_d)$ and $\tilde{\boldsymbol{m}} \sim \mathcal{N}(\overline{\boldsymbol{m}}, C_m)$ yields solutions $\boldsymbol{m}$ of (4) that have mean $\hat{\boldsymbol{m}}$ and covariance $\hat{C_m}$.

To do so, we recall that RTO minimizes (see Equation (4))

$$f(\boldsymbol{m}) = \frac{1}{2}\left|\left| C_m^{-1/2}(\boldsymbol{m} - \tilde{\boldsymbol{m}})\right|\right|^2 + \frac{1}{2}\left|\left| C_m^{-1/2}(\boldsymbol{m} - \tilde{\boldsymbol{m}})\right|\right|^2. \tag{A.5}$$

The minimizer of f, make the gradient of f vanish and thus can be obtained by solving

$$\nabla f = G^T C_d^{-1}(G\boldsymbol{m} - \tilde{\boldsymbol{d}}) + C_m^{-1}(\boldsymbol{m} - \tilde{\boldsymbol{m}}) = 0.$$

Thus,

$$\begin{aligned}\boldsymbol{m} &= (G^T C_d^{-1} G + C_m^{-1})^{-1}(G^T C_d^{-1}\tilde{\boldsymbol{d}} - C_m^{-1}\tilde{\boldsymbol{m}}) \\ &= (I - KG)C_m(G^T C_d^{-1}\tilde{\boldsymbol{d}} - C_m^{-1}\tilde{\boldsymbol{m}}) \\ &= \tilde{\boldsymbol{m}} + K(\tilde{\boldsymbol{d}} - G\tilde{\boldsymbol{m}}).\end{aligned}$$

By choosing $E[\tilde{\boldsymbol{m}}] = \overline{\boldsymbol{m}}$ and $E[\tilde{\boldsymbol{d}}] = \boldsymbol{d}$, we obtain the desired expected value of $\boldsymbol{m}$:

$$\begin{aligned}E[\boldsymbol{m}] &= E[\tilde{\boldsymbol{m}}] + K(E[\tilde{\boldsymbol{d}}] - GE[\tilde{\boldsymbol{m}}]) \\ &= \overline{\boldsymbol{m}} + K(\boldsymbol{d} - G\overline{\boldsymbol{m}}).\end{aligned}$$

Likewise, picking $\mathrm{Cov}(\tilde{\boldsymbol{m}}) = C_m$ and $\mathrm{Cov}(\tilde{\boldsymbol{d}}) = C_d$ yields the desired covariance for $\boldsymbol{m}$:

$$\begin{aligned}\boldsymbol{m} &= \tilde{\boldsymbol{m}} + K(\tilde{\boldsymbol{d}} - G\tilde{\boldsymbol{m}}) \\ &= (I - KG)\tilde{\boldsymbol{m}} + K\tilde{\boldsymbol{d}} \\ \mathrm{Cov}(\boldsymbol{m}) &= (I - KG)C_m(I - KG)^T + KC_dK^T \\ &= \hat{C_m} - \hat{C_m}G^TK^T + KC_dK^T \\ &= \hat{C_m} - (I - KG)BG^TK^T + KC_dK^T \\ &= \hat{C_m} - BG^TK^T + KGBG^TK^T + KC_dK^T \\ &= \hat{C_m} - BG^TK^T + K(GBG^T + C_d)K^T \\ &= \hat{C_m} - BG^TK^T + BG^TK^T \\ &= \hat{C_m}\end{aligned}$$

Intuitively, RTO draws samples that explore the posterior uncertainty by perturbing the minimum of a conventional, deterministic objective function according to the uncertainties described by the likelihood and the prior. More concretely, perturbing the data by C_d explores the range of models that are compatible with the measured data and data uncertainty. In a similar though perhaps less obvious way, the model must be perturbed according to C_m in order to explore the full range of models compatible with the prior. If only the data are perturbed, the resulting model covariance is

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{m}) &= KC_dK^T \\ &\neq \hat{C_m}\end{aligned}$$

which is not the correct covariance. The result of correctly perturbing both the data and the model regularization is an exploration, one optimization problem at at time, of the full

range of models compatible with the measured data and prior assumptions about the model encoded in the regularization term.