Responses to Anonymous Referee #1

> I have read the manuscript with interest and I think that it will be a good contribution to the field of integrated geophysical modelling and inversion. The manuscript is well written and well organized. The authors present an inversion code relying on Monte-Carlo sampling in a Bayesian framework. The theoretical background pertaining to the Parallel Tempered Markov-Chain Monte-Carlo (PTMCMC) that is provided allows a good understanding of the principles behind the implementation.

We thank the referee for their constructive review.

> The code they use is an extension of an existing software, and there is therefore not much information, for instance, about the way they calculate the forward geophysical problem. The manuscript is relatively de-attached from the software the authors introduce, which allows it to remain general and to provide a good introduction to Bayesian and Monte-Carlo techniques. However, I think that it is a little bit too detached from the code itself and more indications as to how users could use Obsidian in practice and to reproduce the work presented would be useful. The example they use to illustrate the methodology is appropriate.

We'll respond to individual suggestions below, but will just point out here that all of the configuration files and data sets we used to generate these solutions are available as part of the repository. We will review the documentation on the repository to ensure that the instructions for running Obsidian with these configurations are straightforward and can be followed without an expert knowledge of the code's inner workings.

> The literature is generally well reviewed and well used but there are a few occurrences where references are miscited or should be added (in particular when it comes to less statistical and more geological considerations). I come back to it where necessary in the detailed comments below.

We appreciate the suggestions of appropriate papers to cite where provided and have included them as applicable.

> This paper is used as a companion paper by Olierook et al. (2019) and is cited multiple times by them. The authors should consider citing Olierook et al. (2019) as an application example.

We agree. The Olierook et al. paper was still in prep at the time we submitted this paper, which is the only reason it hasn't appeared here. It is also still in review at Solid Earth, but at least a reference to the discussion paper can be included here.

> An aspect which is practically missing from the manuscript relates to the computational requirements of inverse modelling using Obsidian v.0.1.2. The model the authors are using as an illustration example seems small and yet I have the impression that carrying out the inverse modelling was relatively computationally intensive. A little bit more information would be welcome, and it would be useful to geoscientists planning to use Obsidian v.0.1.2.

In general the use of parallel-tempering MCMC is already very computationally intensive, and Obsidian was conceived as a code optimized to run on large distributed clusters such as AWS. Although we don't try to hide this – even our most efficient runs use a few CPU-hours per independent sample (see Table 1) – we agree that some additional wording about the computational cost, and what one obtains for having paid that cost, could benefit the paper. We have added a paragraph ("Since only samples...") describing this to section 2.2.

> Does the implementation restrict the modelling of one given property (say, density contrast) to one type of sensors (say, gravimeters)? I am asking this question because of the way equation 9 is formulated. It seems to imply that one physical property cannot be recovered from the joint inversion of two datasets. For instance, this would mean that, in its current version, Obsidian would not support an extension to the joint inversion of gravity anomaly measurements with tensor gravity gradiometry to recover density contrast?

We understand the problem with wording here, and clarify that each Obsidian sensor can in principle respond to any combination of rock properties, and so multiple sensors can respond to the same rock property if desired. If a forward model for a tensor gravity gradiometry sensor were included in Obsidian, nothing would prevent the user from combining it with gravity anomaly. We have revised the text before Equation 9 to refer to "K rock properties necessary and sufficient to evaluate the forward models for all relevant sensors."

Moreover, Obsidian allows the user to formulate multivariate Gaussian petrophysical priors that treat rock properties as correlated, for example between rock density and seismic wave speed – this would then allow different data sets to jointly constrain rock properties even if each one responded to only one rock property. We make this explicit now in the Priors section: "This allows the user to formulate priors that capture intrinsic covariances between rock properties, though of a somewhat simpler form than the petrophysical mixture model of Giraud et al (2017)."

> Title. 'Sampling of [. . .] inversions'. I think that you cannot sample an inversion as it is a process, but that you can do sampling for 3-D inversions.

Changed to "...for 3-D geophysical inversions...".

> P2.l2-3. several works have recognised the issue. Consider adding a few references.

We already have cited a number of references that seem relevant to us here, but have clarified in the text that "the issue" is "[the expense of] acquiring direct observations at depth".

> P2.l6-7. 'gravity, magnetic, and electrical measurements integrate data from the surrounding volume'. This is true for all geophysical methods, even high-frequency seis- mics. You can replace by something like 'All geophysical measurements [...].'

Replaced as suggested.

> P2.l13-14: In the work of A. Tarantola, non-uniqueness is clearly stated. It is one of the limitations of geophysical inversion and mitigating it is one of the motivations for integration and joint inversion as presented in this manuscript. Consider adding a word about non-uniqueness in geophysics to this sentence and perhaps another reference (for instance Sambridge (1998) might be relevant here).

We have added the Sambridge reference as requested.

> P2.l22. 'All input sources of information [. . .] are probability distributions'. This is not the case in all inversion schemes. If this is a general truth you are saying (and I think it is a general truth), and if this is how all inputs are treated in your work/Obsidian, then consider stating it clearly.

We agree that this isn't true of all inversion schemes, but by definition it is true of Bayesian schemes – the posterior depends only upon the likelihood and prior, which are probability distributions, although the way in which probability is expressed is quite flexible. We have changed this sentence to read: "In a Bayesian approach, model elements are flexible but all statements about the fit of a model, either to data or to pre-existing expert knowledge, are expressed in terms of probability distributions."

> P3.l3-4. 'posterior around each local maximum may in these cases significantly un- derestimate uncertainties'. This is a good point and it is often overlooked. Consider adding a reference to support this or an example illustrating this.

We have rephrased to: "Use of the inverse Fisher information matrix to describe posterior uncertainty implicitly assumes a single multivariate Gaussian mode; for posteriors with multiple modes or significant non-Gaussian tails, the inverse Fisher information provides a lower bound on the posterior variance (Cramer 1946; Rao 1945) and may be a significant underestimate."

> P3.l10-11. 'Giraud et al. (2017, 2018) demonstrate an optimization-based Bayesian inversion framework for 3-D geological models, which finds the maximum of the pos- terior distribution (maximum a posteriori, or MAP), and expresses uncertainty in terms of the posterior covariance around the MAP solution; while they show that fusing data reduces uncertainty around this mode, they do not attempt to find or characterize other modes, or higher moments of the posterior'. This is partially true. Giraud et al. (2017, 2018) use uncertainty information and assess the reduction of uncertainty after inversion, and find the maximum of the MAP, but they do not show the posterior explicitely. Giraud et al. (2016) on the other hand, do calculate the posterior covariance matrix.

We agree with this characterization and this sentence now states that "...they do not attempt to find and characterize other modes, and only Giraud et al. (2016) calculate the posterior covariance."

> P6.l9. The shape of matrix Σ might be determined using probabilistic geological modelling (e.g, Wellmann et al. (2010), Pakyuz-Charrier et al. (2018a), de la Varga et al. (2018)). You could add that in the discussion.

This point gets to the heart of how what we are doing differs from previous probabilistic modeling frameworks from the above authors. The uncertainty-propagation framework works well for structural data because the model is a direct interpolation of the data. This would correspond in our case to sampling from the prior, and hence for Σ to take the shape of the prior, making MCMC (as opposed to simple Monte Carlo sampling) unnecessary. The addition of likelihood components for geophysical data, however, may narrow and shift the posterior shape, making naive sampling less efficient and making the optimal proposal shape less intuitive. This was recognized by de la Varga & Wellmann (2016) who recast the problem in terms of MCMC.

We have added a new paragraph in the Introduction to distinguish MCUE from MCMC methodologically, since the need for such distinction comes up again later. In this place in the text, we have also included the sentence: "If constraints from additional data are weak, Σ could take the shape of the prior; if there are no other constraints, as in MCUE (Pakyuz-Charrier et al 2018a,b), sampling directly from the prior may be easier."

> P6.l16. "These authors found that in general updating blocks of parameters simultaneously was inefficient". My impression is that you also refer to inversions schemes using graphic cuts to update the models. If this is the case please state it clearly/briefly.

No, this was not our intention.

> P6.l31. "using information from ensembles of particles", does the comment also extend to inversions using particle swarm optimization? If this is the case please state it clearly/briefly

In this section we are discussing only (Metropolis-Hastings) MCMC sampling schemes, not optimization schemes or other particle-based sampling methods such as sequential Monte Carlo. The sentence now reads: "Many other types of proposals can be used in Metropolis-Hastings sampling schemes, using information from ensembles of particles (as distinct from particle swarm optimization or sequential Monte Carlo; Goodman & Weare 2010)..."

> P6.l26-30. You could consider making it clearer that this is what your version of Obsidian does so that readers/users are not wondering.

We have updated the text to read: "PTMCMC is a meta-method used by Obsidian for sampling..." and have also made explicit in section 2.4 what changes we made between v0.1.1 and v0.1.2 to support this and the Olierook et al 2019 paper.

> P8.l22. The acronym 'IACT' is used only in this place. Please remove.

Fixed.

> P9.l29. I think that the usage of the word 'layer' is a bit confusing from a geological point of view as you later on refer to as an inclusion as a layer, which it is not. Please use more appropriate vocabulary.

We have replaced all occurrences of the word "layer" in this context with "unit".

> P10.l8. Maybe you can state later in the manuscript that your implementation of geological structures is more suited to basin scenarii (and therefore oil and gas exploration cases), and that in hard rock / mining scenarii, different geological modelling approaches can be followed (as you do near the end of the manuscript when refering to gempy).

We do already have such a sentence about Obsidian's world parametrization in the last section of the Discussion; we have added some words about specific applications, as suggested.

> P10.l13. Equation 14. Just for the sake of completeness you may consider to specify what x' and y' are.

The text now reads: "a radial basis function kernel to describe the correlation structure of the surface between two surface locations (x, y) and (x', y')".

> P10.l16. Typo: the bracket needs to be removed.

Fixed.

> P10.l25. Equation 12. Consider adding a short appendix detailing how it is derived.

We now have included a derivation. The reviewer comments for Olierook et al 2019 made a similar request, but that paper is still under review; we have varied the wording accordingly.

P11.l34. – P12.l1. Note that drillhole uncertainty for control points can be modelled (Pakyuz-Charrier et al. (2018b)), as can seismic interpretation (Bond (2015), Schaaf and Bond (2019), Alcalde et al. (2017)).

Noted. The frameworks in these papers seem to be about uncertainty propagation (like MCUE) and so could be used to elicit a prior on geological parameters in the context of fusion with geophysical data, as we do here.

P12.l4-5. I find this discretization a bit coarse. Is it because of the computation cost involved in PTMCMC or due to lack of information or to shorten run time?

Using a finer grid for the mean interpreted reflection horizon would probably not have made much difference to the computational efficiency; we use this grid here in order to reproduce the Beardsmore et al. setup. A finer grid of control points, however, would have dramatically increased the dimension of the problem.

P13.l1. Information entropy has been used in the geosciences after Wellmann and Regenauer-Lieb (2012) introduced it to the field, but it was initially introduced by Shannon (1948). Consider adding this reference, and possibly a brief statement explaining why it is appropriate to use it.

Cited, with the statement "this measure is appropriate to summarize posterior uncertainty in categorical predictions such as the type of rock".

P14.l5-7. Please make this paragraph clearer.

We have expanded this description somewhat: "To maintain the target acceptance rate, the adapted step size approaches the scale of the posterior's narrowest dimension, and the random walk will then slowly explore the other dimensions using this small step size. The time it takes for a random walk to cover a distance scales as the square of that distance, so we might expect the worst-case autocorrelation time for random-walk MCMC in a long, narrow mode to scale as the condition number of the covariance matrix for that mode."

P14.l32. The information about the number of computational hours is relevant only if the specs of the computer used are known. I think that more information about this aspect of the work presented and of Obsidian should be given: does it run on supercomputers, do it scale well? Just a little bit of information on this aspect would be useful to users and would strengthen the paper.

These questions are addressed in McCalman et al. (2014), but we now point the reader to them by adding the following text to the beginning of section 2.4: "Obsidian was designed to run on large distributed architectures such as supercomputing clusters. McCalman et al. (2014) shows that the code scales well to large numbers of processors, by allowing individual MCMC chains to run in parallel and initiating communication between chains only when a PTMCMC swap proposal is initiated. The inversion of Beardsmore et al. (2016) was performed on Amazon Web Services using 160 cores."

We also include technical specifications of the Artemis cluster on which our experiments were run at the beginning of section 3, and state that each run used 32 cores for up to 8 hours of wall time, to provide typical end users with a better idea of the requirements.

P16.l19-25. This paragraph is not very clear to me.

This paragraph is really about the alpha parameter in the sensor noise prior and how it relates to certainty about the noise level. We have rewritten to emphasize this: "The uncertainty on the variance of a sensor is determined by the α parameter in that sensor's prior, with smaller α corresponding to more uncertainty. For example, the gravity and magnetotelluric sensors use a prior with $\alpha = 5$, so that the resulting t-distribution for model residuals in the likelihood has $\nu = 2\alpha = 10$ degrees of freedom. The magnetic anomaly sensor prior uses $\alpha = 1.25$, allowing a residual distribution with thick tails closer to a Cauchy distribution than a Gaussian."

P17.l17: "Suppose that σ is unknown, however, and is allowed to vary alongside [theta]" does it mean you allow heteroscedasticity? If so this needs to be stated.

In this case, all we mean is that the overall scale of homoscedastic errors may not be known, and hasn't been included as part of the dataset. We have made this more specific: "Suppose that σ is not perfectly known a priori, however (but is still assumed to be the same for all points in a single dataset, and is allowed to vary..."

P17.l8. I'm not sure I understand the usage of the term 'fiducial' here.

We mean the original Moomba inversion presented in Beardsmore et al 2016, and have now replaced occurrences of the word "fiducial" with a citation to this previous work.

> P18.l3. "one potential weakness of this approach to balancing sensors". How would that relate to defining the relative weight of the different types of sensors in the joint inversion problem?

The point we attempt to raise in this paragraph is that if the data for a given sensor have real variation beneath the scale of the basic world parametrization to resolve, that variation will be treated by our approach as "noise".

> P18.l10. Equation number is missing.

Fixed.

> P19. The models are shown only in 2D. A 3D view would be welcome.

We have now included some views of the voxelized probability of occupancy for the granite intrusion and basement layers of runs B and D (new figure).

> P20.l24-25. You mention a number of interpolation techniques. Have you tried kriging, as it is widely used in geostatistics?

Our understanding is that kriging is synonymous with Gaussian process regression, so yes, in fact we use it here. We have inserted a reference to the term "kriging" in section 2.4 when we first mention the use of Gaussian processes for depth-to-boundary interpolation.

> P23.l16-17. I am not sure that I understand the meaning of this sentence. Please clarify. By gradients, do you refer to the jacobian matrix? Or am I missing something?

In this case we mean the derivatives of the prior and likelihood with respect to model parameters being sampled. We clarify in this section now that we mean derivatives of the posterior with respect to parameters rather than some spatial derivative, and mention Hamiltonian Monte Carlo (Duane et al. 1987; Neal 2011) and Riemannian manifold Monte Carlo (Girolami & Calderhead 2011) by name as examples of proposals that need derivative information.

> P23.l30. the package proposed by de la Varga et al. (2018) offers the advantage of being open source but it is not the only one performing probabilistic geological modelling. For instance, other works using ideas introduced by Wellmann et al. (2010) such as Pakyuz-Charrier et al. (2018a) also achieves this.

We agree that both Pakyuz-Charrier papers are probabilistic, but as with other similar issues above, we also want to distinguish uncertainty propagation methodologically from sampling of the model posterior. MCUE is a fine solution if the only data are structural, but as mentioned in de la Varga & Wellmann (2016), MCMC sampling of the posterior becomes necessary to fuse structural data with other types.

Besides the abovementioned paragraph in the Introduction describing MCUE as a branch of probabilistic modeling, we have updated the sentence mentioned here to give a more specific description of GemPy's advantages: "The GemPy package developed by de la Varga et al. (2018) makes an excellent start on a more general-purpose open-source code for 3-D geophysical inversions: it uses the implicit potential-field approach (Lajaunie 1997) to describe geological structures, includes forward-models for geophysical sensors, and is designed to produce posteriors that can easily be sampled by MCMC."

Responses to Anonymous Referee #2

> This study explores the influence of various practitioner decisions on MCMC posterior sampler efficiency for a geophysical joint inversion with a layered paramitrization; specifically, the influence several of proposal, prior, and likelihood function options. The tests are well designed and succeed in addressing the questions asked. I personally did not find much of the results and conclusions surprising, most of it could be deduced from purely theoretical grounds. However, the topic is important and this paper gives a good empirical basis from which future geophysical posterior sampling work can draw. On these grounds, I think it deserves to be published.

We thank the referee for their feedback and are pleased to hear that they recommend publication once their comments are addressed.

> Most of the paper is well written and clear, with the introduction being the exception. It seems rushed and the odd use of Bayesian/statistical/probabilistic terminology (in the introduction only) suggests a lack of familiarity. I personally don't see the need for additional detail on the software use and implementation since that is not what the paper is about and if anything the scope should be more contained not expanded.

While a case could be made simply to cite the original software paper (McCalman et al 2014) regarding all such details, the fact that Referee #1 asked for more details suggest that our paper will be more accessible if at least an overview of the code and its performance is provided here.

> The review of the MCMC literature review is extensive and was interesting to read. Complexity of the shape of the posterior is discussed several times but seldom in the context of previous work. For example, the non-linearity and complex correlations of physical parameters for non-unique magnetotelluric inversion is well known, but no overview is given here on that. I believe the discussion section could be improved by relating more to the known properties of the different geophysical forward problems.
>
> I think the discussion section is needlessly bloated. Here the authors go into detail on many topics which the experiments shown here had no bearing on. Various things that could be done or might work are listed here which are in no way related to what the study presented actually did. I strongly suggest rewriting this section to be more on topic.

Our main aim in this work is to flag and address challenges for the uptake of Bayesian reasoning and MCMC in joint inversion methods for 3-D geological models, in ways we hope are accessible both to geoscientists and to statisticians. Some of the citations suggested by Referee #2 involve some quite advanced methods, relating mostly to 1-D non-parametric inversions for single sensor types, and we are happy to acknowledge them. In contrast, Referee #1 focused on probabilistic methods for error propagation in geological models rather than on posterior sampling, and our impression is that posterior sampling for uncertainty quantification is still quite rare in this area because existing MCMC methods are still too costly. The forward model for each sensor contributes to the posterior shape, but so does the prior. Our discussion is a little more than a page long and addresses future directions. The work most directly corresponding to our direction is de la Varga & Wellmann (2016) and de la Varga et al. (2018), whom we acknowledge and cite.

We recognize that this focus may not have come across well in the original Introduction, and believe that our revisions of the Introduction in response to Referee #1's comments make our intended contribution clearer.

> P1L12) What does "improve inversion results" refer to? Most readers would assume that it means a more accurate inversion. Since accuracy of results, compared to reality, is never quantified in this work, I don't see how this claim is backed up. One might argue that if true sensor noise levels are known, uninformative priors on them would only increase chances of their miss-estimation. Counterarguments based on model inadequacy could be raised of course, but these are not things that this study shed light on so please remove this claim.

Removed. If true sensor noise levels are known in detail then we agree that using an informative prior is more appropriate. Information about noise levels, however, is frequently not available in detail for public survey data our end users might want to fuse, nor was it available for the data set we used.

> P1L13-15) I don not see why this claim about using gradient information is in the abstract. The statement is probably true, but this study did not show anything new to support it.

Removed.

> P2L25) It's not clear what is meant by posterior ensembles being a 'gold-standard'.

This is a normative statement from the statistical community. The true "gold standard" would be an analytic form for the exact posterior. MCMC, however, has theoretical guarantees to converge to the target distribution given enough computing time. We have removed the "gold-standard" wording and have updated the text as follows: "The output of a Bayesian method is also a probability distribution (the posterior) representing all values of system parameters consistent both with the available data and with prior beliefs. For complex statistical models the exact posterior cannot be expressed analytically; in such cases Monte Carlo algorithms, in particular Markov chain Monte Carlo (MCMC; Mosegaard 1995, Sambridge 2002) can provide samples drawn from the posterior for the purpose of computing averages over uncertain properties of the system."

> P2L29) Online updating is not necessary or sufficient for optimal for decision-making; these are separate things. The only relevance I can see here is that it could speed up decision making.

We agree this was poorly worded; we are referring to the potential for Bayesian updating and Bayesian optimization for acquisition of additional data. We have updated the text to read: "The inference also can be readily updated as new information becomes available, using the posterior for the previous inference as the prior for the next one. This use of Bayesian updating allows automated decision-making about which additional data to take to minimize the cost of reducing uncertainty (Mockus 2013)."

P3L3) What about overestimating uncertainties?

See our response to Referee #1 on a similar question. This sentence now reads: "Use of the inverse Fisher information matrix to describe posterior uncertainty implicitly assumes a single multivariate Gaussian mode; for posteriors with multiple modes or significant non-Gaussian tails, the inverse Fisher information provides a lower bound on the posterior variance (Cramer 1946; Rao 1945) and may be a significant underestimate."

P3L4) The "no 'one-size-fits-all' solution exists" comment is very important. Perhaps give the reader some direction by citing something (e.g. Wolpert et al., 1997, No free lunch theorems for optimization: IEEE transactions on evolutionary computation, 1, 67-82.)

The sentence is referring to sampling, not optimization. There are several MCMC-related references that make this point and we have chosen a recent one: Green (2015) Bayesian computation: a summary of the current state, and samples backwards and forwards, Statistics and Computing July 2015, Volume 25, Issue 4, pp 835–862. We have also revised this sentence to further highlight the point: "Since the most appropriate sampling strategy may depend ont he characteristics of the posterior for specific problems, sampling methods must usually be tailored..."

P4L21-29) This paragraph should probably lead with the last sentence (lines 27-29). The parts about deterministic inversion reads like an odd tangent and I didn't see the relevance and purpose of it until a second read through.

This is a good suggestion and we have revised the paragraph accordingly: "Although each of these elements has a correspondence to some similar model element in more traditional geophysical inversion literature (for example Menke et al 2018), interpreting model elements in terms of probability may motivate different mathematical choices from the usual non-probabilistic misfit or regularization terms." We also now mention the role of cross-validation in calibrating non-probabilistic regularization terms that do not arise in a Bayesian setting where a prior or hyperprior governs the extent of regularization, and add some citations.

P7) I did not pick up on the fact that all your tests use PTMCMC until the second read-through; this section should probably make that more explicit.

Section 3 intro, paragraph 3 now starts: "All experiments use PTMCMC sampling, with 4 simultaneous temperature ladders... each with 8 temperatures, unless otherwise specified."

P12L1-5) Where the seismic lines used to inform the layer interface Gaussian process variogram?

More information on the construction of the prior used to set up the original Moomba inversion is given in a NICTA technical report (Beardsmore, 2014), which we now cite at the beginning of section 2.5 in addition to the less complete conference papers. The original seismic lines are not explicitly mentioned in any of these sources, nor were we able to learn this from the original authors. We do now provide maps of the locations of sensor readings we used for the three sensors we actually include, in a new figure referenced at the beginning of section 3.

P14L8) What is 'global posterior shape'? is it always defined?

We agree this wording is vague and have updated the sentence to be more specific: "The adaptive (anisotropic) Gaussian random walk (Haario et al. 2001), or aGRW, attempts to learn an appropriate covariance structure for a random walk proposal based on the past history of the chain."

P14) Could you have used the layer Gaussian process covariances directly to create a proposal function, it seems like that is what the CNp effectively does?

The proposal described in this comment is a sort of random walk with the same covariance as the prior. This is the behavior of pCN in the limit of small stepsize. In the limit of the medium-to-large stepsizes taken in the higher-temperature chains, pCN behaves very differently and samples more effectively than a random walk would, especially in high-dimensional spaces, a point we make when we introduce pCN in section 3.1 and made in more detail by Cotter et al. 2013.

> P16L4) "Figure 2 shows that iGRW and aGRW have more trouble travelling between different posterior modes than pCN" Tell us how you deduce this from that figure.

This is a fair point – the trace plots don't obviously support this conclusion. The worst-case autocorrelation and the differences in posterior weight between the two modes among repeat runs under similar conditions do support this conclusion, but we refer to these already in the preceding paragraphs. We have thus rephrased to say: "The different proposals vary in performance when hopping between modes despite the fact that all three proposals are embedded within a PTMCMC scheme..."

> P17) Please mention why MT noise levels were fixed.

We varied MT noise levels as well. We can see how this might not have been clear and so we now say: "the noise prior is set to $\alpha = 0.5$, $\beta = 0.05$ for all sensors (gravity, magnetic, and magnetotelluric)."

> P21) Increasing the amount of data points by interpolation seems like a terrible idea. Why would anyone even attempt it? The observed effects should be obvious. If there are actual examples of Bayesian posterior analysis papers which do this, please cite one; otherwise, this seems like an odd and unnecessary test to include.

We agree that it's a terrible idea, and that nobody who is already thinking probabilistically about the data would be likely to do this. We were thinking specifically about the potential for uncritical use of re-gridded data by end users not already accustomed to Bayes or MCMC, especially when we (the statistician collaborators) learned that re-gridding was common for public geophysical survey data. In retrospect it would have been a fairer test to compare original to re-gridded measurements using actual widely adopted re-gridding methods, or to demonstrate new likelihoods for handling re-gridded potential field data.

This section isn't central to our results, though, and is more about pedagogy than development of new knowledge. At this stage we've decided to remove this section and replace it with a brief warning in section 2.1 ("The implicit assumption behind the use of mean square error...").

> P21L16-19) I don't know about gravity and and magnetic, but Gaussian process likelihood functions have been used for MT and seismic MCMC. Relevant work should be cited here, E.g.:
>
> Agostinetti, N. P., and A. Malinverno, 2010, Receiver function inversion by trans- dimensional Monte Carlo sampling: Geophysical Journal International, 181, 858–872. Bodin, T., M. Sambridge, H. Tkalcic, P. Arroucau, K. Gallagher, and N. Rawlinson, 2012, Transdimensional inversion of receiver functions and surface wave dispersion: Journal of Geophysical Research: Solid Earth, 117. Xiang, E., R. Guo, S. E. Dosso, J. Liu, H. Dong, and Z. Ren, 2018, Efficient hierarchical trans-dimensional Bayesian inversion of magnetotelluric data: Geophysical Journal International, 213, 1751–1767.
>
> Also, there are ways to learn the correlation during sampling: Steininger, G., J. Dettmer, S. E. Dosso, and C. W. Holland, 2013, Trans-dimensional joint inversion of seabed scattering and reflection data: The Journal of the Acoustical Society of America, 133, 1347–1357.

We thank the referee for bringing these papers to our attention; we cite them now in the paragraph in section 2.1 regarding correlations in the likelihood.

> P22L19) "The clearest lesson we can draw ..." I'm not sure why this is the lesson you lead with, in the introduction it was stated as known; almost every MCMC application to geophysics show this and it was not among the questions that your tests were set up to answer.
>
> P22L20) "Our results were sensitive to ..." Each point raised in this sentence will be true for for any difficult posterior sampling problem. This is not a new result and this sentence adds nothing to the manuscript.

Since we view our contribution here as focused on the interaction between problem setup and sampling efficiency, we have replaced this text with a one-sentence introduction to the next paragraph: "Our experiments show concrete examples of how the efficiency of MCMC sampling changes with assumptions about the prior, likelihood, and proposal distributions for an Obsidian inversion, particularly as tight constraints on the solution are relaxed and uncertainty increases."

P22L34) I don't agree with the claim that either of these outcomes are counter-intuitive. Tighter constraints lead to narrower local optima, hence more sampling is needed. Cauchy likelihood functions are more likely to give multi-modal posteriors than Gaussian likelihood functions, even for the most trivial problems (e.g. with just one parameter).

Removed the first part of this sentence re: whether intuitive or not, since this may depend on the reader; added mention of likelihoods ("but relaxing priors or likelihoods may sometimes widen...") to statement about relaxed constraints.

P23L6-10) None of these three dot-point listed statements were informed by the experiments presented in this manuscript. The claims are also obvious and well known.

P24L10) "However, proposals using gradients from auto-differentiation are probably needed to make further progress in this area." This claim, while probably true, is not really backed up by what is in the manuscript. Why is it listed as a conclusion?

Removed. This is really a statement about future work, which we leave in our Discussion section.

P24L14) This is a trivial claim by itself. How can it help design future work. Will the better fit derived from uninformative priors lead to more accurate results in terms of uncertainty estimation. This ties in with my comment for P1L12.

We have updated the text to relate more specifically to a conclusion about what was done, again tying back to the sampling: "Hierarchical priors on observational noise provide a way to capture uncertainty about the weighting among datasets, although this may also make sampling more challenging as when priors on world parameters are relaxed."

P2417) Without some guiding principle for how to do the sub-sampling, this is not useful.

Removed (since the corresponding section has been removed).

TECHNICAL CORRECTIONS:

P2L20) "..., but about uncertainties." Awkward use of terminology, a Bayesian probability is an uncertainty and an assumption. Assumptions are specified as uncertainties quantified by probability distributions.

Changed to: "...not only about expected values or point estimates for system parameters, but about their beliefs regarding the true values of those parameters."

P2L26) "The posterior distribution is a representation of all possible outcomes and hence provides an internal estimate of uncertainty." The world parameterization is the representation of all 'possible' outcomes. What does the 'internal estimate' mean? This sentence is incoherent.

We have reworded this paragraph in addressing the "gold-standard" comment above.

P6L12) Spell out what SGR stands for here.

We define this acronym ("sequential geostatistical resampling") in the Introduction where it is first used.

P6L26-27) Grammar mistake.

Actually a LaTeX mistake; it seems GMD's LaTeX template doesn't support the form of the natbib command we used. We have changed to: "Parallel-tempered MCMC, or PTMCMC [ref], ..."

P13L1) Table 1, what is N? First it was iteration count, then number of layers, then what? Readers shouldn't have to fish through the past 12 pages to find out.

We have relabeled this symbol "Nsamp" to distinguish it from other uses of N in the paper. We have also now included the symbols for each of these quantities in the table caption.

P18L4) Grammar mistake.

Fixed; now reads: "may include systematic residuals..."

Efficiency and robustness in Monte Carlo sampling 3-D geophysical inversions with Obsidian v0.1.2: Setting up for success

Richard Scalzo[1], David Kohn[2], Hugo Olierook[3], Gregory Houseman[4], Rohitash Chandra[1,5], Mark Girolami[6,7], and Sally Cripps[1,8]

[1]Centre for Translational Data Science, University of Sydney, Darlington NSW 2008, Australia
[2]Sydney Informatics Hub, University of Sydney, Darlington NSW 2008, Australia
[3]School of Earth and Planetary Sciences, Curtin University, Bentley WA 6102, Australia
[4]School of Earth and Environment, University of Leeds, Leeds, LS2 9JT, UK
[5]School of Geosciences, University of Sydney, Darlington NSW 2008, Australia
[6]The Alan Turing Institute for Data Science, British Library, 96 Euston Road, London, NW1 2DB, UK
[7]Department of Mathematics, Imperial College London, London, SW7 2AZ, UK
[8]School of Mathematics and Statistics, University of Sydney, Darlington NSW 2008, Australia

Correspondence: Richard Scalzo (richard.scalzo@sydney.edu.au)

Abstract. The rigorous quantification of uncertainty in geophysical inversions is a challenging problem. Inversions are often ill-posed and the likelihood surface may be multimodal; properties of any single mode become inadequate uncertainty measures, and sampling methods become inefficient for irregular posteriors or high-dimensional parameter spaces. We explore the influences of different choices made by the practitioner on the efficiency and accuracy of Bayesian geophysical inversion methods that rely on Markov chain Monte Carlo sampling to assess uncertainty, using a multi-sensor inversion of the three-dimensional structure and composition of a region in the Cooper Basin of South Australia as a case study. The inversion is performed using an updated version of the Obsidian distributed inversion software. We find that the posterior for this inversion has complex local covariance structure, hindering the efficiency of adaptive sampling methods that adjust the proposal based on the chain history. Within the context of a parallel-tempered Markov chain Monte Carlo scheme for exploring high-dimensional multi-modal posteriors, a preconditioned Crank-Nicholson proposal outperforms more conventional forms of random walk. Aspects of the problem setup, such as priors on petrophysics or on 3-D geological structure, affect the shape and separation of posterior modes, influencing sampling performance as well as the inversion results. Use of uninformative priors on sensor noise enables optimal weighting among multiple sensors even if noise levels are uncertain.

1 Introduction

Construction of 3-D geological models is plagued by the limitations on direct sampling and geophysical measurement (Wellmann et al., 2010; Lindsay et al., 2013). Direct geological observations are sparse because of the difficulty in acquiring them,

with basement geology often obscured by sedimentary or regolith cover; acquiring direct observations at depth via drilling is expensive (Anand and Butt, 2010; Salama et al., 2016). Indirect observations via geophysical sensors deployed at or above the surface are more readily obtained (Strangway et al., 1973; Gupta and Grant, 1985; Sabins, 1999; Nabighian et al., 2005b, a). All geophysical measurements integrate data from the surrounding volume, so it is difficult to resolve precise geological constraints at any given position and depth, except where borehole measurements are also available. Determining the true underlying geological structure, or range of geological structures, consistent with observations constitutes an often poorly constrained inverse problem. One natural way to approach this is forward-modelling, where the responses of various sensors on a proposed geological structure are simulated, and the proposed structure is then updated or sampled iteratively (for examples see Jessell, 2001; Calcagno et al., 2008; Olierook et al., 2015).

The incompleteness and uncertainty of the information contained in geophysical data frequently mean that there are many possible worlds consistent with the data being analyzed (Tarantola and Valette, 1982; Sambridge, 1998; Tarantola, 2005). To the extent that information provided by different datasets is complementary, combining all available information into a single joint inversion reduces uncertainty in the final results. Accomplishing this in a principled and self-consistent manner presents several challenges, including: (i) how to weigh constraints provided by different datasets relative to each other; (ii) how to rule out worlds inconsistent with geological processes (expert knowledge); (iii) how to present a transparent accounting of the remaining uncertainty; and (iv) how to do all this in a computationally efficient manner.

Bayesian statistical techniques provide a powerful framework for characterizing and fusing disparate sources of probabilistic information (Tarantola and Valette, 1982; Mosegaard and Tarantola, 1995; Sambridge and Mosegaard, 2002; Sambridge et al., 2012). In a Bayesian approach, model elements are flexible but all statements about the fit of a model, either to data or to pre-existing expert knowledge, are expressed in terms of probability distributions; this forces the practitioner to make explicit all assumptions, not only about expected values or point estimates for system parameters, but about their beliefs regarding the true values of those parameters. The output of a Bayesian method is also a probability distribution (the *posterior*) representing all values of system parameters consistent both with the available data and with prior beliefs. For complex statistical models the exact posterior cannot be expressed analytically; in such cases, Monte Carlo algorithms, in particular *Markov chain Monte Carlo* (MCMC; Mosegaard and Tarantola, 1995; Sambridge and Mosegaard, 2002) can provide samples drawn from the posterior for the purpose of computing averages over uncertain properties of the system. The uncertainty associated with the posterior can be visualized in terms of the marginal distributions of parameters of interest, or rendered in 3-D voxelisations of information entropy (Wellmann and Regenauer-Lieb, 2012). The inference also can be readily updated as new information becomes available, using the posterior for the previous inference as the prior for the next one. This use of Bayesian updating allows automated decision-making about which additional data to take to minimize the cost of reducing uncertainty (Mockus, 2013).

Although Bayesian methods provide rigorous uncertainty quantification, implementing them in practice for complicated forward models with many free parameters has proven difficult in other geoscientific contexts, such as landscape evolution (Chandra et al., 2018) and coral reef assembly (Pall et al., 2018). Sambridge and Mosegaard (2002) point out the challenge of capturing all elements of a geophysical problem in terms of probability, which can be difficult for complex datasets and

even harder for approximate forward models or world representations where the precise nature of the approximation is hard to capture. The irregular shapes and multimodal structure of the posterior distributions for realistic geophysics problems makes them hard to explore. Use of the inverse Fisher information matrix to describe posterior uncertainty implicitly assumes a single multivariate Gaussian mode; for posteriors with multiple modes or significant non-Gaussian tails, the inverse Fisher information provides only a lower bound on the posterior variance (Cramer, 1946; Rao, 1945) and may be a significant underestimate. Moreover, the large number of parameters needed to specify 3-D structures also means these irregular posteriors are embedded in high-dimensional spaces, increasing the computational cost for both optimization and sampling. Since the most appropriate sampling strategy may depend on the characteristics of the posterior for specific problems, sampling methods must usually be tailored to each individual problem and no "one-size-fits-all" solution exists (Green et al., 2015).

(Agostinetti and Malinverno, 2010; Bodin et al., 2012; Xiang et al., 2018).

While 1-D inversions for specific sensor types may use some quite sophisticated sampling methods (Agostinetti and Malinverno, 2010; Bodin et al., 2012; Xiang et al., 2018), the uptake of MCMC sampling appears to have been slower for 3-D structural modeling problems. Giraud et al. (2016); Giraud et al. (2017, 2018) demonstrate an optimization-based Bayesian inversion framework for 3-D geological models, which finds the maximum of the posterior distribution (*maximum a posteriori*, or MAP); while they show that fusing data reduces uncertainty around this mode, they do not attempt to find or characterize other modes, and only Giraud et al. (2016) calculate the actual posterior covariance. Ruggeri et al. (2015) investigate several MCMC schemes for sampling a single-sensor inverse problem (crosshole georadar travel time tomography), focusing on sequential, localized perturbations of a proposed 3-D model (*sequential geostatistical resampling*, or SGR); they show that sampling is impractically slow due to high dimensionality and correlations between model parameters. Laloy et al. (2016) embed the SGR proposal within a parallel-tempered sampling scheme to explore multiple posterior modes of a 2-D inverse problem in groundwater flow, improving computational performance but not to a cost-effective threshold.

The above methods are non-parametric, in that the model parameters simply form a 3-D field of rock properties to which sensors respond. Although this type of method is flexible, parametric models, in which the parametrized elements correspond more directly to geological interpretation, comprise a more transparent and parsimonious approach. Wellmann et al. (2010) describes a workflow for propagating uncertainty in structural data through to uncertainty in the geological interpretation. Pakyuz-Charrier et al. (2018b, a) further develop a Monte Carlo approach to uncertainty estimation for structural and drill hole data, showing the impact of different distributions used to summarize uncertainty in the data. Such approaches are much simplified by the use of an implicit potential-field parametrization of geological structure (Lajaunie et al., 1997), in which the data coincide directly with model parameters; conditioning on further data, and hence use of MCMC methods, is not necessary in this context. de la Varga and Wellmann (2016); de la Varga et al. (2018) recast modeling of structural data in terms of MCMC sampling of a posterior, required in order to fuse structural data with other types of data, including geophysical sensor data. A large-scale 3-D joint inversion with multiple sensors remains to be done in this framework.

McCalman et al. (2014) present Obsidian, a flexible software platform for MCMC sampling of 3-D multi-modal geophysical models on distributed computing clusters. Beardsmore (2014); Beardsmore et al. (2016) demonstrate Obsidian on a test problem in geothermal exploration, in the Moomba gas field of the Cooper Basin in South Australia, comparing their results

to a deterministic inversion of the same area performed by Meixner and Holgate (2009). These papers outline a full-featured open-source inversion method that can fuse heterogeneous data into a detailed solution, but make few comments about how the efficiency and robustness of the method depends on the particular choices they made.

In this paper, we revisit the inversion problem of Beardsmore et al. (2016) using a customized version (Scalzo et al., 2019) of the McCalman et al. (2014) inversion code. Our interest is in exploring this problem as a case study to determine which aspects of this problem's posterior present the most significant obstacles to efficient sampling, which updates to the MCMC scheme improve sampling under these conditions, and how plausible alternative choices of problem setup might influence the efficiency of sampling or the robustness of the inversion. The aspects we consider include: correlations between model parameters; relative weights between datasets with poorly constrained uncertainty; and degrees of constraint from prior knowledge representing different possible exploration scenarios. The methods and software discussed here are also applied to a different geological setting in Olierook et al. (2019), including a new sensor for surface lithostratigraphic measurements.

2 Background

In this section we present a brief overview of the Bayesian forward-modeling paradigm to geophysical inversions. We also provide a discussion of implementing Bayesian inference via sampling using MCMC methods. We then present background on the original Moomba inversion problem, commenting on choices made in the inversion process before we begin to explore different choices in subsequent sections.

2.1 Overview of Bayesian inversion

A Bayesian inversion scheme for geophysical forward models comprises of three key elements:

1. the underlying parametrized representation of the simulated volume or history, which we call the *world* or *world view*, denoted by a vector of *world parameters* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$
2. a probability distribution $p(\boldsymbol{\theta})$ over the world parameters, called the *prior*, expressing expert knowledge or belief about the world before any datasets are analyzed; and
3. a probability distribution $p(\mathcal{D}|\boldsymbol{\theta})$ over possible realizations of the observed data $\mathcal{D}$ as a function of world parameters, called the *likelihood*, that incorporates the prediction of a deterministic forward model $g(\boldsymbol{\theta})$ of the sensing process for each value of $\boldsymbol{\theta}$.

The *posterior* is then the distribution $p(\boldsymbol{\theta}|\mathcal{D})$ of values of the world parameters consistent with both prior knowledge and observed data. *Bayes' theorem* describes the relationship between the prior, likelihood, and posterior:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}. \tag{1}$$

Although each of these elements has a correspondence to some similar model element in more traditional geophysical inversion literature (for example Menke, 2018), interpreting model elements in terms of probability may motivate different mathematical

choices from the usual non-probabilistic misfit or regularization terms. The terminology we use in this paper will be typical of the statistics literature. In other geophysical inversion papers a "model" might refer to the world representation, whereas below we will use the word "model" to refer to the *statistical* model defined by a choice of parametrization, prior, and likelihood. A non-Bayesian inversion would proceed by minimizing an *objective function*, one simple form of which is the mean square misfit between the (statistical) model predictions and the data, corresponding to our negative log likelihood (*assuming* observational errors that are independent and Gaussian-distributed with precisely known variance). To penalize solutions that are considered *a priori* unlikely, the objective function might include additional *regularization* terms corresponding to the negative log priors in our framework; the weight of a non-probabilistic regularization term might ordinarily be optimized by cross-validation against subsets of the data (MacCarthy et al., 2011; Wrona et al., 2018), whereas in our framework we include plausible (possibly vague) constraints as part of the prior distribution. The full objective function would correspond to the negative log posterior, and minimization of the objective function would correspond to maximization of the posterior probability, under some choice of prior.

Indeed, there is considerable flexibility in choosing the above elements even in a fully probabilistic context. For example, the partitioning of information into "data" and "prior knowledge" is neither unique nor cut-and-dried. However, there are guiding principles: the ideal set of parameters $\boldsymbol{\theta}$ is both *parsimonious* — as few as possible to faithfully represent the world — and *interpretable*, referring to meaningful aspects of the world that can easily be read off the parameter vector. Information resulting from processes that can be easily simulated belong in the likelihood: for example, one might argue that the output of a gravimeter should have a Gaussian distribution, because it responds to the mean rock density within a volume and hence obeys the central limit theorem, or that the output of a Geiger counter should follow a Poisson distribution to reflect the physics of radioactive decay. Even processes that are not so easily simulated can at least be approximately described, for example by using a mixture distribution to account for outlier measurements (Mosegaard and Tarantola, 1995) or a prior on the unknown noise level in a process Sambridge et al. (2012). Other information about allowable or likely worlds belongs in the prior, such as the distribution of initial conditions for simulation, or interpretations of datasets with expensive or intractable forward models.

The implicit assumption behind the use of mean square error as a (log) likelihood — that the residuals of the data for each sensor from the corresponding forward model are independent Gaussian — may also not be true if the data have been interpolated, resampled, or otherwise modified from original point observations. For example, gravity anomaly and magnetic anomaly measurements are usually taken at ground level along access trails to a site, or along spaced flight lines in the case of aeromagnetics. In online data releases, the original measurements may then be interpolated or resampled onto a grid, changing the number and spacing of points and introducing correlations on spatial scales comparable to the scale of the smoothing kernel. This resampling of observations onto a regular grid may be useful for traditional inversions using Fourier transform techniques. However, if used uncritically in a Bayesian inversion context, correlations in residuals from the model may then arise from the resampling process rather than from model misfit, resulting in stronger penalties in the likelihood for what would otherwise be plausible worlds, and muddying questions around model inadequacy. If such correlations are known to exist, they can be modeled explicitly as part of the likelihood. For example, autoregressive models are already being used as error models for 1-D inversions of magnetotelluric and seismic data (Agostinetti and Malinverno, 2010; Bodin et al., 2012; Xiang et al., 2018).

Care must be taken in formulating such likelihoods to avoid confusion between real (but possibly unresolved) structure and correlated observational noise.

The inference process expresses its results in terms either of $p(\boldsymbol{\theta}|\mathcal{D})$ itself or of integrals over $p(\boldsymbol{\theta}|\mathcal{D})$ (including credible limits on $\boldsymbol{\theta}$). This is different from the use of point estimates for the world parameters, such as the *maximum likelihood* (ML) solution $\boldsymbol{\theta}_{\rm ML} = \sup_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})$ or the *maximum a posteriori* (MAP) solution $\boldsymbol{\theta}_{\rm MAP} = \sup_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. To the extent that ML or MAP prescriptions give any estimate of uncertainty on $\boldsymbol{\theta}$, they usually do so through the covariance of the log likelihood or log posterior around the optimal value of $\boldsymbol{\theta}$, equivalent to a local approximation of the likelihood or posterior by a multivariate Gaussian. As mentioned above, these approaches will underestimate the uncertainty for complex posteriors; a more rigorous accounting of uncertainty will include all known modes, higher moments of the distribution, or (more simply) providing enough samples from the distribution to characterize it.

The posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ is rarely available in closed form. However, it is often known up to a normalizing constant: $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Sampling methods such as MCMC can therefore be used to approximate the posterior, without having to explicitly evaluate the normalizing constant (the high-dimensional integral in the denominator of Eq. 1). It is to these methods we turn next.

2.2 Markov chain Monte Carlo

A MCMC algorithm comprises a sequence of world parameter vectors $\{\boldsymbol{\theta}^{[j]}\}$, called a *(Markov) chain*, and a *proposal distribution* $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ to generate a new set of parameters based only on the last element of the chain. In the commonly-used *Metropolis-Hastings algorithm* (Metropolis et al., 1953; Hastings, 1970), a proposal $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{[j]})$ is at random either *accepted* and added to the chain's history ($\boldsymbol{\theta}^{[j+1]} = \boldsymbol{\theta}'$) with probability

$$P_{\rm accept} = \min\left(1, \frac{P(\mathcal{D}|\boldsymbol{\theta}')P(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{[j]}|\boldsymbol{\theta}')}{P(\mathcal{D}|\boldsymbol{\theta}^{[j]})P(\boldsymbol{\theta}^{[j]})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{[j]})}\right), \tag{2}$$

or *rejected* and a copy of the previous state added instead ($\boldsymbol{\theta}^{[j+1]} = \boldsymbol{\theta}^{[j]}$). This rule guarantees, under certain regularity conditions (Chib and Greenberg, 1995), that the sequence $\{\boldsymbol{\theta}^{[j]}\}$ converges to the required stationary distribution, $P(\boldsymbol{\theta}|\mathcal{D})$, in the limit of increasing n.

Metropolis-Hastings algorithms form a large class of sampling algorithms, limited only by the forms of proposals. Although proofs that the chain will *eventually* sample from the posterior are important, clearly chains based on *efficient* proposals are to be preferred. A proposal's efficiency will depend on the degree of correlation between consecutive states in the chain, which in turn can depend on how well matched the proposal distribution is to the properties of the posterior.

One simple, commonly used proposal distribution is a (multivariate) *Gaussian random walk* (GRW) step u from the chain's current position, drawn from a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}^{[j]} + \boldsymbol{u}, \qquad \boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3}$$

This proposal is straightforward to implement, but its effectiveness can depend strongly on $\boldsymbol{\Sigma}$, and does not in general scale well to rich, high-dimensional world parametrizations. If $\boldsymbol{\Sigma}$ has too large a scale, the GRW proposal will step too often into

regions of low probability, resulting in many repeated states due to rejections; if the scale is too small, the chain will take only small, incremental steps. In both cases, subsequent states are highly correlated. If the shape of $\mathbf{\Sigma}$ is not tuned to capture correlations between different dimensions of $\boldsymbol{\theta}$, the overall scale must usually be reduced to ensure a reasonable acceptance fraction. If constraints from additional data are weak, $\mathbf{\Sigma}$ could take the shape of the prior; if there are no other constraints, as in MCUE Pakyuz-Charrier et al. (2018b, a), sampling directly from the prior may be easier.

The SGR method (Ruggeri et al., 2015; Laloy et al., 2016) can be seen as a mixture of multivariate Gaussians, in which $\mathbf{\Sigma}$ has highly correlated sub-blocks of parameters, corresponding to variations of the world over different spatial scales. Ruggeri et al. (2015) and Laloy et al. (2016) evaluate SGR using single-sensor inversions in crosshole georadar travel time tomography, with posteriors corresponding to a Gaussian process — an unusually tractable (if high-dimensional) problem that could be solved in closed form as a cross-check. These authors found that in general updating blocks of parameters simultaneously was inefficient, which may not be surprising in a high-dimensional model: for a tightly constrained posterior lying along a low-dimensional subspace of parameter space, almost all directions — hence almost all posterior covariance choices — lead towards regions of low probability. Directions picked at random without regard for the shape of the posterior will scale badly with increasing dimension.

Many other types of proposals can be used in Metropolis-Hastings sampling schemes, using information from ensembles of particles (Goodman and Weare, 2010, as distinct from particle swarm optimization or sequential Monte Carlo), adaptation of the proposal distribution based on the chain's history (Haario et al., 2001), derivatives of the posterior (Neal et al., 2011; Girolami and Calderhead, 2011), approximations to the posterior (Strathmann et al., 2015), and so forth. The GRW proposal is not only easy to write down and fast to evaluate, but requires no derivative information. We will compare and contrast several derivative-free proposals in our experiments below.

The posterior distributions arising in geophysical inversion problems are also frequently multi-modal; MCMC algorithms to sample such posteriors need the ability to escape from, or travel easily between, local modes. *Parallel-tempered MCMC*, or PTMCMC (Geyer and Thompson, 1995), is a meta-method used by Obsidian for sampling multi-modal distributions that works by running an ensemble of Markov chains. The ensemble is characterized by a sequence of $M+1$ parameters $\{\beta_i\}$, with $\beta_0 = 1 > \beta_1 > \beta_2 > \ldots > \beta_M > 0$, called the *(inverse) temperature ladder*. Each chain samples the distribution

$$P_i(\boldsymbol{\theta}|D) \propto (P(D|\boldsymbol{\theta}))^{\beta_i} P(\boldsymbol{\theta}), \tag{4}$$

so that the chain with $\beta_0 = 1$ is sampling from the desired posterior, and a chain with $\beta_i = 0$ samples from the prior, which should be easy to explore. Chains with intermediate values $0 < \beta < 1$ sample intermediate distributions in which the data's influence is reduced, so that modes are shallower and easier for chains to escape and traverse. In addition to proposing new states within each chain, PTMCMC includes Metropolis-style proposals that allow adjacent chains on the temperature ladder, with inverse temperatures β and β', to swap their most recent states $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ with probability

$$P_{\text{swap}} = \min\left(1, \left[\frac{P(\mathcal{D}|\boldsymbol{\theta}')}{P(\mathcal{D}|\boldsymbol{\theta})}\right]^{\beta'-\beta} \frac{P(\boldsymbol{\theta}')}{P(\boldsymbol{\theta})}\right). \tag{5}$$

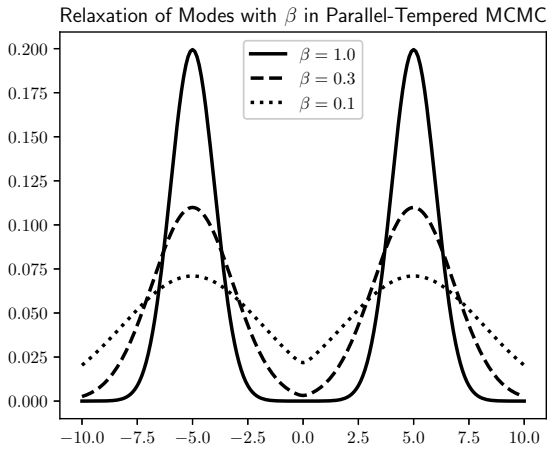

Figure 1. Parallel-tempered relaxation of a bimodal distribution.

This allows chains with current states spread throughout parameter space to share global information about the posterior in such a way that chain i still samples $P_i(\boldsymbol{\theta}|D)$ in the long-term limit. The locations of discovered modes diffuse from low-β_i chains (which can jump freely between relaxed, broadened versions of these modes) towards the $\beta_0 = 1$ chain, which can then sample from all modes of the unmodified posterior in the correct proportions. The temperature ladder should be defined so that adjacent chains on the ladder are sampling from distributions similar enough for swaps to occur frequently.

Figure 1 illustrates the sampling of a simple bimodal probability distribution (a mixture of two Gaussians) via PTMCMC. The solid line depicts the true bimodal distribution, while the broken lines shows the stationary distribution of tempered chains for smaller values of β. The tempered chains are more likely to propose moves across modes than the untempered chains, and the existence of a sequence of chains ensures that the difference in probability between successive chains is small enough that swaps can take place easily.

Since only samples from the $\beta = 1$ chain are retained as samples from the true posterior, and since the time for information about well-separated modes to propagate down the ladder increases as the square of the number of inverse temperatures used, PTMCMC can be extremely computationally intensive. It should be employed only in cases where multiple modes are likely to be present and where capturing the relative contributions of all of these modes is relevant to the application. For problems that have many deep, well-separated modes (e.g. Chandra et al., 2018), swap proposals may provide the only source of movement in the low-temperature chains; such posteriors present next-generation challenges for sampling.

Even without regard to multiple modes, PTMCMC can also help to reduce correlations between successive independent posterior samples. Laloy et al. (2016) use SGR as a within-chain proposal in a PTMCMC scheme, demonstrating its effects on correlations between samples but noting that the algorithm remains computationally intensive.

2.3 Performance metrics for MCMC

Because MCMC guarantees results only in the limit of large samples, criteria are still required to assess the algorithm's performance. Suppose for the discussion below that up to the assessment point, we have obtained N samples of a d-dimensional posterior from each of M separate chains; let $\boldsymbol{\theta}_i^{[j]} = (\theta_{1i}^{[j]}, \ldots, \theta_{di}^{[j]})$ be the $d \times 1$ vector of parameter values drawn at iteration $[j]$ in chain i. Let

$$\hat{\theta}_{ki} = \frac{1}{N} \sum_{j=1}^{N} \theta_{ki}^{[j]}$$

be the mean value of the parameter θ_k in chain i, across the N iterates, and let $\tilde{\theta}_k = \frac{1}{M} \sum_{i=1}^{M} \hat{\theta}_{ki}$ be the sample mean of θ_k across all iterates and chains. Then

$$B_k = \frac{1}{M-1} \sum_{i=1}^{M} (\hat{\theta}_{ki} - \tilde{\theta}_k)^2$$

Further define

$$s_{ki}^2 = \frac{1}{M-1} \sum_{j=1}^{N} (\theta_{ki}^{[j]} - \hat{\theta}_{ki})^2$$

and

$$W_k = \frac{1}{M} \sum_{i=1}^{M} s_{ki}^2.$$

For Metropolis-Hastings MCMC, the *acceptance fraction* of proposals is easily measured, and for a chain that is performing well should be ~ 20–50%. Roberts et al. (1997) showed that the optimal acceptance fraction for random walks in the limit of a large number of dimensions is 0.234, which we will take as our target since the proposals we will consider are modified random walks.

We examine correlations between samples within each chain separated by a lag time l using the *autocorrelation function*,

$$\rho_{lki} = \frac{1}{(N-l)W_k} \sum_{j=l+1}^{N} (\theta_{ki}^{[j]} - \hat{\theta}_{ki})(\theta_{ki}^{[j-l]} - \hat{\theta}_{ki}), \tag{6}$$

The number of independent draws from the posterior with equal statistical power to each set of N chain samples scales with the area under the autocorrelation function or *(integrated) autocorrelation time*,

$$\tau_{ki} = 1 + 2 \sum_{l=1}^{N} \left(1 - \frac{k}{N}\right) \rho_{lki}. \tag{7}$$

A *trace plot* of the history of an element of the parameter vector $\boldsymbol{\theta}$ over time summarizes the sampling performance at a glance, revealing where in parameter space an algorithm is spending its time; Fig. 3 shows a series of such figures for some of the different MCMC runs in the present work.

Gelman and Rubin (1992) assess the number of samples required to reach a robust sampling of the posterior by comparing results among multiple chains. If the simulation has run long enough, the mean values among chains should differ by some small fraction of the width of the distribution; intuitively, this is similar to a hypothesis test that the chains are sampling the same marginal distribution for each parameter. More precisely, the quantity

$$\hat{V}_k/W_k = \frac{N-1}{N} + \frac{M+1}{MN} B_k/W_k \tag{8}$$

provides a metric for convergence of different chains to the same result, which decreases to 1 as $N \to \infty$. The chains may be stopped and results read out when the metric dips below a target value for all world parameters $\boldsymbol{\theta}$. The precise number of samples needed may depend on the details of the distribution; the metric provides a stopping condition, but not an estimate of how long it will take to achieve.

The results from this procedure must still be evaluated according to how well the underlying statistical model describes the geophysical data, and whether the results are geologically plausible — although this is not unique to MCMC solutions. The distribution of residuals of model predictions (forward-modeled data sets) from the observed data can be compared to the assumed likelihood. The standard deviation or variance of the residuals (relative to the uncertainty) provide a convenient single-number summary, but the spatial distribution of residuals may also be important; outliers and/or structured residuals will indicate places where the model fails to predict the data well, and highlight parts of the model parametrization that need refinement.

Finally, representative instances of the world itself should be visualized to check for surprising features. Given the complexity of real-world data, the adequacy of a given model is in part a matter of scientific judgment, or fitness for a particular applied purpose to which the model will be put. We will use the term *model inadequacy* to refer to model errors arising from approximations or inaccuracies in the world parametrization or the mathematical specification of the forward model — although there will always be such approximations in real problems, and the presence of model inadequacy should not imply that the model is unfit for purpose.

2.4 The Obsidian distributed PTMCMC code

For our experiments we use a customized fork (v0.1.2 Scalzo et al., 2019) of the open-source Obsidian software package. Obsidian was previously presented in McCalman et al. (2014) and was used to obtain the modeling results of Beardsmore et al. (2016); v0.1.1 was the most recent open-source version publicly available before our work. We refer the reader to these publications for a comprehensive description of Obsidian, but below we summarize key elements corresponding to the inversion framework set out above, and describe the changes we have made to the code since v0.1.1.

Obsidian was designed to run on large distributed architectures such as supercomputing clusters. McCalman et al. (2014) shows that the code scales well to large numbers of processors, by allowing individual MCMC chains to run in parallel and initiating communication between chains only when a PTMCMC swap proposal is initiated. The inversion of Beardsmore et al. (2016) was performed on Amazon Web Services using 160 cores.

World parametrization: Obsidian's world is parametrized as a series of discrete units, each with its own spatially constant rock properties, separated by smooth boundaries. Each unit boundary is defined by a set of *control points* that specify the subsurface depth of the boundary at given surface locations. The depth to each unit boundary at any other location is calculated using a two-dimensional Gaussian process regression (kriging) through the control points; each unit is truncated against the overlying unit to allow lateral termination of units and ensure a strict stratigraphic sequence (**NEED FIGURE HERE**).

For a world with N units, indexed by i with $1 \leq i \leq N$, each with a grid n_i regularly spaced control points at sites x_i and with K rock properties necessary and sufficient to evaluate the forward models for all relevant sensors, the parameter vector is therefore

$$\boldsymbol{\theta} = (\alpha_{11} \ldots \alpha_{N n_N}, \rho_{11} \ldots \rho_{NK}), \tag{9}$$

where α_{ij} is the offset of the mean depth of the top of unit i at site j, and ρ_{is} is the rock property of unit i associated with sensor s. Taken together, the rock properties for each unit and the control points for the boundaries between the units fully specify the world. This parametrization requires that interface depths be single-valued, not for example permitting the surface to fold above or below. Such a limitation still enables reasonable representations of sedimentary basins, but may hinder faithful modeling of other kinds of structures.

Prior: The control point depth offsets within each unit i have a multivariate Gaussian prior with mean zero and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_i}$. The Gaussian processes which interpolate the unit boundaries across the lateral extent of the world use a radial basis function kernel to describe the correlation structure of the surface between two surface locations (x, y) and (x', y'),

$$k(x, y; x', y') = \exp\left(-\frac{(x-x')^2}{\Delta_x^2} - \frac{(y-y')^2}{\Delta_y^2}\right), \tag{10}$$

and has mean function $\mu_i(x, y)$ that can be specified at finer resolution to capture fine detail in unit structure. The correlation lengths Δ_x and Δ_y could in principle be varied, but in this case are fixed in value to the spacing between control point locations along the x and y coordinate axes, respectively. The rock properties for each unit i, which are statistically independent of the control points, also have a multivariate Gaussian prior, with mean $\boldsymbol{\mu}_{\rho_i}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_i}$. This allows the user to formulate priors that capture intrinsic covariances between rock properties, though of a somewhat simpler form than the petrophysical mixture models of Giraud et al. (2017). The prior for the full parameter vector is therefore block-diagonal,

$$\begin{aligned} P(\boldsymbol{\theta}) &= \prod_{i=1}^{N} P(\boldsymbol{\alpha}_{i\cdot}) P(\boldsymbol{\rho}_{i\cdot}) \\ &= \prod_{i=1}^{N} N(\boldsymbol{\alpha}_{i\cdot}; 0, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_i}) N(\boldsymbol{\rho}_{i\cdot}; \boldsymbol{\mu}_{\boldsymbol{\rho}_i}, \boldsymbol{\Sigma}_{\boldsymbol{\rho}_i}). \end{aligned} \tag{11}$$

Likelihood: The likelihood for each Obsidian sensor s is Gaussian, meaning that the residuals of the data $\mathcal{D}_s$ from the forward model predictions $f_s(\boldsymbol{\theta})$ for the true world parameters θ are assumed to be independent, identically distributed Gaussian draws. The underlying variance of the Gaussian noise is not known, but is assumed to follow an inverse gamma distribution $\mathrm{IG}(x; \alpha_s, \beta_s)$ with different (user-specified) hyperparameters α_s, β_s for each sensor s. This choice of distribution amounts to

a prior, but the hyperparameters α_s and β_s for each sensor are not explicitly sampled over; instead, they are integrated out analytically, so that the final likelihood has the form

$$P(\mathcal{D}_s|\boldsymbol{\theta}) = \prod_{k=1}^{K_s} t_{2\alpha_s}\left(\frac{f_s(\boldsymbol{\theta}) - \mathcal{D}_s)}{\sqrt{\beta_s/\alpha_s}}\right), \tag{12}$$

where $t_\nu(x)$ is a Student's-t distribution with ν degrees of freedom. This distribution is straightforward to calculate, although the results may be sensitive to the user's choices of α_s and β_s; unrestrictive choices (e.g. $\alpha_s = \beta_s = 1$) should be used if the user has little prior knowledge about the noise level in the data. The likelihood including all sensors is therefore

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{s=1}^{S} P(\mathcal{D}_s|\boldsymbol{\theta}), \tag{13}$$

since each sensor probes a different physical aspect of the rock. Obsidian v0.1.1 includes forward models for gravity and magnetic anomaly, magnetotellurics, reflection seismic, thermal, and contact-point sensors for drill cores; v0.1.2 introduces a lithostratigraphic sensor used in Olierook et al. (2019).

MCMC: The sampling algorithm used by Obsidian is an adaptive form of PTMCMC, described in detail in Miasojedow et al. (2013). This algorithm allows for the progressive adjustment of the step size used for proposals within each chain, as well as the temperature ladder used to sample across chains, as sampling progresses. A key feature of the adjustment process is that the maximum allowed change to any chain property diminishes over time, made inversely proportional to the number of samples; this is necessary to ensure that the chains converge to the correct distribution in the limit of large numbers of samples (Roberts and Rosenthal, 2007). The Obsidian implementation of PTMCMC also allows it to be run on distributed computing clusters, making it truly parallel in resource use as well as in the requirement for multiple chains. Obsidian v0.1.2 uses the same methods for adapting the PTMCMC temperature ladder and the sizes of within-chain proposals as v0.1.1, but adds new options for within-chain proposal distributions (see §3.1 below).

2.5 The original Moomba inversion problem

The goal of the original Moomba inversion problem (Beardsmore, 2014; Beardsmore et al., 2016; McCalman et al., 2014) was to identify potential geothermal energy applications from hot granites in the South Australian part of the Cooper Basin (cf. Carr et al. (2016) for a recent review of the Cooper Basin). Modeling the structure of granite intrusions and their temperature enabled the inference of the probability of the presence of granite above 270 $^\circ$C at any point within the volume. The provenance of the original datasets involved, and how they were used to set up the prior for the original inversion, are described in more detail in the final technical report published by NICTA (Beardsmore, 2014)[1], while the results are described in the corresponding conference paper (Beardsmore et al., 2016); we summarize key elements in this section as appropriate.

The chosen region was a portion of the Moomba gas field with dimensions of 35 $\times$ 35 $\times$ 12 km volume centered at -28.1° S, 140.2° E. The volume is divided into six layers, with the first four being thin, sub-horizontal, Permo–Triassic sedimentary layers, the fifth corresponding to Carboniferous–Permian granitoid intrusions (Big Lake Suite), and the sixth to a

[1] https://arena.gov.au/projects/data-fusion-and-machine-learning-for-geothermal/

Proterozoic basement (Carr et al., 2016). The number of layers and the priors on mean depths of layer boundaries were related to interpretations of depth-converted seismic reflection horizons published by the Department of State Development (DSD) in South Australia (Beardsmore et al., 2016). Data used in the original inversion include Bouguer anomaly; total magnetic intensity; magnetotelluric sensor data at 44 frequencies between 0.0005 and 240 Hz; temperature measurements from gas wells; and petrophysical laboratory measurements based on 115 core samples from holes drilled throughout the region. Rock properties measured for each sample include density, magnetic susceptibility, thermal conductivity, thermal productivity, and resistivity.

The original choices of how to partition knowledge between prior and likelihood struck a balance between accuracy of the world representation and computational efficiency. The empirical covariances of the petrophysical sample measurements for each layer were used to specify a multivariate Gaussian prior on that layer's rock properties; although these measurements could be construed as data, the simplifying assumption of spatially constant mean rock properties left little reason to write their properties into the likelihood. The gravity, magnetic, magnetotelluric, and thermal data all directly constrained rock properties relevant to the geothermal application and were explicitly forward-modeled as data. "Contact points" from drilled wells, directly constraining the layer depths in the neighborhood of a drilled hole as part of the likelihood, were available and used to inform the prior, but not treated as sensors in the likelihood. Treating the seismic measurements as data would have dramatically increased computational overhead relative to the use of interpreted reflection horizons as mean functions for layer boundary depths in the prior. Using interpreted seismic data to inform the mean functions of the layer boundary priors also reduced the dimension of the parameter space, letting the control points specify long-wavelength deviations from seismically derived prior knowledge at finer detail: each reflection horizon was interpolated onto a 20×20 grid.

Given this knowledge of the local geology (Carr et al., 2016; Beardsmore, 2014; McCalman et al., 2014), the world parameters for geometry were chosen as follows: The surface was fixed by a level plane at zero depth. The control point grids for the relatively simple sedimentary layers were specified by 2×2 grids of control points (lateral spacing: 17.5 km). The layer boundary for the granite intrusion layer used a 7×7 grid (lateral spacing: 5 km), and also underwent a nonlinear transformation stretching the boundary vertically, to better represent the elongated shapes of the intrusions. Including the rock properties, this allowed the entire world to be specified by a vector of 101 parameters, a large but not unmanageable number.

Figure 4 show horizontal slices through the posterior probability density for granite at a depth of 3.5 km, similar to that shown in figure 9 of Beardsmore et al. (2016), for three MCMC runs sampling from the original problem. While the posterior samples from the previous inference are not available for quantitative comparison, we see reasonable qualitative agreement with previous results in the cross-sectional shape of the granite intrusion.

3 Experiments

To demonstrate the impact of problem setup and proposal efficiency in a Bayesian MCMC scheme for geophysical inversion, we run a series of experiments altering the prior, likelihood, and proposal for the Moomba problem. We approach this variation

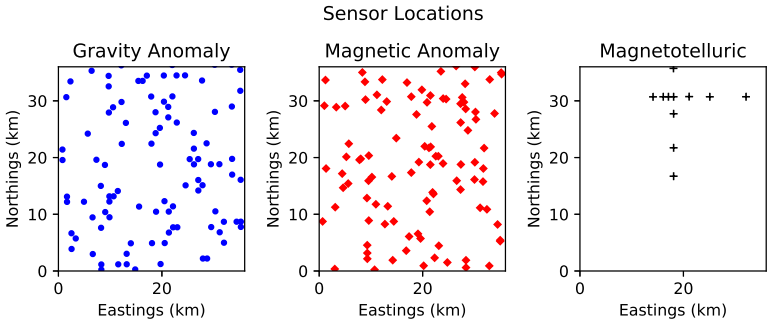

Figure 2. Locations of sensor readings used in the inversions in this paper.

as an iterative investigation into the nature of the data and the posterior's dependence on them, motivating each choice with the intent of relating our findings to related 3-D inversion problems.

The experiments described in this section were run on the Artemis high-performance computing cluster at the University of Sydney. Artemis's standard job queue provides access to 56 nodes with 24 Intel Xeon E5-2680-V3 (2.5 GHz) cores each, and 80 nodes with 32 Intel Xeon E5-2697A-V4 (2.6 GHz) cores each. Each run used 32 cores and ran for up to 8 hours wall time.

The datasets we use for our experiments are the gravity anomaly, total magnetic intensity, and magnetotelluric readings originally distributed as an example Moomba configuration with v0.1.1 of the Obsidian source code. In order to focus on information that may be available in an exploration context (i.e. publicly available geophysical surveys without contact points), we omit the thermal sensor readings, relying on a joint inversion of gravity, magnetic, and magnetotelluric data. Maps of the locations of these sensor readings, referred to the coordinate system of the inversion, are shown in Figure **??**.

All experiments use PTMCMC sampling, with 4 simultaneous temperature ladders or "stacks" of chains, each with 8 temperatures, unless otherwise specified. The posterior is formally defined in terms of samples over the world parameters, so when quantifying predictions for particular regions of the world and their uncertainty (such as entropy), the parameter samples are each used to create a voxelised realization of the 3-D world, and the average observable calculated over these voxelised samples. A quantitative summary of our results is shown in Table 1, including, for each run:

- the shortest ($\tau_{\min}$), median (τ_{med}), and longest ($\tau_{\max}$) autocorrelation time measured for individual model parameters;
- the standard deviations σ_{grav} and σ_{mag}, of the gravity and magnetic anomaly sensor data from the posterior mean forward model prediction, in physical units;
- the mean information entropy $\bar{S}$ (Shannon, 1948; Wellmann and Regenauer-Lieb, 2012) of the posterior probability density for granite, averaged over the volume beneath 3.5 km, in bits (i.e. presence or absence of granite; an entropy of 0 bits means total certainty, while 1 bit of entropy indicates total uncertainty) — this measure is appropriate to summarize posterior uncertainty in categorical predictions such as the type of rock;
- the CPU time spent per worst-case autocorrelation time, as a measure of computational efficiency.

Table 1. Performance metrics for each run, including: best-case ($\tau_{i,\min}$), median ($\tau_{i,\mathrm{med}}$), and worst-case ($\tau_{i,\max}$) autocorrelation times for model parameters; standard deviations σ_{grav} and σ_{mag} of residuals of the posterior mean forward model predictions for the gravity anomaly and magnetic anomaly data; volume-average information entropy $\bar{S}$; number N_{samples} of chain iterates (with each iterate representing a single evaluation of all forward models for a given set of world parameters); and CPU-hours per autocorrelation time (i.e. the computational cost of obtaining a single independent sample from the posterior).

Run	$\tau_{i,\min}$ (/1000)	τ_{med} (/1000)	$\tau_{i,\max}$ (/1000)	σ_{grav} (mgal)	σ_{mag} (nT)	$\bar{S}$ (bits)	N_{samp}	CPU (h) $/\tau_{\max}$	Comments
A	4.3	16.4	67.8	0.4	19.2	0.79	764.5k	10.8	baseline iGRW
A1	4.7	10.7	42.8	0.4	18.5	0.68	1566.5k	8.1	... with $N_\beta = 12$
B	2.1	4.0	28.4	0.5	18.8	0.66	628.8k	5.5	baseline pCN
B1	2.4	4.4	24.3	0.5	20.5	0.62	1166.5k	6.2	... with $N_\beta = 12$
C	1.9	17.4	> 143.2	0.5	20.9	0.57	872.6k	> 19.7	baseline aGRW
C1	2.7	14.1	310.6	0.4	17.1	0.61	2190.2k	53.8	... with $N_\beta = 12$
D	2.3	7.2	54.9	0.8	5.7	0.47	586.6k	11.5	Cauchy likelihood
E	3.0	8.0	> 172.1	0.7	6.4	0.51	669.2k	> 29.0	5 km margin
J	1.6	26.3	115.4	0.8	7.0	0.61	1172.6k	11.0	loosen rock property priors
J2	2.1	7.9	53.8	1.1	9.4		497.7k	14.4	... using 1 top layer only
K	4.2	19.8	64.7	0.5	9.9	0.90	708.8k	9.9	loosen control point priors
K2	3.7	7.7	24.7	0.5	8.4		479.1k	7.4	... using 1 top layer only

3.1 Choice of within-chain proposal

The initial work of McCalman et al. (2014) and Beardsmore et al. (2016) used an isotropic Gaussian random walk (iGRW) proposal within each chain,

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_n + \eta \boldsymbol{u}, \qquad \boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{I}), \tag{14}$$

where η is a (possibly adaptive) step size parameter tuned to reach a target acceptance rate, which we choose to be 25% for our experiments. Each dimension of a sampled parameter vector is "whitened" by dividing it by a scale factor corresponding to the allowed full range of the variable (of order a few times the prior width; this also accounts for differences in physical units between parameters). This should at least provide a scale for the marginal distribution of each parameter, but does not account for potential correlations between parameters. The covariance matrix of the iGRW proposal is a multiple of the identity matrix, so that on average, steps of identical extent are taken along every direction in parameter space. When tuning the proposal, the adaptive scheme tunes only an overall step size applying to all dimensions at once.

The iGRW proposal is the simplest proposal available, but as noted above, it loses efficiency in high-dimensional parameter spaces, and it is unable to adapt if the posterior is highly anisotropic — for example, if parameters are scaled inappropriately or are highly correlated. To maintain the target acceptance rate, the adapted step size approaches the scale of the posterior's

narrowest dimension, and the random walk will then slowly explore the other dimensions using this small step size. The time it takes for a random walk to cover a distance scales as the square of that distance, so we might expect the worst-case autocorrelation time for random-walk MCMC in a long, narrow mode to scale as the condition number of the covariance matrix for that mode.

The adaptive (anisotropic) Gaussian random walk (Haario et al., 2001), or aGRW, attempts to learn an appropriate covariance structure for a random walk proposal based on the past history of the chain. The covariance of the aGRW proposal is calculated in terms of the sample covariance of the chain history $\{\boldsymbol{\theta}^{[j]}\}$:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_n + \eta \boldsymbol{u}, \qquad \boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_n), \tag{15}$$

in which

$$\boldsymbol{\Sigma}_n = \frac{n}{n+a} \operatorname{cov}\left\{\boldsymbol{\theta}^{[j]}\right\} + \frac{a}{n+a} \boldsymbol{I}, \tag{16}$$

where a is a timescale for adaptation (measured in samples). As the length n of the chain increases, the proposal will smoothly transition from an isotropic random walk to an anisotropic random walk with a covariance structure that reflects the chain history.

A third proposal, addressing high-dimensional parameter spaces, is the *preconditioned Crank-Nicholson* (pCN) proposal (Cotter et al., 2013):

$$\boldsymbol{\theta}_{n+1} = \sqrt{1-\eta^2}\boldsymbol{\theta}_n + \eta \boldsymbol{u}, \qquad \boldsymbol{u} \sim P(\boldsymbol{\theta}) \tag{17}$$

with $0 < \eta < 1$ and $P(\boldsymbol{\theta})$ a multivariate Gaussian prior. For $\eta \ll 1$, the proposal resembles a GRW proposal with small step size, while for $\eta \sim 1$ the proposal becomes a draw from the prior. This proposal results in a sampling efficiency that is independent of the dimensionality of $\boldsymbol{\theta}$; in fact, it was developed by Cotter et al. (2013) to sample infinite-dimensional function spaces, arising in inversion problems using differential equations as forward models, where the prior is specified in the eigenbasis for the forward model operator. In our case, the prior incorporates the correlation between neighboring control points in the Gaussian process layer boundaries, so we might expect that a proposal that respects this structure would improve sampling.

Our first three runs (A, B, C) use the iGRW, pCN, and aGRW (with $a = 10$) proposals respectively. All three algorithms give roughly similar results on the baseline dataset. The autocorrelation time for this problem remains very long, of the order of 10^4 samples. This means that $\sim 10^6$ samples are required to achieve reasonable statistical power.

There are nevertheless differences in efficiency among the samplers. The pCN proposal has not only the lowest median autocorrelation, but the lowest worst-case autocorrelation across dimensions. The aGRW proposal has the largest spread in autocorrelation times across dimensions, with its median performance comparable to iGRW and its worst-case performance at least three times worse (it had still failed to converge after over 1000 CPU-hours). Repeat trials running for twice as many samples with 12 chains per stack instead of 8 (Runs A1, B1, C1) produced similar results, although we were then able to reliably measure the worst-case autocorrelation time for aGRW. For all samplers, but most noticeably aGRW, the step size can take a long time to adapt. Large differences are sometimes seen in the adapted step sizes between chains at similar temperatures in different stacks, and do not always increase monotonically with temperature.

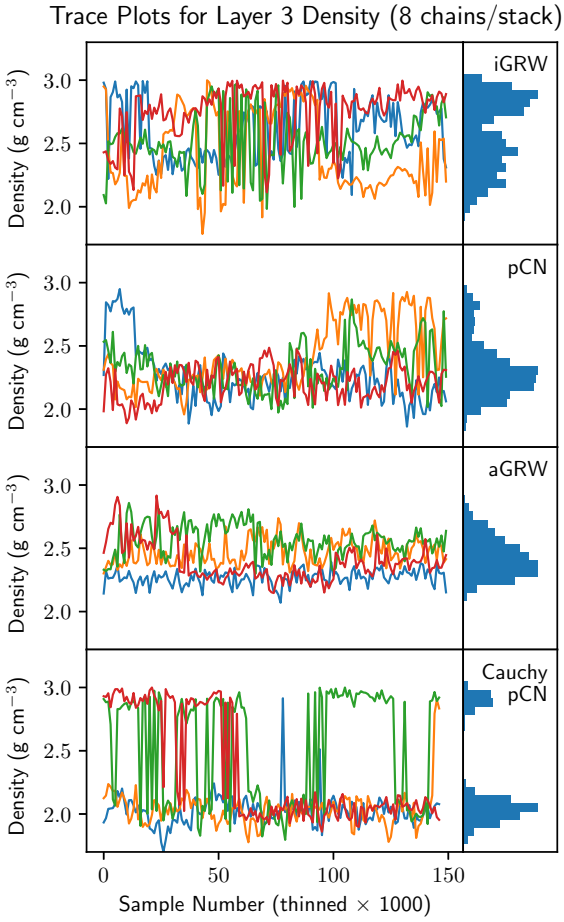

Figure 3. Trace plots (left) and marginal densities (right) for layer 3 rock density as explored by iGRW, pCN, and aGRW proposals, and by a pCN proposal under a Cauchy likelihood (top to bottom). The four colors represent the four different chains.

The differences are shown in Fig. 3, showing the zero-temperature chains from the four stacks in each run sampling the marginal distribution of the rock density for layer 3, a bimodal parameter. The iGRW chains converge slowly, and though they manage to travel between modes with the help of parallel-tempered swap proposals, the relative weights of the two modes are not fully converged and vary between re-runs at a fixed length. Each aGRW chain has a relatively narrow variance and none successfully crosses over to the high-density mode despite parallel-tempered swaps. Only the pCN chains converge "quickly" (after about 70k samples) and are able to explore the full width of the distribution.

These behaviors suggest that the local shape of the posterior varies across parameter space, so that proposals that depend on a global fixed scaling across all dimensions are unlikely to perform well. The clearly superior performance of pCN for this problem is nevertheless intriguing, since for sufficiently small step size near $\beta = 1$, the proposal reduces to GRW.

The different proposals vary in performance when hoppping between modes despite the fact that all three proposals are embedded within a PTMCMC scheme with a relatively simple multivariate Gaussian prior, to which aGRW should be able to adapt readily. We believe pCN will prove to be a good baseline proposal for tempered sampling of high-dimensional problems because of its prior-preserving properties, which ensure peak performance when constraints from the data are weak. As the chain temperature increases, the tempered posterior density approaches the prior, so that pCN proposals with properly adapted step size will smoothly approach independent draws from the prior with an acceptance probability of 1. The result is that when used as the within-chain proposal in a high-dimensional PTMCMC algorithm, pCN proposals will result in near-optimal behavior for the highest-temperature chain, and should explore multiple modes much more easily than GRW proposals.

This behavior stands in contrast to GRW proposals, for which the acceptance fraction given any particular tuning will approach zero as the dimension increases. In fact, aGRW's attempt to adapt globally to proposals with local structure may mean mid-temperature chains become trapped in low-probability areas and break the diffusion of information down to lower temperatures from the prior. A more detailed study of the behavior of these proposals within tempered sampling schemes would be an interesting topic for future research.

3.2 Variations in likelihood / noise prior

In the fiducial Moomba configuration used in Beardsmore et al. (2016), the priors on the unknown variance of the Gaussian likelihood for each sensor are relatively informative. The uncertainty on the variance of a sensor is determined by the α parameter in that sensor's prior, with smaller α corresponding to more uncertainty. For example, the gravity and magnetotelluric sensors use a prior with $\alpha = 5$, so that the resulting t-distribution for model residuals in the likelihood has $\nu = 2\alpha = 10$ degrees of freedom. The magnetic anomaly sensor prior uses $\alpha = 1.25$, resulting in a residual distribution with thick tails closer to a Cauchy distribution than a Gaussian.

If specific informative prior knowledge about observational errors exists, using such a prior, or even fixing the noise level outright, makes sense. In cases where the amplitude of the noise term is not well-constrained, using a broader prior on the noise term may be preferable. When more than one sensor with unknown noise variance is used, identical broad priors allow the data to constrain the relative influence of each sensor on the final results. The trade-off is that a more permissive prior on the noise variance could mask structured residuals due to model inadequacy or non-Gaussian outliers in the true noise distribution.

The idea that such broad assumptions could deliver competitive results arises from the incorporation of Occam's razor into Bayesian reasoning, as demonstrated in Sambridge et al. (2012). For example, the log likelihood corresponding to independent Gaussian noise is

$$\log \mathcal{L} = -\frac{1}{2} \sum_{j=1}^{N_d} \left[\frac{(f_{sj}(\boldsymbol{\theta}) - \mathcal{D}_{sj})^2}{\sigma^2} + \log 2\pi\sigma^2 \right]. \tag{18}$$

Ordinary least-squares fitting maximizes the left-hand term inside the sum, and the right-hand term is a constant that can be ignored if the observational uncertainty σ is known. This clearly penalizes worlds $\boldsymbol{\theta}$ resulting in large residuals. Suppose that σ is not perfectly known *a priori*, however (but is still assumed to be the same for all points in a single sensor dataset), and

is allowed to vary alongside $\boldsymbol{\theta}$: the left-hand term penalizes small (overly confident) values of σ, while the right-hand term penalizes large values of σ corresponding to an assumption that the data are entirely explained by observational noise.

Typical residuals from the Beardsmore et al. (2016) inferences correspond to about 10% of the dataset's full range, so we next perform a run in which the noise prior is set to $\alpha = 0.5, \beta = 0.05$ for all sensors (gravity, magnetic, and magnetotelluric). The corresponding likelihood (with the noise variance prior integrated out) becomes a Cauchy (or t_1) distribution, with thick tails that allow substantial outliers from the core. This choice of α and β thus also allows us to make contact with prior work where Cauchy distributions have been used (B.C Silva and Cutrim, 1989; de la Varga et al., 2018): a Gaussian likelihood with unknown, $\mathrm{IG}(0.5, \beta_s)$-distributed variance is mathematically equivalent to a Cauchy likelihood with known scale $2\beta_s$. The two choices are conceptually different, since in the Gaussian case outliers appear when the wrong variance scale is applied, whereas in the Cauchy case the scale is assumed known and the data have an intrinsically heavy-tailed distribution.

Under this new likelihood the residuals from the gravity observations increase (by about a factor of 1.5–2), while the residuals from the magnetic sensors decrease (by a factor of 3–4). This rebalancing of residuals among the sensors with an uninformative prior can be used to inform subsequent rounds of modelling more readily.

The inference also changes: in run D, a granite bridge runs from the main outcrop to the eastern edge of the modelled volume, with the presence of granite in the northwest corner being less certain. Agreement with run B and with the Beardsmore et al. (2016) map is still good along the eastern edge. The posterior entropy also decreases substantially, due to increase in the probability of granite structures at greater depths (beneath 3.5 km).

The weight given to the gravity sensor is thus an important factor determining the behavior of the inversion throughout half the modeled volume. With weakened gravity constraints, the two modes for the inferred rock density in layer 3 separate widely (see Fig. 3), though the algorithm is still able to move between the modes occasionally. The marginal distributions of the other rock properties do not change substantially, and remain unimodal.

The comparison map for the inversion of Beardsmore et al. (2016) comes from the deterministic inversion of Meixner and Holgate (2009), which uses gravity as the main surface sensor but relies heavily on seismic data, with reflection horizons used to constrain the depth to basement, and measurements of wave velocities (which correlate with density) from a P-wave refraction survey to constrain density at depth. While Meixner and Holgate (2009) mention constraints on rock densities, no mention is made of the level of agreement with the gravity data.

Without more information — a seismic sensor in our inversion, priors based on the specific seismic interpretations of Meixner and Holgate (2009), or specific knowledge about the noise level in the gravity dataset that would justify an informative prior — it is hard to say how concerned we should be about the differences between the deterministic inversion and our probabilistic version. The comparison certainly highlights the potential importance of seismic data, both as a constraint on rock properties at depth and on the geometry of geological structures.

Indeed, one potential weakness of this approach to balancing sensors is model inadequacy: the residuals from the inference may include systematic residuals from unresolved structure in the model, in addition to sensor noise. The presence of such residuals is a model selection question that in a traditional inversion context would be resolved by comparing residuals to the

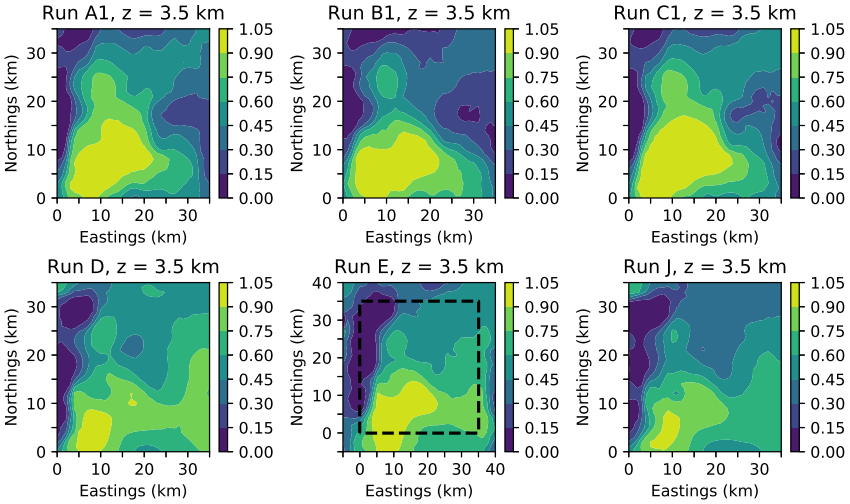

Figure 4. Slices through the voxelised posterior probability of occupancy by granite for each run at a depth of 3.5 km.

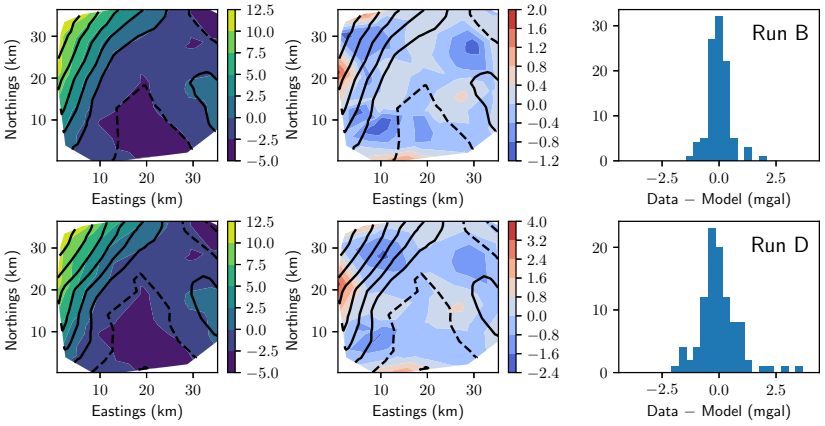

Figure 5. Gravity anomaly at the surface ($z = 0$). In a contour plot (left): filled contours = observations, black lines = mean posterior forward model prediction. Residuals of observations from the mean posterior forward model are also shown as a contour map (middle) and histogram (right).

assumed noise level, but this depends strongly upon informative prior knowledge of the sensing process for *all sensors* used in the inversion. The remaining experiments will use the Cauchy likelihood unless otherwise specified.

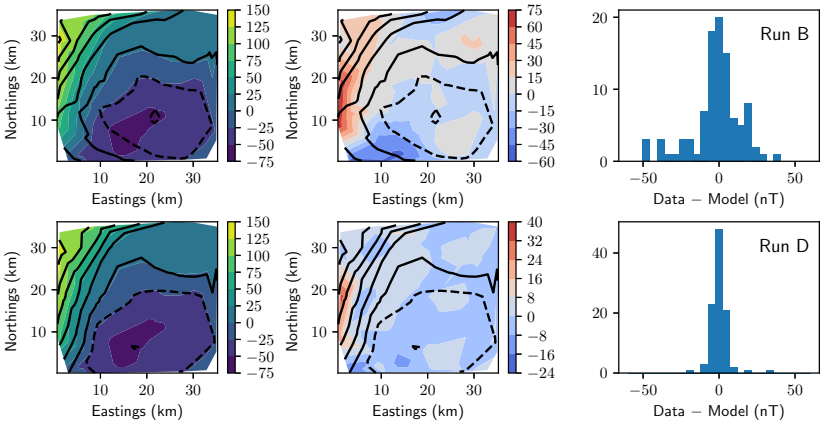

Figure 6. Magnetic anomaly at the surface ($z = 0$) in the same format as Fig. 5.

3.3 Boundary conditions

The boundary conditions Obsidian imposes on world voxelisations assume that rock properties rendered at a boundary edge (north/south, east/west) extend indefinitely off the edges, e.g.

$$\rho_{is}(x < x_{\min}) = \rho_{is}(x_{\min}), \tag{19}$$

$$\rho_{is}(x > x_{\max}) = \rho_{is}(x_{\max}). \tag{20}$$

This may not be a good approximation when rock properties show strong gradients near the boundary. The residual plots shown in Fig. 5 and 6 show persistently high residuals along the western edge of the world, where such gradients appear in both the gravity anomaly and the magnetic anomaly.

For geophysical sensors with localized response, one way to mitigate this is to include in the world representation a larger area than the sensor data cover, incorporating a margin with width comparable to the scale of boundary artifacts, in order to let the model respond to edge effects for sensors with a finite area of response. In run E, we add a boundary zone of width 5 km around the margins of the world, while increasing the number of control points in the granite intrusion layer boundary from 49 (7×7 grid) to 64 (8×8 grid). Neither the model residuals nor the inferred rock geometry substantially differs from the previous run, suggesting that the remaining outliers are actual outliers and not primarily due to mismatched boundary conditions. The autocorrelation time, however, increases substantially due both to the increase in the problem dimension and the fact that the new world parameters are relatively unconstrained, hence poorly scaled relative to the others.

3.4 Looser priors on rock properties and layer depths

In cases where samples of rock for a given layer are few or unavailable, the empirical covariance used to build the prior on rock properties may be highly uncertain or undefined. In these cases, the user may have to resort to a broad prior on rock properties.

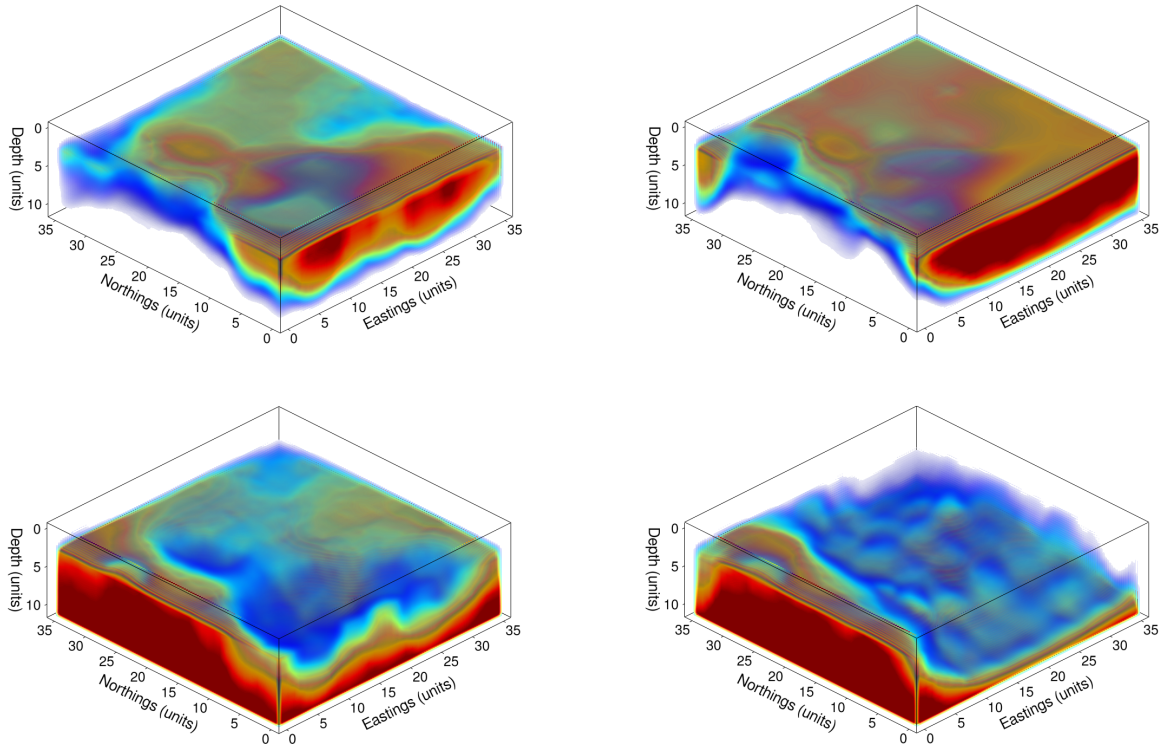

Figure 7. Volume renderings of the posterior mean for runs B (left) and D (right), demonstrating the change in inference at depth. Top: probability of occupancy for layer 5 (granite intrusion). Bottom: probability of occupancy for layer 6 (basement). Red volumes indicate high probability, blue volumes low probability; zero-probability regions have been rendered transparent to make the shape of the region more readily visible.

The limiting case is when no petrophysical data are available at all. Similarly, definitive data on layer depths may become unavailable in the absence of drill cores, or at least seismic data, so that a broad prior on control point depths may also become necessary.

We re-run the main Moomba analysis using two new priors. The first (run J) simulates the absence of petrophysical measurements. The layer depth priors are the same as the Beardsmore et al. (2016) setup, but the rock property prior for each layer is now replaced by an independent Gaussian prior on each rock property, with the same mean as in previous runs but a large width common to all layers:

$$\rho_{is} \sim \mathcal{N}(\mu_{\rho_{is}}, \sigma_{\rho_{is}}). \tag{21}$$

The standard deviations are 0.2 g cm^{-3} (density), 0.5 (log magnetic susceptibility), and 0.7 (log resistivity in Ω m).

The Run J voxelisation shows reasonable correspondence with the baseline run D, though with larger uncertainty, particularly in the northwest corner. In the absence of petrophysical samples, but taking advantage of priors on overlying structure from

seismic interpretations, a preliminary segmentation of granite from basement can thus still be obtained using broad priors on rock properties. Although the algorithm cannot reliably infer the bulk rock properties in the layers, the global prior on structure is enough for it to pick out the shapes of intrusions by looking for *changes* in bulk properties between layers.

The second run (run K) removes structural prior information instead of petrophysical prior information. The priors on rock properties are as in the Beardsmore et al. (2016) setup, but the control point prior for each layer is replaced by a multivariate Gaussian with the same anisotropic Gaussian covariance,

$$\mathbf{\Sigma_\alpha} = \sigma_\alpha \begin{bmatrix} 1.0 & 0.5 & \dots & 0.5 \\ 0.5 & 1.0 & \dots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \dots & 1.0 \end{bmatrix} \tag{22}$$

with $\sigma_\alpha = 3$ km.

Run K yields no reliable information about the location of granite at 3.5 km depth. This seems to be due solely to the uncertain thickness of layers of sedimentary rock that are constrained to be nearly uniform horizontal slabs in Run J, corresponding to a known insensitivity to depth among potential-field sensors. When relaxed, these strong priors cause a crisis of identifiability for the resulting models. Further variations on Runs J and K show that replacing these multiple thin layers with a single uniform slab of ~ 3 km depth (Runs J2 and K2) does not aid either convergence or accuracy, as long as more than one layer boundary is allowed to have large-scale structure.

As mentioned above and in Beardsmore et al. (2016), the strong priors on layer boundaries and locations were originally derived from seismic sensor data. Such data will not always be available, but seem to be critical to constrain the geometry of existing layers to achieve a plausible inversion at depth.

4 Discussion

Our experiments show concrete examples of how the efficiency of MCMC sampling changes with assumptions about the prior, likelihood, and proposal distributions for an Obsidian inversion, particularly as tight constraints on the solution are relaxed and uncertainty increases. Unrealistically tight constraints can hamper sampling, but relaxing priors or likelihoods may sometimes widen the separation between modes (as shown in Fig. 3), which also makes the posterior difficult to sample. Additionally, particular weaknesses in sensors, such as the insensitivity of potential-field sensors to the depth of geological features or to the addition of any horizontally invariant density distribution, can make it impossible to distinguish using those data between multiple plausible alternatives, adding to the irregularity and multi-modality of the posterior.

While any single data source may be easy to understand on its own, unexpected interactions between parameters can also arise. Structural priors from seismic data or geological field measurements appear to play a crucial role in stabilizing the inversions in this paper, as seen by the collapse of our inversion after relaxing them.

Our findings reinforce the impression that to make Bayesian inversion techniques useful in this context, the computational burden must be reduced by developing efficient sampling methods. Three complementary ways forward present themselves:

1. to develop MCMC proposals, or non-parametric methods to approximate probability distributions, that both function in (relatively) high-dimensional spaces and capture local structure in the posterior;

2. to develop fast approximate forward models for complex sensors (especially seismic) that deliver detailed information at depth, along with new ways of assessing and reducing model inadequacy;

3. to develop richer world parametrizations of 3-D geological models that faithfully represent real-world structure in as few dimensions as possible.

All three of the MCMC proposals studied here are variations of random walks, which explore parameter space by diffusion and do not easily handle posteriors with detailed local covariance structure such as the ones we find here. Proposals that can sense and adjust to local structure from the present state require, almost by definition, knowledge of the posterior's gradient (as in Hamiltonian Monte Carlo; Duane et al., 1987; Hoffman and Gelman, 2014; Neal et al., 2011) or higher-order curvature tensors (Girolami and Calderhead, 2011, as in Riemannian Manifold Monte Carlo;) with respect to the model parameters, which in turn require gradients of both the prior and the likelihood (in particular, of forward models).

Taking derivatives of a complex forward model by finite differences is also likely to be prohibitively expensive, and practitioners may not have the luxury of rewriting their forward model code to return derivatives. This is one goal of writing fast emulations of forward models, particularly emulations for which derivatives can be calculated analytically (see for example Fichtner et al., 2006a, b). Smooth universal approximators, such as artificial neural networks, are one possibility; Gaussian process latent variable models (Titsias and Lawrence, 2010) or Gaussian process regression networks (Wilson et al., 2012) are others, which would also enable nonlinear dimensionality reduction for difficult forward models or posteriors. reduction. Algorithms that alternate between fast/approximate forward models for local exploration, on the one hand, and expensive/precise forward models for evaluation of the objective function, on the other, have proved useful in engineering design problems Jin (2011); Sóbester et al. (2014). These approximate emulators give rise to model inadequacy terms in the likelihood, which can be explicitly addressed; for example, Köpke et al. (2018) present a geophysics inversion framework in which the inference scheme learns a model inadequacy term as sampling proceeds, showing proof of principle on a crosshole georadar tomography inversion. A related, complementary route is to produce analytically differentiable approximations to the posterior, built as the chain explores the space (Strathmann et al., 2015; Lan et al., 2016).

Another source of overall model inadequacy comes from the world parametrization which can be viewed as part of the prior. Obsidian's world parametrization is tuned to match sedimentary basins, and is thus best suited for applications such as oil, gas, and geothermal exploration; it is too limited to represent more complex structures, particularly those with folds and faults, that might arise in hard rock or mining scenarios. The GemPy package developed by de la Varga et al. (2018) makes an excellent start on a more general-purpose open-source code for 3-D geophysical inversions: it uses the implicit potential-field approach (Lajaunie et al., 1997) to describe geological structures, includes forward-models for geophysical sensors, and is designed to produce posteriors that can easily be sampled by MCMC. GemPy is also specifically written to take advantage of autodifferentiation, providing ready derivative information for advanced MCMC proposals.

5 Conclusions

We have performed a suite of 3-D Bayesian geophysical inversions for the presence of granite at depth in the Moomba gas field of the Cooper basin, altering aspects of the problem setup to determine their effects on the efficiency and accuracy of MCMC sampling. Our main findings are as follows:

- Parametrized worlds have much lower dimensionality than non-parametric worlds, and the parameters also offer a more interpretable description of the world — for example, boundaries between geological units are explicitly represented. However, the resulting posterior has complex local covariance structure in parameter space, even for linear sensors.
- Although isotropic random walk proposals explore such posteriors inefficiently, poorly adapted anisotropic random walks are even less efficient. A modified high-dimensional random walk such as pCN outperforms these proposals, and the prior-preserving properties of pCN make it especially attractive for use in tempered sampling.
- The shape of the posterior and number of modes can also depend in complex ways upon the prior, making tempered proposals essential.
- Hierarchical priors on observational noise provide a way to capture uncertainty about the weighting among datasets, although this may also make sampling more challenging as when priors on world parameters are relaxed.
- Useful information about structures at depth can sometimes be obtained through sensor fusion even in the absence of informative priors. However, direct constraints on 3-D geometry from seismic interpretations or structural measurements seem to play a privileged role among priors, owing to the relatively weak constraints on depth of structure afforded by potential field methods.

In summary, both advanced MCMC methods and careful attention to the properties of the data are necessary for inversions to succeed.

Code and data availability. The code for version 0.1.2 of Obsidian is available at https://github.com/rscalzo/obsidian/tree/0.1.2-beta. All configuration files for 3-D model runs specified in this paper, together with corresponding datasets, are available in named subfolders of https://github.com/rscalzo/obsidian/tree/0.1.2-beta/examples/scalzo18, and are also provided as supplementary material.

Appendix A: Analytic integration of likelihood over sensor noise prior

The usual mean-square likelihood often used in geophysical sensor inversions assumes the residuals of the sensor measurements from each forward model are independent Gaussian-distributed with some variance σ^2. In typical non-probabilistic inversions, this noise amplitude is specified exactly as part of the objective function. A probabilistic inversion would specify a prior $P(\sigma^2)$ for the probability density of the noise variance, and condition the likelihood $P(y|\theta, \sigma^2)$ on this variance and on the other parameters describing the world.

Unless the noise levels in the sensors are themselves targets for inference, sampling will be more efficient if their values are integrated out beforehand. If the conditional likelihood $P(y|\theta,\sigma^2)$ is independent Gaussian, a point mass prior $P(\sigma^2) = \delta(\sigma^2 - \sigma_0^2)$ results in a Gaussian likelihood $P(y|\theta)$. If some uncertainty about σ^2 exists, specifying the prior $P(\sigma^2)$ as an inverse gamma function $\mathrm{IG}(\sigma^2|\alpha,\beta)$ enables the integration over σ^2 to be done analytically. The parameter α describes the weight of the tail in the distribution of σ^2, while β/α gives a typical variance scale referred to the sample variance of the data; these parameters are then either specified by experts with prior knowledge about the sensors, or are set to uninformative values e.g. $\alpha = \beta = 1$.

For a single observation y, we start with

$$P(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{A1}$$

$$P(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-2(\alpha+1)} \exp\left(\frac{-\beta}{\sigma^2}\right) \tag{A2}$$

(where the mean μ is given by the forward model). Carrying out the integration over the prior proceeds as follows:

$$P(y|\mu,\alpha,\beta) = \int_0^\infty P(y|\mu,\sigma^2) P(\sigma^2|\alpha,\beta)\, d\sigma^2 \tag{A3}$$

$$= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}}\, \sigma^{-2(\alpha+1)} e^{\frac{-\beta}{\sigma^2} - \frac{(y-\mu)^2}{2\sigma^2}} \tag{A4}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi}} \int_0^\infty u^{\alpha-\frac{1}{2}} e^{-\frac{1}{2}\left[(y-\mu)^2+\beta\right)]u}\, du \tag{A5}$$

$$= \frac{\beta^\alpha \left[(y-\mu)^2+\beta)\right]^{-\left(\alpha+\frac{1}{2}\right)} \Gamma\left(\alpha+\frac{1}{2}\right)}{\Gamma(\alpha)\sqrt{2\pi}} \tag{A6}$$

$$= \frac{1}{\sqrt{2\pi\beta}} \frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\Gamma(\alpha)} \left[\frac{(y-\mu)^2}{2\beta} + 1\right]^{-\left(\alpha+\frac{1}{2}\right)} \tag{A7}$$

which is the density for a t-distribution in the normalized residual $\xi = \frac{y-\mu}{\sqrt{\beta/\alpha}}$ with $\nu = 2\alpha$ degrees of freedom.

Author contributions. The study was conceptualized by SC, who with MG provided funding and resources. RS was responsible for project administration and designed the methodology under supervision from SC and GH. RS and DK carried out development of the Obsidian code resulting in v0.1.2, carried out the main investigation and formal analysis, and validated and visualized the results. RS wrote the original draft text, of which all co-authors provided review and critical evaluation.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work is part of the Lloyd's Register Foundation – Alan Turing Institute Programme for Data-Centric Engineering. RS thanks Lachlan McCalman, Simon O'Callaghan, and Alistair Reid for useful discussions about the development of Obsidian up to v0.1.1.

References

Agostinetti, N. P. and Malinverno, A.: Receiver function inversion by trans-dimensional Monte Carlo sampling, Geophysical Journal International, 181, 858–872, https://doi.org/10.1111/j.1365-246X.2010.04530.x, 2010.

Anand, R. R. and Butt, C. R. M.: A guide for mineral exploration through the regolith in the Yilgarn Craton, Western Australia, Australian Journal of Earth Sciences, 57, 1015–1114, 2010.

B.C Silva, J. and Cutrim, A.: A robust maximum likelihood method for gravity and magnetic interpretation, Geoexploration, 26, 1–31, 1989.

Beardsmore, G.: Data fusion and machine learning for geothermal target exploration and characterisation, Tech. rep., NICTA Final Report, https://arena.gov.au/projects/data-fusion-and-machine-learning-for-geothermal/, 2014.

Beardsmore, G., Durrant-Whyte, H., McCalman, L., O'Callaghan, S., and Reid, A.: A Bayesian inference tool for geophysical joint inversions, ASEG Extended Abstracts, 2016, 1–10, 2016.

Bodin, T., Sambridge, M., Tkalcic, H., Arroucau, P., Gallagher, K., and Rawlinson, N.: Transdimensional inversion of receiver functions and surface wave dispersion, Solid Earth, 117, 2012.

Calcagno, P., Chilès, J., Courrioux, G., and Guillen, A.: Geological modelling from field data and geological knowledge: Part I. Modelling method coupling 3D potential-field interpolation and geological rules, Physics of the Earth and Planetary Interiors, 171, 147 – 157, recent Advances in Computational Geodynamics: Theory, Numerics and Applications, 2008.

Carr, L., Korsch, R. J., Reese, B., and Palu, T.: Onshore Basin Inventory: The McArthur, South Nicholson, Georgina, Wiso, Amadeus, Warburton, Cooper and Galilee basins, central Australia, Geoscience Australia, 2016.

Chandra, R., Azam, D., Müller, R. D., Salles, T., and Cripps, S.: BayesLands: A Bayesian inference approach for parameter uncertainty quantification in Badlands, arXiv preprint arXiv:1805.03696, 2018.

Chib, S. and Greenberg, E.: Understanding the Metropolis-Hastings Algorithm, The American Statistician, 49, 327–335, 1995.

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D.: MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster, Statistical Science, 28, 424–446, 2013.

Cramer, H.: Mathematical methods of statistics, Princeton University Press, 1946.

de la Varga, M. and Wellmann, J. F.: Structural geologic modeling as an inference problem: A Bayesian perspective, Interpretation, 4, SM1–SM16, 2016.

de la Varga, M., Schaaf, A., and Wellmann, F.: GemPy 1.0: open-source stochastic geological modeling and inversion, Geoscientific Model Development, pp. 1–50, 2018.

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

Fichtner, A., Bunge, H.-P., and Igel, H.: The adjoint method in seismology: I. Theory, Physics of the Earth and Planetary Interiors, 157, 86–104, 2006a.

Fichtner, A., Bunge, H.-P., and Igel, H.: "The adjoint method in seismology: II. Applications: travel times and sensitivity functionals", Physics of the Earth and Planetary Interiors, 157, 105–123, 2006b.

Gelman, A. and Rubin, D.: Inference from Iterative Simulation Using Multiple Sequences, Statistical Science, 7, 457–472, 1992.

Geyer, C. J. and Thompson, E. A.: Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference, Journal of the American Statistical Association, 90, 909–920, 1995.

Giraud, J., Jessell, M., Lindsay, M., Martin, R., Pakyuz-Charrier, E., and Ogarko, V.: Uncertainty reduction of gravity and magnetic inversion through the integration of petrophysical constraints and geological data, in: EGU General Assembly Conference Abstracts, vol. 18, pp. EPSC2016–3870, 2016.

Giraud, J., Pakyuz-Charrier, E., Jessell, M., Lindsay, M., Martin, R., and Ogarko, V.: Uncertainty reduction through geologically conditioned petrophysical constraints in joint inversion, Geophysics, 82, ID19–ID34, 2017.

Giraud, J., Pakyuz-Charrier, E., Ogarko, V., Jessell, M., Lindsay, M., and Martin, R.: Impact of uncertain geology in constrained geophysical inversion, ASEG Extended Abstracts, 2018, 1, 2018.

Girolami, M. and Calderhead, B.: Riemann manifold langevin and hamiltonian monte carlo methods, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73, 123–214, 2011.

Goodman, J. and Weare, J.: Ensemble samplers with affine invariance, Communications in Applied Mathematics and Computational Science, 5, 65–80, 2010.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P.: Bayesian computation: a perspective on the current state, and sampling backwards and forwards, Statistics and Computing, 25, arXiv:1502.01148, 2015.

Gupta, V. K. and Grant, F. S.: 30. Mineral-Exploration Aspects of Gravity and Aeromagnetic Surveys in the Sudbury-Cobalt Area, Ontario, pp. 392–412, Society of Exploration Geophysicists, 1985.

Haario, H., Saksman, E., and Tamminen, J.: An Adaptive Metropolis Algorithm, Bernoulli, 7, 223, 2001.

Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications, Biometrika, 57, 97–109, 1970.

Hoffman, M. D. and Gelman, A.: The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo., Journal of Machine Learning Research, 15, 1593–1623, 2014.

Jessell, M.: Three-dimensional geological modelling of potential-field data, Computers & Geosciences, 27, 455 – 465, 3D reconstruction, modelling & visualization of geological materials, 2001.

Jin, Y.: Surrogate-assisted evolutionary computation: Recent advances and future challenges, Swarm and Evolutionary Computation, 1, 61 – 70, 2011.

Köpke, C., Irving, J., and Elsheikh, A. H.: Accounting for model error in Bayesian solutions to hydrogeophysical inverse problems using a local basis approach, Advances in Water Resources, 116, 195–207, 2018.

Lajaunie, C., Courrioux, G., and Manuel, L.: Foliation fields and 3D cartography in geology: Principles of a method based on potential interpolation, Mathematical Geology, 29, 571–584, https://doi.org/10.1007/BF02775087, 1997.

Laloy, E., Linde, N., Jacques, D., and Mariethoz, G.: Merging parallel tempering with sequential geostatistical resampling for improved posterior exploration of high-dimensional subsurface categorical fields, Advances in Water Resources, 90, 57–69, 2016.

Lan, S., Bui-Thanh, T., Christie, M., and Girolami, M.: Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian Inverse Problems, Journal of Computational Physics, 308, 81 – 101, 2016.

Lindsay, M., Jessell, M., Ailleres, L., Perrouty, S., de Kemp, E., and Betts, P.: Geodiversity: Exploration of 3D geological model space, Tectonophysics, 594, 27 – 37, 2013.

MacCarthy, J. K., Borchers, B., and Aster, R. C.: Efficient stochastic estimation of the model resolution matrix diagonal and generalized cross–validation for large geophysical inverse problems, Journal of Geophysical Research, 116, B10 304, https://doi.org/10.1029/2011JB008234, 2011.

McCalman, L., O'Callaghan, S. T., Reid, A., Shen, D., Carter, S., Krieger, L., Beardsmore, G. R., Bonilla, E. V., and Ramos, F. T.: Distributed bayesian geophysical inversions, Proceedings of the Thirty-Ninth Workshop on Geothermal Reservoir Engineering, Stanford University, pp. 1–11, 2014.

Meixner, T. and Holgate, F.: The Cooper Basin Region 3D Geological Map Version 1: A search for hot buried granites, Geoscience Australia, Record 2009/15, 2009.

Menke, W.: Geophysical Data Analysis (Revised Edition), Elsevier Ltd, 2018.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E.: Equation of state calculations by fast computing machines, The journal of chemical physics, 21, 1087–1092, 1953.

Miasojedow, B., Moulines, E., and Vihola, M.: An adaptive parallel tempering algorithm, Journal of Computational and Graphical Statistics, 22, 649–664, 2013.

Mockus, J.: Bayesian approach to global optimization: theory and applications, Kluwer Academic Press, 2013.

Mosegaard, K. and Tarantola, A.: Monte Carlo sampling of solutions to inverse problems, Journal of Geophysical Research: Solid Earth, 100, 12 431–12 447, 1995.

Nabighian, M. N., Ander, M. E., Grauch, V. J. S., Hansen, R. O., LaFehr, T. R., Li, Y., Pearson, W. C., Peirce, J. W., Phillips, J. D., and Ruder, M. E.: Historical development of the gravity method in exploration, Geophysics, 70, 63ND–89ND, 2005a.

Nabighian, M. N., Grauch, V. J. S., Hansen, R. O., LaFehr, T. R., Li, Y., Peirce, J. W., Phillips, J. D., and Ruder, M. E.: The historical development of the magnetic method in exploration, Geophysics, 70, 33ND–61ND, 2005b.

Neal, R. M. et al.: MCMC using Hamiltonian dynamics, Handbook of Markov Chain Monte Carlo, 2, 2011.

Olierook, H. K. H., Timms, N. E., Wellmann, J. F., Corbel, S., and Wilkes, P. G.: 3D structural and stratigraphic model of the Perth Basin, Western Australia: Implications for sub-basin evolution, Australian Journal of Earth Sciences, 62, 447–467, 2015.

Olierook, H. K. H., Scalzo, R., Kohn, D., Chandra, R., Farahbakhsh, E., Houseman, G., Clark, C., Reddy, S. M., and Müller, R. D.: Bayesian geological and geophysical data fusion for the construction and uncertainty quantification of 3D geological models, Solid Earth Discuss., in review, https://doi.org/10.5194/se-2019-4, 2019.

Pakyuz-Charrier, E., Giraud, J., Ogarko, V., Lindsay, M., and Jessell, M.: Drillhole uncertainty propagation for three-dimensional geological modeling using Monte Carlo, Tectonophysics, 747-748, 16–39, https://doi.org/10.1016/j.tecto.2018.09.005, https://linkinghub.elsevier.com/retrieve/pii/S004019511830310X, 2018a.

Pakyuz-Charrier, E., Lindsay, M., Ogarko, V., Giraud, J., and Jessell, M.: Monte Carlo simulation for uncertainty estimation on structural data in implicit 3-D geological modeling, a guide for disturbance distribution selection and parameterization, Solid Earth, 9, 385–402, 2018b.

Pall, J., Chandra, R., Azam, D., Salles, T., Webster, J., and Cripps, S.: BayesReef: A Bayesian inference framework for modelling reef growth in response to environmental change and biological dynamics, arXiv preprint arXiv:arXiv:tba, 2018.

Rao, C. R.: Information and the accuracy attainable in the estimation of statistical parameters, Bulletin of the Calcutta Mathematical Society, 37, 81–89, 1945.

Roberts, G. O. and Rosenthal, J. S.: Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms, Journal of Applied Probability, 44, 458–475, 2007.

Roberts, G. O., Gelman, A., Gilks, W. R., et al.: Weak convergence and optimal scaling of random walk Metropolis algorithms, Annals of Applied Probability, 7, 110–120, 1997.

Ruggeri, P., Irving, J., and Holliger, K.: Systematic evaluation of sequential geostatistical resampling within MCMC for posterior sampling of near-surface geophysical inverse problems, Geophysical Journal International, 202, 961–975, 2015.

Sabins, F. F.: Remote sensing for mineral exploration, Ore Geology Reviews, 14, 157 – 183, 1999.

Salama, W., Anand, R. R., and Verrall, M.: Mineral exploration and basement mapping in areas of deep transported cover using indicator heavy minerals and paleoredox fronts, Yilgarn Craton, Western Australia, Ore Geology Reviews, 72, 485 – 509, 2016.

Sambridge, M.: Exploring multidimensional landscapes without a map, Inverse Problems, 14, 427–440, https://doi.org/10.1088/0266-5611/14/3/005, 1998.

Sambridge, M. and Mosegaard, K.: Monte Carlo methods in geophysical inverse problems, Reviews of Geophysics, 40, 2002.

Sambridge, M., Bodin, T., Gallagher, K., and Tkalcic, H.: Transdimensional inference in the geosciences, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371, 20110 547–20110 547, iSBN: 1364503X (ISSN), 2012.

Scalzo, R., Kohn, D., OĆallaghan, S., McCalman, L., and Simpson-Young, B.: rscalzo/obsidian: 0.1.2-beta, https://doi.org/10.5281/zenodo.2580422, 2019.

Shannon, C. E.: A mathematical theory of communication, Bell System Technical Journal, 27, 379–423, 1948.

Sóbester, A., Forrester, A. I. J., Toal, D. J. J., Tresidder, E., and Tucker, S.: Engineering design applications of surrogate-assisted optimization techniques, Optimization and Engineering, 15, 243–265, 2014.

Strangway, D. W., C. M. Swift, J., and Holmer, R. C.: The application of audio-frequency magnetotellurics (AMT) to mineral exploration, Geophysics, 38, 1159–1175, 1973.

Strathmann, H., Sejdinovic, D., Livingstone, S., Szabo, Z., and Gretton, A.: Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families, in: Advances in Neural Information Processing Systems, pp. 955–963, 2015.

Tarantola, A.: Inverse problem theory and methods for model parameter estimation, vol. 89, siam, 2005.

Tarantola, A. and Valette, B.: Generalized nonlinear inverse problems solved using the least squares criterion, Reviews of Geophysics, 20, 219–232, 1982.

Titsias, M. and Lawrence, N.: Bayesian Gaussian Process Latent Variable Model, Artificial Intelligence, 9, 844–851, arXiv: 1309.6835 ISBN: 978-1-4503-1285-1, 2010.

Wellmann, J. F. and Regenauer-Lieb, K.: Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models, Tectonophysics, 526, 207–216, 2012.

Wellmann, J. F., Horowitz, F. G., Schill, E., and Regenauer-Lieb, K.: Towards incorporating uncertainty of structural data in 3D geological inversion, Tectonophysics, 490, 141 – 151, 2010.

Wilson, A. G., Knowles, D. A., and Ghahramani, Z.: Gaussian Process Regression Networks, International Conference on Machine Learning, p. 17, arXiv: 1110.4411, 2012.

Wrona, T., Pan, I., Gawthorpe, R. L., and Fossen, H.: Seismic facies analysis using machine learning, Geophysics, 83, O83–095, 2018.

Xiang, E., Guo, S. E., Liu, J., Dong, H., and Ren, Z.: Efficient hierarchical transdimensional Bayesian inversion of magnetotelluric data, Geophysical Journal International, 37, 81–89, 2018.