Comments from Reviewer 1:

1. Line 25: I do not quite see how a physical relationship predicts noise distributions. The noise distribution is, in fact, that part of the data that the physical relation cannot predict, bascially by definition.

We agree that the noise distribution cannot be predicted. We have deleted noise distributions. Now we say "a physical relationship is defined that predicts data that would be recorded for any particular set of model parameters…".

2. Line 27: I wonder why you restrict yourself to imaging problems. Your method should be applicable to a much broader range of inverse problems, beyond the imaging class.

We agree that the method can be applied to a broader range of inverse problems. We restrict ourselves to imaging problems in the manuscript mainly because we have so far only applied the method to imaging problems.

3. Line 29: This is not a particularity of geophysical inverse problems but of practically any inverse problem.

Thank you for the comment. We have changed to "inverse problems" instead of "geophysical inverse problems".

4. Line 33: I think this sentence is basically a contradiction in itself because finding an optimum just does not mean that you have actually solved an inverse problem! Also, note that people are often more careful and do not find an optimum but merely a set of model parameters that fits the data to within the noise. Otherwise, you risk fitting the noise.

We agree that finding an optimum does not mean an inverse problem is completely solved. We note that this is just the standard way how we deal with inverse problems. To be clearer, we have changed to:
"Solutions to an inverse problem are often found by seeking an optimal set of parameter values that minimizes the difference or misfit between observed data and model-predicted data to within the data noise".

5. Line 36: Isn't code and software the same thing?

We have deleted "and software".

6. Line 56: style: applied to applications

Thank you for the comment. We have changed to "applied to a range of geophysical inverse problems".

7. Line 59: I do not think this is the real problem. THE problem is the poor scaling of Metropolis-Hastings. The number of samples that you need to draw in order to obtain an independent sample grows as n^2, where n is the model space dimension. Hence, when you go to higher dimensions, you simply need too many samples to achieve convergence.

We agree. We have changed our description to "However, the algorithm becomes inefficient in high

dimensional space because of poor scaling due to its random walk behaviour".

8. Line 61: I think this formulation is not so clean. In fact, many of the methods that you list afterwards are actually McMC methods.

We have changed to "In order to solve Bayesian inference problems more efficiently, a variety of more advanced methods have been introduced to geophysics, …".

9. Line 69: I think this statement is a kittle diffuse. What means "large-scale" or "extremely high"? Not sure such sentences are useful.

Thank you for the comment. To be clearer, we have changed our description to: "Bayesian solutions to large scale problems (e.g., those involving thousands of parameters to be estimated) remain intractable because of their unaffordable computational cost due to the curse of dimensionality".

10. Line 74: I never quite understood why this particular measure is chosen. Can this be explained in a few words? Is it mere computational convenience?

Yes, you are right. It is mainly because KL divergence is easy to estimate computationally. We have added this: "One commonly-used measure of the difference between the pdfs is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1975) as it is easier to estimate computationally than other measures."

11. Line 77: The logic of this argument is not clear. Why exactly is optimisation more efficient than sampling?

Thank you for the comment. It is not theoretically clear that optimisation is more efficient than sampling. But from a practical point of view, many studies have demonstrated that optimisation is more efficient than sampling in Bayesian inversion (Bishop 2006; Blei et al., 2017). Also, traditional optimisation-based methods are more efficient than sampling-based methods in solving inverse problems. Of course, we note that this optimisation is different from optimisation in variational inference, but they share some common characteristics.

To be clearer, we changed our description to:
"The method has been demonstrated to be computationally more efficient and more scalable to high dimensionality in some class of problems…"

12. Line 80: This statement is too strong. Of course, mini-batch approaches work with McMC provided that the batch is large enough to represent the full dataset. McMC methods are pretty forgiving when it comes to this kind of sloppiness.

We agree. But as you said this requires the batch to be large enough to represent the full dataset, whereas variational inference does not have such strong requirements. On the other hand, we also note that this will inevitably reduce accuracy of McMC methods, which is the most important feature of McMC methods when compared with variational inference. We have changed our description to "the same strategy cannot easily be used for McMC…"

13. Line 82: Also here I think you are a bit too harsh to McMC, which can be trivially parallelised by running multiple chains; either just in a sloppy way or with parallel tempering.

We slightly disagree with this comment. We agree that McMC can be parallelised by running multiple chains. However, in principle we would then require that every chain converges, and in such cases parallelisation does not help much. In addition, as our experience much computation is spent in the burn-in period which does not contribute to the final samples, so parallelisation across chains contributes limited improvements to the overall efficiency.

14. Line 84: style again: applied to applications

We have changed to “applied to a range of geophysical inverse problems”.

15. Line 87: No superlatives in scientific text!

We have deleted “extremely”.

16. Line 92: Used for what?

We have changed to “Kucukelbir et al. (2017) used a Gaussian family in variational inference to create a method called …”

17. Line 102, particles

Done.

18. Theoretical background section, I really appreciate the authors' effort to make the theory understandable. Still, I afraid that this section is too condensed for people who do not know the method yet. However, expanding this overview also does not make much sense. A difficult issue.

Thank you for the comment. As also suggested by Reviewer 2, we have expanded the description to give a more detailed overview of the method. See the new Theoretical background section.

19. Line 130, This is the crux of the whole problem, which, I think, deserves a bit more attention. How do you know a priori what a useful family is? It seems like this connects to the No-Free-Lunch Theorem: The method becomes efficient only when you chose a family that is actually useful, which is something you can only do when you have an a priori idea about the solution.

We agree. The family plays an important role in variational inference as it determines the accuracy of the approximation. In practice we can only choose some simplified family that provides the information that we seek (such as a specific closed-form approximation to the posterior distribution) or such that the optimisation problem can easily be solved, (as occurs, for example, when using a Gaussian family). So you are right, this connects to the No-Free-Lunch theorem. For example, a mean-field family or a Gaussian family is easy to solve, but they cannot provide a good approximation; whereas using samples (particles) to approximate the posterior pdf can be more accurate, but the problem will be more difficult to solve. We have added this information on page 6:

“In variational inference, the choice of the variational family Q is important because it determines both the accuracy of the approximation and the complexity of the optimisation problem. A good choice should be rich enough to approximate the posterior pdf accurately or at least provides the information that we seek, but simple enough such that the optimisation problem is tractable. Different choices of family may also allow different types of algorithm to be developed.”

20. Line 135, This is not easy to see, given that q in eq. (3) is arbitrary. Can you explain?

The nonnegativity of KL divergence is not obvious. It requires mathematical proof. Since this is not our main purpose, we do not prove it in the manuscript.

For your information, a simple proof is as follows:
Use the fact that $\ln a \leq a - 1$ for all $a > 0$,
$-KL(q\|p) = -\sum_x q(x) log\frac{q}{p} = \sum_x q(x) log\frac{p}{q} \leq \sum_x q(x)\left(\frac{p}{q} - 1\right) = \sum_x [p(x) - q(x)] =$
$\sum_x p(x) - \sum_x q(x) = 0$

21. Line 142, Now I do not understand your previous argument anymore. If the evidence is treated as a constant, then you can actually directly minimise equation (3).

Yes, minimise equation 3 is equivalent to maximize ELBO. The only difference is that we cannot calculate the actual value of equation 3 (only up to a constant), but can compute the value of the ELBO. From the optimisation perspective, they are essentially the same because a constant does not affect the optimisation procedure.

22. Line 156: Does this mean that you only consider one Gaussian? This suggests that you only try to find the posterior mean covarience.

Yes, you are right. In this method we only use one Gaussian, and it only finds the posterior mean and covariance.

23. Line 273: This statement is far too general to be universally true! It is not difficult to come up with code examples that runs as fast as C++ code.

We agree. We have changed to "… which suffers from slow execution for computationally intensive numerical simulations".

24. Line 276: Dask, Please explain what that is.

We have added explanation of Dask. Now we say:
"… we use a Python library called Dask, which is designed for parallel and distributed computing, to parallelize the forward computation at the sample (particle) level."

25. Line 296: What is the traveltime of a velocity? Velocities do not travel.

We have changed to "Travel times associated with group velocity at different … ".

26. Line 305: Is this not a contradiction to what you wrote above? You wrote that gradients from the traveltime tomography example are computed with ray tracing. However, now you use fast marching for the forward problem, meaning that you should obtain gradient from the adjoint fast marching code.

Thank you for the comment. We use the fast marching method for the travel time field calculation, and then obtain gradients by tracing rays through the travel time field. This is one of the standard ways to calculate gradients (see Rawlinson and Sambridge, 2004 – cited in the manuscript). We note that the adjoint method can also be used to calculate gradients. From our experience, the two methods produce similar results.

27. Figure 3: I am surprised by these results. There is absolutely no data coverage east of the island. Therefore, the standard deviations should be much larger, equal to the standard deviation of the prior.

Thank you for the comment. Outside of the island, the standard deviation is around 0.93 which is the standard deviation of the prior as expected. On the east side of the island just off the coast, although no seismometer is deployed, there are rays that travel through those areas (see details in Galetti et al., 2017), and consequently the standard deviation is smaller than that of the prior. We have added this information on page 16:

"In the offshore areas the standard deviation is around 0.93 which is the standard deviation of the prior as no ray path goes through these regions. By contrast, on the east side of the island just off the coast, although no seismometer is deployed, there are rays that travel through those areas (see details in Galetti et al., 2017), and consequently the standard deviation is smaller than that of the prior."

28. Line 310 – 319: It is difficult to avoid the impression that this is a comparison of apples and oranges. Why did you choose all of these different setups, and what makes them comparable in a meaningful way?

We agree that it is difficult to compare different methods in a completely fair way. However, it would still be useful to compare them on a specific problem so that practitioners can at least have an idea of the character of each method and their required computational cost. To provide a relatively fair comparison, for each method we followed the best practice from other studies. For example, for ADVI we used a standard Gaussian as the starting point and used the ADAM optimizer (Kucukelbir et al., 2017). For SVGD, we used 500 initial particles and updated them until the mean and standard deviation models become stable. For the effects of different number of particles on the results, see the discussion in Zhang et al., (2021). For sSVGD, when the average misfit value across all particles becomes stationary (i.e., post burn-in), we started to collect samples. The number of particles is selected by trial and error so that we can use a minimum number of particles to generate accurate results. We have added the information in the text:

"For ADVI, we started the method with a standard Gaussian distribution in the unconstrained space and performed 10,000 iterations at which point the misfit value ceases to decrease using the ADAM optimisation algorithm. For SVGD, we generated 500 particles from the prior distribution and updated them using equation (17) for 3,000 iterations at which point the mean and standard deviation models became stable. For sSVGD, we started from 20 particles generated from the prior distribution and updated them using equation (22) for 6,000 iterations after an additional burn-in period of 2,000 iteration, after which the average misfit value across all particles became approximately stationary."

29. Line 333: See my comment above. Why is this a meaningful statement, given that it is not obvious how these different setups can be compared. In fact, this touches upon another very important issue: Convergence! How can convergence of all these implementations be assessed? How is a user who does not understand all the methodological details supposed to decide on a specific setup for a specific problem? Is the statement in this sentence not just saying that the different methods have not converged to the same extent?

Although it is difficult to fully assess the convergence, it can at least be estimated in several ways in practice (see our discussion in point 28). In addition, we also extended each method for many more iterations, to check that the results showed similar features. As a result, we are confident that the main features of the solution from each method have essentially converged.

30. Line 359: See comment above. This suggests that you do not actually compute the exact derivatives of the forward problem.

See our discussion in 26. It might be true that the adjoint method would provide more accurate derivatives, but from our tests using rays to compute derivatives is sufficient. In addition, ray tracing is computationally more efficient.

31. Line 376: There is no such thing as a true model. It should better be called input model.

Thank you for the comment. Since the input model here is the model we proposed as the "ground truth", we prefer to call it "true structure" as input model does not really sound like ground truth. We have changed to "true structure" in the manuscript.

32. Line 399: "because of lower resolution in those areas", I do not quite understand the logic of this statement.

In the deep part (> 1.5 km) and close to the sides, the data are less sensitive to the velocity, and consequently there is higher uncertainty in those areas. As a result, the mean model is less similar to the input (true) model. We have changed our description to:

"In the deep part (> 1.5 km) and close to the sides, the mean models appear to be less similar to the input model because the waveform data are less sensitive to the velocity structure in those areas."

33. Figure 4: "10 sources …", style: sentence should not start with numbers

Done.

Comments from Reviewer 2:

1 General comments

I found this manuscript interesting to read, and the authors' initiative to release their Variational Inversion Package (VIP) will certainly benefit to the geosciences community.

We thank the reviewer for the positive feedback.

My main comment is about the writing of Section 2, about Variational Inversion (VI) theory. That is, as a non expert of VI methods, I've found Sect. 2 difficult to read and fully understand. Indeed, I think that the authors assume unduly prior knowledge, on VI methods, from the lay-reader. This seems to contradict the initial motivation of this paper: to release a VI package to help spreading/widening the use of VI methods in the community. That is, the authors write in the introduction: "VI has not been widely used in geophysics because the method is not easily accessible to non specialists". I then encourage the authors to make further efforts to explain the basics of VI, in this companion paper with respect to the VIP codes. And to refrain from jumping so fast from one equation to another. More precisely, without having a close look at other (more complete technically) papers (for example the authors's paper Zhang et al 2023), it is not easy to follow all the equations here. From the point of view of possible future practitioners of VI, this may be a little discouraging. Therefore, I think, it would benefit the paper to explain the VI methods in a much more pedagogical way, in Section 2.

Thank you for the comment. We have expanded Section 2 so that the theory is more clearly explained. See details in the new Section 2. Note that we still omitted detailed mathematical derivations in several places as they are not important for practitioners to understand and use the method.

2 Other comments (listed in no preferential order)

1. Lines 33–44: Though VI methods are suited to tackle non-linear inverse problems, they may become computationally (too?) costly when facing large scale 3–D tomographic problems, because of the 'curse of dimensionality', and are also dependent on the choice of prior information on the model solution. For completeness, the authors could then also mention in the introduction that progress was recently made for solving large-scale linear(-ised) 3–D tomographic problems using the SOLA-Backus-Gilbert inversion (Zaroli, 2016, 2019). That is, SOLA tomography seeks local-average properties of the 'true' Earth model, accompanied with information on resolution and uncertainty. It strictly avoids using any a priori information/constraints on the model itself —hence avoiding related bias effects (Zaroli, Koelemeijer, Lambotte, 2017). Moreover, it is not necessary to discretise the infinite-dimensional model space —hence avoiding discretisation-related artefacts.

Thank you for the comment. We have added this information:
"To overcome these issues, the SOLA-Backus-Gilbert inversion method has recently been applied to large scale linearised tomographic problems. This method evaluates the weighted average of the true model parameters and provides both resolution and uncertainty estimates (Zaroli, 2016; Zaroli et al., 2017). In addition, the method does not require regularization and can be conducted in a parameter-free way which avoids bias caused by parameterisation (Zaroli, 2019). Unfortunately, the method is only developed for linear problems; since most Geophysical problems are significantly nonlinear, our goal is to provide methods that estimate solutions and uncertainties for that class of problems."

2. Lines 61–66: Maybe the authors should also mention HMCLab (Zunino et al., 2023), a framework for solving inverse problems using the Hamiltonian Monte Carlo method, and also mention the work on Gaussian process models by Valentine and Sambridge (2020).

We have added these references in the main text.

3. Line 134: Eq and p(m, dobs) should be defined explicitly here (I note that Eq is defined in Zhang et al 2023 (but not in this paper) and p(m, dobs) is only defined later, at line 153).

We have added the definition:

"…where logp(m, dobs) is the joint distribution of model m and data dobs. The expectations are calculated with respect to the known pdf q, and we have used Bayes' theorem to expand the posterior pdf p(m|dobs)."

4. Line 229: Matrix → Matrices.

Done.

5. Line 230: if set → if we set.

Done.

6. Line 251: Taking account → Taking into account.

Done.

7. Line 275: complied → compiled.

Done.

8. Lines 300–301: It would be useful to add the total number of data.

We have added this information:

"In this study we use a total number of 401 travel time measurements at 10 s period."

9. Figure 3: The models (except ADVI) all look somehow very noisy/patchy. Why are there non-negligible values in the offshore areas, where there is no data information (i.e., no ray paths)?

This is probably because the number of samples are not sufficient to represent the distribution in the far offshore areas. In fact, the distribution in those areas is equal to the prior distribution, which spans a broad region of the high dimensional parameter space. As a result, the limited number of samples that we can use is not sufficient to explore it and approximate the posterior pdf accurately. We have added discussion on this:

"In far offshore areas…the results obtained using sSVGD and MH-McMC exhibit more heterogeneous structures, which probably indicates that the two methods have not converged sufficiently. These areas are only loosely constrained by the data (or not at all) and hence have large posterior uncertainties requiring many more randomly generated samples in order to explore and represent the posterior distribution accurately compared to areas with tighter constraints from the data."

10. Fig. 3: It could be useful to plot the terrane boundaries (those shown in Fig. 2b) in Fig. 3.

Done.

11. Line 326: tend → trend.

Done.

12. Lines 327–328: Annotation 4 depicts very high uncertainties (~ 0.9 km/s). Why are these uncertainties so high here, while the local data coverage seems pretty good and the model variations seem to be quite smooth?

This is likely because few ray paths go through this area due to its low velocity (it is surrounded by high velocities). As a result, this area has high uncertainties. This is probably also the reason why the high

uncertainty is clearer in the gradient-based methods as the gradient is zero in this area. We have added this in the text:

“In addition, the East Irish Sea (annotation 4) shows high uncertainties. This is probably because few ray paths go through this area due to its lower velocity, and consequently the area is not well constrained by the data.”

13. Lines 333–334: “...other maps provide more detailed information” —yes, but the authors should recognise that these maps also seem to be more contaminated by noise, for example in the offshore areas where there is a poor data coverage.

We agree. See our comments in 9. We also note that for areas that are well constrained by the data, the results should be more stable as in those areas we do not require such a large number of samples to approximate the posterior distribution.

14. Lines 338–341: As most (global) tomographies suffer from strongly uneven data coverage, would the non-convergence issue be a serious problem in practice, for those applications?

We do not think this will be a serious problem as those areas with poor data coverage would have high uncertainties, and therefore the structure would not be reliable. As a result, the structure would probably not be interpreted.

15. Line 353: demonstrates → suggests.

Done.

16. Line 410: resolution → data coverage.

Done.

17. Fig. 6: Why is there so much difference between the marginal distributions at the second well log obtained using SVGD and sSVGD? (See the white circles in Fig. 1 of this review.)

This is because SVGD and sSVGD are different methods. sSVGD is an McMC method which adds a random noise term to the dynamics of SVGD, which makes the method explore the space more broadly. By contrast, SVGD is a deterministic method which requires a large number of particles to approximate the posterior distribution in a high dimensional space. As a result, for a relatively small number of particles (e.g., 500 as we used in this study) the method can underestimate uncertainty as all the particles fall within the high probability area. This is the reason why the marginal distributions obtained using sSVGD has broader distributions. This can also explain the biased results obtained using SVGD in the deeper part (> 1.5 km). However, we note that sSVGD might also produce biased results because of discretisation error.

18. Fig. 6 also shows that SVGD produces better results than sSVGD, since it fits better the true velocity (red) profile. (See Fig. 1 of this review.) So why are the authors claiming (in lines 422–423) that “SVGD can produce biased results for high dimensional problems”?

See the comments above. From a Bayesian perspective, fitting better the true velocity does not mean the result is more accurate because this does not mean the estimate of the posterior distribution is more accurate. Put another way, when we do not know the true solution we require that it lies in a positive probability region of our estimate of the posterior. In the deeper part of the model (> 1.5 km), SVGD clearly produces biased results as the true velocity (red line) lies outside of the range of values with significantly positive probability. By contrast, the results obtained using sSVGD include the true velocity value with nonzero probability almost everywhere.

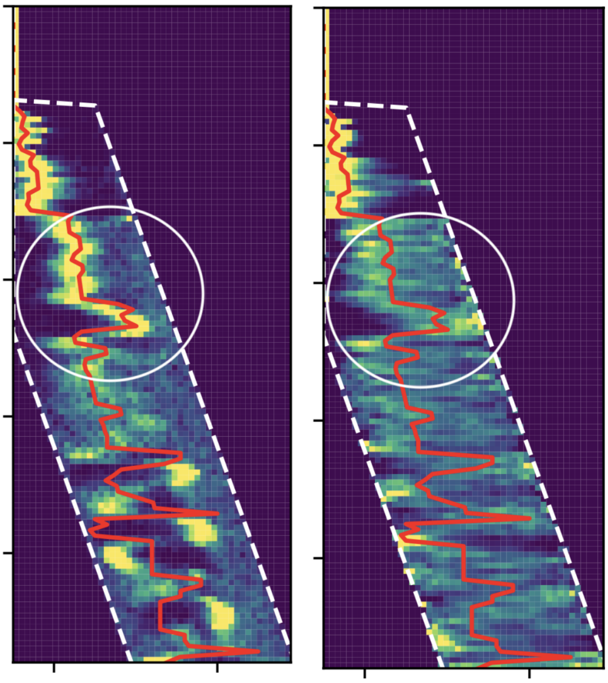

Figure 1: Zoom-in on the the marginal distributions at the second well log obtained using (left) SVGD and (right) sSVGD. The white circles mark significative differences between the two VI methods—what is the reason for this? And why SVGD seems to fit better the true velocity profile (red line), while the authors claim that SVGD can produce biased results?