Response to Reviewer 1

Comments by Reviewer 1 in blue, answers in black.

It is generally a well-written manuscript, however, there are a few issues to be highlighted. The authors state that the focus is on the methodological part which is pretty similar to Fichtner et al. and Krisher et al. except for inverting for density. The method part may be shorten to avoid repetitions. The main difference compared to most of the previous FWI studies is the simultaneous inversion of density with wavespeeds. However, there are also other examples in literature from exploration to lithospheric studies where density is also inverted, but they are not cited. It is worthwhile to have a complete literature review, and it would also be more insightful to compare the results of density inversions to those previously published tomographic studies.

We feel that shortening the methodological part would risk deteriorating the repeatability of the work, and therefore respectfully prefer to keep it in the current format. While we have attempted to invert for density, we did not overly stress this particular aspect, because of the inconclusive results regarding this parameter in our study. Nevertheless, we have expanded the section discussing the density model and its limitations, referencing some relevant literature and discussing possible avenues to move ahead this field. This is now discussed in a separate section (Section 6.4).

On the other hand, the message of the paper needs a bit of clarification. The authors very briefly discuss the potential reasons for the failure of retrieving density in conclusions. However, it is not clear if density can indeed be retrieved or not in FWI. If not, is it better not to invert it, or is there any advantage of inverting for it despite the trade-off between seismic parameters? It would be helpful tp extend these discussions considering the results of other studies.

Especially in the case of exploration seismology, results regarding density seem to be encouraging. This has all to do with the fact that information derived from impedance contrasts can be used much more efficiently thanks to the availability of reflection data. Other factors include the controlled-source aspect of such studies and the source and receiver density. In favourable set-ups, such results may be used on lithospheric scales as well – especially if higher frequency data is used. In our case, however, the data are clearly transmission-dominated, meaning that reflection data are virtually absent. This is very similar in other applications at this spatial scale. Nevertheless, as we have shown in Blom et al. (2017), including density as a free parameter is important to avoid contamination of the reconstructed velocity distributions. A more in-depth discussion of density is included in Section 6.4 of the paper.

Some further detailed comments are as given below:

Page 2, line 40, last sentence: The authors state that not only P and S wavespeeds but also density can be constrained by full-waveform inversion. However, the results of the study seem not promising to support this statement. The authors discuss the potential reasons, but they do not discuss much what should be done to retrieve density or what the strategy in future studies should be. Will taking full anisotropy and attenuation into account in inversions help retrieve the density model correctly?

We agree that this was a deficiency of the original manuscript. This is improved in the revised version. Though we were optimistic at the beginning of this study, we now see our ability to constrain 3-D density much more critically. Taking other parameters, including more complex anisotropy or attenuation, into account is unlikely to help. This is because any increase in the total number of model parameters can only act to decrease resolution. The actual crux of the problem is that sensitivity to density is so small that the recovered density model is strongly affected by unavoidably subjective regularisation. In this sense, a way forward would be to replace the current deterministic inversion by a fully probabilistic inversion that does not require regularisation. Then, at least, one would obtain complete and honest information about our state of knowledge on 3-D density inside the Earth. Research in this direction is underway [e.g., Gebraad et al., 2020. "Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo", JGR, in press]

Page 5, line 115: Parameterization is an important decision in FWI to capture the physics of the medium. It is not clear if the authors are inverting for only SV and SH wavespeeds for radial anisotropy or also the anisotropic parameter \eta.

The anisotropic parameter \eta is not inverted for. The more parameters are included in an inversion, the less well constrained it becomes. \eta is constrained by P-waves with variable incidence angle, for which we have too little data to constrain it properly. More broadly speaking, there is currently no 3-D heterogeneous model of \eta at regional to global scales. (This is different in exploration, where the necessary data may indeed be available.) We have formulated more clearly in the paper which exact parameters are inverted for (Section 4.1).

Why was not the amplitude information used in the misfit, which may be more sensitive to density variations?

Amplitudes are so sensitive to local velocity variations and source magnitude, that the risk of accidentally polluting the inversion would be too strong. Some amplitude information is used in the misfit indirectly: the relative amplitudes within a window. This does give valuable additional constraints on all parameters.

The manuscript nicely summarizes the strategies considered in the inversion algorithm. However, it is not described how the crust was considered in simulations and inversions, which can have a significant effect on the constructed model. Besides, focusing on a subdomain of the model of Fichtner et al. by performing ~95 iterations, one would expect higher resolution in the model, which also deserves investigation of depths shallower than 60 km where some interesting crustal features may be observed.

The initial crust was implemented in the form of a velocity gradient, meaning that the initial model has no sharp crust-mantle discontinuity. This approach is widely used in regional- to global-scale tomography [e.g., Fichtner & Igel, 2008; French et al., 2014], and it essentially served two purposes: (1) The absence of a discontinuity facilitates the construction of a finite-element mesh, and in particular, it avoids the presence of very small elements. (2) An initially smooth transition between crust and mantle allows the data to actually modify its sharpness. Thus, in the final model, the velocity gradient from crust to mantle is as sharp as required by the data. The disadvantage of this approach lies in a limited interpretability of the final model in the sense that no sharp separation between crust and mantle exists. (To illustrate resolution at shallower depth, we added to the Supplementary material a new depth slice at 50 km depth for the spike tests.) The shortest periods in the simulations are 28 seconds, which limits the amount of detail present in the final model – irrespective of the number of iterations.

Section 4.3: It would be good to mention the smoothing and/or pre-conditioning strategies if applied.

In summary, the preconditioning amounts to the 99th percentile clipping, removal of information close to the edges, and smoothing. The latter two are informed by the frequency band (and resulting typical wavelengths) of the wave propagation in each iteration, and become smaller / finer as the inversion goes to shorter periods. The exact parameters used for each iteration are explained in full detail in the Supplementary Material, and we have added a clarification in the manuscript text referring to this.

Section 4.4: To my understanding, the starting model has multi-resolution (smooth global model, inverted European model, and higher-resolution Anatolian model), and a smoothed version of it is used. What is the resolution or the degree of smoothness of the chosen model to start iterations?

Indeed the starting model has several levels of resolution. However, no additional smoothing was applied prior to the inversion presented in this work. We have added information on the resolution lengths of the starting model.

Figure 2: There are quite some small-scale variations (smaller than those in velocity models) in the starting model of density. How was density constructed in the starting model? How good is it to start with?

In the starting model, density was not treated in any specific way, although it was updated along with the other parameters and also smoothed strongly. Because of this, density in the starting model has no particular interpretative value.

Figure 6: Why is the histogram split into two parts? Why is there a gap around the zero phase shift?

This is a result of the way the adjoint source computation is implemented for the time-frequency misfit. Internally, a decision algorithm attempts to filter out the windows in which the traces are too dissimilar to allow for a meaningful comparison. This algorithm happens to include a criterion based on a division by the maximum absolute phase shift within a TF window. This would be a very small number if the traces are very similar, and the resulting criterion big, which results in rejection. In other words, the window has a high risk of being (erroneously) rejected if the traces are very similar. This is of course a very crude algorithm, but the effect of this can be offset by slightly adapting the windows. As a result, however, the windows with a near-zero phase shift disappear from the distribution which is, as a result, very much non-Gaussian.

Figure 10: Looks like there is a strong anti-correlation between P- and S-wavespeed models (i.e., the S-wavespeed model shows all slow wavespeed in continents, and fast wavespeed in oceans at shallower depths whereas the P-wavespeed model is predominantly fast). There is also a sharp boundary in the P-wave model on the Eastern part. How do the authors interpret these anomalies? Are these features also observed similarly in other tomographic models?

We agree with the reviewer that at first sight there does appear to be an anticorrelation, however upon closer inspection this turns out to be not actually the case – the P-wave model is fast in general. The sharp boundary in the P-wave model shows the boundary between the non-updated and updated parts of the model domain, i.e. where there is data coverage and where there is none.

Figure 15: Looks like the Gaussian anomalies used in the spike-tests of the P-wave model are larger than those of S-wavespeeds and density. Is it a plotting issue (or illusion?) or any specific reason for choosing it to be like that? It would be more insightful to show the spike-test results at other depths as well.

Indeed, there seemed to be a plotting issue. The figure has been adapted, and the supplementary material now contains further slices at 50, 300, and 500 km depth.

The authors have performed ~95 iterations, plus five additional iterations for every parameter during the spike tests. It is a large number of iterations. How expensive is each iteration? Or what is the overall computational cost?

In Table 1, the column *nx*ny*nz*nt* in is a measure of computational cost per simulation. As the computational cost of waveform simulations scales to the fourth power of the maximum frequency, the bulk of the computational cost is in the final iterations; the first 50 iterations therefore represent less than ¼ of the total computational cost. Giving an absolute number of wall-clock time or GPU time is less informative, because it is very much dependent on the machine the simulations are run on. The same holds for e.g. (estimates of) energy consumption. The cluster that was used in our case was upgraded halfway through the work, which means that 'before' and 'after' values would not be comparable.