Response to Reviewer 2

Comments by Reviewer 2 in purple, answers in black.

(1) The authors claim that they also include radial anisotropy in their inversion, however, I cannot find any results for radial anisotropy in this manuscript. I strongly recommend the authors to present these solutions in order to make the statement consistent.

Information on the different Vsh and Vsv models is already presented in Figure 8 in the paper.

(2) It seems to me that the ETH group has published a number of papers on the same region (Europe) using the same approach (FWI). Why publishing another one or performing another inversion for the same target? What is the long term goal of this group for imaging the European upper mantle using FWI? Just including more data and publishing another paper once a while is not a good scientific exploration as far as I consider. For instance, I can list some papers which are similar to this manuscript. Can the authors provide some statements about significant improvements of this work in comparison to their previous work? Such as (1) Fichtner and Villasenor, 2015, EPSL; (2) Rickers et al, 2013, EPSL; (3) Fichtner et al, 2013, EPSL.

First, we note that science generally has two components: the development of new methods, and their application to new data. Both modes of operation are legitimate and needed to ensure progress, and one is not *per se* better than the other.

Second, we would like to point out that none of the publications listed above covers the region studied here. Fichtner & Villasenor (EPSL, 2015) considered the Western Mediterranean, with a focus more than 2000 km away from the current one. Rickers at al. (EPSL, 2013) studied the North Atlantic, with a focus on Iceland and Jan Mayen, more than 3000 km to the north of this study. Finally, Fichtner et al. (EPSL, 2013) focus on the Anatolian micro-plate and the North Anatolian Fault zone, to the east of the current study region. The work of Fichtner et al. (GJI, 2013), not listed above, covers all of Europe and Western Asia. Though this study includes the Eastern Mediterranean, it uses only around half of the amount of data for a model volume that is more than 100 times larger. Thus, Fichtner et al. (GJI, 2013) operated at a completely different scale in all respects.

This said, it is clear that the present study is entirely independent from the ones we did before, and it focuses on a region that had not been studied with the help of full-waveform inversion. Given the geologic relevance of the Eastern Mediterranean, we believe that our work is therefore more than justified.

(3) In fact, From Figures 7 and 14, I am surprised to see that predicted seismograms from their starting model (Fichtner et al, 2013) have such large mismatches in comparison to observed waveforms, even for this long-period frequency band. If their previous model is already good enough, we should not see such significant differences.

As explained in the response to the previous question, Fichtner et al. (GJI 2013) used a significanly smaller amount of data for a model volume that was orders of magnitude larger. Furthermore, the shortest period in Fichtner et al. (GJI 2013) was around 50 s, i.e., around twice as large as the shortest period used here. In summary, it is not at all surprising that the Eastern

Mediterranean part of the 2013 model does not explain the shorter-period data at much smaller scale considered here. This high misfit immediately serves as another motivation to carry out this work – clearly the existing models are insufficient.

(4) The authors just present their own tomographic model. In fact, the Mediterranean region is well studied by using either travel time tomography or even full waveform inversion by other groups. The authors ignored the same endeavor towards FWI imaging for the European crust and upper mantle by other groups. I wonder how their current model compares with other previous tomographic or FWI models? As far as I can tell from their Figure 12, the resolution for subducting slabs underneath the Mediterranean is quite low in comparison to previous travel time tomography (Wortel and Spakman, 2000, Piromallo and Morelli, 2003) or FWI models (Zhu et al, a series of studies). I strongly suggest the authors to make some comparisons with these previous studies, in order to convince the readers about the advantages or disadvantages of their current inversion.

We have added additional figures to the Supplementary Material, where we compare our final model to the Utrecht model UU-P07 (Amaru, 2007; this model is the successor of the model used in Wortel & Spakman, 2000). There is a clear correspondence between the models, which is encouraging given the fact that we compare a travel-time P-velocity model and a waveform S-velocity model. As is visible from these figures, the models have similar levels of detail as our final model – something similar holds for the models of Piromallo & Morelli (2003) and Zhu et al (2015). It may have appeared that there is more detail in the models mentioned by the reviewer, because those are usually viewed on a larger scale – in particular UU-P07, which is a global model. The advantage of zooming in, as we do, is that there is a lot more information, and as a result, e.g. the amplitudes of the anomalies are stronger.

(5) The authors claim that their model has good resolution down to the mantle transition zone. However, in their horizontal cross sections (Figure 10), I only see results down to 300 km. I suggest they to put more slices, say from 400 to 700 km.

We do not claim they have good resolution down to the mantle transition zone – only in some parts is the model updated down to the mantle transition zone. This is already summarised in Figure 8 of the paper, and we have now added further depth slices through both the final model and the resolution tests in the Supplementary Material.

(6) For vertical cross sections (Figures 12 to 13). I don't see any reasons to put slices with absolute velocity here. Since they will be dominated by 1D structures, in particular for the mantle transition zone, which involves velocity jumps across 410 and 660. I suggest only show relative velocity perturbations in their right panel.

One of the largest added benefits of using waveform tomography is the fact that we get access to absolute velocities. Any thermodynamical interpretation relies heavily on absolute velocity – see for example Deschamps & Trampert, 2003; Cobden et al, 2008, 2009, 2018. For this reason we prefer to show both types of cross-section.

(7) In fact, for the Hellenic subduction (Figure 13), numerous previous studies have demonstrated that the Hellenic slab can penetrate the 660 km discontinuity down to the lower mantle (from either travel time tomography or FWI), however, the authors' result seems different from previous conclusions. Do we need to re-evaluate this consensus, or the authors' model does

not have enough resolution to resolve it?

We agree with the reviewer that the initial formulation here was unfortunate; by no means did we intend to imply that the Hellenic slab must terminate near the MTZ! We have rephrased this part of the text to clarify this.

(8) For waveform comparisons, such as Figures 7 and 14, the authors only show comparisons between observed and predicted seismograms for a couple of station-source pairs. However, as pointed out by the authors, there are numerous seismograms. I am not surprised to see some improvements for some selected pairs of earthquakes and stations. How about other pairs? Since there are many stations in the authors' dataset, why not plot all three components (vertical, radial and transverse) for most data and predictions, just like common-shot gathers as did by exploration seismologists. I think that will be a better way to convince the readers that their model can really explain data, instead of just intensionally or randomly selecting some paris with good match. In addition, I suggest the authors to change color for predicted seismograms from the starting model, such as blue or other colors different from the current waveforms.

We have experimented with this. In the end we find that a comparison of single traces give the most insightful results, as is commonly done in waveform tomography. A record section, for our setup, results in the interference of too many different phases that are a result of significant 3-D structure. A record section may be more suitable in a situation where all receivers are located roughly in a line, as is the case for e.g. Beller et al (GJI, 2017). As a compromise we have included figure 6 which shows a very detailed misfit development.

(9) Figure 3, why changing iteration numbers for different frequency bands? In order to demonstrate that their model can fit data for different frequency ranges, I suggest just compare the starting one with their current model (100th iteration).

Figure 3 is meant as a demonstration of the multiscale method, and shows that the chosen points to include higher frequency content are suitable. This is why we prefer to show the iterations as given in the figure.

(10) Figure 4, it seems to me that the color of panel d is highly saturated, then they can see a strong banana kernel, how about using some reasonable colors. In addition, do they need to add additional weighting coefficients in order to use both body and surface waves for this inversion?

The colour scale in panel d is intentionally made the same as the colour scale in panels e and i. This oversaturation is the only way to make the kernels for all windows visible. This is intended as an illustration of the power of the windowing strategy: because of the significant phase shift visible in the body waves, the gradient in vp is much stronger for this window than for the surface wave window. This information would be lost almost entirely if a single window were taken for both arrivals. No additional weighting is included – this happens automatically through windowing.

(11) Figure 11, I suggest to plot iso-surface for relative velocity perturbation instead of absolute values, which is a typical way to examine subducting slabs.

With waveform tomography, we have a method that comes with the luxury of giving us absolute velocities. We are taking advantage of that because it directly feeds into thermochemical interpretation – this is the reason that we show these. Moreover, compared to the anomalies, the background variations are relatively minimal, so showing relative instead of absolute velocities will not make a very large difference.

(12) In section 6.2, in order to evaluate their model, the authors chose additional 6 earthquakes as an independent dataset. As far as I can tell, this is a very small dataset to convince the readers that their model can explain waveforms that are not used in the inversion. I strongly suggest them to increase the size of this independent dataset.

We used events that cover the study area well. We would understand the argument if the events would cluster, but they don’t. Doing more would just inflate the manuscript and result in unnecessary additional computational cost.

Some minor modifications: [...]

We have implemented the suggestions given by the reviewer.