1. An advantage of this study in comparison with many other previous works is the application of an adaptive mesh that allows 400m spatial resolution along the subduction interface. It would be useful to include a bit more information about how the resolution changes over time and space. Does the domain of the finest resolution follow the changing location of the interface, for instance? The grid adapts throughout the simulation, keeping the finest resolution where the spatial gradients of fields are highest. The manuscript has been updated accordingly. So yes, the finest resolution follows the changing location of the interface.

2. Can you specify how the "thin weak layer" is built up to decouple the plates? Which rheology is used? The rheology of the subduction interface has a large impact on the stress transfer between the plates, also controlling back-arc deformation. Therefore, please discuss your subduction interface rheology, viscosity and compare it with previous studies. We decreased the maximum composite viscosity and the friction coefficient of yielding strength of the weak layer, and kept the diffusion, dislocation and Peierls creeps of the weak layer the same as that of the mantle material. We have edited the description of 'weak layer' in Section 2.1. (Lines 60-61)

3. The main method needs more justification: the location of the volcanic arc would be connected to the thermal regime of the overriding plate and the underlying mantle, as well as driven by the slab dip angle. How realistic is it to assume a fixed hot region for representing the volcanic arc? The hot-region is not fixed, but is just an initial thermal condition. A hot region is clearly an approximation to an arc, hence why we have investigated a range of initial conditions. Furthermore, why the 1300 K isotherm is not elevated in the "arc" region? The 1300 K isotherm is elevated in the "arc" region, but may not be that obvious. We have added zoom-in figures of the hot region in figure 4 (a) (c) (e) (g) (i). Does the chosen arc location fit the relationship of slab dip and arc location discussed by Ha et al. 2023 G3? Thank you for suggesting this paper. Yes, we have many cases which fit the geometry shown in Figure 9 of Ha et al. paper. We have run a series of models varying the arc-trench distance from 100 to 1000 km, which covers the range shown in this paper. Our slab dips vary from around 40 to 70 degrees, and are also similar to those in Ha et al. 2023. The authors should also reflect on previous works, where the volcanic arc formed self-consistently driven by the gradual hydration of the mantle wedge and modelling melt extraction. There are particular works, that addressed their role on upper plate rifting: e.g. Corradino et al. 2022 Sci. Rep.; Baitsch-Ghirardello et al. 2014 Gondwana Res. Thank you for letting us know about these works, we have added them to the text (lines 27-29 in the revised manuscript).

4. The authors use an ocean-ocean setup. Do the findings of the study applicable to a continental overriding plate? Thank you for this comment, this is a limitation of our research. As we mentioned in Limitation (Line 341-346 in original manuscript), we point out that a continental crust in the OP differs from our oceanic setup. Since crust is weaker than mantle and continental crust is thicker than oceanic crust, then we might expect continental lithosphere from a compositional perspective to be less viscous than oceanic lithosphere but thermally continental lithosphere can be older and hence colder and therefore could be expected to be more viscous than oceanic lithosphere. We would need to model this case to make definitive statements as to which is most important (composition or temperature), but it is reasonable to expect that the trends will be similar. Also since the OP weakens from its base, the nature of the near surface crust might be less critical. Therefore, while our findings might, to some extent, also be applicable to a continental overriding plate, we prefer to highlight this as a limitation.

5. I think the authors should better reflect on the limitation of using a 2D setup. The poloidal mantle flow is overpredicted in such 2D models, and there seems to be a broad consensus that in nature, back-arc extension is connected to the toroidal component of the mantle flow: e.g. McCabe 1984 Tectonics, followed by many modelling papers. This is connected to the 4th conclusions points, that back-arc extension is caused by the poloidal mantle flow. This is right, but rather a model limitation, thus I suggest moving it out from the conclusions. Thank you for this comment. Yes, toroidal flow plays an important role. The reason why we wrote about poloidal flow in the Conclusion is to note the importance of the size of the (poloidal) wedge flow cell and how this flow acts on the OP. It highlights the effect of poloidal flow. We think that the 2D setup is a significant limitation of this work, which we highlight in the Limitations sections.

6. As for the asthenosphere-lithosphere coupling: how would different mantle thermal gradients affect back-arc extension? Can you show a viscosity profile? Would a different profile, for instance, by assuming a different mantle thermal gradient or grain size evolution affect the coupling between the plate and underlying mantle? This should be mentioned at least in the discussion. Yes - it is possible that a different viscosity profile would change the coupling slightly, but you can see that there is a dramatic change in viscosity from the lithosphere to the asthenosphere, and therefore we do not expect changes that would change our conclusions. We present not just a profile but the whole viscosity field in figure 2b. The thermal structure though evolves according to the equations of physics and is not a free variable. We have added a brief comment in the discussion 4.1 – 'The detail of the effect of basal drag will depend upon the asthenosphere-lithosphere coupling, which is self-consistently solved for in our work which involves a thermal lithosphere but the basal drag could differ slightly in reality with more complex lithospheres.'

7. ln. 125: this means a subduction velocity larger than 11 cm/yr. Is it in agreement with observations and reconstructions? This is important, because the velocity of the induced poloidal flow would have similar values (in a 2D model) and this is linked to the potential coupling with the overriding plate. I suggest showing a plot on the relation between the modelled subduction velocity and upper plate lithosphere thinning over time. Yes, the peak subduction velocity in our models (it is worth noting that this lasts for only a short period of time - hence this period would be hard to identify on the seafloor of the related plate) is high compared to frequently quoted subduction velocities. The thinning occurs over a very short period of time, frequently close to the time of peak subduction velocity. For most of the early part of the simulations the plate thickness is constant or thickening very slightly by cooling. From animations (not presented, but happy to include in supplementary information if advised) of our simulations, it is clear that the OP thickness stays virtually constant and the subduction velocity increases as subduction progresses. Then in cases with extension the OP plate thins very quickly. Rather than present a plot of such simple behaviour we have described it in the text – 'We note that the SP sinking velocity in this short time period increases to a very high value (with local peak of > 10 cm/yr, and a lower peak when averaged peak over 1 Myr [geologicall observable timeframe], as expected, as the length of subducting slab increases, before decreasing to nearly steady values of < 2 cm/yr once the slab approaches and enters the more viscous lower mantle.' (Lines 133-136 in the revised manuscript.)

8. "Horizontal extensional force can be ignored as a cause of Extension in our models": this statement needs more attention, I suggest. Kinematically the retreating slab drives the divergence of the overriding plate. Extensional deformation will be localized along

the rheological weakest location. It is either along the imposed thermal weakness or the location overlying the mantle upwelling, connected to the return flow. Edited in Section 4.1 (Revised manuscript).

9. Instead of providing a list of a selected previous modelling papers (Line no. 15-17), it would be better and more useful to group them, which previous models contributed to which aspect of back-arc extension: e.g. analogue vs numerical, 2D vs 3D, assumed hydration and melting or not, used Newtonian or more complex rheologies, used spontaneous or forced subduction initiation, etc. Thank you for this suggestion. We have edited and grouped these papers by the method of simulation: analogue vs numerical models (Line 15-17 in the revised manuscript).

10. Subduction initiation would have an impact on the formation of an arc and also on the style of upper plate deformation (cf. Stern 2004, EPSL). In understand that this is not the primary topic of this manuscript. However, if one assumes spontaneous SI, by the time the leading edge of the slab reaches the prescribed 200 km, a back-arc spreading center could have been already formed. Thank you for this comment. Yes, this is a good point, but as you say it is not the primary topic of this manuscript. The work in this manuscript clearly would not be relevant to cases where a back arc spreading centre has already formed before the leading edge of the slab has reached 200 km. It is unclear whether this would be the case in spontaneous SI. Since our work does not address this case, we do not comment extensively, but accept it would be an excellent avenue of future research. We have added this text 'including for example missing out the initiation stage of subduction.' in the manuscript at lines 386-387 in the revised version, to remind the reader of this limitation.

11. The authors write that the overriding plate region in the close vicinity of the trench record compression before rifting. I don't think this is the artifact or due to the mentioned coarse time stepping. In our previous models (Balazs et al. 2022; Corradino et al. 2022) we visualized the stress field and the orientation of the principle stress axis and also found this compressional stress accumulation on the forearc region driven by vertical suction (resulting horizontal compression) of the slab. But, when the slab starts rolling back, of course, extensional deformation will be localized along the rheologically weakest part of the overriding plate, in your case, that is this region, where the "arc" was defined. Yes, we agree that the compression we have found near the trench before the extension is not due to the mentioned coarse time stepping and it is real and not an artifact. We note though that in Fig.10 in the original manuscript, the location where the EH happens recorded compression, but in fact we had incorrectly marked the extension location. It is marked correctly in Fig.11 (in revised manuscript) and is in a region under extension.

12. The statement in the introduction, that the nature of the overriding plate has not been extensively studied or the majority of the models listed above include a homogeneous OP is not the case. Just a few recent example: Wolf et al. 2019 JGR Solid Earth, Yang et al. 2021 G3. In our two papers on this topic: Balazs et al. 2022 Tectonics and Corradino et al. 2022 Sci. Rep., we particularly addressed the role of inherited structures, the formation of a volcanic arc and the possible locations of back-arc rifting. We have edited the text in Introduction (Lines 20-30 in the revised manuscript).

13. fig. 1: The location of the "arc" region is drawn above the slab in the zoomed image, while it is drawn as laterally shifted in the larger image. Thank you for pointing out this error. It has now been corrected.

14. no. 297-299: “The high negative buoyancy and strength of an older SP encourage a higher trench retreat rate and a stronger mantle flow (Garel et al., 2014), so that the flow is strong enough to break at the far-field location before the weak zone is broken. Under such circumstances, the models show EF mode.” In fact, when the plate is too old and strong it resists to bend, therefore there is an optimum age, cf. Di Giuseppe et al. 2009 Lithosphere. Thank you for pointing it out. Yes, a plate will be hard to bend when its age is very old, although our tested ages of the SP are not old enough to get this phenomenon. Since the SP age on Earth is younger than our maximum tested SP age, we didn’t mention this point. We have added to the text in lines 328-329 (revised version) mentioning this possibility - ‘It is possible that at very old age the slab might resist bending leading to another mode.’

15. The text can be improved, for instance, this is not an optimal way of references: “and other references”. The convention, the authors use to explain Complete Thinning and Spreading as “Extension” is misleading and not necessary. In the discussion, it is particularly challenging to follow the reasoning. I suggest simply using well established terms: when talking about strain: divergence or rifting, for processes spreading, post-rift relaxation, etc. Some sentences might be also simplified, like this: “The model goes to rift when the basal drag wins out, but thermal healing is always efficient because all models showing Extension show it healing after a few Myrs of Extension as well.” We have decided to keep with our simple classification system of 2 states, Extension or No Extension. We note that Complete Thinning is Incipient Extension, and to make this clearer we have used the term Complete Thermal Thinning / Incipient Extension in Figure 3 and the text where we define these two states to make things clearer and hopefully less misleading.

16. To limitations: eclogitization? Partial melting: significantly drops viscosity and increases roll-back. Thank you, we have added these limitations in the ‘Limitations’ section. We have edited the text in Limitations as follows: “Similarly, we did not consider eclogitization, which would affect SP.” “The process of forming an arc involves partial melting which can be expected to lower the viscosity and density in local regions. There is also a possibility for important feedback between the arc and subduction that we cannot capture in these models.”

17. fig. 2: it is hardly possible to see the stress values in the overriding plate. This part should be enlarged and zoomed. Please indicate the horizontal and vertical scale in the figures. Thank you for the suggestion. We have added zoomed-in figures.

18. fig. 10: this figure should be placed in a supplementary material, and here you might rather show the models stress field and velocity field just before and after rifting. Yes, with the old figure 10, which was in error, we agree that a figure showing the model stress field and velocity field just before and after rifting would have helped show how this could be the case (such a result could have been correct - due to the possibility of non-uniform stress state with depth through the lithosphere) and was an excellent suggestion. Now that we have discovered the error in our original figure 10, (now figure 11) the updated correct figure provides a simple explanation of the stress state, and we prefer to keep it and not introduce more complex figures and text.