Reply to Referees'comment on the manuscript “Can subduction initiation at a transform fault be spontaneous?” by Arcay et al., submitted to Solid Earth Discussion.

Comments from Referees are in italic and underlined. Our response is given in normal characters, while modifications in the revised manuscript are indicated using bold characters.

Comments from Reviewer # 2

General comments:

The study presented in this manuscript addresses the issue of spontaneous subduction initiation via a parametric numerical study. The question of how subduction initiates is of great importance in geodynamics, as it touches on the core of plate tectonics. To date, this study is the most comprehensive parameter study I have seen. By varying a large number of material and model parameters that may have a potential impact on the occurrence of subduction initiation, the authors delineate the physical parameters that result in old plate sinking (OPS). Results show that spontaneous subduction initiation OPSe at a transform fault is very unlikely at present Earth conditions. This result is not entirely new, as the difficulty of initiating subduction has already been pointed out by other authors (e.g. [McKenzie, 1977; Cloetingh et al., 1989; Mueller and Phillips, 1991]). With the exception of [Mueller and Phillips, 1991], these previous studies did not specifically address subduction initiation in a transform fault setting. The amount of numerical models that have been conducted for this study and the wealth of information about the influence of different parameters that have been investigated add an important new perspective on the issue of spontaneous subduction initiation and make this manuscript suited for publication in Solid Earth.

We greatly appreciate the careful and constructive review made by the Reviewer and warmly thank her/him for the work done to comment our manuscript.

The introduction is structured in a clear manner. In the model setup section, I would suggest some rearrangements to make it more concise (see comments below). Most importantly, a large fraction of the results is described in the model setup section. I strongly suggest moving this description to a separate results section.

We agree with Reviewer 2’s comment. As answered to Reviewer 1’s request regarding the same issue, we awkwardly removed the Latex command \section{Results} during the writing process. It has been re-inserted:

p. 12 l. 17: “**3. Results**”

Note that the numbering of the following subsections is thus completely modified.

The results are presented in a two-fold manner: first, all simulations that do not exhibit OPS are described in detail. Several different regimes are identified. After that, all simulations that exhibit OPS are described and two different modes are identified. After that, model limitations are explored and results are compared to natural examples. The structure of sections 2-4 where model setup, descriptions of the results and discussion are mixed makes it at times hard to follow the paper and should therefore be improved.

Please see our response to the previous comment. The Results section is now clearly separated from the model section.

Additionally, the language needs improvement, as sentences are often phrased in a confusing manner.

We have sent the manuscript during the revision process to a professional website of scientific English editing (www.aje.com). We enclose the Editing Certificate provided by AJE (certificateAJE_Arcay_et_al.pdf). Please consult the file that compares the previous version and the revised manuscript (maintext_diff.pdf) to evaluate the corrections, since we cannot reproduce here all the corrections that have been made regarding the language.

I think that the manuscript would benefit greatly from an additional section (to be included after the Introduction) that explains the basic physics/mechanics of OPS, similar to what is done in the study by

[Mueller and Phillips, 1991]. In my opinion, this would make it much easier for the reader to understand the influence of the different physical parameters that have been varied in this study.

We thank the Reviewer and agree with him. We add a new section, starting p. 7, l. 27 "**2.4 Parametric study derived from force balance**", located after the main subsections describing our numerical setup, and before the section detailing the ranges of investigated parameters. The goal of this new subsection is two-fold. It first states the first-order force balance that may drive the evolution of our simulations. It then explain how the different forces may vary as a function of the different tested parameters. In addition, in this section we introduce and explain the two kinds of parameter investigation that we have done: either a rather systematic exploration of the 6 physical properties depicted in Fig. 3, or a more limited set of experiments for some additional experiments. We think that this new section significantly helps the understanding of our modeling strategy and the results we obtained. The content of this new subsection is:

p. 8, l. 11-p. 9, l. 7: "**The first order forces driving and resisting subduction initiation at a transform fault indicate which mechanical parameters would be worth testing to study OPS triggering. Without any external forcing, the unique driving force to consider is (1) the plate weight excess relative to the underlying mantle. Subduction is hampered by (2) plate resistance to deformation and 15 bending; (3) the TF resistance to shearing; and (4) the asthenosphere strength, resisting plate sinking (e.g., McKenzie, 1977; Cloetingh et al., 1989; Mueller and Phillips, 1991; Gurnis et al., 2004). To unravel the conditions of spontaneous subduction, we vary the mechanical properties of the different lithologies forming the TF area to alter the incipient subduction force balance. The negative plate buoyancy (1) is related to the plate density, here dependent only on the thermal structure and plate age A (Sect. 2.2) since we do not explicitly model density increase of metamorphised (eclogitized) oceanic crust. Nonetheless, we 20 vary the crust density, ρc, imposed at the start of simulation along the plate surface to test the potential effect on plate sinking. We also investigate how the density of the weak layer forming the interplate contact, ρT F , which is not well known, may either resist plate sinking (if buoyant) or promote it (if dense). The plate strength and flexural rigidity (2) are varied in our model by playing on different parameters. First, we test the rheological properties of the crustal layer both in the brittle and ductile realms, by varying γc and Eac (Eqs. 2 and 4). Second, the lithospheric mantle strength is varied through the mantle 25 brittle parameter, γm, that controls the maximum lithospheric stress in our model. Third, we vary the lateral extent (Lw) of the shallow lithosphere weakened domain, related to the crust alteration likely to occur in the vicinity of the TF.We study separately the influence of these 6 mechanical parameters (ρc , ρT F , γc , Eac , γm , Lw) for most plate age pairs. The TF strength (3) is often assumed to be quite low at the interplate contact (Gurnis et al., 2004; Gerya et al., 2008). We thus fill the TF "gouge" with the weak material (labeled 1 in Fig. 2) and, in most experiments, set it as γTF =5×10−4. In some experiments, we replace the weak material filling the TF gouge by the more classical oceanic crust (labeled 3 in Fig. 2) to test the effect of a stiffer fault. In that case, γT F = γc = 0.05 and Lw = 0 km: the TF and both plate surfaces are made of gabbroic oceanic crust (Table 3). Note that when γc = γT F = 5 × 10−4 , the weak layer and the oceanic crust are mechanically identical, and the weak layer then entirely covers the whole plate surface (Lw =1100 km). Similarly, as the activation energy Eac is the same for the oceanic crust and the weak material, assuming a low ductile strength for the TF is equivalent to covering the whole plate surface by the weak layer (setting Lw =1100 km). Apart from the 6 main physical properties that are repeatedly tested (Sect. 2.5), we perform additional experiments for a limited number of plate age combinations to investigate a few supplementary parameters. In this set of simulations, we vary the asthenosphere resistance competing against plate sinking (4), either by changing the asthenospheric reference viscosity at the lithosphere base or by inserting a warm thermal anomaly simulating an ascending plume head (Fig. 2). We also test the influence of the lithosphere ductile strength that should modulate plate resistance to bending (2) by varying the mantle activation energy, Eam. At last, we further explore the TF mechanical structure (3) by imposing an increased width of the TF weak gouge, and different thermal structures of the plate boundary forming the TF.** "

The authors do a good job deciphering the impact of each investigated parameter on OPS, but I think the combination of different parameters is most likely as important. A section which explains the potential interaction between forces resisting and promoting OPS at the outset of the paper could be used to discuss the interplay between the different parameters.

We think that the interplay between the investigated parameters is addressed in the different regime diagrams

depicted in Fig. 6, which display the modeled tectonics as a function of the parameter combination. Note that we have revised the regime diagram 6e, in which OPS is actually modelled, but in very narrow plate age intervals. To take into account the Reviewer's comment, we have modified the text at the end of Section 4.3:
p. 21, l. 26-31: "**To achieve OPS, the cursors controlling the plate mechanical structures have been tuned beyond the most realistic ranges ("yellow" domain, Fig. 3) for 2 parameters at least, and beyond reasonable values for at least one parameter ("red" domain, Fig. 6e to h). Nevertheless, combining different unlikely ("yellow") parameter values (for ρ_TF and L_w) does help to achieve OPS for slightly less extreme mechanical conditions, as one parameter only has to be pushed up to the unrealistic ("red") range (ρ_c, Fig. 6e). Note however that the plate age intervals showing OPS are then extremely narrow (A_y <3 Myr, A_o <25 Myr) and are not consistent with the 3 potential candidates of natural OPS.**"

Specific commments:

Abstract
In my opinion, the study does not really represent a completely "new" exploration of the spontaneous subduction initiation concept, as there have been quite a few numerical studies looking at subduction initiation at transform fault. What distinguishes this study from other studies is the extent of investigated parameters.
We have modified the corresponding sentence at the beginning of the abstract:
p. 1, l. 1-2: "**We present an extensive parametric exploration of the feasibility of "spontaneous" subduction initiation, i.e, lithospheric gravitational collapse without any external forcing, at a transform fault (TF).**"

Introduction:
The introduction is well written and concise. It contains both information on natural candidates for spontaneous subduction initiation as well as an overview of existing numerical studies. In section 1.2 I am missing references to [McKenzie, 1977; Cloetingh et al., 1989; Mueller and Phillips, 1991]. In particular, [Mueller and Phillips, 1991] should be referenced.
These references have been added in the new subsection presenting the subduction force balance 2.4, rather than in the section 1.2 of the introduction that focuses on the modeling of spontaneous subduction because some of the references quoted by the Reviewer dealt with subduction initiation under compression (which we exclude from our study):
p. 8, l. 13-15: "S**ubduction is hampered by (2) plate resistance to deformation and bending; (3) the TF resistance to shearing; and (4) the asthenosphere strength, resisting plate sinking (e.g., McKenzie, 1977; Cloetingh et al., 1989; Mueller and Phillips, 1991; Gurnis et al., 2004).**"

Model Setup:
2.1: As this is a numerical paper, I would suggest stating the governing equations in the beginning for completeness. Personally, I also prefer the numerical description not to be the first part of the model setup, as the numerical code is simply a tool to solve the governing equations for a given model. For this reason, I would suggest to move the description of the numerical solution (method, number of tracers, resolution) to the end of the Model setup section (maybe after section 2.4.) and focus on the governing equations including the rheology. In my opinion, it would also be good to include a description of the boundary conditions (they are only depicted in fig.2).
The numerical code used in this study has been used in Arcay et al., 2005; 2006; 2007a,b; ...2017; so we do not think that it is necessary to give the details of every equation, that are very common in mantle convection modeling and were already presented. We had specified that we used the extended Boussinesq approximation. To follow the Reviewer's piece of advice, we have moved the description of numerical aspects at the end of the Model setup section, in a new subsection: "**2.6 Numerical code and resolution**" (p. 12, l. 3-16).
Moreover, we have moved the description of the mechanical boundary condition along the box bottom from Fig. 2 caption to the main text:
p. 7, l. 28-30: "**When the box bottom is open, a vertical resistance against flow is imposed along the box base, mimicking a viscosity jump 10 times higher than above (Ribe and Christensen, 1994; Arcay,**

2017).”

As the thermal boundary conditions imposed in this study are very classical and have been previously described several times, we think that Fig. 2 is sufficient to present other mechanical and thermal boundary conditions.

I was also missing a description of how density is computed in the model, which should be added in the model setup section (potentially together with the governing equations).

We have added the equation of state giving density, at the beginning of the Model setup section:

p. 5, l. 5-9 : “**Density (rho) is assumed to be temperature- and composition-dependent:**

rho(C, T) = rho^ref (C)(1 − alpha *(T − T_s))   (1)

where rho^ref is the reference density at the surface, C is composition (mantle, oceanic crust or weak material; Sect. 2.3), alpha is the thermal expansion coefficient, T is temperature, and T_s is the surface temperature (Table 2). For the mantle, rho^ref_m is fixed to 3300 kg.m −3 , while ρ ref for the oceanic crust and the weak material is varied from one experiment to another (Sect. 2.4).”

In geodynamical models, it is also common to introduce viscosity cutoffs to avoid numerical problems. Were any cutoffs used here? If yes, this information should also be included.

There is no minimum cutoff in viscosity. We have specified the use, or not, of cutoffs at the end of Section 2.1:

p. 6, l. 14-15 : “**Note that the brittle behavior acts as a maximum viscosity cutoff. Regarding strain rate, a minimum cutoff is set to 2.6×10^−21 s^−1 , but no maximum cutoff is imposed.**”

2.2 :

p.6, l.6: ... overestimates a bit ... What is "a bit"? This seems to be a vague statement. Could you provide numbers?

We have removed this expression and detailed a bit the discrepancy between the half-plate cooling model and surface observations:

p. 6, l. 22-28**: “However, the HSC model, as well as some variations of it, such as the global median heat flow model (GDH1, Stein and Stein, 1992), have been questioned (e.g., Doin et al., 1996; Dumoulin et al., 2001; Hasterok, 2013; Qiuming, 2016). Indeed, such conductive cooling models predict too cold young oceanic plates (by ∼100 to 200oC) compared to the thermal structure inferred from high resolution shear wave velocities, such as in the vicinity of the East Pacific Rise (Harmon et al., 2009). Similarly, worldwide subsidence of young seafloors is best modeled by taking into account, in addition to a purely lithosphere conductive cooling model, a dynamic component, likely related to the underlying mantle dynamics (Adam et al., 2015).**”

p.6 ,l.8: Where does the factor 0.75 come from? Is there a reference that compares the heat flow from such models to observations?

We have more justified the use of a corrective factor equal to 0.75. It is based on two independent studies of plate cooling. The first one is the new model of plate cooling proposed by Grose & Afonso (2013), showing that when the hydrothermal circulation close to the mid-ocean ridge (MOR) and the insulating effect of the oceanic crust are included in the thermal model, predicted heat flows are reduced by 75% with respect to the GDH1 model by Stein and Stein (1992). The second study is a numerical parametric study of early small-scale convection (SSC), triggered as soon as the plate is older than 5 Myr, by Buck & Parmentier (1986), which shows that to account for the thermal effect of SSC partly balancing the conductive cooling from above, the plate thicknesses predicted by the half-space cooling model must be corrected by a factor close to 0.75 (between 0.64 and 0.80) to obtain the simulated lithospheric thicknesses. We have detailed these observations in the text:

p. 6, l. 28-p. 7, l.17 : “**Recently, Grose and Afonso (2013) have proposed an original and comprehensive model for oceanic plate cooling, which accurately reproduces the distribution of heat flow and topography as a function of seafloor age. This approach leads to young plates (<50 Myr) 100 to 200°C hotter than predicted using the HSC 6and Parsons and Sclater (1977) models, especially in the shallowest part of the lithosphere. This discrepancy notably comes from, first, heat removal in the vicinity of the ridge by hydrothermal circulation, and, second, the presence of an oceanic crust on top of the lithospheric mantle that insulates it from the cold (0°C) surface and slows down its cooling and**

thickening. Taking into account these two processes reduce the surface heat flows predicted by the GDH1 model by 75 % (Grose and Afonso, 2013). Our study focus on young oceanic plates that are the most frequent at TFs (Ay <60 Myr, Table 1), but we cannot simply reproduce the complex cooling model proposed by Grose and Afonso (2013). Therefore, we calculate lithospheric thicknesses zLB (A) as 0.75 of the ones predicted by HSC.Plates warmer than predicted by the HSC model are consistent with the hypothesis of small-scale convection (SSC) occuring at the base of very young oceanic lithospheres, i.e., younger than a threshold encompassed between 5 and 35 Myr (Buck and Parmentier, 1986; Morency et al., 2005; Afonso et al., 2008). An early SSC process has been suggested to explain short- wavelength gravimetric undulations in the plate motion direction in the central Pacific and east-central Indian oceans detected at plate ages older than 10 Myr (e.g., Haxby and Weissel, 1986; Buck and Parmentier, 1986; Cazenave et al., 1987). Buck and Parmentier (1986) have shown that the factor erf^−1(0.9) ∼ 1.16 in Eq. 5 must be replaced by a value encompassed between 0.74 and 0.93 to fit the plate thicknesses simulated when early SSC is modeled, depending on the assumed asthenospheric viscosity. This is equivalent to applying a corrective factor between 0.74/1.16 ∼0.64 and 0.93/1.16 ∼0.80, and we set here the lithospheric thickness z_LB as 75% of the ones predicted by HSC. Between the surface and z_LB (A), the thermal gradient is constant. ”

p.6,l.9: Assuming a constant temperature gradient between the surface and z_lb seems to be at odds with the assumption of half space cooling, which was used to determine the lithospheric thickness. How do you justify the use of such a thermal gradient? As the temperature field will have a significant impact on the viscosity structure of the lithosphere, assuming such a thermal gradient will result in an overall stiffer lithosphere, which could potentially have a large impact on OPS.

The model proposed by Grose & Afonso (2013) is not purely based on the half-space cooling model, as aforementioned, and produces lithospheric thermal structures that are significantly hotter than predicted by the models of Parsons & Sclater (1977) and of Stein & Stein (1992), by 100 to 200°C. A thermal state hotter than predicted by the half-space cooling (HSC) model has been also suggested by the analysis of shear wave velocity structure in the vicinity of some MORs, as quoted above. We thus chose to take into account this warmer state of young oceanic lithospheres in our modeling, which seems to be more realistic as it includes the thermal effects of both hydrothermal circulation and insulation by oceanic crust formation. As Grose & Afonso’s model is quite complex and not easy to reproduce, we choose to set a constant thermal gradient. It is true that as a consequence we alter the mechanical structure of the cooling plate, that may be then hotter and thus softer (and not stiffer, to our mind) than if the HSC model had been used.

2.3:

eq.(1) Is there a particular reason why you chose the Byerlee criterion instead of a Mohr-Coulomb criterion?

The brittle behavior simulated using equation (2) allows for modeling a yield stress depending on the lithostatic pressure (rho g z), instead of the normal stress. This is more convenient in Christensen’s code which does not directly solve the pressure field (see our answer to the Reviewer’s comment on the former page 7, eq. 3). However, the brittle parameter Gamma is computed as a function of the friction coefficient f_s, which is the actual ratio between shear stress and normal stress on the brittle fault. Gamma is instead the ratio between the tectonic horizontal stress and the vertical pressure, as explained in Section 2.5.1.

p. 6,l.26: Could you add a reference to justify the way you approximate the brittle strain rate?

The reference has been added:

p. 6 , l. 4-5 : “**The brittle deviatoric strain rate is computed assuming the relationship (Doin and Henry, 2001): ε̇ = ε̇_ref (tau/tau_y)^n_p ,...**”

p.7, eq.(3): Is there a particular reason why you use the lithostatic pressure in this equation and not the total pressure?

We have explained this choice in the new subsection 2.6 “Numerical code and resolution”,

p. 12, l. 11-12 : “**Note that using the lithostatic pressure in Eq. 4 is here numerically safer than computing the total pressure, which is not directly solved by Christensen’s code.**”

2.4:

I would suggest to use "Model geometry" instead of "box composition" in the title.

This subsection presents how the different compositions are distributed within the simulation box at the start of simulation. It corresponds to the description of both the geometry and the different materials that are simulated. Hence the title of the subsection (now 2.3) has been changed to:
p. 7 l. 32: “**2. 3 Lithological structure at simulation start**”.

As the choice of test parameters is of particular importance in this study, I would also suggest to merge the description of the model geometry together with the description of the initial thermal structure and merge the choice of tested physical properties with section 2.5.

We agree with the Reviewer that the justification of the choice of parameters should be presented separately. Please see below how we have modified it. Nevertheless, we prefer to have two distinct sections to present the composition distribution and the initial thermal state that requires a more detailed discussion (see above).

When it comes to the description of the investigated physical parameters, I was missing a bit the motivation for the specific choices made. For example, why did you choose the density of the TF as a parameter to be investigated? Is there any field evidence for such variations?

We did not find any reference to accurately assess the TF (transform fault) density, as explained in the text. The TF might vary from a composition mainly crustal close to the surface, to a much more mafic composition at depth. That is the reason why we varied the TF density between these 2 end-member values.

Also, I was wondering why the properties influencing the ductile strength of the lithospheric mantle were not considered at all here. As the lithospheric mantle makes up a large part of both the old and young plate, I would expect that it may have a significant impact on OPS. I am aware that this would add a large number of additional parameters to the existing study. For this reason, I think it is important to clarify why only the brittle parameter was changed for the lithospheric mantle and not any other parameters. I m aware that some of this motivation is given later in specific subsections, but while reading the manuscript, these questions arose for me when reading section 2.4.

We did test the ductile strength of the lithospheric mantle, though not systematically, by varying the reference mantle viscosity at plate base (by modifying the asthenosphere viscosity), and by varying the activation energy (E_a, Eq. 4) for the mantle, keeping constant the asthenospheric strength. We recognize that among the numerous experiments that we have performed and presented in the text, the reader may have some difficulties to notice these simulations: they were briefly summed up in the former Section 2.8.4 (p. 16, l. 7-9 in the initial manuscript). To correct it, we now announce these tests at the end of the new section 2.4 in which we justify the choice of the parameters that we have investigated:
p. 9, l. 4-6: “**We also test the influence of the lithosphere ductile strength that should modulate plate resistance to bending (2) by varying the mantle activation energy, Eam.**”

For this reason, I would suggest to remove the description of the choice of tested physical properties from section 2.4 and merge it with section 2.5.

We have indeed removed the description of the choice of tested physical properties from Section (now labelled) 2.3. ‘**Lithological structure at simulation start**’. It is still not merged with Section 2.5, but explained in the dedicated section 2.4, ‘**Parametric study derived from force balance’**, p. 8.

2.5.2:

p.9,l.5: You mention here that densities are a function of temperature in the model. This should be mentioned in the model setup section.

We have added the density dependence in temperature at the beginning of the model set-up section:
p. 5, l. 5-9: “**Density (rho) is assumed to be temperature- and composition-dependent:**
rho(C,T)= rho_ref(C)(1−alpha(T −Ts)) (1)
where rho_ref is the reference density at the surface, C is composition (mantle, oceanic crust or weak material; Sect. 2.3), alpha is the thermal expansion coefficient, T is temperature, and Ts is the surface temperature (Table 2). For the mantle, rho_ref is fixed to 3300 kg.m−3, while rho_ref for the oceanic crust and the weak material is varied from one experiment to another (Sect. 2.4)”.

2.5.3.

p.9,l.8-10. This sentence should be rewritten as it was very hard to read. I understand that you rescale the activation energy to account for the changed value of the stress exponent. I may have missed it, but I did not find any corresponding expression in Dumoulin et al. (1999). If I read correctly, they also use a different form of the rheological law. Could you therefore clarify how the activation energy rescaling is done?

We have detailed the way we rescale the activation energy used for the oceanic crust layer:

p. 11, l. 9-18: "**The most realistic interval for the crustal activation energy E_a^c can be defined from experimental estimates E_a^exp for an oceanic crust composition. Nonetheless, E_a^exp are associated with specific power law exponent, n, in Eq. 4, while we prefer to keep n = 3 in our numerical simulations for the sake of simplicity. Therefore, to infer the E_a^c interval in our modeling using a non-Newtonian rheology, we assume that without external forcing, mantle flows will be comparable to sublithospheric mantle convective flows. The lithosphere thermal equilibrium obtained using a non-Newtonian rheology is equivalent to the one obtained with a Newtonian ductile law if the Newtonian E a is equal to the non-Newtonian E a multiplied by 2/(n + 1) (Dumoulin et al., 1999). As sublithospheric small-scale convection yields strain rates by the same order of plate tectonics (∼ 10^−14 s^−1 , Dumoulin et al., 1999), this relationship is used to rescale the activation energies experimentally measured in our numerical setup devoid of any external forcing. We hence compute the equivalent activation energy as follows: E_a^c = (n + 1) × E_ a^exp /(n_e + 1), where n_e is the experimentally defined power law exponent.**"

p.10,l.2: "Still as a weakening mechanism..." Which weakening mechanism do you refer to? Why would a weakening mechanism imply a low activation energy?

We have replaced this awkward sentence by the more accurate following ones:

p. 11, l. 26-29: "**Nevertheless, a low plate ductile strength promoted by a thick crust has been suggested to favor spontaneous subduction initiation at a passive margin (Nikolaeva et al., 2010). We choose to not vary the crust thickness but to test in a set of experiments the effect of a very low crustal activation energy instead (equal to 185 kJ.mol −3 , Fig. 3e).**"

2.5.4

p.10, l.10: Here, an introduction to first results is given. As this subsection still belongs to the Model Setup section, I feel that a new section "Results" is needed.

The Latex command \section{Results} has been removed by mistake during the writing process, while it was exactly put at the place suggested by Referees 1 and 2. It has been re-inserted:

p. 12 l. 10: "**3. Results**"

Please note that consequently the numbering of the next subsections is thus completely modified.

2.6:.

In general, this section is well structured. In terms of comprehensibility, I would suggest to use the exact same terms for the different regimes as are used in fig.4. This would make it easier to relate the description in the text to the figure.

We have taken into account the Reviewer's suggestion when the terminology in Section (now) 3.1 was different from the one used in Fig. 4:

p. 12, l. 30: "**Second, we observe the YP ductile dripping...**".

p. 13, l. 7: "**Fourth, the YP sinking is triggered in some models...**"

p. 13, l. 13: "**Fifth, in one experiment, a double subduction initiation is observed:...**"

p. 13, l. 17: "**Sixth, the vertical subduction of the YP initiates...**". Please note that the adjective "vertical" has also been added in Fig. 4-6.

I would also suggest to color code the boxes (or something equivalent) in figure 4 according to the percentage of simulations that show the respective behaviour.

We think that a color coding of the panels in Fig. 4 would make the Figure rather hard to read. We did not discuss the relative proportions of each simulated behavior, because during the modeling process we were focusing on the conditions of OPS triggering, by tuning parameters to obtain it, which has lead to a biais in the parameter space exploration. Nevertheless, the reader may get an estimate of these percentages by looking at **the new Table S2** in the Supplementary material (p. 9-15), that presents our simulations as a function of the obtained tectonic regime. Table S2 has been color-coded depending on the simulated regime.

2.7:
p.11, l.13: Here the authors correctly state that OPS occurs when driving forces overcome resisting forces. Is there any way to estimate those forces beforehand for all simulations? As you have all the input, I think a rough estimate should be possible. Doing so would in my opinion add a very important aspect to the paper, as it would give us a better insight into the physics of the OPS problem. An estimation of those forces following the lines of [Mueller and Phillips, 1991] should be enough here.

We indeed tried to derive a simplified but quantified force balance from our experiments before submitting the initial version of the paper. We found that even a first order force balance in agreement with our modeling results was not easy to establish. However, the forces acting in subduction initiation at a TF are now presented to the reader in the new section 2.4 (p. 8-9).

p.11, l.30: "...very probably..." should be replaced with "most likely".

The correction has been made:
p. 15, l. 11: "**This swiftness most likely comes from...**".

2.8.2:
p.15, l.2: "... is supposed to be localized..." I think the authors rather mean "... is localized...".

The correction has been made:
p. 17, l. 12: "**The results presented in Sect. 3.3.1 are obtained when the weak material is localized at the TF only.**"

p.15, l.3: "... crust weakening laterally spreads out away from the TF..." I did not quite understand what the authors mean here. The sentence sounds as if they include a kind of weakening process in the models, which is not the case. I think the authors are referring to different simulations where they vary L_w? In this case, they observe a switch from YP vertical subduction to a gravitational instability.

We have modified the sentence:
p. 17, l. 12-14: "**Assuming that the weak material laterally spreads out away from the TF (L_w > 0 km), the mode of YP vertical subduction switches to YP sinking by gravitational instability**."

In this case, I think not only the extent of weakened crust plays a role, but also the chosen upper boundary condition (free slip), which inhibits plate sinking. The authors shortly discuss this issue in section 3.3. However, I think it has to be taken into account here that the mechanical impact of the weak crustal material may be overestimated due to the choice of the upper boundary condition. I think it would be enough to run a single simulation with "sticky air" to see if this is the case or not.

We have performed the tests suggested by the Reviewer for one plate age pair. The modified numerical set-up, as well as the obtained results, are detailed in the Supplementary material (**end of Section S3** and **Fig. S6**) and summed up in **the new Section 5.1.2** (p. 22). We detail this point below, by responding to a next comment about the free surface condition (comment on the Section formerly numbered '3.3 p.18, l. 23').

2.8.4.
Here the authors state that an additional weakening of the lithospheric mantle is required to allow for OPS. This is a very important point in my opinion, as it highlights the importance of the lithospheric mantle in this process. Could the additional weakening not also arise from a weaker ductile rheology?

We agree with the Reviewer, ductile mechanisms likely to decrease the lithospheric mantle must be considered. We have developed this discussion in a new section ("**5.1.4 Weakening of the oceanic mantle lithosphere**" p. 23-24). Please see below our answer to the Reviewer's comment on (the former) Section 3.3, on the sentence previously p.18, l.19.

p.16, l.8: Here it is stated that some simulations were also run with a lowered activation energy for the lithospheric mantle. I may have missed it, but I could not find any reference to the supplementary beforehand. I think it would be helpful to state before that a large number of additional simulations were run to test other physical parameters and that you chose to only focus on some of them.

We recognize that these additional experiments were hard to notice for the reader. They are now announced in Section 2.4, presenting our modeling strategy:

p. 9, l. 4-6: "**We also test the influence of the lithosphere ductile strength that should modulate plate resistance to bending (2) by varying the mantle activation energy, E_a^m.**"
These simulations are then presented in Section 3.3.4:
p. 18, l. 20-23: "**Moreover, we test different means to lower the OP rigidity. For four plate age pairs for which OPS aborts (5 vs 35, 7 vs 70, 7 vs 80 and 7 vs 90), we decrease the mantle ductile strength by lowering the activation energy E_a^m (Table 2) but keep constant the mantle viscosity at 100 km depth and the mantle brittle parameter (Gamma_m =1.6). We find that lowering E_a^m instead of the mantle brittle parameter is much more inefficient for obtaining OPS (Table S1).**"

2.8.5
It is interesting that a plume-like thermal anomaly does not trigger any OPS in the simulations presented here, but seems to be a very important process in other studies (e.g. [Burov and Cloetingh, 2010] [Crameri and Tackley, 2016] [Stern and Gerya, 2017] and others). Is it potentially related to melting processes (which are not modeled in the simulations presented here?) I think this issue is worth discussing.
We have detailed the discussion of the effect of a hot thermal anomaly on spontaneous subduction in the new Section 3.3.6:
p. 19, l. 7-16: "**The hot thermal anomaly never trigger OPS in our modeling, contrary to other studies, even if we have investigated large plate age contrasts (2 vs 40, sim. S17j, and 2 vs 80, S18k) as well as small age offsets and plates younger than 15 Myr (Table S1). To obtain a successfull plume-induced subduction initiation, it has been shown that the plume buoyancy have to exceed the local lithospheric (plastic) strength. This condition is reached either when the lithosphere friction coefficient is lower than ∼ 0.1 (Crameri and Tackley, 2016), and/or when the impacted lithosphere is younger than 15 Myr (Ueda et al., 2008), or when a significant magmatism-related weakening is implemented (Ueda et al., 2008) or assumed (Baes et al., 2016) in experiments reproducing modern Earth conditions. We hypothesize that if the mantle brittle parameter was sufficiently decreased, we would also achieve OPS by plume head impact. Besides, lithosphere fragmentation is observed by Ueda et al. (2008) when the plume size is relatively large in relation to the lithosphere thickness, in agreement with our simulation results showing the dismantlement for a significantly young (A y =2 Myr) and thin lithosphere.**"

2.8.6.
This is a very interesting section, as you list additional parameter that might have an influence on OPS, but did not turn out to have a first order effect. Together with the results from section 2.8.4. ,this indicates that the strength of the lithospheric mantle may be crucial in enabling OPS. For this reason, I think the potential effect of mantle rheology should be discussed more, e.g. with respect to other rheologies such as low temperature plasticity. Additionally, the hinge may be weakened by e.g. grain size reduction and thus a switch to diffusion creep could potentially help to initiate OPS. I am not saying that you should run additional simulations, but a more detailed discussion would be nice to highlight this issue. What you could do is to extract the effective viscosity in the hinge, which should be affected by brittle failure for low values of gamma_m. This should give you an estimate of the effective strength of the lithospheric mantle that is needed for OPS. You could then discuss which processes or parameters other than brittle failure could result in such effective viscosity values.
We have followed the Reviewer's suggestion. A new section has been added in the Discussion ("**5.1.4 Weakening of the oceanic mantle lithosphere"** p. 23-24). We have first derived a rough estimate of the mantle strength reduction necessary to achieve OPS:
p. 23, l. 29-32: "**A first-order estimate of the necessary mantle weakening is computed by comparing cases showing OPS to those in which OPS fails (Sect. S5 in the Supplementary material). The mantle weakening allowing for OPS is low to moderate for young plates and high plate age offsets (strength ratio ≤35), and larger when the plate age contrast is small (strength ratio ∼280).**"
We have detailed this estimate in the new Section S5 ("**Amount of lithospheric mantle weakening to model**") in the Supplementary material (p. 23, l. 35-p. 25, l. 9).
In the main text, we then discuss to which extent this weakening could be reached through different mechanisms:
p. 23, l. 32-p. 24, l. 2: "**One may wonder if such mantle strength decreases are realistic. Different mechanisms of mantle weakening may be discussed, such as (1) low-temperature plasticity (Goetze and Evans, 1979), that enhances the deformation of slab and plate base (Garel et al., 2014), (2) creep**

by grain-boundary sliding (GBS), (3) grain-size reduction when diffusion linear creep is activated, or fluid-related weakening."

We finally explain that these different weakening processes may not be activated in the setting of spontaneous subduction at oceanic TFs:

p. 24, l. 2-16: "**Peierls'plasticity limits the ductile strength in a high stress regime at moderately high temperatures (<1000°C, Demouchy et al., 2013) but requires a high differential stress (>100 to 200 MPa) to be activated. Similarly, GBS power law regime (2) operates if stresses are >100 MPa, for large strain and low temperature (<800°C, Drury, 2005). In our experiments, the simulated deviatoric stress is generally much lower than 100 MPa (Sect. S5 in the Supple. material). Consequently, implementing Peierls and/or GBS creeps in our model might not significantly change our results. Indeed, both softening mechanisms would not be activated and would thus not promote OPS in experiments failing in achieving it. Grain-size sensitive (GSS) diffusion linear creep (3) can strongly localize deformation at high temperature (e.g., Karato et al., 1986). In nature, GSS creep has been observed in mantle shear zones in the vicinity of a fossil ridge in Oman in contrast at rather low temperature (<1000°C, Michibayashi and Mainprice, 2004), forming very narrow shear zones (<1 km wide). However, the observed grain-size reduction of olivine**
is limited to ∼0.2-0.7 mm, which cannot result in a noticeable viscosity reduction. A significant strength decrease associated with GSS linear creep requires additional fluid percolation once shear localization is well developed within the subcontinental mantle (e.g., Hidas et al., 2016). The origin of such fluids at great depth within an oceanic young lithosphere is not obvious. Furthermore, GSS-linear creep may only operate at stresses <10 MPa (Burov, 2011), which is not verified in our simulations (Section S5 in the Supple. material)."

The end of Section 5.1.4 (p. 24, l. 17-26) corresponds to the second part of the former section 4.1 (Model limitations.)

3 Analysis

I was not sure why you started a new section here, as you continue to describe model results. I would therefore merge this section with the description of previous model results.

We partly agree with the Reviewer. Some authors prefer the interpretation of results to be done in the Discussion, while many modelers rather consider that the interpretation of simulations, that can easily be verified by looking at the different obtained mechanical fields, does belong to the Results section. As a compromise, we found an intermediate solution by presenting our interpretation in a "Analysis" section, distinct from both the Results and the Discussion sections.

3.1:

Judging by the title, the question of which parameters result in OPS is the main focus of the manuscript. Therefore sections 3.1 to 3.3 are in my opinion the most important results sections. For this reason, I would suggest to not refer to figures in the supplementary only, but to move some figures from the supplementary to the main part of the manuscript to better illustrate the distinction between mode1 and mode2 OPS.

We prefer to keep the main text of the article as concise as possible. We are afraid that the reader gets lost and that our 'take-home message' becomes less clear if additional figures are included in the main part of the paper.

3.3:

I liked that this section summarizes the different parameters and classifies them into resisting and promoting OPS. As suggested above, I would move part of this discussion to a separate section after the introduction where the basic physics/mechanics of the OPS process are explained (following the lines of [Mueller and Phillips, 1991]).

We have followed the Reviewer's suggestion, as detailed previously in this letter (**new Section 2.4** p. 8).

p.18, l.19: the necessity of a low brittle yield strength in the mantle is discussed her. In my opinion, weakening of the lithospheric mantle does not necessarily have to occur via brittle failure, but may also be due to different weakening processes, such as shear heating, grain size reduction and/or fluid infiltration. Additionally, a different creep mechanism such as low temperature plasticity could be crucial to weaken the lithospheric mantle. However, I think that this discussion should take place in the actual discussion section

and not here.
We have considered three mechanisms of mantle weakening among the ones suggested by the Reviewer. Please see our reply to the Reviewer's comment on the section formerly labeled 2.8.6. We have written a subsection in the Discussion focussing on lithospheric mantle weakening (“**5.1.4 Weakening of the oceanic mantle lithosphere**”, p. 23-24).

p.18,l.23: The free surface/free slip discussion should also be moved to the discussion section. Moreover, I am not really convinced by the arguments here that a free surface/sticky air approach would result in similar results. It is true that models with a weak crust and a free slip upper boundary condition show similar kinematics compared to models with a free surface/sticky air layer. However, I have the feeling that the importance of the strength of the crust is overestimated in the models shown here, as it not only resists bending, but also has to decouple the plate from the upper boundary. As a stick air layer is relatively simple to implement, a few simulations should be enough to show whether this is correct or not.
The tests suggested by the Reviewer have been performed. They are detailed in Section S3 (p. 22 l.32 – p. 23 l. 10) and illustrated by Fig. S6 in the Supplementary material:
“At last, the influence of the mechanical boundary condition at the box top is investigated. A free-slip condition inhibiting any vertical motion is prescribed in all the simutations presented before, whereas it has been shown that a free surface condition allowing for vertical deflection at the plate surface could strongly promote subduction initiation (Crameri et al., 2012b; Crameri and Tackley, 2016). We test how the implementation of a sticky air layer enabling for the free plate surface deformation could modify the OPS triggering modeled in our study by comparing the critical crustal brittle parameter that must be imposed to achieve OPS, with and without a free surface. Simulation S26a (Table S1) is chosen, since the plate age pair 5 vs 40 is just right above the threshold necessary for OPS triggering when the mechanical parameter set is the one displayed in Fig. 6-5 for (γ_c = 0.0005). We first perform 3 additional experiments to accurately estimate the threshold in crustal brittle parameter without free surface, γ c f ree slip (Simulations S26ai, S26aii and S26aiii, Table S1) and find that γ_c^free_slip ∼ 0.0025 (0.0001 ≤ γ_c^f ree_slip < 0.005).
Next, new experiments are run in which a thin low viscosity layer is inserted at the surface of the simulation box, 5 km thick (Fig. S6). This low viscosity layer is assumed to be made of water (density of 1000 kg.m^−3) as the transform faults considered in this study are all oceanic. Therefore, this low viscosity layer is dubbed a "sticky water layer" (SWL). The rheological parameters of the SWL are tuned to minimize its viscosity (E_a = 0 kJ/mol, γ_SW L = 5 × 10 −4 for instance) so that ν_SW L ∼ 3.8 × 10^11 Pa.s. Crameri et al. (2012a) have shown that, to correctly reproduce a true surface boundary condition, the SWL properties must enable to verify: C_Stokes 1, where C Stokes is the ratio between the pressure difference at the box surface and the vertical stress resulting in the surface deflection. For a Stokes flow, C_Stokes writes as (Crameri et al., 2012a):
C_Stokes =(1/16) (Δρ/ρ^ref_m) (H_0/H_SWL)^3 (ν_SWL/ ν_mantle) (1)
where Δρ is the slab density contrat (= αρ_ref ΔT~150 kg.m^-3), H_0 is the simulation box height (Table 2), H_SWL is the SWL thickness, ν_SWL is the SWL viscosity and ν_mantle is the mantle viscosity. By recalling that ν_mantle = ν_asth = 2.74 × 10^19 Pa.s (caption of Table S1), the SWL viscosity allows for verifying the required condition (C_Stokes ∼ 5.39 × 10^−5).
A short preliminary run is performed with the reference brittle parameter of the oceanic crust (γ c = 0.05) during 20 kyr to let the transform fault topography equilibrate (Fig. S6a). The crust brittle parameter is then varied between 0.0005 and 0.05 (Simulations S26f to S26fvi, Table S1). We find that γ_c ^free_surf ace ∼ 0.0175 (0.01 ≤ γ_c^f ree_surface < 0.025, Fig. S6b and c). The threshold in γ_c allowing for OPS is thus decreased by a factor ∼ 7 when the free surface is simulated for the plate age pair 5 vs 40.”
These extra experiments are summed up in the new Section 5.1.2 “**Free slip vs free surface condition**” in the main manuscript:
p. 22, l. 23-p. 23, l. 2: “**One may argue that the necessity of decoupling propagation close to the surface by shallow softening is related in our modeling to the absence of free surface (e.g., Crameri and Tackley, 2016). We test it by seeking for the threshold in the crustal brittle parameter allowing for OPS for one plate age pair 5 vs 40 (sim. S26a in Table 3) as a function of the mechanical boundary condition imposed at the box top, either free-slip without vertical motion or free surface, mimicked by inserting a “sticky water" layer (see the Supplementary material Sect. S3 and Fig. S6). For the selected plate age**

pair, the threshold in crustal brittle parameter turns out to increase from 0.0025 without free surface to ∼0.0175. Hence, the necessary crust weakness that must be imposed to model OPS may be overestimated by a factor ∼7. This result agrees with previous studies showing that the free surface condition promote the triggering of one-sided subduction in global mantle convection models (Crameri et al., 2012). Nevertheless, note that the threshold enabling OPS when the free surface is taken into account may still be an unlikely value, since it is close to the limit of the extremely low range of the crust brittle parameter (”red" domain, Fig. 3).”

We have limited the experiments including a sticky water layer to one plate age pair only, because our preliminary experiments performed with a sticky material layer mimicking a free surface behavior were suggesting that the issue would benefit from a numerical resolution study, which is beyond the scope of the present additional experiments (the numerical resolution used in all other simulations having been studied in details and validated in Arcay, 2017).

3.4:

p.19, l.10: Actually, the initiation process can be very fast in models without a prescribed weak zone when elasticity is included, as elastic stresses within the lithosphere are released at initiation (see e.g. Thielmann & Kaus (2012)). However, these simulations studied subduction initiation under compression, thus it is not clear if the same would happen for the model geometry used in this study.

We agree with that the effect of elasticity on the speed of the OPS initiation is not so easy to unravel. We have therefore modified the text:

p. 24, l. 16-20: “**Nonetheless, the potential effect of elasticity on the OPS kinetics is not clear. On the one hand, including elasticity could slow down OPS initiation by increasing the threshold in the strength contrast, as aforementioned. On the other hand, the incipient subduction has been shown to remain as fast as modeled in the present study in elasto-visco-plastic models testing different modes of subduction initiation (Hall and Gurnis, 2003; Thielmann and Kaus, 2012; Baes et al., 2016).**”

p.19, l.13: I do not completely agree here that elasticity only plays a minor role in the OPS process. [McKenzie, 1977] did show that elasticity may play a major role in this process, although his assumptions may have overestimated the impact of elasticity (see also discussion in [Mueller and Phillips, 1991]). As the models in Farrington et al. (2014) already start with a downward pointing slab, the initiation of free subduction is not fully included in their model, which is why I think it is difficult to draw any definite conclusions for the initiation of OPS from their simulations. Their study shows however, that the stress field in the hinge of the subducting plate is significantly altered if elasticity is included, in particular close to the surface. To me, this indicates that the importance of crustal parameters, in particular the brittle parameter of the crust may be overestimated when elasticity is not considered.

We perfectly agree, this point was exactly what we intented to suggest (see above and the initial version of our manuscript.

However, this is just a hypothesis and only further studies could shed more light on this issue. In any way, I don’t think that the influence of elasticity should be dismissed.

It was not our intention. We have even balanced a bit more our interpretation in the revised version of the end of Section 5.1.3:

p. 23, l. 24-26: “**However, if elasticity might compete against subduction initiation by limiting the localization of lithospheric shearing, it may also help incipient subduction through the following release of stored elastic work (Thielmann and Kaus, 2012; Crameri and Tackley, 2016)**”.

Discussion

4.1 This section is clear. I would add the discussion points from previous sections here.

We have followed the Reviewer's piece of advice. The first part of the Discussion, '5.1 Model limitations' is now made of 4 subsections. Among them, we have put the influence of the mechanical boundary condition at the surface of the simulation box (p. 22, “**5.1.2. Free slip vs free surface condition**”), and the factors favoring high velocities during the initiation process, including the discussion on elasticity ((p. 23, “**5.1.3 Initiation swiftness and influence of elastic rheology**), that were both before discussed in the Results section. We now discuss the potential of different weakening processes to reach the amount of softening necessary to model OPS in a separated subsection ((p. 23, “**5.1.4 Weakening of the oceanic mantle**

lithosphere).

4.2 This section is also clear. The high plate velocities observed in the simulations after subduction initiation are indeed quite large and may be a result of the chosen mantle rheology. However, as this manuscript is focused on the subduction initiation stage, I feel that this topic has to be left for future work. As it is anyway still debated whether the Yap subduction zone initiated at 20 Ma or whether it initiated earlier, it is reassuring that the simulations do not support its spontaneous initiation. I also do not find it surprising that subduction initiation due to OPS is not very probable, as earlier studies had also already hinted at this.

We thank the Reviewer for his constructive comment. We agree that the difficulty to initiate spontaneous subduction has already been partly addressed, however our goal is to better delimitate the parameter ranges enable the process, to show how narrow, extreme and hard to match they are.

Conclusions

The conclusions sum up the main results of this study quite well. Although it may seem to be a negative result, I think it is very important to show that OPS is not easy to achieve at present day conditions (within the model assumptions).

We agree with the Reviewer as we consider this point to be the main result of our study. It was and remains the meaning of the last sentence of the paper:

p. 26, l. 26-27: "**We finally conclude that the spontaneous instability of the thick OP at a TF is an unlikely process of subduction initiation in modern Earth conditions.**".

I would also add that the results highlight the importance of weakening processes within the lithospheric mantle, as these may significantly contribute to the occurrence of OPS.

We have added a sentence to recall this point in the conclusion:

p. 26, l. 19-20: "**Our study highlights the predominant role of a lithospheric weakening to enlarge the combination of plate ages allowing for OPS**".

Tables

Table 3: Would it be possible to group the different simulations according to the resulting deformation regime? I think this would make it easier to grasp the influence of the different parameters.

We thank the Referee for his suggestion, indeed such a Table will greatly help the reading. We have built a complementary Table (Table S2 in the Supple. Material) that compiles our experiments as a function of the simulated tectonic regime, which is highlighted using different colors. We still keep Table S1 to rank our simulations as a function of the simulated plate age pair, which we think is also necessary for the paper reading.

Additional references

-Burov, E. B., and S. Cloetingh (2010), Plume-like upper mantle instabilities drive subduction initiation, Geophysical Research Letters, 37, L03309.

-Cloetingh, S., R. Wortel, and N. Vlaar (1989), On the initiation of subduction zones, Pure and Applied Geophysics.

-Crameri, F., and P. J. Tackley (2016), Subduction initiation from a stagnant lid and global overturn: new insights from numerical models with a free surface, Prog. in Earth and Planet. Sci., 3(1), 30, doi:10.1186/s40645-016-0103-8.

-McKenzie, D. (1977), The initiation of trenches: a finite amplitude instability, Island Arcs, Deep Sea Trenches, and Back-Arc Basins, 1, 57–61.

-Mueller, S., and R. Phillips (1991), On the initiation of subduction, JGR, 96, 651–665.