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 # 1

General Comments

Arcay et al. present a parametric study of spontaneous subduction initiation, using numerical models, with a view to constraining the conditions required for this process to occur. This work comes at the perfect time. "Spontaneous" subduction initiation is being used as a mechanism to explain many features of the rock record around subduction zones in many recent studies, in a number of scientific areas. A full parameter study of the dynamic feasibility of this mechanism has not yet been undertaken and as such, I believe this work to be very important and will be useful to many. The modelling has been undertaken carefully and rigorously. The authors have checked a number of modelling assumptions that they have made to see whether they would influence their results, and discussed the others. As such, I believe the results of this study are robust. The majority of the conclusions drawn towards the end of the manuscript summarise the results well and are fair.
However some are perhaps too strong in places (specifically with regards to the Izu-Bonin-Marianas system, or IBM). I am not convinced that this study implies that subduction initiation via older-plate-sinking is impossible, simply that is has highlighted that it requires very particular conditions and perhaps a mechanism to weaken the top of the sinking/bending plate at it progresses. I agree that the study of all recent subduction initiation events implies that all (but one) do not fit the spontaneous model (especially given the specifics of how spontaneous initiation occurs in this study) and that this is a significant observation. However, the IBM remains a stand out for many reasons and it seems likely to me that this is because the IBM is the only example of spontaneous subduction initiation in this set. It would explain why fore-arc-basalt is only found at the IBM for one. What this study has done for me is put hard limits on what conditions must have been like at the time of initiation at the proto IBM, rather than the other way around.
It has also demonstrated how dynamically unlikely, and therefore rare, this type of event must be. These are still very powerful conclusions. This is actually what you glean as a reader from reading the current abstract already, so this is good. The careful consideration of the limits of "reasonability" of the parameter space is something that is of particular note in this paper: it would be great to see such a method adopted in all geodynamic parameter studies! A particular criticism I would have is towards the language used through the manuscript. This makes the manuscript difficult to read and my fear would be that it would put many people off attempting to do so (a shame when the science is good). I am not able to go through and correct the grammar, word choice and sentence structure throughout the entire manuscript as this is a lot of work. There were also a few places where I found it difficult to assess the science due to confusing use of language (I was brought close to suggesting that my revisions are "major" because of this). I would strongly recommend that the authors seek help from a native English speaker or a professional translator.

Given the importance of this work, and the care with which it has been undertaken, I would recommend this manuscript for publication, provided the comments below are at least considered and the language is corrected throughout.

We thank Referee 1 for his very careful review, his positive comments and his numerous constructive suggestions. 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, as we cannot reproduce here all the corrections that have been made regarding the language.

Specific Comments (by section)

Abstract

The abstract contains everything that I believe it should and is structured well. However, like the rest of the paper, it suffers from the confusing use of language. Some examples just from the abstract: "We propose a new exploration of the concept of "spontaneous" subduction" – "We present a parametric exploration of the feasibility of "spontaneous" subduction initiation"? "in recent subduction initiations" – "from recent subduction initiation events"? "The basic parameters to simulate OPS are" - "The parameters which exert the strongest control over whether OPS is feasible or not are. . .."? Etc.

We have modified the text to correct our English:

- p. 1, l. 1: "**We present an extensive parametric exploration of the feasibility of "spontaneous" subduction initiation**""
- p. 1. l. 2: "**from recent subduction initiation events at a TF… »**
- p. 1, l. 13-14: "**The parameters that exert the strongest control over whether OPS can occur or not are…** "

In addition: "We find that all mechanical parameters have to be assigned extreme values to achieve OPS, that we consider as irrelevant" – It seems in the paper than the parameters don't all simultaneously have to be set to extreme values?

Indeed, our results show that, simultaneously, one parameter at least must be set to an unrealistic value (belonging to the "red range" in Fig. 3) and two parameters at least must be chosen in the infrequent interval ("yellow range"). We have revised Fig. 6 accordingly. We have slightly moderated our conclusions summed up in the abstract and in the conclusion section:

-p. 1., l. 15-17: "**We find that at least one mechanical parameter has to be assigned an unrealistic value and at least two other ones must be set to extreme ranges to achieve OPS, which we do not consider realistic."**
- p. 26, l. 31-32 **: "OPS occurs (…) only if the initial mechanical setup is adjusted beyond reasonable limits for at least one key thermomechanical parameter."**

Also "irrelevant" would imply that this is an unimportant result, when it is really one of the key results of this paper! Is this what the author intends to write here? I would argue that it is very relevant.

We thank Referee 1 for his suggestion. "Irrelevant" has been removed and replaced by a more appropriate expression:

p. 1., l. 17 : **"…, which we do not consider realistic."**

Introduction

This introduction is a very thorough overview. In terms of content, I have very little to add. Just one small comment/question: is it uncontested that the forearc of the IBM has been consumed by subduction erosion? There are many studies which assume otherwise. I would perhaps reword this part to reflect this. Figure 1 is clear to me.

The reviewer says that there are many studies which assume otherwise. Since no references have been provided by the reviewer about studies that do not support tectonic erosion in the IBM margin, it is difficult to answer. Lallemand (2016) describes in a chapter all the pieces of evidence supporting margin loss by tectonic erosion along that subduction zone according to many authors (Hussong and Uyeda, 1981; Bloomer, 1983; von Huene and Scholl, 1991; Lagabrielle et al., 1992; Mitchelll et al., 1992; Fryer et al., 1999, 2006). Quickly summarizing the situation: (1) the IBM trench is devoid of any trench fill today and probably in the past since it has always been fringed by few volcanic islands, (2) 17 to 41 Ma volcanic rocks supposed to belung to the former arc have been reported near the trench, (3) it has been shown that the forearc has subsided by more than 2 km since about 40 Ma, (4) dismantlement of the margin is attested by the numerous fractures and even serpentinite diapirs. As 5 papers were already quoted in the initial text (p. 2 l.2 : Natland and Tarney, 1981; Hussong and Uyeda, 1981; Bloomer, 1983; Lallemand, 1995, and l.3 : Lallemand, 2016), we think it is sufficient.

Model Setup

2.1 Does the code have a name?

No, U.R. Christensen did not give a name to his code.

Fig 2: It might be good to put the meaning of symbols used (Lw(Ao) etc.) and it might be good to label isotherms (perhaps just in the inset?).

The isotherm labels haved been added in the inset of Fig. 2, and the correction regarding the L_w definition has been made:

Fig. 2 (p. 5, l. 3-4 in the caption): "**Lw is the width at the surface of the younger plate and of the older plate (aged of Ay and of Ao Myr, respectively) over which the oceanic crust is assumed to have been altered and weakened by the TF activity.** "

Why is the 1400 isotherm so irregular? Is this the initial condition?

The irregular depth of the 1400-K isotherm reflects the small-scale convection existing in the initial conditions. This was indicated in the initial manuscript (p. 7, l. 4-6 in the revised manuscript). We have verified that this initial mantle thermal state was not affecting the OPS process (l. 6-7).

2.2 Is the method of using a conductive lid with a constant thermal gradient really valid for young plates? Is the value of 0.75, for the "overcooling" correction grounded in anything? If it is then it is probably worth mentioning.

We have detailed this point a bit more in the revised manuscript :

p. 6 (l. 28)-7 (l. 17) : **"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). 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 0.75 of the ones predicted by HSC. Between the surface and z_LB (A), the thermal gradient is constant.

2.5 I like the summary figure 3. These are not all the parameters varied however. Would it be possible to encompass the fact that the asthenospheric temperature, width of thermal step and the presence of a plume were also tested here for completeness?

If other parameters were locally tested, the main parametric study only encompasses the 6 parameters mentioned in Figure 3 so we choose to keep this representation to match the parameters range given in Figures 4-5-6. (Note that we did not test the asthenospheric temperature, but the asthenospheric viscosity.)

Nevertheless, we add a new section (2.4 "**Parametric study derived from force balance**") to take into account a comment from Referee 2, to justify the choice of the 6 parameters, in which we explain that, apart from these 6 main parameters, we also test a few additional parameters, and explain why:
p. 9, l. 1-7: "**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 (iv), 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 (ii) by varying the mantle activation energy, E_a^m. Eventually, we study the TF structure impact by exploring a few different widths of the TF weak gouge, also testing different thermal structures of the plate boundary forming the TF.** "

2.5.1 Gamma c is close to 0.08, not 0.8 using this equation.
The Reviewer is referring to data computed in Sect. 2.5.1 p. 9, l. 14 (revised manuscript). Using in Equation 5 (previously numbered 5) Lambda=0.5, f_s=0.6 and rho=2920 kg/m3, we obtain Gamma_c ~ 0.7. To obtain Gammas_c~0.8 (0.766 exactly), one has to use: Lambda=0.45 and rho=3300 kg/m3 instead. We thank Referee 1 for his checking. We have corrected it and clarified the value assumed for Lambda:
p. 9 l. 20-21: "**Assuming high pore fluid pressure in the oceanic crust (λ ≥ 0.45), γ c from Eq. 5 is then close to 0.8 (Fig. S1).**"

However, forgive me if I am wrong, but I do not see where the term (1-rho_w/) comes in. Anderson theory of faulting gives us:
Δσ_xx = [2f_s (p_lith – p_w)] / ((1+f_s^2)^0.5 - f_s)
and λ=p_w/p_lith so surely
(Δσ_xx)/p_lith =(2f_s (1-λ)) / ((1+f_s^2)^0.5 - f_s)
This also makes more sense to me when thinking about the mantle, where you would expect no pore fluid so you rightly use λ=0. In your current equation, why should _w play any role in this case?
We deeply thank the Reviewer for his careful revision. Lambda was awkwardly labeled as the "pore fluid pressure ratio" while in our actual definition it should be labeled the "pore fluid pressure coefficient", the pore fluid pressure p_w writing as: p_w = g*z*((1-Lambda) rho_w + Lambda * rho)), to have p_w= g*z*rho_w when Lambda=0 (hydrostatic pressure) and p_w= g*z*rho when Lambda=1 (lithostatic pressure). That is the reason why the factor (rho_w-rho) appears in Eq. 5. We have clarified the definition of Lambda, that was previously missing:
p. 9, l. 16-17:"...**where λ is the pore fluid pressure coefficient, ρ w is the water density, and p w is the pore fluid pressure, assuming that p_w = rho_ w gz if λ = 0 and p_w = rho gz if λ = 1**".

Moreover, Referee 1 is perfectly right saying that, if fluid is absent, it is incorrect to use Equation 5. If fluids are absent (leading to p_w=0 Pa), the equation used to compute Gamma must be: Gamma = 2f_s / ((1+f_s^2)^0.5 – f_s) (in agreement with the Reviewer's statement). We have corrected it by adding the former equation (labeled 6) and by indicating the condition in terms of fluid pressure allowing for using either Eq. 5 or Eq. 6:
p 9, l. 13-14: "**Gamma = 2f_s (1-Lambda) (1-rho_w/rho)/ ((1+f_s^2)^0.5 – f_s) if p_w /= 0 Pa**
Gamma = 2f_s / ((1+f_s^2)^0.5 – f_s) if p_w=0 Pa "
but also by modifying Fig. S1 in the Supple. Material displaying the relationship between f_s and gamma when p_w=0 to account for this correction. Accordingly, we have also modified the text dealing with the Gamma estimated for the mantle:
- p. 9 l. 30, p.10 l. 1-2: "**To simplify, we suppose the pore fluid pressure p_w to be very low, close to zero, assuming that the lithospheric mantle is dry in absence of any previous significant deformation.**"
- p. 10, l. 2-7 "**The coefficient of internal friction from Eq. 6 for a dry mantle decreases from f s = 0.65 (Byerlee, 1978) to f s ∼ 0.35 or 0.45 if peridotite is partly serpentinized (Raleigh and Paterson, 1965; Escartìn et al., 1997), leading to γ m between 2.8 and 0.8. However, assuming γ m = 2.8 would lead to an extremely high lithospheric strength (∼1 GPa at only 11 km depth) since our rheological model neglects other deformation mechanisms. We thus restrict the maximum Gamma_m to 1.6, which has been shown to allow for a realistic simulation of subduction force balance for steady-state subduction**

zones (Arcay et al., 2008). The most likely interval for Gamma_m is eventually [0.8-1.6] (Fig. 3b)."

Of course using this line of reasoning assumes an interconnected fault network within the material considered. I do not see a problem with this (in the crust at least) as the author is searching for the lower bound limit here, but I think that this is worth stating.
We have taken into account this warning and added it in the text:
p. 9, l. 25-27: "**Note that relationship between the presence of fluid and its effect on the effective brittle strength (Lamda value) depends on the fault network and on the degree of pore connectivity, which may be higly variable (e.g. Carlson and Herrick, 1990; Tompkins and Christensen, 1999).**"

With regards to explaining the low brittle parameters for the mantle, see my comments below discussing Peierl's creep.
The Reviewer's comment has been taken into account in the revised discussion (new Sect. 5.1.4, see our response in the present letter p. 9).

2.5.3 The author has made the effort to correct for the fact that different studies use different stress exponents but not corrected for the fact that different studies use different rheological prefactors. These prefactors effectively normalise each flow law and as such, the activation energy and rheological prefactor cannot be thought of as independent. In general, experimental flow laws with higher activation energies have lower rheological prefactors and vice-versa. Therefore here, the author is likely significantly over-estimating the variability in experimental flow laws (a better way of doing this is to take all the experimental flow laws one wishes to consider and finding their average and standard deviation and using these as bounds for example). If the author has applied a form of normalisation, either to ensure a constant upper mantle viscosity (which I know is commonly done) or with the original experimental flow laws in mind, then there is no problem, although I would ensure that this is made clear in the text. Side note: I see that the effect of the crustal activation energy is very limited in the results section, so if running these models again is necessary, but difficult, then perhaps it is worth leaving out the investigation of activation energy?
This is a relevant remark since pre-factors are also variable from one flow law to another, as mentioned by the Reviewer. To simplify, we have chosen in this paper to explore the variability to only one parameter for crustal rheology (activation energy) as a proxy for other sources of rheology variations (e.g. chemistry, fabrics, grain size). We deem it relevant to maintain a wide range for the crust activation energy (hence for crustal rheology) since the amount of decoupling through the subducting crust is crucial for subduction dynamics. Adjusting the pre-exponential factor would have possibly reduced the range of crustal ductile strength, while we thought more adequate to explore the largest interval.

2.5.4 The last paragraph would be a good introduction to a whole new section as from here on in as it seems like the rest of section 2 is now results and not model setup. I would personally just call this section "Results".
The Latex command \section{Results} has been removed by mistake during the writing process, while it was exactly put at the place suggested by the Referee. It has been re-inserted:
p. 12 l. 17: "**3. Results**"
Note that the numbering of the next subsections is thus completely modified.

2.6 I would make clear that the 65% are non-OPS. The last "almost OPS" mode paragraph is very confusing. It would be better to say that "in 40% of models which appear to start to show OPS behaviour, freeze up within. . .." Or something similar, rather than talk about these models as if they are proper OPS.
The sentence has been modified to make it clearer. Additionally, the number of simulations has been changed (5 simulations that were not useful to mention have been removed, including 1 OPS case, while 10 new experiments have been performed to respond to a request from Referee 2 regarding the free surface boundary condition, made of 7 OPS and 3 non-OPS cases):
p. 12, l. 25-26: "**This large simulation set shown in Fig. 4 represents ~73% of the 302 experiments presented in this study, which do not show a clear OPS.**"

2.8 Fig 6: This regime diagram is great. It might be useful to have points on the diagram corresponding to the actual models run.

We initially thought of depicting the experimental points, but it made the sketches inserted to illustrate the different regimes difficult to handle. We thus did not modify the regime diagrams depicted in Fig. 6 but add two additionnal figures in the Supplementary data showing all the experiments used to define the boundaries delimiting the different regimes, without the sketches (see the new Fig. S2 and S3 p. 17-18 in the Supple. Material). These additional figures are quoted in the main text:
- caption of Fig. 6, p. 16, l. 3-4 counted from the bottom: " **The corresponding experiments are displayed in Fig. S2 and S3 in the Supplementary material.**"
- p. 20, l. 29-31: "**The boundary between OPS and the absence of subduction can be defined for a normal mantle brittle strength γ m = 1.6 (Fig. 6f) using simulations in which OPS aborts (such as simulations […..], Fig. S3 in the Supplementary material).**"

I feel the individual sections below would benefit from having their own regime diagram where the parameter being looked at has one of the axes (eg. For 2.8.4 it would be good to see how the critical mantle brittle parameter varies visually). However, I do understand that having hundreds of regime diagrams is not useful and it is difficult to put them together for such multi-dimensional results.
We agree with the Reviewer's last comment and prefer to not multiply the regime diagrams. Fig. 6 was the best compromise that we found, to make the presentation of our results as concise and clear as possible.

2.8.2 "The aforementioned results are obtained when crust weakening is supposed to be localized at the TF only (Lw =0 km)." Some of the non-OPS mode examples in figure 4 clearly have Lw>0. . .. Does "aforementioned" just refer to section 2.8?
We have modified the corresponding sentence to make it clearer:
p. 17, l. 12: "**The results presented in Sect. 3.3.1 are obtained when the weak material is localized at the TF only (L_w =0 km).**"

2.8.3 What is Lw in this case?
In this cas, L_w=1100 km. This has been added twice:
- p.17, l. 30: "**OPS can initiate for numerous plate age pairs if the whole crust is mechanically weak (L w = 1100 km, Fig. 6f),...**"
- p. 18, l. 1-2: "**To determine the threshold in γ c allowing for OPS, we choose a high plate age offset, 2 vs 80, the most propitious for OPS (keeping L_w = 1100 km).**"

2.8.4 This is a great point. There is another mechanism that would help facilitate plate bending and that is Peierls'creep. Including this mechanism may have a similar effect to decreasing the mantle friction coefficient. This is perhaps a point for the discussion, but I think it is worth bringing up.
Thanks for this remark. The discussion about the mechanism able to weaken the lithospheric mantle has been moved to the Discussion section (subsection 5.1.4, "**Weakening of the lithospheric mantle**"). We have included the Reviewer's comment:
p. 23, l. 33-p. 24, l. 8: "**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. 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. [...] In our experiments, the maximum deviatoric stresses 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.**"

2.8.5 I find the result that changing the ductile strength of the crust and TF has little effect unsurprising as these regions are most likely to deform in a brittle manner in the case of subduction initiation.
We came to the same conclusion, as it was written in the former Sect. 2.8.6 (now Sect. 3.3.7. p. 20 l. 1-2): "We here verify that the fault gouge weakening, governed by the soft material brittle properties, is independent of temperature and, at first order, is independent of the fault activity in our 2D setup."

This and the fact that the only time that changing the activation energy has any effect is when the plates are effectively crustal plates, would indicate to me that changing the ductile behaviour of the mantle, and not the crust, would have the larger effect and is the more worthwhile investigating. If it comes to re-running models, then I would consider looking at this instead (although I should say that there is technically nothing wrong with it as it is!).

We had already considered this point in the former version of the manuscript, by investigating both the asthenosphere viscosity and the effect of the mantle activation energy. In the revised version, the tests regarding the asthenosphere viscosity are now announced and justified in Sect. 2.4:

p. 9, l. 1-4: "**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 (iv), 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).**"

Regarding the investigation of the mantle activation energy influence, it was mentioned p. 16, l. 7-8 in the initial manuscript (Simulations S25b, c, d; Sim. S32b, c, d; Sim. S33b, c, d; S34c, d, e in Table S1). We admit that this point was extremely briefly explained and could easily be missed by the reader. To correct it, first these tests are announced at the end of the Sect. 2. 4 (that has been added):

p. 9, l. 4-6: "**We also test the influence of the lithosphere ductile strength that should modulate plate resistance to bending (ii) by varying the mantle activation energy, E_a^m.**"

Second, we detail a bit more these experiments in Sect. 3.3.4 (formerly 2.8.4):

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).**"

The plume head having little effect is a very interesting result, particularly as many people invoke the influence of plumes to catalyse spontaneous subduction initiation. I know this section is short, but I would say it deserves its own heading.

We have followed the Reviewer's piece of advice and made a separated section (Sect. 3.3.6) that is more developed:

p. 19, l. 3 -p. 20, l.16: "**3.3.6 Plume-like thermal anomaly**"

The thermal anomaly simulating an ascending plume head below the TF produces effects very similar to those of a reduced E_a^c : no effect if plates are older than 2 Myr, YP dismantlement if A y =2 Myr and if the crust is dense (rho_c =3300 kg.m −3). Otherwise, for a normal crust density, a short stage of YP vertical subduction occurs after plume impact (2vs10, simulation S15h). 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 successful 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 have also achieved 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 It would be good here to emphasise that the brittle parameters were inverted for models which originally displayed OPS, and then do not after the inversion. It took me to read the supplement to understand this.

We have added this point in the revised version:

p. 20 l.12-14 :"**We first test the necessity of the fault softness to simulate OPS by inverting the oceanic crust and TF respective brittle parameter for models that originally displayed OPS (thus by setting for**

the inversion experiments: Gamma_T F =0.05, while Gamma_c =0.0005)."

Likewise, it would be good to emphasise that the models being looked at when increasing the fault width, originally did not demonstrate OPS.
We have added this point in the revised version:
p. 19 l. 24-26: "**We next wonder if OPS (when not modeled) could be triggered by widening the fault gouge from the surface to the bottom of the fault (domain 1 in Fig. 2) by setting the fault width to 20 km instead of 8.3 km in experiments that did not initially show OPS**."
We have also noticed that the width of the weak fault was not mentioned in Table S1. We have corrected it (Simulations S22t, S37r and S37s).

The fact that OPS occurs independent of the width of the step change in thermal profile is a very interesting result!
We thank the Reviewer for his positive regard. We have added this point in the conclusion section:
p. 26, l. 18-20: "**In addition, we find that neither the thermal structure and blurring of the transform fault area nor a plume head impact are able to affect OPS triggering in our modeling setup.**"

Analysis
3.1 Surely the important criterion for mode 2 to occur is for the younger plate to be weak enough to stretch or break and therefore move with the sinking older plate? What was it that led the author to believe that it was more to do with coupling to the asthenosphere? Is there an aspect of the model set-up which means that the YP is always free to move? If there is a reason then it would be good to clarify this in the text.
The mechanical condition at the YP surface as well as along the YP vertical segment on the right-hand side of the simulation box is always free-slip (Fig. 2). When OPS occurs, either in mode 1 or in mode 2, we find that the YP must be able to deform in all cases, either to allow for the asthenosphere upwelling in the vicinity of the TF in OPS-mode 1 (Fig. 5a), or to be stretched as a result of the OP hinge retreat in mode 2 (Fig. 5b, c, though the stretching area does not appear in the close-up). Indeed, YP spreading/stretching systematically occurs in mode 2 (with a spreading center located ~ 150 to 300 km away from the TF), in spite of the free-slip boundary condition. As a consequence we do not think that a difference in YP strength can explain the switch from mode 1 to mode 2. That is the reason why we thought that the difference in OPS-behavior comes from a difference in the degree of lithosphere-asthenosphere coupling, as suggested by the analysis of viscosity profiles (Sect. S4 and Fig. S8 in the Supplementary material, quoted p. 21 l. 15).

3.2 Apart from the wording, this section is clear.
The language has been corrected. For instance:
p. 21, l. 7: "ageing" has been replaced by "**aging**"
p. 20, l. 29: "**a normal mantle brittle strength**"
p. 21, l. 4:"**(separately considering the cases...)**"
p. 21, l. 9: "**the conditions that are the most propitious for OPS...**"

3.3 I am glad that the author discusses the free surface here. Perhaps this is the point at which Peierl's creep could also be mentioned?
The influence of a free surface is now presented and discussed in a new subsection in the Discussion (**5.1.2: Free slip vs free surface condition**, p. 22), as additionnal tests including a sticky "air" layer have been performed to answer to Referee 2's comment. The Peierls'creep mechanism is evoked in the new Section 5.1.4 p. 24 dealing with the different processes that could produce a mantle weakening:
p. 23, l. 33-p. 24, l. 8: "**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. 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. [...] In our experiments, the maximum deviatoric stresses 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."

I would also argue that the concluding sentence here is quite strong. As the rest of this section alludes to, the primary parameter that needs to be tuned "beyond a reasonable value" is the width of the weak layer at the top of the model.

We have moderated the sentence and added a new one:

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 (ρc, 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.**"

Please note that we have also modified Fig. 6, moderated our conclusion in Sect. 6, and at the end of the abstract, as already underlined in the present letter (see our response to Referee 1's comment p. 2 in this letter).

The process suggested by Dymkova and Gerya 2013 surely offers a mechanism by which this weakening could happen? I personally see this result emphasising the need for such a weakening mechanism, rather than suggesting that OPS is impossible.

We discuss the necessary amount of mantle weakening required to achieved in the new Section 5.1.4. "Weakening of the lithospheric mantle" p. 24. Quotation of Dymkova and Gerya's paper has been moved to this section. A mechanism able to soften the lithospheric mantle indeed strongly promotes OPS. In the revised version, we have estimated the amount of strength reduction that should be applied 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.)

3.4 Again, apart from the wording, this section is clear.
The language has been corrected, for instance
p. 23 l. 7: “**has also been**” (instead of “has been also...”)
p. 23 l. 9: “**similar to...**” (instead of “ close to...”)
p. 23. l. 11: “**due to...**” (instead of “thanks to...”)

However, I would add that Reagan et. al 2019 has suggested that subduction initiation really did occur in 0.5-1 Myrs at the IBM (given the very short duration of fore-arc basaltic magmatism).
Please see our response to Referee 1's comment on Sect. 4.2.

Discussion
4.1 Spontaneous initiation would also be easier in 3D simply due to the extra degree of freedom. For example, the model by Zhou et al. 2018 suggests that the sinking plate is able to sink in one place initially and then subduction initiation propagate away from this point. This takes far less energy than requiring that the whole older plate sink along the entire transform fault simultaneously (what is effectively modelled when modelling in 2D).
The effect of a 3D setup with respect to a 2D setup may depend on the mode of subduction initiation that is considered, either by propagation or by “nucleation”, that is, initiation strictly speaking. Along strike-propagation is likely easier than initiation strictly speaking, and cannot be modeled in 2D. We think that Zhou et al. have modeled subduction initiation at the spreading center then propagation away from it by affecting older and older plates. In our study we focus on subduction initiation “nucleation”, and not propagation, as a function of the considered plate age pair. We have modified the text to clarify this point:
p. 22 l. 17-21: “**Finally, one may argue that a 3D setup would intrinsically facilitate OPS propagation at a transform fault. Plate sinking might initiate at the location where the offset in plate thickness is maximum (in the vicinity of a ridge spreading center) and then propagate away from this point (Zhou et al., 2018). However as we focus on subduction initiation strictly speaking and not on subduction propagation, the use of a 2D setup should remain meaningful to unravel the conditions of spontaneous sinking for a given plate age pair, considering apart the problem of the transform fault slip.**”

I agree that permeability through the mantle is likely lower than Dymkova suggest, and this is a very good point to raise here (although I do not think that it necessarily negates my comment for section 3.3). Another feature common to models of initiation, not included in this study, is a strain history dependent rheology (damage). I do not actually think that it would affect the results of this study significantly, though I would say that it is worth a mention at this stage.
The Reviewer might refer to a grain-size reduction process. This weakening mechanism is now discussed in Sect. 5.1.4:
p. 24, l. 6-16: “**Grain-size sensitive (GSS) diffusion linear creep (3) may strongly localize deformation by mantle rock softening at high temperature (e.g., Karato et al., 1986). In nature, grain-size reduction in mantle shear zones in the vicinity of a fossil ridge has been observed 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 may occur thanks to 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. Moreover, GSS-linear creep may operate only at stresses <10 MPa (Burov, 2011), which is not verified in our simulations.** ”

4.2 This section is clear. However, the conclusion of Reagan et. al 2019 (the most recent study informed by the most recent drill core data) is actually that subduction initiation must have occurred very rapidly (<1 Myrs). In this case, the modelling results presented in this paper are not at odds with subduction initiation

having been spontaneous at the IBM.

The results of Reagan et al. (2019) indicate that a few core and submersible samples, located on the inner slope of Izu-Bonin Trench off Bonin islands, show a remarkable short time period of 50-52 Ma for both the full eruption of the « forearc basalts » and the oldest « boninites », all younger boninites being considered as altered or reheated. Then, the authors interpret these data as evidences for near-trench seafloor spreading forming basalt then boninite within less than 2 my after subduction initiation. We personally consider that their conclusion provides one possible scenario based on petrological/geochronological data but it exists at least another scenario (our prefered one) where the present-day sample area was initially located far from the trench and was brought near the trench after margin's removal. The same « forearc basalts » were drilled at site U1438 in the Amami-Sankaku Basin (Arculus et al., 2015), which was in back-arc position at time of subduction initiation, with an estimated age of 51 to 64 Ma (most probable 55 Ma according to the authors). It is highly probable that these FAB samples belung to the same oceanic basin which opened normal to the initial transform fault that will further evolve into a subduction zone (Lallemand, 2016). This is incompatible with the model shown in Reagan et al. (2019) involving a spreading axis parallel to the trench which has never been observed in any forearc in the world!

However, I do agree that post-initiation velocities in the models presented in this paper are unrealistically high once subduction is established and this remains an issue. I would argue that these unrealistically high velocities are, at least in part, the result of 2D modelling: in 3D the subduction zone would "unzip" more gradually. The plate that has not yet started sinking would prevent the sinking part from reaching such high velocities.

We have taken into account the Reviewer's comment in the revised manuscript:

p. 23, l. 13-15: "**Moreover, such unrealistically high velocities at plate sinking onset may result at least in part from the 2D setup since, in a 3D setup, the along-strike propagation slows down the initiation process; however, speeds of hinge retreat remain significantly high (between 13 and 20 cm/yr in Zhou et al., 2018).**"

4.3 I particularly like how this study presents "failed" or "aborted" subduction initiation events as existing on a spectrum with the successful ones. If anything this could be emphasised more!

We thank the Reviewer for his very positive comment.

Conclusions

The conclusions are well structured and summarise all the key points. I would perhaps also mention a few of the other strong conclusions that can be drawn from this study: the thermal blurring having no effect; an incident plume having little effect etc.

We have modified the conclusions:

p. 26, l. 18-20: "**In addition, we find that neither the thermal structure and blurring of the transform fault area nor a plume head impact are able to affect OPS triggering in our modeling setup. Our study highlights the predominant role of a lithospheric mantle weakening to enlarge the combination of plate ages allowing for OPS.**"

The only other recommendation I would have is that the second to last sentence is worth rewording/softening; especially as it would seem that the geological record is not necessarily at odds with the catastrophic mode simulated in this study (see general comments).

We do not think that the IBM subduction zone can be considered as a "spontaneous" subduction initiation, since we do not agree with the interpretation of geological records of subduction initiation at IBM (see for instance our response to Referee 1's comment regarding subduction erosion in the Introduction section, p. 2 in this letter, and our response to the comment on the previous section 4.2 on the preceding page). We estimate that our reasoning is sufficiently detailed and justified in Section 5.2 (former Section 4.2) to not modify our conclusion.

Tables

All three tables are very valuable. If feasible, Table 3 would really benefit from colour-coding (given its scale!) although I am aware that this may not be possible.

We have built a second Table to compile our experiments as a function of the plate deformation that is

simulated (Table S2 in the Supplementary material). To help the reading, Table S2 is color-coded as a function of the simulated behavior. We think that this complementary Table should help the reading. This new Table is quoted in the main text:
p. 12, beginning of the "Results" Section (l. 21-22): "**The experiments are compiled as a function of the plate age pair imposed at the TF in Table S1, while they are ranked according to the simulated deformation regime in Table S2.**"

References quoted in the response to Referee 1 but not in the manuscript:

- Fryer P, Wheat CG, Mottl MJ (1999) Mariana blueschist mud volcanism: implications for conditions within the subduction zone. Geology 27:103–106
- Fryer P, Gharib J, Ross K, Savov I, Mottl MJ (2006) Variability in serpentinite mudflow mechanisms and sources: ODP drilling results on Mariana forearc seamounts. Geochem Geophys Geosyst 7:8. doi:10.1029/2005GC001201
- Grose, C., 2012. Properties of oceanic lithosphere: revised plate cooling model predictions. Earth Planet. Sci. Lett. 333–334, 250–264.
- Lagabrielle Y, Sizun J-P, Arculus RJ et al (1992) The constructional and deformational history of the igneous basement penetrated at Site 786. In: Fryer P, Pearce JA, Stokking LB (eds) Proc. Ocean Drilling Program, Sci Results, Vol 125. Ocean Drilling Program, College Station, Texas, pp 263–276
- Mitchell JG, Peate DW, Murton BJ, Pearce JA, Arculus RJ, van der Laan SR (1992) K-Ar dating of samples from sites 782 and 786 (Leg 125): The Izu-Bonin forearc region. Proc Ocean Drill Program Sci Results 125:203–210.
- von Huene R, Scholl DW (1991) Observations at convergent margins concerning sediment subduction, subduction erosion, and the growth of continental crust. Rev Geophys 29:279–316