We thank the editor and reviewers for very helpful comments, which we think have helped to improve our manuscript. We have tried to address all of them as detailed below.

Editor's comments:

This is a very impressive piece of work on a technically difficult but biologically very relevant question. The model clarifies the role different components of epistasis on inbreeding depression and on the evolution of selfing. The introduction replaces this model in among previous studies that have addressed this subject.
I am very much willing to recommend this preprint but it would be great to use the comments from both reviewers to improve the manuscript. The paper is beautifully written but very dense. There are two main parts (inbreeding depression and evolution of selfing) but many subcases. I don't know if it is feasible but it would be great to provide a way (an additional figure?) to summarize some aspects of your work to the less theoretically-oriented readers. You may want to try provide a schematic description of some features of your model. Visualizing the different components of epistasis would be useful given their importance in your model. Also it would be great to see what it means for the fitness landscape to assume that epistasis is fixed or distributed. You may want to introduce sooner in the text the deleterious mutation rate (not explicitly mentioned before line 276).

We have tried to better organize the Results section (following the comments of reviewer 2). We have also added the new Figure S1 presenting the different components of epistasis and the new Figure 3 which summarizes the main results. We now introduce the deleterious mutation rate on line 130.

Reviewer 1's comments:

The approach is based on the QLE approximation, and is similar to Barton (1995), who looked at the much simpler problem of evolution of recombination with random mating. My main comment is that a bit more could be done to relate these results to the earlier analysis. Specifically:
- epistatic coefficients are assumed of order eps^2, relative to directional coefficients, which are of order eps. This is necessary if the variance components are to be of the same order, and arises because there are L^2 pairwise coefficients, but only L directional coefficients. This emerges in S1 but is not explained in the main text.

Our analysis doesn't assume that epistasis is much weaker than selection: in fact, if this was the case, additional terms would have to be included to describe the overall effect of epistasis on the evolution of selfing. This is now explained in the last paragraph of the discussion in response to another comment below. However, we argue that in this case the effects of epistasis should be negligible relative to the effects of inbreeding depression and purging. Under Gaussian stabilizing selection, directional selection and pairwise epistatic coefficients are of the same order; the benefit of selfing through purging scales with sbar U (eq. 39), while selection for selfing due to LD scales with U^2/n (eq. 50), which are not necessarily of different orders. We now precise that the a_U,V coefficients are assumed to be of the same order of magnitude in the general derivations (lines 162-164, 514-516).

- Deleterious mutations are assumed rare, but this assumption is not made consistently: sometimes, terms like (1-p*j) appear, which should be close to 1; conversely, epistatic terms necessarily involve p*j*p_k, which would seem to be second order. I think that the analysis is actually OK, but it needs a more careful explanation.

In the uniformly deleterious scenario we assumed that pj stays small, and neglect terms in pj^2 (this is now explicitly stated on lines 190-191, 317-318, 465). Terms in pj*pk are not neglected because they lead to terms in U^2 after summing over all pairs of loci (and these are not necessarily negligible relative to the terms in U obtained by summing terms in pj). In the stabilizing selection scenario we also assume that the deleterious allele (allele 1 if pj<0.5, allele 0 if pj>0.5) stays rare, by assuming that (1-2pj)^2 stays close to 1 (this is now precised on lines 225-227).

Specific points:
Eq 1 - Would random variation in selfing rate make any difference. I guess not, but it is not immediately obvious.

We have added a sentence saying that environmental variation in selfing rates shouldn't change our results (the variance in selfing rate should then by multiplied by the heritability of selfing), lines 119-122.

Eq 7 - Maybe this will come later, but there should be a comment on the consequences of setting up selfing rate as a quantitative trait, rather than looking at invasion of a specific allele. I suspect that the qualitative outcome is the same, but again, this is not obvious. (see line 220 also)

We have added a sentence confirming this (lines 113-119).

198 - This choice of a generalised function leads to higher-order epistasis, and so will require some approximation.

We now state that Q>2 leads to higher-order epistatic terms, which we do not compute (lines 239-241).

238 - Write l*sigma=10 here*

Done (line 280).

274 "the different terms in (19) contribute multiplicatively" is ambiguous (though I can guess what it might mean)

We have tried to better explain this in Supplementary File S2 (lines 23-36).

*276 - Worth commenting on why, under these assumptions, only the mean # of bad mutations (n*d) matters.

This is now explained on lines 324-328.

**279 - One should explain the four terms in (25), which should be possible. The crucial assumption that deleterious alleles are rare deserves more prominence. It is not clear that this assumption is used consistently, since one would expect second order terms to arise from a*j and a*j,j; the last two terms are necessarily second order. It may be that the n^2 factor makes the latter two p^2 terms more important than the neglected second-order contributions to the first two terms, but that needs attention.

We now give more explanation on these terms (lines 333-338), and explain that the effect of epistasis on aj, aj,j is taken into account (lines 340-342).

350 - In (33) it seems that p is no longer assumed small. Clarify when that assumption is used, and when not!

It is not, because it is a result from the general model (under stabilizing selection pj may be close to 1).

- In (36) do i, j still refer to loci for selfing & for fitness, respectively?

Yes, this is now explicitly mentioned (lines 470-471).

- The independence of (46) from Q is surprising. Why is this so?

We now give an explanation on lines 628-631.

472 - This statement may reveal a fundamental difficulty. Can the a of different order be assumed to be of the same order? (as it were; cf Barton 1995, and above)?

As we now explain on lines 514-516, the different a_U,V coefficients are assumed of the same order.

- Can one interpret the various components in terms of the change in fitness mean and variance due to recombination, etc? (see 648)

We now show that when epistasis is neglected, the strength of selection for selfing due to purging can be expressed in terms of the fitness increase following a single generation of purging by selfing (lines 449-462, 713-716). It is possible that the term of eq.43 can be related to the change in mean fitness due to recombination, but empirically it would probably be difficult to separate this term from the purging effect just described.

664 How does the advantage of outcrossing mediated by Hill-Robertson effects appear here? This involves terms of order a*j a*k a*jk which do not appear here; they may have been neglected because of different assumptions about the order of terms. This is quite confusing.*

The term in aj ak ajk (favoring recombination and thus outcrossing under negative epistasis) would appear if epistasis was assumed weak, but we argue that in this case the effect of epistasis should remain negligible (relative to the effects of inbreeding depression and purging), at least as long as the selfing rate is small. This is now discussed in the last paragraph of the Discussion.

*484 - Here and elsewhere, can one consistently include the small term a*j in the denominator, whilst neglecting corrections of this order elsewhere?

Adding this term allows one to obtain expressions that do not diverge as recombination tends to zero, and that can thus be integrated over a linear map (this is explained on lines 158-171 in Supplementary File S3).

636 - cite Otto on the importance of variance in epistasis across pairs of loci. (I think she made tis point in a n Annual Review with Feldman, arguing that it necessarily weakened selection for recombination)

Done (line 756).

- It may be unwise to rely on colour in the figures - this makes it hard for people to read them offline.

We would prefer to keep the colours as the figures may be difficult to read otherwise.

S1 - Is there a numerical check that these selection coefficients are correctly calculated?

It is not clear to us how to check the expressions for a_U,V coefficients; however we checked that one recovers the results of Abu Awad & Roze 2018 from these expressions (in Supplementary File S2).

S5 - The selfing rate is the average of the two alleles, yet in the text, it is defined as their sum.

In the simulation program with uniformly deleterious alleles the selfing rate is indeed given by the average of the two alleles at the modifier locus; however this should not affect the results regarding the stability of outcrossing and the equilibrium selfing rate.

Reviewer 2's comments

Organisation of methods and results

The authors present a general expression of the fitness of an individual, Eq. (8), followed by three specific fitness functions, Eqs. (9), (14) and (15) (respectively uniformly deleterious alleles, Gaussian stabilizing selection and non-Gaussian stabilizing selection). In the Results sections, (in particular Effects of epistasis on inbreeding depression, and Effects of epistasis on the evolution of selfing) the authors derive results for the general fitness case and specific fitness functions in turn, however jumping between these treatments can become confusing. I would therefore suggest the following.

• In Supplementary File 1, the analysis is split into four sections (the first untitled); “General fitness expression”, “Uniformly deleterious alleles”, “Gaussian stabilizing selection”, “Non-Gaussian stabilizing selection”. I found this section separation useful, and believe it should be replicated in all the appropriate methods/results sections.

We followed the suggestion of the reviewer.

• In line with my previous comment, it would be helpful to stick to a common naming convention for each of the fitness functions. Currently in the appendix, they have the titles above. In the section "General expression for fitness, and special cases" they have no explicit names. In the following text the first fitness function is variously called "unconditionally deleterious alleles with fixed epistasis", "uniformly deleterious alleles with fixed epistasis" and "the fixed epistasis model". On first reading it can be confusing as to whether these terms are referring to additional assumptions/conditions on the fitness function or not.

We tried to stick to the same denominations.

• It would be instructive to see the motivation for considering the different fitness functions stated clearly on their introduction. For instance, while Eq.(8) is general, it contains a large number of free parameters and it is clear that fitness functions that allow us to parametrise the model more simply are useful (although this could also be stated explicitly). However what is the significance of the different functions considered? Is it motivated primarily by the desire to relate these results to previous results, the demonstrate the generality of the insights derived, etc? Could any of these be understood as being more or less reasonable in a biological sense?

We tried to explain more the motivation for using these different fitness functions (lines 166-168, 194-196, 221-224, 235-239): in the uniformly deleterious alleles scenario there is no variance of epistasis, under Gaussian stabilizing selection there is a positive variance of additive-by-additive epistasis, but additive-by-dominance and dominance-by-dominance epistasis stay negligible, while in the non-Gaussian scenario all components of epistasis become of the same order (and are variable across pairs of loci).

• One of the key points throughout the paper is that the form of Eq.(8) is useful for deriving general results that disentangle the various forms of epistasis, but that each of the fitness functions can be "mapped" to the form of Eq.(8) by assuming that some parameters are small, conducting a Taylor expansion in these parameters and comparing prefactors of D_{U,V} to infer a_{U,V}. This point is clear in Supplementary File 1, but would also be straightforward (and useful) to explain in the main text.

We now say that on lines 165-168.

• Similarly to my third point, when introducing the section "Evolution of selfing in the absence of epistasis", a little motivation for where we're headed might be useful (e.g. "In the following section we will consider the effect of epistasis on the evolution of selfing. However first it is useful as a point of comparison to understand how the model dynamics would behave if epistasis were to be ignored").

We have added a new sentence saying this at the start of this section (lines 395-398).

Queries on analytical results

While the extensive supplementary materials do a great job of thoroughly explaining the calculation, there are a few areas in the main text that I feel could do with more discussion (or perhaps more explicit signposting to relevant sections of the supplementary files).

• I was unclear about what we were assuming about the deleterious mutation rate, U. On line

272, we assume that deleterious alleles stay rare in the population (which I assume holds when U is small) however on line 311 we see this transition where when U gets sufficiently large inbreeding depression moves from being increased to decreased by epistasis. In Figure 1 we have U \approx 0.25.

We now introduce U earlier (line 130) and explain that we assume that the per locus mutation rate u is small (we do not make any assumption on U).

• L272 "Equation 23 assumes ... that the different terms of equation 19 contribute multiplicatively to δ (which often yields better approximations than the additive which often yields better approximations than the additive expression)."
I wasn't sure what this sentence meant mathematically, and also unclear as to why the additive expression would "often" yield better approximations. What does "often" mean here? Is this a heuristic?

We now provide more explanation in Supplementary File S2 (lines 23-36).

• L388 "The prediction for the case of unlinked loci (which often yields better approximations than the additive obtained by setting ρ ij = 0.5 in equation 36) actually gives a closer match to the simulation results than the result obtained by integrating equation 36 over the genetic map. This may stem from the fact that equation 36 overestimates the effect of tightly linked loci."
Is it obvious that Eq.(36) would overestimates the effect of tightly linked loci? Is there any way of determining which of the approximations employed would be causing this?

It is not obvious, but we think that it probably stems from the QLE approximation (which assumes that changes in allele frequencies at loci controlling selfing are small relative to recombination rates), this is now explained on lines 478-481.

• L416-427 [Equation 39] "…, which increases as Q increases in most cases".
Is it clear that Eq.(39) typically increases with Q? How do I see this?

We now better explain this (lines 508-509).

• It's not immediately clear where Eqs. (10-12) come from – I'm sure an extra line or two could make this more obvious.

We have added more explanations (lines 183-186).

• L479-483 "The term in the second line of equation 47 shows that negative additive-bydominance or dominance-by-dominance epistasis between deleterious alleles increase the benefit of selfing, by increasing the efficiency of selection against deleterious alleles in homozygous individuals."
This is difficult for someone not very familiar with the model to read off directly from the equation. As such it probably warrants a longer breakdown of the meaning here (especially given that this is referred back to below Eq.(50)).

We have tried to better explain this equation (lines 564-572).

• Figures – Somewhere I'd like to know which assumptions used in the analytics are leading

to disagreement between the analytic theory and the results of simulations.

Although it is difficult to identify exactly the causes for these discrepancies, we now propose hypotheses (lines 364-366, 479-481, 602-605).

• In particular in Figure 1D (which should probably be plotted over a smaller range of \delta) it appears that the inbreeding in the simulations responds in the opposite direction to that predicted by the analytics as \beta is varied. Does this hold for all parameters? In the main text, the authors state that "Remarkably, the increased purging caused by negative epistasis almost exactly compensates the decreased fitness of homozygous offspring, so that inbreeding depression is only weakly affected by epistasis in this particular model, for the parameter values used in Figure 1". What do other parameter combinations do? Perhaps some additional testing of the robustness of the analytic results to changes in the parameters would be useful in the Supplementary Figures.

We prefer to keep the same range on fig 1D in order to better show the contrast with the other figures. We have added new results for different parameter values (figs S2-S4), showing that the same result holds.

• L368 – minus sign missing in front of \delta V_{\sigma}

Corrected

• L403 The fitness effect of a heterozygous mutation at locus j in an optimal genotype should be added to the parameter/variable list.

Done