We thank the AE and reviewer for the feedback. We have tried to improve in all the areas indicated, including adding extensive new analyses. Review comments in **bold**, our replies in plain text.

**Associate Editor**

**Comments to the Author:**

**Dear Tracey-**

**I've received one review of "Potential survival of some, but not all, diversification methods". I apologize to the authors for the long delay but this ms was only very recently transferred to me, and I thought it fairest to the authors to make a recommendation on the basis of the one review we have in hand. I also read the ms in detail.**

**On the whole, this manuscript makes some interesting points that will be helpful for the field and R1 also makes a similar observation.**

**That said, I have substantive reservations about 3 features of the manuscript. The first involves the target audience, the second a substantive scientific issue but a major claim of the ms, and the third is a place where I think the authors need to either expand a particular analysis or drop it from the ms. I recommend that we reject this version and invite the authors to submit a substantially revised version that addresses my points below (at least in part).**

**(1) Target audience: this is written very generally (a good thing), presumably towards people unfamiliar with the area and maybe not even with diversification models. The manuscript uses concepts that are very general (eg seismograph analogy) but at the same time uses concepts and assumptions that are in fact highly specialized and inaccessible to that same audience (e.g, casual mentions of congruence classes, pulled rates, the details underlying figure 3 etc).**

**I personally think some of the metaphors and examples (coin toss, seismographs) are missing the mark a little and that there is some actual "hard" scientific content here that would make this a more impactful contribution. For my part, I would prefer that - if they retain issues relating to my points (2) and (3) below - that they will need to tailor this a bit more to readers who are familiar in a general sense with this class of models. I have additional thoughts on language etc in the minor comments area.**

The AE is correct here in that we tried, as much as possible, to generalize the arguments for the broad readership at Evolution. This involved using metaphors to help provide readers, even those unfamiliar to the specifics of diversification analyses, to get a mental picture of the concepts we are discussing. As we state below, we’ve tried to minimize some of our uses of colorful language, in particular our use of the “seismograph” metaphor. Note we have retained it in the opening, because we feel it works in that particular context. However, we have removed it elsewhere. We have also provided explicit definitions of items like “pulled rates” and “congruent classes” as parentheticals. We apologize for this oversight in the original submission.

**(2) Claim that SSE models are robust: A major conclusion of this Perspective is that SSE models do not have the same problems as methods that rely on LTT models. I am not convinced that this is the case, even as I accept the authors' arguments that SSE models use more information and that there are \*some\* areas of parameter space where SSE-type models would do better. But on its own, the argument that SSE models use more information (L327) does not justify the statement that they are immune.**

**To illustrate, consider the following taxon-homogeneous (but possibly time-varying) models: (1) constant-rate birth-death, with high relative extinction (CRBD), versus (2) a zero-extinction but exponentially increasing speciation rate (ZEAS) that is LTT-indistinguishable from CRBD. These 2 models are something of a worst-case scenario under the Louca & Pennell (LP) scenario. A 3 parameter ZEAS model can give nearly identical likelihood to the 2 parameter CRBD model. Even a 2 parameter ZEAS can give similar likelihoods to CRBD.**

**So: LP would claim that you can't reject the ZEAS model because it's in the same congruence class and LTT-indistinguishable from CRBD. But presumably, MiSSE returns only something consistent with CRBD and - though I haven't tried MiSSE - I would be surprised if the confidence intervals on parameters truly spanned this parameter space. And put another way, if you simulate data under the ZEAS model that is LTT-indistinguishable from CRBD, I predict that you will get CRBD-like results from MiSSE.**

**Maybe I'm wrong, but this seems like the key point to address before arguing that SSE models are not sensitive to the LP issue. And if MiSSE gives CRBD-like results for a true ZEAS model (LTT-indistinguishable from CRBD with high relative extinction), then it seems not appropriate to claim that SSE models are immune to the LP issue. Keep in mind also that much of LP focused specifically on variation in rates through time under taxon-homogeneous scenarios (but time-varying) again questioning how SSE models would solve the issue.**

**Thus, I would need - minimally - to see this specific issue clearly addressed before I could conclude that SSE models are largely immune to LP-type non-identifiability. The simple case is to find the generating ZEAS models that are LTT identical to CRBD w/ high relative extinction, then show that MiSSE or equivalent SSE models return the right answer.**

We have added this analysis. MiSSE found very little evidence for heterogeneity across taxa and did ok estimating diversification rate, but overestimated both speciation and extinction rates.

**(3) Figure 3 analyses: I feel like this is an extremely important point, one that the field desperately needs (since we keep publishing analyses of temporal predictors of diversification, always finding significant correlations with temperature etc). My personal feeling, based on some exploration of this issue, is that this is another pre-HiSSE SSE scenario, where an awful lot of published literature is not going to hold up. That said, I think the current ms does not do justice to this point for 2 reasons.**

**First, it's a critical take on previously published research (and I strongly agree with the authors here about their interpretation in Fig 3). However, including a critical assessment of previous work as a minor analysis with no details (as in Fig 3 legend and text) seems a poor choice for presenting a criticism of work by other researchers. Second, and more importantly, this point is likely to be missed by the vast majority of people who need to see this result.**

We hasten to point out that we reached out to Condamine about our preprint and its connection to his work in November 2021 asking for feedback. We exchanged pleasantries, but received no feedback on the work. It doesn’t mean our criticisms are fair or well done, of course.

**Thus, I recommend either doing a proper analysis to show that these methods do indeed have exceedingly high false positive rates (e.g. detection of correlations with all sorts of temporally autocorrelated independent variables) and making it a major point, or eliminating it from the ms. I very much have a preference for the former as it's really important for the field to get this message (am I emphasizing enough how important I think this is?!). But I also think it's not good practice to embed "throwaway" criticisms of other researchers' work as a peripheral point, which is sort of how fig 3 comes across at the moment.**

We have now included an estimate (no doubt an underestimate) of the uncertainty in the reconstructed rates, by developing a procedure to iterate over many possible rates. We now indicate this with large black bars in Fig. 3. We think this drives home the point about the uncertainty (and it includes substantially more uncertainty at the tips than we expected). We’ve tried to soften the language a bit so it doesn’t come across a criticism of the Condamine et al. paper but instead a demonstration of the issues even good work will have problems with, hopefully scaring off others from doing this sort of analysis.

**### Other points**

**L26: thousands seems likely to be an overestimate**

We’ve modified this to say, “Hundreds, possibly thousands.” We do suspect, based on a cursory look on google scholar, that the number of empirical studies since Nee et al (1994) is substantial.

**L27-29: But this would be precisely the point that Louca and Pennell would dispute.**

We agree, but we are simply trying to summarize and introduce the problem as a means of appealing to the broad readership at *Evolution*. We’ve modified the text to be clear that if and only if the rates are constant that estimating birth and death rates separately is possible.

**L32-36: Not to be pedantic but I think researchers use these to reconstruct the history of diversity, not to reconstruct lines like the pen of a seismometer - this analogy seems to miss the mark to me as takes attention away from the actual goal. Put another way, the usual goal of an analogy is to cast a more obscure issue in more familiar terms. But probably more people who read this article are at least as familiar with the idea of diversity-diversification-etc through time as they are with the stuff that is plotted on a seismograph!**

We understand that the AE does not like this particular analogy. It was actually inspired by the Louca and Pennel (2020) study, where each simulation scenario they tried looked exactly like what you see from a seismograph. For the sake of appeasing the AE and the reviewer here, we have removed reference to “seismographs” throughout most of the manuscript. However, here we have chosen to retain to help readers have a visual picture of what many studies are trying to accomplish when estimating diversification rates through time.

**L40-42: I am not sure they call everything into question - even they suggest that recent speciation rates might be OK.**

In our experience at least, there is a general feeling among many friends, colleagues, practitioners, reviewers, and even Associate Editors, that estimating diversification rates from molecular phylogenies is no longer valid. This perception of “all is lost” comes largely from, in our view, Pagel’s unhelpful essay that accompanied the Nature paper. We’ve added this citation to this sentence.

**L42: Given that this ms is written very generally (target audience nonspecialists), I think it is premature to introduce "congruent models" without a clear definition/explanation.**

Fair point. We have added a sentence clarifying what is meant by congruent models.

**L53-58: I guess this gets to the heart of the issue but this aspect of LP is what is so maddening, because nearly all nested sets of statistical models would seem to have this same property if we aren't worried about complexity. E.g., I can prove that diametrically opposed topologies are consistent with any DNA sequence alignment (have identical likelihood) if I am free to slap whatever pattern of branch-specific rate variation I would like onto the tree. The "in some cases" qualifier, also not clearly described in LP (!!!), is a big, big qualifier.**

We agree. The frustrating thing about these sorts of models is that congruence seems fairly likely in this space, less so in tree inference.

**L106-138 etc. I guess this is OK, but honestly it doesn't really contextualize the issue for the general reader, in my opinion. Many more researchers have thought about speciation and extinction rates from phylogenies than about the problems of estimating trends from extant data (yes, trends are widely studied, but nearly always with paleo data). As such, it feels like me like it is trying to contextualize LP results using an example that is actually more obscure, and less intuitive, than just getting to the details of LP itself.**

We see the AE’s point here, but our motivation was more to emphasize other areas where these congruency issues exist. Especially with the ancestral state reconstruction aspect: no one would fit a trend BM model to coeval data, but our assumptions about ancestral states assume this model.

**L147-161: I admit that I did not find this example especially clear. And of course, wouldn't presetting the 5% lower or 20% be problematic since it uses information that would normally require 2 estimated parameters?**

We removed the constant slope models, which in practice have two free parameters. We now only focus on the models with a single parameter each.

**L164-167: This is actually a bit more complex, I think, because it mixes an ordered state model (sequence of flips) with the standard unordered model that students use to study coin toss experiments. Hence, these are not really analogous to destroying the concept of estimating the probability of heads. Especially so given that (hopefully) most students have also been exposed to principles of model complexity and would recognize that fitting a linear trend of p(heads) to the coin toss sequence would involve more parameters and thus would not, in fact, explain the data as well as the simple binomial with 1 parameter.**

The remaining models we include are all truly single parameter models. And they could be reasonable under some conditions: if we’re trying to predict the probability of not having fatal lung cancer by age, for example, a model where the rate is assumed to be 1 at birth and decreasing exponentially might be better than a binomial model with a constant rate through age.

**L211-213: Again, the seismograph analogy fails for me. Maybe this works better for others, but it takes me a lot of mental effort to squeeze inference about diversification onto the sort of output that comes from a seismograph.**

We have removed the seismograph analogy from this section.

**L213: An example of where I think the target audience is missed: you haven't clearly explained what the pulled rates are, so it toggles between an article that is written for very basic audience (coin flips, seismograph) with advanced concepts that are not explained (congruence, pulled rates).**

Good point. We have added a sentence clarifying the meaning of “pulled diversification” rates.

**L227: I get the MJ2015 analogy "curvature of biodiversity time" on a very superficial level, but what does this actually mean? I think it's not a good metaphor, to be honest. The concept in physics refers to the fact that space and time are a 4-dimensional entity that has certain non-identifiability properties in general relativity. This is not true for biodiversity-time, in any meaningful way to me. Biodiversity (however measured) has a specific association with any given reference time, and time is an independent variable in our non-relativistic frame of reference.**

We have rephrased and buried the citation rather than focus on it. We agree it’s not a great metaphor, but we think it’s important to cite here.

**L219, figure 2. I don't understand this figure from the legend itself. The last sentence seemed to me to be exactly the opposite of the point I thought I was supposed to take away (fewer events per given time bin early in time). And the figure is about binned events per 10 my time interval, so unclear also why it discusses single 10 my edges. The text discussion on L238-251 makes this much clearer, but based on the figure's position in the ms I read the figure legend first (and I'm still not sure I follow the wording in the figure legend).**

We have added clarification to the Figure legend. We thank the AE for pointing out the lack of clarity here.

**L270-294: These are very important points. I am not sure that the right place for them is to be buried in a general POV manuscript like this one. I am fairly confident that the authors findings in Figure 3 about the Condamine results (and many other articles that have used these methods) reflect a deeply problematic failure of those methods to adequately account for autocorrelation in either the independent variable or the rates themselves. Thus, secular trends in diversification also appear to correlate with all manners of temporally autocorrelated phenomena: the number of published studies now claiming that the Zachos delta-18 oxygen curve (temperature proxy) correlates with diversification is becoming quite large but I fully expect that most of this is the same as the simpsons effect in Figure 3.**

See response above.

**L336: But sister clade contrasts cannot reliably detect differences in net diversification rate. See Rabosky (2009, Ecology Letters, "Ecological limits and diversification rate: alternative paradigms") for discussion. They cannot distinguish between differences in net diversification, and differences in clade-specific carrying capacities. Whatever one's prior beliefs on the plausibility of those models, they are not distinguishable from these types of data alone.**

By “net diversification rate” we meant the overall net accumulation of species over time, not speciation-extinction per se. For example, BCO and JMB had the same savings at one point; if one of us has a million bucks more now, it doesn’t matter if it’s accumulated exponentially or was from a single burst and then constant. But we’ve rephrased so it’s clear we’re talking about comparing the net number of species on each side, regardless of whether a difference is due to different carrying capacities, speciation, extinction, mass extinction, etc.

**L333 etc: I am still struck by the fact that MiSSE gives very similar results to BAMM. Perhaps this should at least be mentioned given that the other methods are mentioned and that BAMM is the only one of that set to which MiSSE has actually been compared (following Vasconcelos et al 2021).**

Good point – we’ve added BAMM to the list of similar models to MiSSE and CLaDs..

**~Dan Rabosky**

**Associate Editor, Evolution**

**Reviewer(s)' Comments to Author:**

**Reviewer: 1**

**Comments to the Author**

**Among the papers responding to Louca & Pennell (2020), I think the interesting new points here are to encourage reporting uncertainties (in both parameter estimates and models) and showing how trait-associated diversification (SSE) models use tree information differently. I came away with an unclear impression of whether the authors' goal is to help people be comfortable with models having weaknesses and use them responsibly (earlier sections) or to advocate giving up on models (later sections).**

We have reworked the conclusion to better clarify our goals with this commentary. After reading through the manuscript carefully we realize that it was not quite clear at the end.

**Some specifics:**

**I wasn't quite sure what to take from the extended coin flip example. We should only consider "reasonable" models, but that isn't even possible for diversification questions? It did make me wonder, though, if we could observe a phylogeny for the same clade 10 million years later, would the LP model congruence problem vanish?**

See comment above.

**I puzzled over Fig 2 for a while. In the caption, saying "reconstructed branching events" at first would make it clearer what an "event" is. Making the rainbow semi-transparent, or just a small bar above the tree, would make it clearer the tree is continuing under it. Also, if the numbers at the bottom are My, there is no edge ending 300 Mya, so "3 vs 30 Mya" would seem more relevant.**

Yes, we understand that the figure legend is confusing. We have added clarity to it and we hope it is now easier to understand.

**Pages 11-13: Would there be any improvement by using longer time intervals deeper in the tree?**

There would be less imprecision due to having more data; other issues of rate estimation would still persist.

**Line 201: "Ancestral rate reconstruction" seems clever wrt "trait", but I don't really think of the rate-over-time "seisomgraph" as an ancestral reconstruction. If different lineages had different rate-over-time functions, that would seem to me more like an ancestral reconstruction, but such a model would perhaps not suffer all these problems because it is not "lineage homogeneous." Also an organizational issue: most of this section is about the somewhat different topic of reporting uncertainties in rate estimates and models.**

It is estimating rates that ancestors had – it’s unlike much ancestral state reconstruction in that it’s based on time slices rather than inheritance down the tree, but it’s just one extreme: things like BAMM or MiSSE allow reconstructions of diversification down the tree based on inheritance, and we’ve worked on models that have time slices and clade inheritance.

**If the authors really want to keep Fig 3, it would help to include the LTT plot so we can see what these functions fit to. And to set a good example, the rate estimates over time should include uncertainties. But the yellow and red lines don't seem so relevant after the earlier discussion about how it's useful if models are a priori well-motivated.**

Good point about uncertainty. We added uncertainty to each of the lines depicted. As for the discussion about well-motivated models, we see the reviewer’s point, but the lines reinforce the point that models that are more complex will almost provide at least a decent fit our messy, complex empirical data, whether or not they actually produced the data.

**Fig 4: The LTT plot is defined to be the same for all trees, so it would be more informative to use that space to better show the likelihood surfaces for each tree (the tiny insets, I guess). Also, perhaps the trees could be plotted differently (node ordering? not circular?) to make the balance differences visually apparent. The red-blue colors in the lower panels are not explained.**

We have added clarity to the figure legend regarding the meaning of the red-blue colors.

**"What are we really learning anyway?" seems like a great thoughtful place to wrap up. But rather than making me reflect on important questions and how they should be approached (or not approached), this section just seemed to advocate not using either simple or complex models. The authors' main point may be in the Conclusion: "We can certainly learn about diversification processes from trees, but we need to recognize that what we can understand largely relates only to the surviving tips." How about discussing what those latter questions are, and what would be productive ways to go about answering them?**

We thank the reviewer for pointing out the lack of clarity here. We have substantially revised the conclusion to clarify our position.