**Editor's Decision MAJOR REVISIONS**

I have now received three comprehensive reviews on your interesting manuscript. All of them acknowledge that this is a well-written piece that adds a useful tool to the PCM literature.  
  
You'll see that the comments revolve around clarifying a few aspects of the writing - especially the Discussion -, but also come suggestions on how to improve the simulations. I tend to agree with them that the Discussion, especially its first half, needs to be heavily reformulated. R2 also has important suggestions on how to estimate analytically the parameter alpha from the data, instead of numerically using insights from previous work. All reviewers point to the need to reformulate Fig. 3.  
  
I look forward to reading the revised version of the manuscript.  
  
[# PeerJ Staff Note: It is PeerJ policy that additional references suggested during the peer-review process should \*only\* be included if the authors are in agreement that they are relevant and useful #]

**Comments from the reviewers**

Reviewer: James Boyko

**Basic reporting**

The text is clear. All figures and tables are easily understandable and add to the quality of the text. Raw data and code is provided and accessible. I have two points to consider on the topic of literature references, both of which are minor and occur in the discussion.  
  
1) Line 368. “...we normally assume that this process is constant over all the branches and nodes of our reconstructed tree...” I don’t think this statement is fair or necessarily true. Hidden Markov models (HMMs) have seen widespread use in discrete character models and PCMs more broadly (e.g., SSE models). For example, work related to one groups implementation of HMMs (Beaulieu et al. 2013, Beaulieu and O’Meara 2016, Boyko and Beaulieu 2021) have amounted to over 1000 citations since 2013 (with half being exclusively to discrete character models without diversification dynamics). Furthermore, even Pagel’s correlated models can be considered as a version of rate heterogeneous models (as discussed in Boyko and Beaulieu 2022 in the context of leading to high false positive rates). In my opinion, this suggests widespread use of rate heterogeneity for the evolutionary modeling of discrete characters. I would suggest being clear that our "null" hypothesis is one of a constant process, though I'm not sure this suggestion would capture the original sentiment of the sentence.

**The reviewer makes a reasonable point. I would argue that a constant-rate M*k* model is the overwhelmingly *predominant* way that discrete character evolution is studied, particularly if you consider that fitting a character model is often just the first step on our way to other types of analyses, such as marginal ancestral state estimation or stochastic mapping. That said, it was not my intention to give short shrift to hidden Markov models, which the reviewer knows I consider to be an important category of method. I have now adjusted the text in a couple of different places (including in the highlighted section) to reflect this general comment re: HMMs.**

**A second, related point by the author regards Pagel’s (1994) correlated model “as a version of rate heterogeneous models.” I agree wholeheartedly, and this perspective was already included in the introduction. Specifically, on lines 65-69 (of the previous submission) we wrote “*The Mk model has already been extended in a number of interesting ways to permit rate heterogeneity over time, among clades, or as a function of another trait that has been mapped along the nodes and edges of the phylogeny. For instance, Pagel (1994) described a clever manner in which an Mk-type model can be used to approximate the interdependent evolution of two different binary traits, in which the condition of character one influence the transition rates of character two and/or vice versa.*”**  
  
2) Line 371. I would remove “totally new” from this sentence as the discrete gamma distributed Mk model has been the standard within phylogenetics for quite some time.

**The reviewer makes a good point, but there’s actually an even more closely related study that makes our contribution even less “totally new” that we were not aware of when writing up the draft of this manuscript, which is Wu et al. (2008; doi:10.1016/j.ympev.2007.06.020). Wu et al. (2008) describe a model for nucleotide sequence rate variation both among sites (according to a Γ distribution), but *also* among edges. Wu et al. (2008) focus only on phylogeny & divergence time estimation, but it is very clearly related to the model & method of our article which is for phenotypic traits on a fixed tree. We have now cited Wu et al. (2008) and pointed out the similarity of the two different models, where appropriate.**  
  
Pagel, M., 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London. Series B: Biological Sciences 255, 37–45. <https://doi.org/10.1098/rspb.1994.0006>  
Beaulieu, J.M., O’Meara, B.C., Donoghue, M.J., 2013. Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. Systematic Biology 62, 725–737. <https://doi.org/10.1093/sysbio/syt034>  
Beaulieu, J.M., O’Meara, B.C., 2016. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. Syst Biol 65, 583–601. <https://doi.org/10.1093/sysbio/syw022>  
Boyko, J.D., Beaulieu, J.M., 2021. Generalized hidden Markov models for phylogenetic comparative datasets. Methods in Ecology and Evolution 12, 468–478. <https://doi.org/10.1111/2041-210X.13534>  
Boyko, J.D., Beaulieu, J.M., 2022. Reducing the Biases in False Correlations Between Discrete Characters. Systematic Biology syac066. <https://doi.org/10.1093/sysbio/syac066>

**Experimental design**

The article is within the aims and scope of the journal. The research question is well defined and certainly fills a knowledge gap. I believe this gamma distributed discrete model will see widespread use and is an exciting addition to the PCM toolkit. A rigorous simulation study is conducted and methods are clear.  
  
Despite an overall excellent simulation design, I have some minor points that I'd like to see addressed.  
  
1) Line 217. You state that the alpha estimate is upwardly biased due to optimizations hitting the upper limit during optimization. However, it is clear from Figure 3 that the median value is actually downwardly biased for most simulations.

**This is fair. In the original submitted manuscript, we did only 20 simulation per simulation condition. We have now increased this to 100. We also give the geometric mean in addition to the median, quartiles, range, and outliers. As a general observation about our findings, median and geometric mean alpha showed very little bias when generating alpha was less than or equal to 1 or 2. For example, median estimated alpha across 100 (rather than 20, but using otherwise identical conditions) simulations per condition was 0.23, 0.47, 0.96, and 1.34 for generating alpha of 0.25, 0.5, 1, and 2, respectively. Geometric mean estimated alpha under the same conditions was 0.23, 0.48, 1.03, and 2.24. For the highest generating values of alpha (4 and 8), the median estimated alpha had relatively low bias (4.82 and 7.01, respectively), while geometric mean alpha was off the charts (e.g., 85.7 and 183.4, but would’ve been higher still if a higher bound for the upper limit of optimization was chosen). We don’t find this to be particularly concerning because very high alpha corresponds to low or no rate heterogeneity across edges, which will result in an essentially flat likelihood surface, equal to the likelihood of the homogeneous Mk model, across a broad range of alpha values.**  
  
If you remove optimizations that hit the upper bound (which is a reasonable step to take considering in most empirical studies, those model fits would not have sufficient support to continue with) is your estimate still upwardly biased?

**We’re not sure that we agree this is a reasonable step and I’m not quite sure what the reviewer is getting at. The upper bound for optimization is arbitrary – it could’ve been 10 or 100 or 1,000 or 10,000 or more – and if we have one or more simulations in which our optimizations that hit this bound (and thus the estimate is set to the upper bound), then our (geometric) mean across simulations and estimates is going to be downwardly biased relative to the true mean, simply by virtue of the fact that the true ML solution is above the upper bound! If we excise the estimate from our data, this downward bias would increase, not decrease.**

I mention this because an upward bias is, in my opinion, a more desirable outcome than a downward bias since the upward bias results in more conservative modeling results. It is concerning to me that the median values of alpha=2,4,8 all have a similar median value and that their upward bias seems driven by outliers.

**The similarity of the median estimates for generating alpha greater than or equal to 2 seems to have been an artifact of only doing 20 simulations per simulation condition. When we re-did these simulations, but with 100, the pattern goes away and the median increases more consistently with alpha. We agree in some ways that an upward bias is “more desirable” than a downward bias. In general, we found no or minimal bias for alpha <= 1 or 2, and then a downward bias in the median but not the geometric or arithmetic mean for generating values of alpha >= 2. In spite of this evident “downward” bias (of the median), only 7 of 100 Gamma models fit significantly better than a homogeneous Mk model for alpha = 4, and only 3 of 100 for alpha = 8. (This on 2,000 taxon simulated trees!) This does not, in our opinion, suggest that the method is overly liberal and tend to reject a homogeneous process when rate heterogeneity is, in fact, low. Hopefully this helps to assuage the reviewer’s reasonable concern.**

2) Line 228. I believe that simulating under values of alpha other than 0.5 would be interesting from a power perspective. You’ve already demonstrated that estimation is good when analyzing relatively large number of taxa for alpha=0.5 (section 2.2). However, the more problematic values of alpha are when rate heterogeneity is weaker (Fig 3). This is what we had found in Boyko and Beaulieu (2021) when conducting a similar simulation study in the context of hidden Markov models. Specifically, we saw that the signal for rate heterogeneity comes from the magnitude of the difference between rate classes. Of course, the gamma distributed Mk model does not have explicit rate classes in the way an HMM does, but the results presented here are consistent with what we had seen. Given this is known to be a problematic region of parameter space, it is worth running power simulations at other levels of alpha to determine power as a function of signal (alpha) and data points (number of taxa). I believe that this may provide useful empirical guidelines for how many taxa are needed to consider looking for rate heterogeneity. It also seems like it would be relatively straightforward to generate these additional power results and to add them to Figure 4 as separate boxplots and lines.

**This is a reasonable suggestion. In our previous manuscript version, we undertook a power analysis for alpha = 0.5. Here, we have re-did this but setting alpha = 0.25 and alpha = 2. (alpha > 2 is basically rate homogeneity, so this seemed less interesting.)**

3) Line 274. This is a very minor point and I leave it to the authors discretion as addressing this point may serve to add confusion for most readers. Nonetheless, I had a bit of trouble understanding your metric upon first reading. I think it could be beneficial to call this the dot product in addition to your description. The dot product also conducts an element wise product and summation as you describe, but does not divide by the number of states. However, this division is unnecessary in the case of known ancestral states since the only element from the estimate contributing to the score is the state which matches the simulated ancestral state. E.g., if you have c(1,0,0) as the simulated value and c(0.33, 0.33, 0.34) as the estimated ancestral state, the dot product of these vectors will be 0.33 (same as the average described in the text). I agree with the authors that discerning between uncertainly wrong and certainly wrong estimates are not particularly relevant to their study.

**Validity of the findings**

The impact and novelty is clearly stated. The data and statistical tests are robust. Conclusions are clearly stated, but, in my opinion, the discussion tends to focus too much on repeating the findings of the study and doesn't necessarily expand on how this is different from existing implementations of rate heterogeneity such as HMMs. For example, one reason I'm excited about the gamma Mk introduced here is that it has fewer parameters than HMMs which may make it more applicable to small datasets.  
  
Table 1. Is it straightforward to include confidence intervals or a measure of uncertainty around alpha? If so, I believe that should be included in the table and perhaps briefly described in the text.

**Additional comments**

Overall, I believe that this is an excellent paper and the method introduced is an important expansion of the PCM toolkit. I'm excited to see this used in other papers and any potential expansions of this work. The paper itself is written to be clear and concise, with mathematical details outlined in an impressively approachable way.

Reviewer: Michael May

**Basic reporting**

The paper is well written in approachable and accessible language, building up from a basic model of constant rates to the more complicated model of Gamma-distributed rates. For the most part, the authors cite appropriate literature, though I have a comment about the language used to refer to Mk models, as well as a comment about the relationship between the authors' model and what is known as a "no common mechanism" model in molecular phylogenetics. These comments are rather technical, and I expound on them below.

**Experimental design**

While not theoretically groundbreaking (much of the idea derives from long-standing tools in molecular phylogenetics), my opinion is that the method is a very solid and valuable contribution to the comparative phylogenetic toolkit---I can certainly imagine recommending it to students and colleagues. The authors perform simulations to show that the method has high power to reject a constant-rate model when the true model is that edge-rates are Gamma distributed, and that failing to model rate heterogeneity worsens estimates of ancestral states. The simulations are well described at a level necessary to understand and recreate them, and I am generally convinced that the method behaves well. However, I think it would be valuable for the authors to: 1) characterize the false positive rate to accompany their power analysis, and 2) assess the performance of branch-rate estimates.  
  
## False positive rates  
  
It might be valuable to accompany the power analysis (section 2.3) with a false positive analysis. A high-powered method may just be too eager to reject the null model, so it is pretty normal to pair a power analysis with a false positive analysis. I think the authors should simulate some data similiar to section 2.3 but when the rates are constant, and check how often they reject the constant model. I am not particularly suspicious that the method will misbehave: I've played around with the implementation a bit (not at the scale of a simulation study) and it does seem to reject the Gamma model almost always when the constant-rate model is true. However, users will not know that (or know to consider it), and will not make the effort to check whether the method has sensible rejection behavior when the constant-rate model is true.

## Branch-rate estimates  
  
The authors present estimates of branch-specific rates for their empirical case study. Of course, I think branch-specific rates are something that users will be very interested in estimating. However, the authors have not actually assessed the accuracy of branch-specific rate estimates. Because of the relationship to the parsimony model (described in the NCM comment, below), I am somewhat concerned that branches will basically be in a high rate category when there is a parsimony change on the branch, a low category when there is no parsimony change, and an intermediate category when there are equally parsimonious reconstructions that place the changes on different branches.  
  
This intuition of course is based on thinking about unordered CTMC models, where the parsimony reconstruction will imply either 0 or 1 changes on a branch. It might be different for an ordered model, since parsimony might still be able to reconstruct multiple changes on a branch. However, while the authors' simulations and empirical data analysis are for ordered models, I assume they don't intend for the method to only be used with ordered models?  
  
Another concern I have is that branch-rate estimates may have a huge amount of uncertainty because there is only one character, and figures like figure 6 could be hiding a lot of uncertainty.  
  
Given that branch-rate estimates are at least presented as a utility of this method, I think it would make sense to assess the performance of those estimates. I would like the authors to show how well they work, both in terms of accuracy and variance. Additionally, I think it would be nice (but not required) to know whether branch-specific rate estimates are just a function of the parsimony number of changes on the branch.

**Validity of the findings**

I have no reason to doubt the validity of any of the findings. The supplementary repository includes all the data and code required to recreate all of the results in the study.

**Additional comments**

Revell and Harmon describe a new method for allowing rates of discrete-trait evolution to vary across edges in a phylogeny. Specifically, they allow the rate of each branch to be drawn from a Gamma distribution with a shape parameter that is estimated from the data. I believe is a strong relationship between this model and what is known as the "no common mechanism" model in molecular phylogenetics, and therefore indirectly related to parsimony. This is not a problem per se, but there are actually properties of this model that the authors may wish to exploit/discuss. Additionally, I am somewhat concerned about the authors' implementation of ancestral state reconstruction. I go into more detail below, then end with minor additional comments and typographical comments.  
  
## Gamma distributed rates and the No Common Mechanism model.  
  
I think the authors can compute transition probabilities analytically rather than numerically, which would lead to much faster and more accurate results.  
  
The model the authors propose is very similar to what is called the "No Common Mechanism" (NCM) model. This model was first introduced by Tuffley and Steel (1999) in a maximum-likelihood context, but was later investigated in a Bayesian context by Huelsenbeck et al. (2008), "A Bayesian Perspective on a Non-parsimonious Parsimony Model". A feature of the NCM is that each site in an alignment has a separate set of branch lengths (but a shared tree topology), which is not particularly relevant to the present method because there is just one character. However, what is relevant is that Huelsenbeck et al. show that transition probabilities can be calculated analytically when the prior on branch lengths is Gamma. Specifically, for a Gamma distribution with α = β, the transition probability matrix is:  
  
P(Q, alpha) = (I - Q / alpha) ^ -alpha  
  
This is used in in the context of unrooted phylogenetic trees so there is no rate and time distinction, so the thing that is Gamma distributed is the product of rate and time. However, if time is known (as in the authors' approach) and the rate is gamma distributed, Q can be rescaled by the edge length t straightforwardly:  
  
P(Q, t, alpha) = (I - t \* Q / alpha) ^ -alpha  
  
In other words, the transition probability can be computed analytically, rather than numerically enumerating rate categories and summing. Some corresponding R code is:  
  
computeTransitionProb <- function(Q, t, alpha) {  
  
I <- diag(1, nrow(Q))  
S <- (I - t \* Q / alpha)  
e <- eigen(S)  
P <- e$vectors %\*% diag(e$values^(-alpha)) %\*% t(e$vectors)  
  
return(P)  
  
}  
  
(This could be written more efficiently by storing an eigen decomposition of Q and performing all the operations on the eigen values, ie e$vectors %\*% [1 - t \* e$values / alpha]^(-alpha) %\*% t(e$vectors).)  
  
I think the authors could potentially (if they wanted) speed up their method by an order of magnitude (or more) and get exact solutions rather than numerical approximations. A downside of the exact solution is that it is not straightforward to reconstruct the marginal distribution of rates on a branch analytically. Perhaps the authors could adopt a hybrid approach: when computing likelihoods, do full analytical calculations; when estimating branch rates, discretize the rates for the focal branch but use analytical calculations for the remaining branches. This way they could still get empirical Bayesian estimates of the branch rates without expensive numerical computations on the entire tree. My understanding is that phytools is not intended to be the fastest program in the world. I am sympathetic to that, and will be perfectly happy if the authors choose not to adopt this recommendation.  
  
Another relevant property of the NCM is that it effectively behaves like a model-based parsimony method. That is, it tends to infer trees and state reconstructions that are equivalent to the corresponding parsimony reconstructions. The reason this happens for the Bayesian NCM is shown in the appendix of Huelsenbeck et al. These results may not be perfectly applicable, because the NCM doesn't include a branch length t (separate from the branch rate r), and because the results are specifically for a Jukes-Cantor model. However, I suspect it is possible that when the authors' Gamma model is favored over a constant-rate model, the ancestral states will be very similar to the parsimony reconstruction. I'm not suggesting that this is a good or bad thing: it is just a property of the model! Nonetheless, it might be worth a mention or small investigation.  
  
## Ancestral states  
  
Regarding ancestral-state estimates, the authors say that (line 268): "For each fitted Mk and Γ model of each simulated tree and dataset, we undertook marginal ancestral state estimation following Pagel (1999; Yang 2006)." Pagel 1999 proposes two alternative approaches for estimating ancestral states ("local" and "global"), so "following" that paper is a bit ambiguous. Additionally, Pagel and Yang actually have different perspectives on how to estimate ancestral states. The procedure the authors describe sounds more consistent with Yang (Pagel's the "global" estimate), which involves first estimating the parameters of the model marginalizing over all states at all nodes, then conditioning on those parameters to estimate the (marginal) distribution of states at a given node. Pagel prefers what he calls the "local" estimate, which involves fixing the state at a node and then estimating the model parameters, and comparing the resulting likelihoods. (Both methods have joint counterparts.)  
  
To understand what the authors were actually doing, I looked into the phytools code. I believe they are doing something more like Yang, because they first fit model parameters then estimate states. However, I think there is a deeper problem with the authors' estimation procedure. According to both Pagel and Yang, the marginal distribution of the state at a given node depends not only on the data that derives from that node, but all of the data at the tips of the tree. This is consistent with the authors' statement that (line 270) "at each node computing the probability of the data given each state of the character, conditioning on the tree and fitted character model, but integrating across all possible states at all other nodes of the tree", as long as we interpret "probability of the data" meaning the probability of "all of the data". However, what appears to be implemented in phytools:::marginal\_asr\_gamma is different: this procedure just uses the state likelihoods at a node (which are only the probability of the data descending from that node) to compute the state probabilities, effectively discarding the information from all the other tips in the tree. The result is not the marginal posterior probability of states, strictly speaking. Later in the paper, the authors refer to ancestral states as (line 326) "scaled likelihoods (ancestral states)", which seems consistent with what is implemented but is hard to understand theoretically.  
  
Including information from all the tips involves more than a single execution of the pruning algorithm. Yang proposes a re-rooting method that applies for time-reversible CTCMs (which the authors are using), but this method assumes the process is stationary at the "root" (which is violated by the Fitzjohn 'nuisance' root frequency). Of course, it's possible that I am misinterpreting the phytools function here. Can the authors clarifying what theory and algorithms they are using to estimate ancestral states?  
  
## Minor comments  
  
- line 46: "the Mk model has now been elaborated to permit an arbitrary number of different rates of character transition between levels"  
  
My understanding of Lewis 2001 is that he \_defines\_ the Mk model as a generalization of the Jukes Cantor model to accommodate an arbitary number of states. Specifically, he says (page 915): "The Mk model is a generalized JC69 model, the latter representing the special case of k = 4 (the JC69 model could thus be referred to as the M4 model)". The authors use the term much more generically, to allow essentially all forms of the rate matrix. I think makes Mk essentially redundant with "continuous-time Markov chain", so I wonder why we would even need the Mk designation.  
  
- line 81: "In the present article, we’ll describe a new extension of the Mk model"  
  
I think most phylogeneticists conventionally separate the form of the rate matrix (in this case, Mk) from the rate of evolution/branch length. For this reason, I don't actually perceive this method to be an extension of the Mk model, but rather a relaxed morphological clock model.  
  
- line 82: "from a (normalized) Γ distribution"  
  
The use of "normalized" here is a bit confusing. First, it is not explained in this paragraph what the normalization is or why it is necessary, so I think it could actually be omitted here and brought up later as a technical detail. Second, I think the usual interpretation of "normalized" with regard to probability distributions is that they integrate to 1, which is always true for the Gamma distribution. My understanding is that "normalized" here actually refers to the fact that the mean of the distribution is constrained to be 1. Third, the authors do not describe \_why\_ the distribution needs to have a mean of 1.  
  
- line 127: "We then subdivide the range of our relative rates, r (which could theoretically vary from 0 to ∞, the range of a Γ-distributed random variable), into m evenly spaced bins"  
  
I suppose evenly spaced here means the bins have equal probability inside of them? This is implied by the 1/m weighting in equation (4), but it is a bit confusing because the rates (breakpoints/midpoints) are not evenly spaced.  
  
- line 151: "As we can compute the probability of any data pattern on the tree for a given shape parameter (α), number of rate categories (m), and transition matrix Q, we can likewise find the set of such values that maximize this probability – in other words, the Maximum Likelihood (ML) estimate."  
  
This makes it sound like m is something to estimated? But later when the authors mention Bayesian inference, they talk about priors on α and Q, and in all of their data analyses, they assume a fixed value of m. Personally, I don't think m is a parameter in the statistical sense, it's just a numerical detail of a numerical integration step, and I think the authors would agree with this perspective. I think it would be nice if the authors clarify this point, maybe being explicit that as m -> ∞, the discrete approximation converges to the integral of the continuous Gamma distribution.  
  
- line 162: "Low values of α correspond to high edge-wise rate heterogeneity, whereas high values of α, relatively little rate variation from branch to branch of the tree"  
  
It is a bit odd that this bit of information does not come earlier, for example when the Gamma distribution is introduced in paragraph starting line ~98.  
  
- line 210 (and throughout): "total of ten optimization iterations"  
  
I'd recommend using a word other than "iteration" here, which might be confused for the number of optimization steps you take in a given optimization run. Independent optimization runs are not sequential, so it is confusing to describe them as iterations.  
  
- figure 3: It's hard to judge how many simulations are hitting the upper bound because all dots have the same x-coordinate. Consider using a stripchart/jittering outliers?  
  
- line 225: "none of 20 simulation replicates met ∆log-likelihood criterion for statistical significance at the 0.05 level"  
  
What is the statistical test? I assume it's a likelihood-ratio test, as described in the subsequent simulation study.  
  
- line 239: "We then computed a P-value based on a likelihood-ratio test with one degree of freedom: the difference in the number of estimated parameters between our two models"  
  
The likelihood ratio test assumes the simple model corresponds to the general model with a parameter value \_inside\_ of parameter space. In this case, the simple model is actually on a boundary (α -> ∞), so the likelihood ratio test is not strictly correct. The likelihood ratio is distributed as a mixture of chi square distributions, one with 0 degrees of freedom and one with 1 degree of freedom. See Goldman and Whelan 2001 "Statistical Tests of Gamma-Distributed Rate Heterogeneity in Models of Sequence Evolution in Phylogenetics" for more details.  
  
- line 378: "Conversely, as α -> 1 / ∞ (but in practice for α >> 1 / ∞) the Γ model of this paper converges on a two-rate process in which some branches evolve with a very high rate, and most not at all."  
  
Convergence has many technical meanings and I find this statement unconvincing. Intuitively, when ∞ is very very small, there will be lots of branches with effectively zero rate, but the branches that have effectively non-zero rate will have rates very different from each other. I would not describe this as a two-rate process. (As an aside, what's the difference between α >> 1 / ∞ and α > 1 / ∞)?  
  
## Typographical comments  
  
- line 33: "This model might take a variety of flavors, all with its own parameters and attributes"  
  
All with their own?  
  
- line 58: "Imagine that a scientist studying this trait is interested in ancestral character conditions in the clade and fits a homogeneous rate model: since this is (by far) the most common assumption of ancestral character evolution"  
  
I'm confused by this colon.  
  
- line 173: "Finally, third, "  
  
It feels a bit awkward to have two transitional adverbs/enumarators here.  
  
- line 371: "totally new model"  
  
Is the model totally new? I think the application is novel, but the model itself is just a relaxed clock, which have a long history phylogenetics (see comments above re: NCM model).

Reviewer: William Gearty

**Basic reporting**

The overall reporting of this manuscript was excellent. The text was clear and unambiguous (except for two places mentioned below). The manuscript has a great introduction that includes a well-rounded summary of existing, related methods and the biological meaning and necessity for the proposed model. I thought the structure of the manuscript worked well, with relevant results within each individual section rather than all together at the end. The figures were well composed and clear. I only have two minor comments:  
  
1. On lines 219-220, the authors note that the pattern was most noteworthy for alpha = 8. However, upon looking at Figure 3A, it appears that the upward bias is most pronounced for alpha = 2, with a higher quartile/whisker and similar outliers. It is possible that there are more outliers with values of 1000 for alpha = 8, but this is not apparent in the figure. I would suggest that the authors modify the figure in a way that better illustrates what is currently written in the text (which I assume is accurate and should therefore be preserved) or modify the text to better reflect what can be seen in the figure (which I assume is misleading and should be changed).  
  
2. On lines 225-227, the authors state that "none of the 20 simulation replicates" cross the statistical significance threshold of 0.05 when alpha = 8. However, Figure 3B appears to show several outliers that cross the threshold for alpha = 8. I imagine that I am just not understanding the text (perhaps the authors mean that there are fewer than 20 replicates that cross the threshold?). I would therefore suggest that the authors reword this sentence to make it clearer what the point is, especially given the values shown in Figure 3B.

**Experimental design**

I thought the experimental design of the study was well defined and meaningful. It is very clear how this method fills a methodological and knowledge gap, and the authors clearly demonstrate that it performs fairly well. The authors have included sufficient detail about the methods and have also included code in a supplementary repository in case readers would like to replicate the simulations and analyses presented within the text. I appreciate that the authors can not simulate and test all permutations of the variables of interest, but I do have one request. Often in phylogenetic comparative methods, there is an observation/concern that overparameterized models better fit data simulated/generated under a simpler model because of the inherent noise of these types of models. Therefore, I would encourage the authors to develop a complementary set of simulations to those in section 2.3 in which the data are simulated under a rate homogeneous model and then both model types are fit to the simulated data to formally test the false positive rate of the new rate-heterogeneous model.

**Validity of the findings**

All findings appear to be robust and statistically sound. All code and data have been provided in a GitHub repository. The conclusions are clearly stated, with well-reasoned and supported statistical and biological insights.