Reviewer 1

This analysis, summarized in Fig 1 is intriguing, non-trivial, and to me convincing—though I am not as convinced that the subsequent analysis fully gets to the bottom of where this pattern comes from. I have some questions and comments which I hope are useful.

> *We very much appreciate the reviewer's help in improving our manuscript.*

Overall, I think the manuscript would benefit from:

- a clear description of the model assumptions

> *We do this now more clearly in the introduction first in lines XXX and again re-affirm our assumptions in line XXX (around where we cite mayr1965systZool)*

- some more discussion of possible explanations for the observed patterns

> *XXXX*

- a clearer unwrapping of the proposed connections to punctuated equilibrium (and other evolutionary hypotheses discussed).

> *Following the suggestions of Reviewer 3, we have removed "punctuated equilibrium" as a literal explanation for our results. This seminal concept in the field is still relevant, however, and we expand our discussion of it, as well as other concepts such as the Red Queen, in the discussion in lines XXXXX*

If I understand correctly, the authors assume that all orders undergo Gaussian fluctuations around an equilibrium size. While the evidence for this presented in Fig 1 is convincing, do the authors rule out other distributions for the way orders change through time? For example, is it clear that these orders do not undergo net diversification or extinction over time? I did not see this ruled out in preference for the equilibrium hypothesis, and yet it seems like it would be reasonable to include a term for net diversification or extinction over these timespans. If the authors did allow for this, would the equilibrium model still be favored, and would this change any of the subsequent analysis?

I suppose another way to ask the question would be: what happens if the authors do try to fit independent birth and death rates to the time series for each order?

> *The reviewer is indeed correct that net diversification or extinction could be considered. In our analysis we explicitly did not consider either, and did so by zero-centering all timeseries. This means we focus on fluctuations about any possible trend toward net diversification or extinction. This zero-centering is made clear in the methods, and also introduced earlier (lines XXXX; near 'To focus attention on the variance') to help readers understand our approach. In the same section we define "equilibrial" in the statistical sense and make that will be our working definition*

Related, am I understanding that the authors use the fit of beta_k for an order as synonymous with the diversification rate of that order? For example, discussion on p11:

"A Gamma distribution, while consistent with multiple processes (e.g., (40)), could result from evolution of diversification rates across an adaptive landscape that promotes niche conservatism and punctuated exploration of niche space."

This might be confusing, because readers may think of diversification rate as being a net diversification rate, which here is assumed to be zero, IIUC.

> *Thanks for pointing out this inconsistency, you are indeed correct that this could confuse readers. The β_k should be interpreted as the inverse variance in diversification as defined as (originations - extinctions). The variance of the difference between these two random variables will increase as* both *speciation and extinction rates increase. We have replaced "rate" with "volatility," where appropriate—including in the title—to signify the variance of origination - extinction, and clearly defined volatility on line XXXX (where we ref suppLimitDist). We have further updated our discussion on XXXXXX to more accurately reflect the idea that evolution across an adaptive landscape can change the volatility of diversification by changing the magnitude of speciation and extinction rates.*

I would assume it matters when plotting fluctuation sizes in e.g. Fig 1 what is the timescale between measurements. There is some discussion of this in the SI, but I'd like to understand if the variation in timeslices is enough to affect these analyses.

> *We have made clear earlier on XXXX (where we say '49 standard timebins') that we use 11 MY time bins as used previously in the literature (e.g. Alroy 2008 and 2010).*

Do the authors acknowledge that there is any room for bias in the classification of orders? Mayr's principle of balance has been invoked before to explain power law patterns in evolutionary history (e.g. in the number of species per genus, etc). I found myself wondering if this might influence where this Gamma distribution for the fitted beta's comes from. For example, I would like to have seen if there is a clear relationship between fitted beta and the mean order size. Is it possible that the distribution of mean order sizes underlies the distribution in order fluctuation sizes?

> *We were also concerned with the potential arbitrariness of taxonomies. We add discussion of this in line XXX (around yule1925). This is also the motivation for our permutation experiment in which we maintain the size of families, but permute the identities of their constituent taxa. This experiment shows that real families contain biological information relevant to their diversification dynamics, i.e. there must be genuine characteristics that make the constituent taxa of real families arise and go extinct according to a similar underlying process. We have also added in the supplement an analysis of the correlation between genus richness in families and their associated β_k values. This is referenced in our motivation for the permutation experiment (around figSupp:betaByRich).*

As stated in the text and shown in S4, the analysis when performed at the level of classes and not orders, seems to fail. I'd like to see an equivalent of Figure 1 for classes, even if placed in the SI. Specifically, does this break down because the distribution of class fluctuation sizes is not Gaussian (maybe itself is already fat tailed, before aggregating across classes?). Or does this break down because the distribution of betas at the class level is not Gamma. The authors show us the resulting P(x) doesn't work, but it isn't clear which ingredient breaks down. (Apologies if I missed this analysis)

> *This would indeed be an interesting analysis and we have now included it in the supplement. XXXXXXX*

I found a few statements related to niches and evolutionary process a bit opaque, and think that readers would benefit from unwrapping/explaining the authors' thinking further. E.g. from L115

"specifying the probability that a given clade, chosen at random, will occupy a region of niche space characterized [by] β_k"

It is not obvious to me why niche space is characterized by beta, the size of fluctuations. I wouldn't really know how to start interpreting it in that way. Do the authors mean in some sense the potential number of

niches within an order is characterized by beta, and orders are occasionally saturating this potential, and at other times (for unstated but probably plausible reasons) dropping below it? Or do the authors regard beta as a trait, that applies to the whole order, and the group of orders together fill out a niche space? It seems like an unusual interpretation of niche, as surely the species within an order are occupying a huge variety of niches. But maybe the authors have some very aggregated definition of niche in mind. This to my mind is all left vague, and it comes up again in discussion:

"Specifically, if β_k values are associated with a clade's physiological and life history traits, and those traits evolve via Ornstein-Uhlenbeck-like exploration of an adaptive landscape, the resulting stationary distribution of β_k will be Gamma (40,41)."

I guess it is not clear to me what it means to say 'a clade's physiological and life history traits', as within an order surely these will be highly variable. DO the authors mean that beta_k might be related to an average of various important traits of lineages wihtin a clade? Again, this connection is not adequately spelled out for readers to form an opinion.

> *We agree that our treatment of "niche" was prime for misinterpretation and confusion, also noted by Reviewer 3. We have opted to replace the term "niche" with "region" or "location" in "adaptive space" following the adaptive landscape idea of Simpson (1953). We pay careful attention to this idea and our definition of it in lines XXXXX. We also acknowledge the confusion around whether the β_k define locations in "niche" space, or if the location leads to a specific β_k value. To clarify this, we have added to the text (XXXXXX) that we interpret each β_k to be an emergent trait of a clade, emerging from its unique set of ecological strategies, and thus could be characteristic of a region of "niche" space, but does not define that region. (around where we cite erwin2012)*

In summary, I think the authors have found a very striking pattern, which has the potential to aid our understanding of fluctuations in diversity. But I would like to understand better what are the potential explanations of this pattern, and I did not find the connections to evolutionary processes invoked clear enough to be convincing yet.

> *We thank the reviewer for the helpful feedback and hope that or revision gives them the desired insight into potential explanations.*

Reviewer 2

Overall, I found the paper to be interesting and the analysis clearly explained. I came to it not knowing anything about superstatistics. Although the analysis is tidy and the approach may well be promising, to be honest, I did not see strong advantages over existing approaches. Fundamentally, the analysis dissects variation in diversity fluctuations across groups, which can be accomplished, for example, through a random-effects ANOVA. You'd get a test of significance of the grouping and could use standard approaches to assess relative support for grouping at order, class, or other levels. To be clear: I don't object to the approach taken, and it may well have advantages as applied in other contexts. From this manuscript, however, I was not able to discern them. And, the finding of macroevolutionary differences among higher taxa is uncontroversial.

We appreciate the reviewer's kind words. A random effects ANOVA could certainly be used to evaluate the support for grouping fluctuations by different taxonomic groups, or not at all, but indeed as the reviewer points out, "differences among higher taxa is uncontroversial." Our goal was not to re-iterate this near truism, instead our goal was to evaluate whether this near truism is in fact sufficient to account for patterns in the fossil record (i.e. fat tails) which up to now seemed to require complicated mechanistic explanations (which we briefly survey in the introduction). The advantage of using the machinery of superstatistics is that it formalizes this simple conjecture—different higher taxa experience different diversification dynamics and these differences, when aggregated, produce the seemingly unusual behavior in the system as a whole. Two additional advantages are: 1) that the success of superstatistics in the fossil record allows us to cast biodiversity questions in abstract terms that connect to other disciplines—from the social sciences to physics—which could help us gain further insight into the possible mechanisms leading to common dynamics across disparate systems; and 2) that superstatistics, unlike ANOVA, does not impose a distributional assumption across groups. This means we can find the best fitting distribution and again learn about the possible mechanisms responsible for the interesting patterns observed by evaluating which mechanisms are consistent with the between group distribution. In our case we discover a Gamma distribution which points to mean-reverting processes such as Ornstein-Uhlenbeck evolution.

In summary ANOVA is a statistical model to understand the relevance of different groupings; its goal is not to predict macroscopic distributions across large systems. Superstatistics is a theory which predicts those patterns and can be used to explore the processes generating them, as well as tangentially perform ANOVA-like tasks of confirming which groupings are sensible.

We clarify the intentions of our superstatistical approach XXXXXXXX

Perhaps some of the benefit of the superstatistics comes from the prediction that fluctuations should be Gaussian within groups, whereas the fluctuations become more fat-tailed with increasing heterogeneity in rates?

This is indeed consistent with the spirit of our thinking.

On this point, I did wonder about the error component of observed fluctuations. My sense is that such errors will be substantial, especially for intervals with few extinctions/originations, and also Gaussian in form, which potentially confounds process interpretations. In any case, it is also possible to assess rate homogeneity under conventional approaches, for example by comparing a model with uniform rates within groups to a model with rate heterogeneity (there are various ways one could do this).

We are also concerned about "error," specifically from biased sampling throughout the fossil record. We account for this using a new method based on binomial sampling and publication bias that we layout in the methods section.

One could certainly evaluate if there is rate heterogeneity by competing rate homogeneous and rate heterogeneous models, but again, this is not our interest—our interest is in the consequence of rate heterogeneity on the macroscopic pattern of fluctuations in taxon richness throughout the fossil record. We seek to clarify this intention with new wording in the introduction XXXXXXXXX

I would also say that I think there are places in which the statistical results are interpreted too strongly in terms of macroevolutionary process. One example is in the claim of "equilibrium" within orders. This sounds biologically significant, until it becomes clear that the term is being used in its sense in statistical physics, which can be generated by a wide range of macroevolutionary processes, including constant-rates birth-death, diversity dependence, and neutrality. These span most of of the scenarios considered seriously in macroevolutionary studies.

We certainly agree with the reviewer. We have introduced the statistical mechanical idea of equilibrium sooner so as to avoid confusing readers of our intentions and interpretations (lines XXXX) (where we first say 'zero-centered')

Another example concerns their interpretation of uniform regimes within orders, and thus that it is only rare events – the formation of new orders – that break out of these regimes. This general dynamic may be at play, but it seems perhaps to reify the model too much. Alternatively, it is possible that biological and macroevolutionary similarity decrease smoothly with phylogenetic distance and this explains why orders and other higher taxa are more coherent than randomly formulated taxa. It is not clear to me that the analyses done can distinguish these possibilities. It is noteworthy that no analysis is presented with families as the focal higher unit (only orders and above). This may be for practical reasons (some families have low generic diversity), but it showing that families do worse than orders would support the scenario favored in the manuscript (although there may be issues of sample size and statistical power).

We have now extended our analysis to families and find, very interestingly, that family-level dynamics are also well captured by superstatistics, but the goodness of fit at the family level is just slightly not as good compared to the family level. As the reviewer indicates, this adds credence t idea. We discuss this in greater depth in lines XXXX

Reviewer 3

I reviewed an earlier version of this manuscript for another journal, and I was ill-pleased to see that my review was largely ignored.

First and foremost, we would like to thank this reviewer for their thorough critique, both for this submission and the previous version. We indeed deeply considered their detailed comments and we are grateful for this opportunity to respond, which was not afforded in the previous review process. Many of their conceptual points we heed at face value; their methodological reservations we thoroughly consider and used to better present our existing methods for dealing with sampling differences. We in fact do not ignore sampling, nor do we use shareholder quorum subsampling (SQS) to account for it. Instead we present a new method. We retain our method in favor of adopting Reviewer 3's alternative suggestion because 1) we nor the other reviewers find fault with this method (Reviewer 3 did not comment on our approach to accounting for sampling differences, though we would welcome their thoughts); and 2) the proposed alternative approaches contain their own assumptions and biases which we deem inappropriate for our analysis. We explain our reasoning in more detail below.

1. The hypotheses that the authors wish to test are best framed in terms of richness: i.e., numbers of taxa. However, what they really are examining are shifts in diversity: a more abstract concept (see Hurlbert 1971 Ecology 52:577, which I probably should have cited in my prior review.). Subsampling routines such as SQS compare the diversity of two collections, not the richness. As Hurlbert showed 40+ years ago, if two collections have the same richness but different distributions of commonness, then subsampling routines will find more taxa from the collection with the more even distribution of finds. Ecologically, that collection is more diverse because there are more relatively common taxa. However, the richness is no different: and the models being tested here concern richness. (There are other concerns with the SQS model: see Hannisdal et al. 2017 Proc. R. Soc. B 284: for a discussion of some; but also see Close et al. 2018 MEE 9:in press)

We apologize for imprecise wording; we are indeed using fluctuations in richness *and not* diversity*. We have changed all instances of "diversity" to "richness" where appropriate. We are also not making use of SQS in any analysis. Instead we use our own method for accounting for sampling differences (which is detailed in the Methods section). The comparison of our sampling standardization routine to SQS in the supplement is intended only to show that our more efficient approach compares favorable to the more established (but computationally and informatically wasteful) SQS method.*

It should be noted, however, that the paleontological literature tends to conflate richness and diversity. Indeed diversity *dynamics is used to describe Sepkoski's curve of genus* richness*. A recent R package* `divDyn` *(https://cran.r-project.org/web/packages/divDyn/vignettes/handout.pdf) reports SQS, among other similar estimates, as within bin* richness*. In some sense this is fine: richness is one of many diversity estimates (for example richness is the Hill number with alpha = 0). But ecologists do indeed reserve special, different meanings for richness versus diversity. We initially embraced the paleontological literature's conflation of diversity and richness, but we agree that erring on the side of clarity is preferred and we have made the needed edits.*

2. Because of this, SQS is not the panacea for the sampling issue. I went on at some length in my last review about the importance of sampling for this. My suspicion is that the authors thought "we used SQS to standardize for sampling: why the reviewer miss that?" I obviously was not clear: the problem is that to get at differences in richness, we need to get at differences in sampling. One could modify Foote's approaches (e.g., Foote 2001 Paleobiol. 27:602; 2003 J. Geol. 111:125; 2005 Paleobiol. 31:6, 2007 Paleobiol. 33:261) as I suggested in the last review. One could also use methods such as TRiPS (Starrfelt & Liow 2016 Phil. Trans. Royal Soc. London Ser. B. Biol. Sci. 371:20150219), although one should also be concerned that the exponential distribution of commonness leads methods like TRiPS to overestimate richness (see Wagner & Marcot 2013 MEE 4:703). Ultimately, to test the how/why of richness fluctuations, some sort of approach like this is necessary.

We sympathize with the reviewer for feeling that their valid concerns about sampling were not addressed. Indeed we have not ignored this central statistical issue—we developed our own method to account for differences in sampling (lines 231–254 in the original submission; see section "Correcting for imperfect and potentially biased sampling" in the current submission). This approach is rooted in the simple idea of binomial sampling. We take this "three timer correction" (as it is often referred to) and go one step further to correct for sampling differences introduced by publication bias of novel taxa.

The methods developed by Foote and proposed by the reviewer are among the best in the field, and we are long time admirers of Prof. Foote. However, these methods were developed for stratigraphic ranges only—which fail to take full advantage of the wealth of occurrence data provided by the PBDB, as noted repeatedly by Alroy (2008; 2010)—and cannot account for publication bias. As such we choose not to implement them.

The TRiPS method suggested by the reviewer employs a binomial likelihood to quantify sampling. Because the sufficient statistic of a binomial likelihood is simply the observed proportion of successes (or failures) we have indeed mimicked this approach by estimating the observed proportion of "Lazarus" genera (i.e. those that have failed to be detected between subsequent intervals of positive detection) for a given time period and using this observed proportion to estimate the expected number of taxa actually present in that time period. This expected value is simply the number of observed taxa divided by the estimated sampling probability (to confirm this one need only recall that the mean of a binomial distribution is $E[X] = np$ thus n—the latent number of taxa—is $n = E[X]/p$ assuming the observed number of taxa is on average near the expectation). The details of our approach are laid out in the methods section.

3. As before, I think that "niche" is the wrong word for the ecological level that the authors are

discussing. They really are discussing something closer to guilds. As I noted before, there is a rich literature initiated by Richard Bambach about guilds and guild assignments. Much of this data can be procured from the Paleobiology Database, too. For example, to get information on the trilobite genus Ceraurus, go to: https:// www.paleobiodb.org/classic/checkTaxonInfo?taxon_no=21472 and click on "Ecology and taphonomy." This provides much of the information to which the authors refer to as affecting "niches" (really, guilds), and they form the bases of Bambach's guild models. One can also use API: the link http://www.paleobiodb.org/data1.2/taxa/list.csv?base_name=Phacopida&show=parent,ecospace,taphonomy,etbasis will provide this information for all genera within the Phacopida. I have provided R-code for doing this below: I strongly encourage the authors to take advantage of these data! Alternatively, the data used in papers such as Bush et al. (2007 Paleobiol. 33:76) could be used.

It is indeed a great resource that Bambach's guild designations are available through the PBDB. We detail our utilization of these data in our response to the next comment.

We also find it necessary to confront the semantics of "niche" versus "guild." We concede that the literature is over-populated with different uses for both terms, many overlapping and likely conflicting between disciplines (e.g. ecology, evolution, paleobiology). To avoid adding further to this confusion we opt to not use either of the terms and instead replace all instances of "niche" with "adaptive landscape," defining carefully what we mean by that term at the first instance of its use (lines XXXX) (where we first say 'macroecological')

4. Following point 3, I am very skeptical that orders are going to be homogeneous ecological units. Indeed, one of the explanations for why groups like orders initially radiated is that they were able to become ecologically diverse.

We have taken this reviewer's advice to investigate the ecological composition of different taxonomic levels. This analysis reveals that families contain on average one ecological guild hypercube, families contain on average 2, and classes and phyla contain many more. This analysis is in the supplemental section "Guild composition of higher taxa" and referenced in the main text at lines XXXXX.

5. As before, I think that the authors do themselves a major disservice using the word "punctuated," although they at least have taken "punctuated non-equilibrium" out of the title. Paleobiologists have taken to using "pulsed" to describe events that are clearly distinct from "background" evolutionary time. Now, in plain English, there is not much difference: but given that "punctuation" has become so strongly linked to a particular speciation model (or set of speciation models), it is best to leave that word alone.

Agreed, and it was thanks to this reviewer's first round of comments on our manuscript that we changed our title for the better. In our current revision we have replaced all instances of "punctuated" with "pulsed," where appropriate.

Some other comments.

Figure. 1. There still is no indication of what these orders are.

The figure now shows families and their names are indicated

SQS Analyses: Again, there is no mention of what the coverage levels were, or how exactly they were chosen.

We have specified the coverage which was set at 0.75 to reflect a similar average richness as produced with rarefaction and our new bias correction method, this is made clear in the supplement in the section "Evaluation of sampling bias correction methods".

As I noted last time, I think that this paper offers a really interesting methodological advancement. However, I also think that this will be ignored because of objections to the theoretical parts of the paper. If the authors can redo the analyses in the ways that I suggested here and in the prior review, then I think that there is a much better chance of both the methods and interpretations gaining some traction.

We greatly appreciate the reviewer's support and their commendable efforts to help us improve out paper. We also acknowledge their reservations. The approach we choose is to further clarify and detail our new method to account for sampling differences. We do not specifically adopt the approaches suggested by the reviewer for the reasons discussed above; however, we go to new lengths to make clear our methodological reasoning.

Finally, here is some R-code that the authors can use to download ecological data from the Paleobiology Database. Just enter (say) accio_ecologic_guild_data(taxon="Phacopida") and you get a tab-delimited text file with the ecological data that the PaleoDB has for phacopid genera.

Thanks for sharing! We have now made our code to interface with the PBDB API more accessible online and in the supplement.

```
clear_na_from_matrix <- function(data, replacement = "") {
    for (i in 1:ncol(data)) {
        if(sum(is.na(data[, i])) > 0) {
            duds <- (1:nrow(data))[is.na(data[, i])]
            data[duds, i] <- replacement
        }
    }
    return(data)
}

accio_ecologic_guild_data <- function(taxon, taxon_level = "genus,subgenus",
                                      output_type = ".txt") {
    http <- paste("http://www.paleobiodb.org/data1.2/taxa/list.csv? base_name=",
                  taxon, "&rank=", taxon_level,
                  "&show=attr,app,ecospace,etbasis", sep="")
    fetch <- RCurl::getURL(http)

    taxon_guilds <- utils::read.csv(text = fetch, header = TRUE,
                                    stringsAsFactors = TRUE)
    taxon_guilds <- clear_na_from_matrix(taxon_guilds, "")
    return(taxon_guilds)
}
```

Bush, A. M., et al. 2007. Changes in theoretical ecospace utilization in marine fossil assemblages between the mid-Paleozoic and late Cenozoic. Paleobiol. 33:76-97. (10.1666/06013.1)

Close, R. A., et al. 2018. How should we estimate diversity in the fossil record? Testing richness estimators using sampling#standardised discovery curves. MEE 9:in press. (10.1111/2041-210X.12987)

Foote, M. 2001. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. Paleobiol. 27:602 - 630. (10.1666/0094-8373(2001)027<0602:ITPOPO>2.0.CO;2)

Foote, M. 2003. Origination and extinction through the Phanerozoic: a new approach. J. Geol. 111:125 - 148. (10.1086/345841)

Foote, M. 2005. Pulsed origination and extinction in the marine realm. Paleobiol. 31:6-20. (10.1666/0094-8373(2005)031<0006:POAEIT>2.0.CO;2)

Foote, M. 2007. Extinction and quiescence in marine animal genera. Paleobiol. 33:261-272. (10.1666/06068.1)

Hannisdal, B., et al. 2017. Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record. Proc. R. Soc. B 284:20170722. (10.1098/rspb.2017.0722)

Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577 - 586. (10.2307/1934145)

Starrfelt, J. and L. H. Liow. 2016. How many dinosaur species were there? Fossil bias and true richness estimated using a Poisson sampling model. Phil. Trans. Royal Soc. London Ser. B. Biol. Sci. 371:20150219. (10.1098/rstb.2015.0219)

Wagner, P. J. and J. D. Marcot. 2013. Modelling distributions of fossil sampling rates over time, space and taxa: assessment and implications for macroevolutionary studies. MEE 4:703 - 713. (10.1111/2041-210X.12088)

Reviewer 4

This is a very interesting manuscript, worthy of consideration for publication in Science Advances. The primary finding is that diversity fluctuations with and between orders, but not within and between classes or phyla, follow a relatively simple compound statistical distribution the implements a superposition of statistics, called "superstatistics". The idea of superstatistics is relatively straightforward, effectively compounding the distributions of random variables from several different distributions to account for the observed data.

One of the reasons the manuscript is interesting is because one might reasonably expect no (relatively) simple statistical characterization would hold for such a complex process as the evolution of 500+ million years of marine animal life. The other primary reason it is interesting is that it holds at just one taxonomic level, but not for the other two tested. Thus, this suggests a higher taxonomic structure to biodiversity dynamics. There is some precedence for this idea (e.g., a coupled logistic equation can be made to fit Sepkoski's family level data, but not his genus level data), but this is the first time that it has been laid out concisely and strongly supported. The fact that the superstatistics only holds for orders is also important because had the superstatistical description held at all taxonomic levels it might have indicated that the finding is relatively trivial (in the way that finding that a whole bunch of morphometric measurements, say height, weight, and arm length, all followed the normal distribution, would be trivial). Thus, I feel this is worth publishing in Science Advances.

We thank the reviewer for their interest and support.

However, it appears to have been written in the flush of (deserved) over-excitement – it needs some work to maximize its impact, as well as its accuracy, especially given that none of the authors are well-versed in the paleontological literature (or at least so I infer, given the way the manuscript is written [see below]).

Most important suggestions for improvement

1) A major shortfall of the paper is the fact that the manuscript appears to have been written 4- 5 years ago, judging by the fact that the data download from the Paleobiology Database (PBDB) was in 2013, and that the most recent citation is from 2014 (1 only), with just one citation from 2013 as well. The PBDB is still growing at a reasonable rate, and so: (i) the analysis needs to be re-run with 2018 data.

We certainly acknowledge this shortcoming and we have completely re-done our analysis with a download of the PBDB from November 2018. We have also made our analysis completely reproducible and well-documented so any interested person can easily re-run our analysis at any future date with even more expanded versions of the PBDB.

And, (ii) the text needs to updated to accommodate the steady rate of publications on equilibrial/non-equilibrial diversity dynamics. In my opinion none of these un- cited more recent papers undermine the fundamental contribution of this manuscript, but a new consensus centering on time-varying equilibrial dynamics is beginning to emerge over the previous more polarized equilibrial/non-equilibrial debate (e.g., see the 2016 Phil Trans Roy Soc B volume edited by Quental and Ezhard).

We also acknowledge that newer literature needs to be engaged. We have done just that throughout the manuscript and greatly appreciate the reviewer's helpful suggestion of the Phil Trans Roy Soc B edited volume.

2) The paper needs to make a 'bigger' deal of the fact that the superposition of statistics approach does not fit the phylum and class-level data. I strongly recommend: (i) highlighting this in the text, as well as (ii) bringing the phylum and class figure panels from the Supplemental Information into the main text.

We have revised Figure 2 to show all taxonomic levels (now including family as well), and have emphasized this finding in the results and discussion (XXXXX)

3) Related to the point made immediately above, the analysis should also be performed at the family level, or if that is not possible, that needs to be explained why. Why stop at the ordinal level?

We have included families. And indeed the new results are very interesting and incorporated throughout the revision.

Intermediate suggestions for improvement

The paper tries, perhaps as it should, to discuss, or at least mention, all the related areas that their analysis might be relevant to. This is understandable, given the background of the authors (who I surmise come from a tradition where advances are considered important to the extent that they identify deep universal principles). But I feel they over-reach in many small ways that, rather than enhancing the paper, detract from it.

4) This 'over-reach' begins in the abstract, where they claim that "three universal properties of macro-evolution .... are sufficient to explain the previously unaccounted for fat-tailed form of fluctuations in the Phanerozoic. There are five issues with this statement: (i) The ability to describe diversity fluctuation with superstatistics does not constitute an explanation, even though it has implications for what an explanation might be. (ii) The results only pertain to orders, so it is not 'universal' in the way that the abstract and title implies. (iii) It only pertains to the marine fossil record – who knows what the terrestrial record would show. (iv) As they note, they assume (using the words"likely driven by") that"niche conservation" and "punctuated adaptive radiation" are sufficient, with the heterogeneity of diversification rates (do they actually mean diversity fluctuations, given that they do not measure rates, per se?) to explain the fat-tail form of the fluctuations – but they are simply assuming that the superstatistical fit to the ordinal data is driven by "niche conservatism" and "punctuated adaptive radiation", and offer nothing to support (not unreasonably) these claims. Thus, it is an over-reach to claim that they have explained three properties of macroevolution, as to me their sentence implies. They have found a surprisingly simple description of the fluctuations of just the orders, and while they can account for their data with the ideas of niche conservatism, with the assumption of non-overlapping

niches across orders, they have not explained it. (v) The term punctuated adaptive radiation is problematic – later in the text they evoke Gould and Eldridge's 'punctuate equilibrium', but that theory was developed right down at the species level, and has nothing to do with adaptive radiations, per se. My guess is what they are trying to say is that the pattern they find is consistent with each order having a coherent ecological domain, with some sort of transition that corresponds to new orders that also places those new orders in a new ecological domains. However, the transitions need not be fast (punctuated), nor does the filling of the new ecological space have to be fast to explain their statistical result (i.e., the orders don't have to radiate, per se).

So, I think they need to re-word as much of the manuscript as they can, sticking to what the super-statistics results actually are and what they actually mean. And then, in the discussion, they can worry about positing connections between the phenomenology they have discovered to other ideas.

we have re-worked much of the paper to avoid over-reach, including major over-hauls of the abstract and introduction where the reviewer found specific issue with niche conservatism and punctuated evolution.

5) Lines 36-38. The casual (sounding?) dismal of escalation (Vermeij, [13]) is poor. The Madin et al. paper is weak, and has not gained traction. Moreover, the evidence of escalation is deeply supported, even if there hasn't been a successful attempt to describe it statistically (I suspect the primary reason it has not been so described is that it consists of many, many, separate events, which, due to the nature of escalatory dynamics, leads to a long-term ratchet in the diversity dynamics). So, this needs to be re-written with a little more nuance.

Revised for added nuance and fairness to the literature; see lines XXXX (around where we cite vermeij2008 now)

6) Line 231. This section is labeled: "Three-timer and publication sampling correction", yet in the SM the authors claim that their results are pretty robust whether sample standardization is used or not. Further, the three-timer method of Alroy (date) has been 'improved' upon by Alroy (2014), and Alroy (2015), while there has been another approached developed by Foote (2016). So, I think section has to be labeled to encompass what the authors have actually done more broadly; the way it is written undersells the robustness of their results, and pays a sort of homage to Alroy that will undermine the credibility of the paper (there is nothing wrong with using the three-timer method, but the section has to be written with a broader view).

This section has been re-titled "Correcting for imperfect and potentially biased sampling" and contains a superior description of the method and a more nuanced contextualization in the most recent literature on this subject.

7) Line 344. It is stated that the data are available though the PBDB – but where are the data you analyzed available (especially, after all that cleaning)? To be frank, this statement comes across as lazy and dismissive.

This statement has been extended to ensure that any reader can reproduce our results.

Minor suggestions for improvement

I have not been exhaustive here, there are too many little places where the text could be improved, but here are some suggestions:

Lines 20-23. It is stated in the abstract that that "Its success opens up new research directions to better understand the universal nature of non-equilibrium dynamics across disparate systems of interest from

societal to physical to biological." I don't see that the authors' (understandable) excitement about the greater potential of superstatiscal approaches (i.e., societal, physical) has any relevance with respect to this application to the paleontological record. It is distracting. If they really feel the need to say it, it should go elsewhere in the paper.

removed

Line 23. I think the authors might mean "punctuated by innovations that lead to new orders".

changed accordingly

Line 26. 'extant' tends to mean those living today, so maybe say "in the number of taxa alive at any given time'.

changed to "richness of taxa"

Line 30. In the line before, 'distribution' is singular, so should 'tail' be singular on this line?

changed to "fat-tailed distribution"

Line 31. I'm not sure any one has tried to actually 'predict' unusually large fluctuations in diversity. This sentence feels like it needs a re-write.

changed to "Understanding the fat-tailed nature of these fluctuations. . . "

Line 101-102. It is stated that "The β_k thus represent the inverse variances of homogeneous origination-extinction processes, . . . ". Of what? Genus richness?

yes, genus richness has been specified

Lines 107-111. The discussion of the possible relationship between the Gaussian fluctuations in genus richness and neutral-like process (in contrast to the largely dead idea of SOC [self organized criticality]) is distracting, and should be moved to the discussion.

moved to Discussion

Line 116. Insert "by" after characterized? And why not remind us of what β_k means in the paleontological context?

done

Lines 116-118. The speculation that "[t]he form of this stationary distribution could shed light on . . . .." is distracting, and should be moved to the discussion. Part of the reason it is distracting is because for an evolutionary biologist there is already a rich literature on the processes that lead to different clades exploring different regions of eco-space (adaptive landscapes), a literature that is usually couched in terms of the specific properties of the taxa and the environment. So, this statement, rather than advancing the field, I think just undermines the credibility of the authors, largely through the failure in their knowledge of the literature of the paleontological record and its interpretation.

moved to the discussion

Line 224-225. What exactly does latest Cenozoic mean? The Holocene, the Neogene, the last 10 million years?

we now specify that this means the most recent ~11my bin as advanced by Alroy 2008

Line 240–242: It is stated: "To eliminate further bias due to preferential publication of novel taxa we divide observed number of genera per order per time period by the expected number of genera given publications in that time period". I don't understand this hypothesized bias – please explain!

This is a concern first proposed by Alroy (2010). We agree with the principle that researchers might show preference for publishing on new groups, thus leading to a possible correlation between number of publications for a given interval and number of taxa. We find only very weak evidence that more publications lead to more taxa regardless of taxon identity and time interval. We report this weak correlation and have now added a citation to Alroy (2010) to further justify the possibility of publication bias.

Line 257: Is some punctuation missing here?

this sentence has been rewritten

Between line 348 and 349: I assume you mean commutative property of the Poisson distribution where the text says "communicative property of the Poisson distribution"

corrected

Lines 358–360. This is a sloppy sentence – in this paper the authors claim that much of the raw signal in massive fossil datasets, at least signals regarding fluctuations, are not artifacts of sampling, as has been proposed before [48]. But surely this depends on the type of signal being measured, so the conclusions of this paper need not be at odds with the conclusions of ref [48].

this section has been removed

Figure

Figure 1: (i) Shown are three curves, for trilobites, gastropods, and ammonoids. The figure caption states that these are within order fluctuations, so please give the name of the exemplar order. This will help the casual reader to see that you have not done the analysis at the class- level, which is what the icons will evoke. (ii) I presume the red curve is a group of nautiloids, not ammonites, given that the groups persists into the early Paleogene – please confirm, or correct the position of the curve.

this figure now contains different clades and their names are made clear