Dear Dr. Chase,

We are writing to resubmit the manuscript “Empirical abundance distributions are more uneven than expected given a statistical baseline”, by Renata Diaz, Hao Ye, and S. K. Morgan Ernest, for consideration as a Synthesis article. We are grateful to the reviewers for their insightful and constructive comments, which we believe have helped strengthen the manuscript considerably. In response to the reviewers’ questions, we have implemented a set of changes that we hope makes the manuscript more technically robust and more interesting and accessible to a general ecological audience. We have added two resampling analyses to explore how the tendency for rare species to escape detection during field sampling, and sampling error in the shape of the SAD more generally, affect our results. We have also added three metrics for describing deviations between observed and randomly-generated SADs, which provide a more nuanced picture of the overall magnitude of observed deviations, how deviations vary over large ranges of species richness and total abundance, and how statistical deviations relate to ecological properties such as the number of rare species in a system. We situate these new results in a more extensive discussion of complexity science as it relates to the SAD and have expanded our discussion of how to interpret deviations from a statistical baseline in ecological terms. We hope this helps illustrate the ecological significance of our results and lights the way forward for new applications of this approach.

Thank you very much for your time and consideration,

Renata Diaz, Hao Ye, and S. K. Morgan Ernest

Response to reviewer comments

Referees' comments to the author(s):

Referee: 1

Comments for the Authors

This manuscript addresses an important issue in ecology: how do we tease apart ecological processes from statistical artifacts? It adds substance to previous suggestions that deviations of observed macroecological patterns from the predictions made by purely statistical models can help identify truly essential ecological mechanisms.

Using the species abundance distribution as a test metric, the authors convincingly analyze a large number of data sets and reach useful conclusions concerning both the pattern of deviation of observation from the feasible set model and the important role of system size. Appropriately they have selected two features of the distribution, evenness and skewness, to use in in their analyses.

The authors mention in the Discussion the potential importance of their assumption of indistinguishability. This is indeed important. A paper that addresses the role of unique versus indistinguishable species and individuals in a statistical model that analyzes a different but related kind of feasible set approach is:

Zhang, Y. & Harte, J. “Population dynamics and competitive outcome derive from resource allocation statistics: the governing influence of the distinguishability of individuals”. Theoretical Population Biology 105:52-63(2015).

It would be nice if the authors could address (speculate on), in the Discussion, the issue of how alternative assumptions about uniqueness might influence evenness and skewness.

We agree that the issue of unique vs. indistinguishable species warrants further exploration. Without extensive resampling, it is difficult to know precisely what impact different alternative scenarios may have - especially over the range of variation in S, N, and the ratio of N to S. However, to help readers think about this issue, we have added a passage in the Discussion [lines 503-515] working through an example of how labelling species order might be expected to modify the characteristics of the feasible set (and therefore the deviations we detect). We

selected this example because it is relatively simple to explain and has a clear ecological interpretation; going forward, we agree that systematically exploring the consequences of different assumptions about the distinguishability of species and individuals is an important next phase for this line of work.

Another topic that might warrant some discussion is the ratio of N to S. Do the authors see any pattern in the relationship between deviation from feasible set and that ratio?

The importance of average abundance (N/S) is most apparent for the subset of our datasets, and specifically the Gentry dataset, with high species richness and extremely low average abundance (N/S < 3). These communities actually show deviations in the opposite direction to most other communities (e.g., they are highly even, have low skewness, etc). We have added a section of the discussion highlighting these results and exploring possible statistical and biological drivers [431-449]. Because examining these effects shifts the focus to include deviations where the observed SAD is both very high and very low relative to the distribution of sampled values, we have shifted to the use of 2-tailed confidence intervals for these shape metrics.

A few minor things:

Line 46. The first sentence in the Introduction is a bit awkward. I would write: “How the total number of individuals in a community is divided among …” (note: “the number” is singular.)

We have updated this sentence [63-64].

Line 51. The SAD is a hollow-shaped curve when x is plotted against y. The authors should state explicitly what x and y are. Those “in the know” of course know what is meant but the average reader may wonder. After all there are axes on which SAD plots are not hollow!

We have added this clarification when the SAD is defined [70-71].

Line 143-145. If you exclude systems with S = N or S = N+1, and systems with S = 1, then how can you have an N as small as 3 (line 145)?

We have corrected these ranges, to ranges of S from 2-250 and N from 4-40714 [178].

Referee: 2

Comments for the Authors

I very much appreciate the approach that Diaz et al. are taking in this manuscript; namely that to infer mechanism ecologists must not focus on the general shape of universal patterns derived from statistical mechanics, but rather the *deviations* from these patterns. I like the use of Locey & White approach to generate expected distributions. So in all I think this can be a very useful contribution.

However, that being said I think the work could be made more effective. My issues are:

(1) the used assemblage datasets are taken as being without sampling variance; i.e. there is no consideration of the fact that the makeup of the individual datasets themselves are driven by stochastic process. Thus the abundances are taken as fixed and used to generate each feasible set. But, this assumption is invalid. As anyone who has generated assemblage data knows, there are any number of stochastic processes that underlie dataset generation. Thus, if one goes back and resamples a site, the actual abundances will vary each time. And, often sampling biases complicates this further (for instance the 'big tree bias' of McCune & Menges 1886). The fundamental issue, then, is whether the relatively subtle differences seen here fall inside or outside of the cloud of responses obtained when allowing abundance lists to vary. Without this I simply cannot be certain how robust the given results are.

I think what needs to happen is to use some type of resampling approach (e.g. monte carlo simulations, jackknife analysis) to begin quantifying the inherent variability present within abundance datasets. Only after we know the amount of noise generated by sampling variation can we know if the observed results fall outside of this range and are trustworthy.

Sadly, I see this resampling analysis to be foundational to what is presented here; without it we simply can't know how to interpret the results, no matter how many datasets are included. Simply carpet bombing with highly variable and untrustworthy data will not lead to robust results.

We agree that sampling uncertainty in empirical data is an important consideration for the SAD. We implemented two resampling analyses to explore how major types of sampling uncertainty affect the deviations we detect.

First, we reasoned that field sampling could systematically result in an under-estimation of species richness, especially for rare species. We used species richness estimators to estimate the true number of species in our communities and added any "missing" species as rare species (abundance = 1). These adjusted SADs were consistently more extreme relative to their feasible sets than the raw data.

Second, we used a jackknife subsampling scheme to explore how sampling error across the entire SAD, and not just for rare species, affects the shape of the SAD and any deviations that may occur. For every raw SAD, we drew 10 subsampled SADs, each with 60% of the individuals from the original SAD drawn at random and without replacement. This introduced variation in shape between the subsampled and the raw SADs, but yielded subsamples that were still large enough for this analytical approach to be appropriate. In general, the subsampled SADs showed less pronounced - but still detectable, and still highly unlikely - deviations from their feasible sets. The full methods for both of these analyses are at lines 181-213 and results are at lines 387-398, 421-423, and in Appendix A7.

(2) Nowhere did I see any indication of just how much variance is being accounted for in these deviations from expected. Is it 1% of total? 10%? This may seem trivial, but this helps the reader assess just how important these potential ecological controls are. Remember the old adage that statistical significance is not the same as biological significance! Your deviations may be statistically significant (and almost cannot help to be given your sample size!) but if they only explain 0.1% of the variance then it may be biologically and ecologically irrelevant - and likely swamped by sampling variation.

We did not attempt to describe the overall magnitude or effect size of deviations in the initial manuscript, because the values for skewness and evenness vary so widely over the ranges of S and N in our datasets that it is nonsensical to report deviations in those terms. However, we agree that this is an important consideration, and we have introduced an additional measure to describe the overall magnitude of the differences between observed SADs and their feasible sets.

Following Locey & White (2013), we identified the central tendency of the feasible set for a given S and N as the SAD with the lowest average dissimilarity to all other samples from that feasible set. For an overall measure of how different observed SADs are from their feasible sets, we calculated the degree of dissimilarity between observed SADs and the central

tendencies of their feasible sets. To then test whether the observed degree of dissimilarity was highly unlikely given the feasible set, we compared the observed dissimilarity to the distribution of dissimilarity scores comparing all samples from the feasible set to the central tendency. Observed SADs are consistently more dissimilar to the central tendencies of their feasible sets than the majority of samples from their feasible sets, with dissimilarity scores from 1.5 to 9.7 times greater than the average for samples from the corresponding feasible sets. The full methods for this measure are at lines 286-309 and results are at lines and 352-259.

We have also added some additional clarification regarding the reasoning behind using percentile scores for these analyses and what statistical significance means in this context [lines 293-309]. We use percentile scores because the actual ranges for the summary metrics vary widely over S and N and are therefore not comparable, but the percentile scores can be used to make comparisons between SADs with very different S and N. For a single community, a percentile score of 99 means the observed deviation has a 1% chance of coming from the sampled distribution at random. Repeated many times, it is expected that an event with a 1% probability will occur about 1% of the time, and this is why we compare the percent of observed extreme percentile scores to the percent that would be expected at random. If no deviations are present, aggregating over a large number of communities will not increase the proportion of extreme percentile scores.

(3) No rationale is given for why Simpson's Index is solely used to document evenness. The issue is that Simpson weights its scores on the most abundant species - thus those that are out on the right tale of the distribution (Peet 1974). And I don't see why one would want to do that. Not that Shannon is perfect either, given that it *also* down-weights the rarest species. But the issues are large enough that you really need to carefully defend your choice, and perhaps also make sure the results are robust across methods.

While we initially chose Simpson's index simply because it is widespread and generally understood by ecologists, we appreciate the concerns associated with using any single diversity metric to summarize the SAD. In addition to the dissimilarity metric discussed above, we have added two more summary metrics - Shannon's index, and the proportion of rare (abundance = 1) species - to our analyses. Our results are generally robust across the different metrics (Table 1).

(4) This is minor but the background literature for this topic really needs to be broadened. The current manuscript reads as if these ideas only date back to the late 2000's. There are only 2 cited papers that predate 2007, and one of these is Jaynes treatise on Statistical Mechanics! This does a deep disservice to a number of voices in the ecological community. Why no

Preston citations - given that he generated the first statistical mechanics approach to SADs? Especially given that in his 1981 paper shows that SAD shape falls within the realm of universality? And, why no discussion of complex systems and the statistical mechanisms underlying ecology? This is essential to your basic premise. How can the ideas in Brian Maurer's 1999 "Untangling Ecological Complexity" not be considered? Or my 2007 Ecology Letters paper with Jim Brown: it even falls within your apparent 2007 cutoff for citations!

We completely agree that these ideas are important inspiration and context for this analysis. Taking advantage of the longer Synthesis format, we have expanded the first three paragraphs of the introduction to provide more context on approaches from complexity and statistical mechanics as they apply to macroecology and the SAD in particular [lines 51-94] and have added specific references to important background literature as part of that expansion.

I do think this work can be an important contribution. But first we need to know just how profound sampling variability is. And we need to know that the deviations are not only significant, but non-trivial. Sadly, I think this means going back to the computer and coding some additional analyses. Because without this you have not made the firm logical foundation upon which we can interpret these results.

Jeff Nekola, Masaryk University

Referee: 3

Comments for the Authors

It was a pleasure to read the paper titled “Empirical abundance distributions are more uneven than expected given their statistical baseline”. The authors use over 22,000 sampled communities and feasible set sampling to assess SADs within and among the different communities. They develop some interesting methods, such as a Breadth Index, by investigating skewness and evenness of a given community, comparing it with what would be expected by chance (statistically) by resampling the feasible set. I agree with the authors that ecological processes act upon SADs, but this is often in combination with the statistical artefact of the hollow curve.

In McGill et al.'s 2007 Ecology Letters paper (https://doi.org/10.1111/j.1461-0248.2007.01094.x) they highlight the following: "Collect as large a sample size as possible. As described, we do not know what a good sample size is, but it clearly at least in the 100s and quite possibly in the 1000s of individuals." I think that this paper makes a good push to understanding what a 'sample size' might be, in terms of the number of individuals as their definition of a community. I am not aware of much work in this space, so in this regard I find it very well done and novel! I will note that the authors do limit themselves in this work by focusing only on the feasible set as an approach to define the statistical baseline of a SAD. However, it was refreshing for them to clearly highlight this limitation of their work (lines 292-303), and I don't see this as a problem, but commend them for their clarity of their scientific advance.

The paper was well written, easy to follow, and to me is statistically sound. I do think it is within the scope of Ecology Letters, and will make a great contribution to the literature surrounding SADs. All this said, I have two broad comments/reservations about the manuscript that I think the authors should consider and I believe could improve the manuscript. I don't see these as 'deal-breakers', but I do think that if addressed, the manuscript will be more widely applicable.

1.) Currently, while it is interesting and novel – the pitch of the manuscript is rather narrow, focused on the statistical deviations of the feasible set. It is applicable to anyone quite familiar with the SAD literature, but I fear it is not broadly relevant. Basically, I think the readership of such a paper is rather 'niche'. The only reason I highlight this as I don't think it would be overly difficult to better place this paper into a broader scope. One suggestion would be to move past the statistical deviation and try to better link this with the ecological/biological interpretations inherent within SADs (i.e., the proportion of rare, or common, species in a community). It is implied that this work shows our ability to say something about 'small' communities is limited, but what does that mean in practice? Are all previous SAD findings from small communities invalid and are indeed statistical artifacts? Are we thus more likely to have rare species in larger communities? If we want to characterize a community, ecologically, what does this work mean we should keep in mind and how does it influence our interpretations of a SAD? The authors hint at this in lines 260-263, but I think this could be substantially expanded to highlight what these findings mean ecologically a bit more. I guess what I'm trying to say is what does 'detect deviations' mean in practice? Again, I don't think this is too overly difficult. As an example, the authors define skewness and evenness in line 189. But this definition is strictly statistical. Here, an understanding of what skewness/evenness means in terms of ecology would go a long way to better interpret these results. This is one example, and there are plenty of areas through the paper that by inserting a few sentences/explanations, could help this link with ecological interpretation (another example could be line 314 where the authors highlight the role of ecological processes). If the authors can make this more concrete, I feel that the manuscript could be more widely understood and more suitable for the broad readership of Ecology Letters (and probably cited as well).

We appreciate the reviewers' perspective and have worked on broadening the manuscript in a number of ways to appeal to a wider audience. In response to reviewer 1, we broadened the introduction and discussion to link our results more directly with the history of the SAD and complexity science in particular [lines 53-94]. Building upon reviewer 3's comments, we have added additional metrics and discussion intended to strengthen the ecological interpretation of our results. We were particularly excited by the suggestion to include the proportion of rare (abundance = 1) species in a community, and have woven this in throughout the manuscript as an example of an attribute of the SAD that can be used to compare observed communities to the baseline (lines 104-113). We have also expanded our discussion of how biological processes may interact with mathematical ones to generate the observed results, including unusual results for the number of rare species for a subset of communities (lines 431-450).

2.) The authors' main finding is that small communities have a reduced ability to detect deviations (Lines 40-42). They highlight more in depth in the discussion (Lines 286-287) that perhaps about 50-100 individuals "may indicate a general range of values below which we have diminished power to detect deviations". I agree with this point and follow their logic. I also don't think that a 'hard threshold' needs be determined and such 'general guidelines' are a scientific advance at present. However, it isn't immediately clear to a reader how this was determined. While I am not opposed to the decomposing of the datasets in Figure 2/Figure 3 and think this makes logical sense, I do believe that there is a key figure missing to immediately illustrate the point made in the abstract and throughout the paper. I would envision something that doesn't consider the different datasets and is a scatter plot with the 'size' of the community (for N~22,000) on the y-axis and then some measure of statistical power on the x-axis (I guess this would be your breadth index as defined?). This should then show some 'dip' in power around the size of 50-100 individuals, and generally increasing power with a greater number of individuals. Of course, I'm sure there are plenty of ways to clearly illustrate this. I think such a figure would immediately highlight what is, to me, one of your key findings.

We have added a figure to the main text (Figure 4) showing how the narrowness of the feasible set - defined either as the dissimilarity of the samples in the feasible set to the central tendency of the feasible set, or using the breadth index calculated for particular summary metrics - varies with the size of the feasible set (number of possible SADs for a given S and N). In general, smaller feasible sets are more variable and often much less narrow than large ones, but there is not an obvious community size threshold below which feasible sets become extremely broad.

We believe there are several reasons for this. First, the narrowness of the feasible set depends on the ratio of N to S, in addition to the actual values of N and S and the number of possible SADs in the feasible set. For example, if the ratio of N to S is very small, all possible SADs will be very similar to each other and the feasible set will be more narrowly defined than is usual for feasible sets of that size - this is most obvious for the dissimilarity metric for the Gentry dataset (Figure 4). Second, the combinations of N and S in our datasets are unevenly distributed among the different datasets and represent a patchy, uneven subset of the possible N x S state space (Figure 1), which means they are poorly suited for exploring thresholds or systematic rules.

Finally, there is considerable variability in the breadth indices that we believe derives from the highly irregular underlying distributions of summary metrics for samples from the feasible sets. For example, the sampled distributions for skewness often have long tails either to the left or right; and the sampled distributions for the number of rare species have many ties and 0 values.

Combined, these factors mean this analysis is suitable for describing a general relationship between N and S, the narrowness of the feasible set, and possible effects on statistical power, but clearly a more extensive exploration will be needed to establish more specific thresholds or rules. We have updated our language throughout the manuscript to emphasize these considerations (lines 133-150; lines 463-496; lines 521-534).

Two minor comments:

1.) I would have liked to see the link to the data/code available with submission. I went to try and find it, mostly to try and think about #2 a bit more in depth than above, but I then saw it said 'after' acceptance. Not sure what Ecology Letters policy is on this, but just a comment.

This was an unintentional omission! The code for the main analysis is at www.github.com/diazrenata/scadsanalysis; the code for the sampling algorithm is at www.github.com/diazrenata/feasiblesads. These repositories are archived on Zenodo at https://doi.org/10.5281/zenodo.4711104 *and* https://doi.org/10.5281/zenodo.4710750*, respectively. These links have been added to the main text (line 158).*

2.) I think some of the details from Appendix S3 describing the neat-looking GitHub feasible set sampling could be moved to the main text (space permitting) as I think this is important to the paper.

We have expanded the explanation of the sampling algorithm in the methods (lines 231-261).

****

Editor's comments to the author(s):

Editor

Editors Comments for the Author(s):

Three highly qualified reviewers have now carefully read the work, as have I. As you will see, all of the reviewers very much appreciated the work you have accomplished, but at the same time have some rather critical concerns and comments that should be carefully considered and implemented in a substantial revision. The most important comments, from my perspective, can be seen most clearly in the reviews by referee 2 and 3. First, I agree with the comment that the manuscript is often written, especially in the introduction, in a highly technical and specialized way. The SAD literature is broad and deep, and a more 'ecological' focus on that, in addition to the more specialized arguments, would certainly broaden the paper's impact.

We have broadened our framing of the paper to include more depth on complexity and statistical mechanics as they relate to macroecology and the SAD in particular (especially lines 50-94), and have added additional metrics and interpretation to illustrate the ecological implications of the deviations we detect (lines 268-319; lines 416-450).

Second, reviewer 2 seems to bring up a rather important point about issues of sample variance, especially when small samples are taken (ref. 3 also indicated some concerns in this direction). Some of the suggestions, such as resampling, seem rather reasonable, even though it would take a fair amount of work, this would clearly improve confidence in the results. I would strongly suggest the authors take on this task head on, albeit a potentially 'big one', as it seems that in moving forward in SAD analyses, we do need to do a better job at exploring error and variance, especially when samples are small.

We agree that this is an important consideration. We have implemented two resampling analyses to explore how nondetection of rare species, and sampling variability more generally, affect our results (see response to Reviewer 2, and lines 199-213, 388-398, and 421-424, and Appendix A7).

****