**Overview of response (cover letter content?)**

[We are writing to resubmit the article … for consideration as a Review & Synthesis paper…]

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 magnitudeof 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.

**New results overview (for Morgan and Hao)**

**Rarefaction resampling**

Rationale: Sampling in the field often results in an underestimation of species richness, especially of the richness of rare species, because rare species are likely to escape detection. We explore how non-detection of rare species affects our results by calculating the estimated true number of species and adding any “missing” species as rare species with abundance equal to 1.

Specific methods: We estimate the true number of species by computing the bias-corrected Chao and ACE estimates for an observed SAD (as implemented in vegan::estimateR), adding one standard error to each estimate, and taking the mean. This yields a generous estimate of species richness. If the estimated species richness exceeds the sampled number of species, we add the missing species each with abundance 1. We then run the adjusted SADs through the computational pipeline as normal.

Results: Adjusting in this fashion consistently increases the deviation between the observed SAD and its feasible set. Compared to raw SADs, adjusted SADs are even more skewed, less even, have lower Shannon diversity, have greater absolute deviations from the rest of the feasible set and have a larger proportion of rare species than expected given their feasible sets. These results are not surprising, because increasing the number of rare species without changing the rest of the distribution naturally increases unevenness, etc. See <https://github.com/diazrenata/scadsanalysis/blob/feb-rev/analysis/rev_prototyping/rarefaction.md>

**Jacknife resampling**

Rationale: There may also be sampling error in the abundances of all the observed species in a sample, not just the rare ones. To explore how variability across the entire distribution affects our results, we resample a subset of the observed individuals and analyze these resampled SADs.

Specific methods: For an observed SAD, we generate 10 resampled SADs by randomly sampling 60% of the individuals in the observed sample without replacement and without regard to species. We use 60% of the observed number of individuals to introduce substantial room for sampling variability without reducing the size of most communities (and, accordingly, the size of their feasible sets) beyond the realms where we believe our approach to be appropriate. We run each resampled SAD through the computational pipeline. For each observed SAD, we take the mean summary statistic – percentile score, etc – across the 10 resamples as the “resampled” statistic. Because resampling in this fashion increases computational effort up to tenfold, we used this resampling procedure on the entirety of the Mammal Communities, Misc. Abundance, and Gentry datasets, but only on a random sample of 300 (of ~2000) BBS sites and 2000 (of ~20,000) FIA sites.

Results: The proportion of extreme results is generally lower in resamples than in the raw values, but resampling does not change the direction or the significance of the effect except for the number of singletons for MCDB and BBS. See <https://github.com/diazrenata/scadsanalysis/blob/feb-rev/analysis/rev_prototyping/jacknife_results.md>

**“Proportion off”**

Rationale: Reviewer 2 asked about effect size, which doesn’t make sense to calculate for metrics like skewness and evenness, because they vary widely in their range, mean, standard deviation, etc. across large variation in S and N. We can compute a coarse metric of difference, the proportion of individuals allocated to species of different abundances across two SADs. We can then ask whether observed SADs are more different from their FS than elements of the FS are from each other, and by how much more.

Specific methods: See GitHub for an illustration of the metric. It’s the area of difference between two SADs divided by the total area under one SAD. It ranges from 0-1. I think of it as the proportion of individuals “mis-allocated” between two SADs. An increase of .1 means that one pair of SADs has an additional 10% more of the total N allocated to different species than does another pair.

Results: Observed SADs are consistently more different from the FS than elements of the FS are from each other. We can report an estimate of how much - for example, among communities in the MCDB for which we find a significant deviation in the dissimilarity metric, the score for the observed SAD is on average 0.1 higher than the 95th percentile of scores for elements of the feasible set. See <https://github.com/diazrenata/scadsanalysis/blob/feb-rev/analysis/rev_prototyping/additional_metrics_results.md#proportion-off>

**Number of rare (abundance = 1) species**

Rationale: This is an attribute that ecologists, and specifically ecologists who study SADs, care about in particular. It’s more intuitive to interpret than “an increase in skewness”, and one can calculate an effect size off of it. Reviewer 3 suggested it.

Specific methods: Calculate the number of singletons in each draw from the feasible set and compare this distribution to the number of singletons observed. You can also calculate how many more (or fewer) singleton species are in the observed relative to sampled vectors.

Results: BBS, MCDB, and Misc. Abund have higher numbers of rare species than their feasible sets, although the difference is weak for BBS (7%). FIA has no effect. Gentry has dramatically too **few** rare species relative to the feasible set.

This could be because Gentry also has a very low ratio of N/S. It also behaves oddly for other metrics (Shannon, skewness, and evenness).

The effect also gets stronger with increasing N/S – if you filter to communities with N/S > 10, you find significant deviations across the board.

BBS has on average about an increase of 6% in the % of singleton species, compared to the 95th percentile of the feasible set, for communities with a deviation. MCDB and Misc. Abund average 20% and 16%. See <https://github.com/diazrenata/scadsanalysis/blob/feb-rev/analysis/rev_prototyping/additional_metrics_results.md#number-of-singletons>

**Shannon diversity**

Rationale: Reviewer 2 specifically mentioned Shannon’s as an alternative diversity metric that weights rare vs. common species differently than Simpson’s evenness.

Specific methods: Calculate Shannon diversity and analyze it the same as evenness or skewness.

Results: Shannon diversity is consistently low across all datasets. Interestingly, 30% of Gentry sites are also extremely *high* in Shannon diversity. These tend to be the ones with the lowest ratio of N to S. This also shows up in Gentry results for skewness, evenness, and the number of singletons – it reverses the trend seen across most other datasets as N gets closer to S. See <https://github.com/diazrenata/scadsanalysis/blob/feb-rev/analysis/rev_prototyping/additional_metrics_results.md#shannon-diversity>

**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 can visually compare the maximum likelihood lognormal, Poisson, and exponentials to a handful of feasible sets and comment – broadly – as to whether they appear to be more or less even than the feasible set samples. I think there is a strong need to do the sampling to say anything more. My suggestion is to make the rough comparisons, explain why the sampling is necessary for more speculation, and say that this is a clear next step for this work.

My thoughts on why the sampling is necessary (briefly): 1) though some sets of assumptions converge to functional forms, comparing the “in the limit” functional approximation isn’t necessarily appropriate given the way we see the feasible set varies over S and N and the fact that many or most ecological communities may not be “in the limit”; 2) the functional approximation describes only the central tendency and not the underlying distribution of forms; this underlying distribution is critical to detecting deviations with confidence; 3) you can’t directly compare the goodness-of-fit of the functional approximations to the deviation from the feasible set. (The issue is that if you sample from a functional approximation, like a lognormal, most of your samples will not have N and S equal to the original SAD. If you then filter to just the ones that do, you’ve actually filtered the distribution to the area of best-overlap between the theoretical and the actual distribution.) There is a strong need to actually implement the sampling if you are to get comparable results.

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?

Yes, and I think there’s room to talk about this more. This is most pronounced for a subset of communities from Gentry where N/S is very low. These communities are the most species rich in our data and have an N/S ratio up to about 3. These communities often show the reverse trend of that seen for most of the other datasets: they are unusually *even*, have too *few* rare species, etc. While we can’t rule out that this is due to something specific to Gentry, I think this is an opportunity to talk more about the ratio of N to S as another statistical attribute that affects the feasible set and in turn may affect our expectations and what we should find surprising about what we observe in nature.

Adding this will probably shift some emphasis from “how the feasible set differs for large v. small communities” closer to “how the attributes of the feasible set modulate if, and what sorts of, deviations we detect” and bolsters an argument for explicitly considering statistical baselines for the communities and community attributes one is interested in.

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.)

Sure!

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!

Absolutely.

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)?

Good catch, the actual minimum is 4. (I think those numbers were left from when we were talking about pre-sampling and post-sampling filtering separately).

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.

This is a very good question; we have implemented two resampling strategies and discuss the results. Generally, accounting for nondetection of rare species *strengthens* deviations, while resampling subsets to explore sampling error across the whole distribution weakens, but does not wholly obscure, the trends we find for the full datasets. See the new results (attached/below).

(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.

It doesn’t make sense to calculate an effect size for skewness or evenness (or Shannon diversity), because the range/mean/standard deviation for these metrics varies over large changes in S and N. This is why we use the percentile score to make our comparisons. I think this comment indicates that we should explain this more clearly in the text.

We have also added additional metrics that do allow for an estimate of the magnitude of the deviation: the proportion of rare (abundance = 1) species, and a metric of dissimilarity between two SADs that does not depend on S and N. Where these metrics show deviations between the observed SADs and the feasible set, we note the magnitude of these deviations. See the new results section.

Finally, the comment about sample size – and the concern about carpetbombing with a large number of datasets, above - may indicate a need to be more clear about what statistical significance means in this context. For a single community, a percentile score of 99 means that that community is *very* unlikely to have emerged at random from the feasible set (but there’s still a 1% chance, which is why we ask whether we are finding extreme values too often over large numbers of communities). Having a very large community may increase our confidence that even a small absolute deviation is highly unlikely, because we have greater confidence in the *expected* value. In this context, even a small shift is meaningful because it would be very hard to achieve at random. However, having a large *number of communities* does not increase our odds of finding deviations if there are none or if the communities are too small to make distinctions. This is, in fact, what we think is going on in FIA – there are 20,000 communities, and no signal!

(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.

We have added Shannon’s, and two more metrics. The results are generally robust across metrics.

(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!

Absolutely, and well taken. I take this as a nudge to reorient the introduction and end of the discussion to draw more deeply on complexity science/statistical mechanics as it relates to macroecology and the SAD. I think this will work well with the additional space and possible adjustment in framing for a “Synthesis” paper (rather than a Letter).

A possible narrative for start of the intro:

* The macroecological approach to community ecology synthesizes observations from numerous communities to identify common phenomena that may reflect general processes
* The SAD is one of the most widely studied phenomena in this space, but efforts to use it to diagnose biological process have faltered due to the ubiquity of the hollow curve
* As has been observed for decades, there’s a strong possibility that mathematical factors influence the emergent form of the SAD.
* This invites us to borrow thinking from statistical mechanics to frame our expectations for the form of the SAD and what we find surprising about nature.
* This brings the focus to detecting and explaining deviations between observation and what emerges from solely mathematical processes.

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).

This is well taken, and I think a nudge to add both new results and additional points in the discussion.

New results:

* Adding an analysis of the proportion of rare (abund = 1) species. Generally, there are too many rare species in observed distributions (but only weakly so for BBS, and Gentry has too few!) This is a more intuitive finding, in ecological terms, than “a general increase in skewness”.

New discussion points:

* Adding the number of rare species for something more ecologically tangible
* The commentary that the patterns for the number of rare species, and several other metrics, reverse for Gentry, and specifically for Gentry when N/S gets small. The data do not support a lot of inference here, but there’s room for commentary about what might be odd about Gentry or these communities. Speciose communities might be more prone to nondetection of rare species? Something about these communities – tropical forests - might be forcing high evenness?
* Emphasizing that this approach is probably most powerful if you have a specific attribute in mind – such as the number of rare species – which you can then compare directly between the feasible set and the observed SAD. Five metrics, including holistic ones like skewness and evenness and targeted ones like the number of rare species, giving consistent results is a strong indication that observations are unlike the random expectation, but there is nuance in which metrics deviate how strongly and for which communities. If you have one in mind, you can be more targeted and concrete in your interpretation.

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.

A figure showing how the size of the feasible set corresponds to the breadth index is in the supplement (S6), but we can move it to the main text. The Synthesis allows for up to 10 figures.

I think in doing so there will need to be some discussion of how to interpret it. There is a decline, but there isn’t a clean threshold. I think this is in part because the size of the feasible set derives from both S and N, and that the ratio matters in addition to the actual size of the FS, and in part because we have an uneven and truncated array of S and N represented in our datasets (and indeed in what we can reasonably sample). *But*, even absent these considerations, I think it’s inherently a messy trend. Hence talking about it in general, rather than “hard threshold”, terms.

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.  
  
I forgot to link to it in the ms! Happy to, obviously.

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.

Space permits!

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

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.

I agree. I suggest reframing the intro (see response to R2) and including the number of singletons to talk more about the ecological interpretation of these results in the results and discussion (see response to R3).

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.

Agreed, and done! See response to R2 and new results from resampling.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*