1 **Title:** Empirical abundance distributions are more uneven than expected given their statistical baseline

2 **Running title:** SADs deviate from statistical baselines

3 **Author names and affiliations:**

4 Renata M. Diaz*[1], Hao Ye[2], S. K. Morgan Ernest[3]

5 [1] School of Natural Resources and Environment, University of Florida, Gainesville, Florida, USA.
6 renata.diaz@weecology.org; *corresponding author

7 [2] Health Science Center Libraries, University of Florida, Gainesville, Florida, USA. haoye@ufl.edu

8 [3] Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, Florida, USA.
9 skmorgane@ufl.edu

10 **Address for correspondence:**

11 Renata M. Diaz
12 110 Newins-Ziegler Hall
13 PO Box 110430,
14 Gainesville, FL 32611-0430
15 renata.diaz@weecology.org
16 Phone: (352) 846-0643
17 Fax: (352) 392-6984

18 **Statement of authorship:** RMD and SKME conceived the analysis; HY devised the algorithm to sample the
19 feasible set, reviewed the coded implementation, and wrote the explanatory vignette; RMD conducted
20 the analyses and wrote the first draft of the manuscript; all authors contributed substantively to revisions.

21 **Data accessibility statement**: All data used are available publicly via Zenodo and figshare. Upon
22 publication, all code and data will be archived and made publicly available via Zenodo.

23 **Keywords:** Species abundance distributions; feasible set; combinatorics; macroecology; constraints

24 **Conflict of interests:** The authors declare no conflicts of interest.

25 **Type of article**: Letter

26 **Word counts:**

27 Abstract: 147
28 Main text: 4195
29 No text boxes

30 **Number of references:** 25

31 **Number of figures, tables, and text boxes:** 4 figures; 1 table; 0 text boxes

32

33   Abstract

34   The prevalence of the species abundance distribution's hollow-curve shape across many communities is

35   frequently assumed to reflect ecological processes structuring communities. However, this hollow curve

36   can also emerge as a statistical phenomenon of dividing a particular number of individuals into a given

37   number of species. While the hollow curve may be a statistical artefact, ecological processes may induce

38   subtle deviations between empirical species abundance distributions and their statistically most probable

39   forms. Examining ~22,000 communities, we found that empirical species abundance distributions are

40   more skewed and uneven than their statistical baselines. However, small communities – with few species

41   or individuals – exhibit poorly-resolved statistical baselines, thereby reducing our capacity to detect

42   deviations. The extraordinarily skewed and uneven nature of empirical species abundance distributions

43   provides new avenues for testing ecological theory, while the issues posed by small communities

44   illustrate the limitations of statistical baselines for studying ecological patterns in small samples.

Introduction

46   The distribution of how the total number of individuals in a community are divided among the species in

47   that community, or the species abundance distribution (SAD), is one of the few ecological patterns whose

48   shape is so consistent that it is often considered an ecological law (Lawton 1999, McGill et al. 2007).

49   Across varied ecosystems and taxa, the species abundance distribution is dominated by a few very

50   abundant species and a larger number of increasingly rare species, generating a distinctive hollow- or J-

51   shaped curve (Fisher et al. 1943). Community ecologists have used the SAD to test numerous theories

52   intended to determine which biological processes are most important for structuring assemblages of

53   species, by comparing theoretical predictions for the SAD to observed SADs (McGill et al. 2007). However,

54   this approach has proven inconclusive because many theories predict similar shapes for the SAD (McGill

55   et al. 2007), and even experimental manipulations generate little variation in the shape of the SAD (Supp

56   and Ernest 2014). After decades of attention to its consistent pattern and potential as a theoretical

57   benchmark, the utility of the SAD for assessing the processes structuring ecological communities remains

58   unclear.

59   Accumulating evidence suggests that statistical constraints may actually generate the most striking

60   feature of the species abundance distribution – the hollow curve (White et al. 2012, Locey and White

61   2013, Frank 2019). Indeed, power-law or log-series distributions (i.e. hollow curves) emerge from many

62   distinct generative processes for abundance distributions generally (Frank 2009, Frank 2019). In the

63   specific case of the SAD, approaches using both statistical mechanics (i.e. the Maximum Entropy Theory

64   of Ecology (METE); Harte et al. 2008, Harte 2011) and combinatorics (i.e. 'the feasible set'; Locey and

65   White 2013) generate realistic hollow curves via the random division of the total number of individuals in

66   a community, N, into the total number of species present, S. Given that these stochastically generated

67   SADs are excellent empirical fits to those seen in nature (Harte 2011, White et al. 2012, Locey and White

68   2013), failing to account for these considerations may have led us to focus on a distinctive but relatively

69   uninformative aspect of the SAD.

70   If SADs are statistically inclined to be hollow curves, it is no surprise that we have struggled to use the

71   hollow curve to definitively identify the ecological processes that shape SADs. However, this does not

72   necessarily mean that the SAD cannot be biologically informative. Biological factors may introduce subtle,

73   but meaningful, deviations between the shapes of observed SADs and the shapes of the SADs expected

74   due to the mathematical constraints imposed by S and N, which we hereafter refer to as the "statistical

75   baseline" (Locey and White 2013, Harte and Newman 2014). Thus, it may be possible to use these

76   *deviations* to detect strong ecological processes or evaluate theories (Harte and Newman 2014, Xiao et al.

77   2016). If the vast majority of mathematically achievable SADs for a community share a similar shape, an

78   empirically observed SAD that deviates even slightly from this statistical baseline is unlikely to have

79   emerged at random (Locey and White 2013). Such a deviation is therefore likely to be the signature of a

80   non-random – i.e., biological – process operating on the relative abundances of species (Harte and

81   Newman 2014). It may be possible to evaluate ecological processes and theories based on how well they

82   predict these *deviations* between observed SADs and their statistical baselines.

83   Successfully interpreting SADs in this fashion depends on our capacity to detect and quantify deviations

84   between empirical observations and statistical baselines, which requires metrics and computational

85   approaches that allow us to quantify and interpret whatever deviations may exist. Here, we build upon

86   the combinatoric approach developed by Locey and White (2013) to define and explore the statistical

87   baselines for SADs. For a given N (total number of individuals) and S (total number of species), there exists

88   a finite (but potentially very large) set of possible distributions of individuals into species. Collectively, this

89   set of possible SADs is referred to as the feasible set, with each possible SAD constituting a single element

90   of the set. If an observed SAD is drawn at random from the set of mathematically possible SADs, it is likely

91   to have a shape similar to the shapes most common in the feasible set. The feasible set can therefore be

92   used as a statistical baseline for assessing whether observed SADs deviate from what is likely to occur

93   merely due to mathematical constraints, and to explore how the statistical baseline varies over across

94   ranges of values of S and N (Locey and White 2013).

95   The specificity, or vagueness, of the expectations derived from the statistical baseline is critically

96   important for disentangling the aspects of the SAD that are likely generated by statistical constraints from

97   those generated by other processes. If the vast majority of mathematically possible SADs are similar in

98   shape – generating a very specific, narrowly defined statistical baseline – then even small deviations

99   between an observed SAD and this baseline can signal the operation of ecological processes. However, if

100   many different shapes occur with more even frequency in the feasible set, the statistical baseline is less

101   specific and less well defined, and our sensitivity for distinguishing biological signal from statistical

102   constraints is greatly reduced. In general, a poorly defined statistical baseline is more likely to occur when

103   the size of the community, in terms of S and $N,$ is small, because in such cases there may be too few

104   possible SADs in the feasible set for a particular shape to emerge as the most common, and therefore

105   most likely, shape. When this occurs, we have reduced confidence that even an observation that deviates

106   from the statistical baseline did not emerge at random from the relatively restricted pool of possible

107   outcomes (Jaynes 1957). This general concern has been acknowledged in efforts to comparing ecological

108   observations to statistical baselines (Harte 2011, White et al. 2012, Locey and White 2013) but there has

109   not yet been a quantification of these effects for the SAD or an identification of the range of community

110   sizes most strongly affected. Because ecologists study the SAD for communities varying in size from the

111   very small – S and $N < 5$ – to the enormous – S and $N \gg 1000$ – identifying the community sizes for which

112   we can and cannot confidently detect deviations from the statistical baseline is necessary to appropriately

113   contextualize our interpretations.

114   Here we use the feasible set to define statistical baselines for empirical SADs for 22,000 communities of

115   birds, mammals, trees, and miscellaneous other taxa. We then compare *observed* SADs to their

116   corresponding statistical baselines and evaluate 1) if the shapes of observed SADs consistently deviate

117   from their statistical baseline, 2) how the specificity of the statistical baseline varies over ranges of S and

118   N, and 3) whether this variation appears to be associated with variation in our capacity to detect

119   deviations between observations and the corresponding baselines.

120   **Methods**

121   *Datasets*

122   We used a compilation of community abundance data for trees, birds, mammals, and miscellaneous

123   additional taxa (White et al. 2012, Baldridge 2015, Baldridge 2016, data from Baldridge 2016). This

124   compilation consists of cleaned and summarized community abundance data for trees obtained from the

125   Forest Inventory and Analysis (Woudenberg et a al 2010) and Gentry transects (Phillipes and Miller 2002),

126   birds from the North American Breeding Bird Survey (Sauer et al. 2013), mammals from the Mammal

127   Community Abundance Database (Thibault et al. 2011), and a variety of less commonly sampled taxa

128   from the Miscellaneous Abundance Database (Baldridge 2015). Because characterizing the random

129   expectation of the SAD is computationally intractable for very large communities, we filtered our datasets

130   to remove communities with more than 40720 individuals, which was the largest community we

131   successfully analyzed. This resulted in the removal of 4 communities from the Miscellaneous Abundance

132   Database. We further filtered the FIA database. Of the 103,343 communities in FIA, 92,988 have fewer

133   than 10 species. Rather than analyze all these small communities, we randomly selected 10,000 small

134   communities to include in the analysis. We also included all FIA communities with more than 10 species,

135   which added 10,355 FIA communities to the analysis and resulted in a total of 20,355 FIA communities.

136   Finally, for sites that had repeated sampling over time, we followed White et al. (2012) and Baldridge

137   (2016) and analyzed only a single, randomly selected, year of data, because samples taken from a single

138   community at different time points are likely to covary. It should be noted that our analyses include data

139   from the Mammal Community Database and Miscellaneous Abundance Database that were collected

140   over longer timescales and cannot be disaggregated into finer units of time. We also removed from our

141   analyses any communities with only one species, or for which N = S or N = S + 1, because these

142   communities have only one mathematically possible SAD. Our final dataset consisted of ~22,000

143   communities with S and N ranging from 2 to 250 and 3 to 40714, respectively (see Figure S1 in Supporting

144   Information). Details and code for the filtering process can be found in Appendix S2.

145   *Generating the statistical baseline*

146   We use the concept of the "feasible set" to establish a statistical baseline for the SAD (Locey and White

147   2013). For a given number of individuals *N*, there is a finite number of unique ways to partition those

148   individuals into *S* species. The complete set of these unique partitions is the feasible set. Because, in this

149   approach, neither species nor individuals are distinguishable from each other, partitions are unique if and

150   only if they differ in the number of species that have a particular abundance (Locey and White, 2013).

151   Operationally, this means that for *S = 3* and *N = 9*, the species abundances *(1, 3, 5)* and *(2, 2, 5)* count as

152   distinct partitions, but *(1, 3, 5)* and *(3, 1, 5)* are only one element of the feasible set because they each

153   contain one species with an abundance 1, 3, and 5, respectively, and they differ only in the *order* of the

154   numbers. In the absence of justification for additional assumptions regarding the distinguishability of

155   species and/or individuals, we adopted this simple set of assumptions that has previously been shown to

156   generate realistic statistical baselines (Locey and White 2013).

157   While it is possible to list all possible partitions in the feasible set for small *S* and *N*, the size of the feasible

158   set increases rapidly with *S* and *N*. Therefore, characterizing the statistical properties of the feasible set

159   for large *S* and *N* can be computationally intensive. This renders it necessary to draw samples from the

160   feasible set, rather than enumerating all of its elements. Unbiased sampling of large feasible sets is a

161   nontrivial computational problem that has constrained previous efforts in this vein (Locey and White

162   2013). We developed an algorithm to efficiently and uniformly sample feasible sets even for large values

163   of S and N. In brief, the algorithm takes a generative approach to sampling the feasible set. Individuals are

164   allocated one species at a time, beginning with the least abundant species. At each step, the number of

165   individuals to allocate for the current species is determined at random, with probability based on the

166   number of feasible sets with that specific abundance for that species, and conditional on the individuals

167   that have already been allocated. For example, if we have 3 species and 7 individuals, the least abundant

168   species can have an abundance of 1 or 2. Allocating 1 individual to the least abundant species allows for

169   the SADs (1, 1, 5), (1, 2, 4), and (1, 3, 3), and allocating 2 individuals to the least abundant species means

170   the only possible SAD is (2, 2, 3). We therefore allocate 1 individual with probability 3/4, and 2 individuals

171   with probability 1/4. If, at the first step, we allocated 1 individual to the least abundant species, the

172   second species can have an abundance of 1, 2 or 3 with equal probability, because there is exactly 1

173   possible SAD with each of these possible abundances for the first two species. This process is continued

174   until all individuals have been allocated. We implemented this algorithm in an R package, available on

175   GitHub at www.github.com/diazrenata/feasiblesads; details of the sampling methodology are available in

176   Appendix S3.

177   For every community in our database, we drew 4000 samples from the feasible set to characterize the

178   distribution of statistically probable shapes for the SAD. We filtered the 4000 samples to unique

179   elements. For small values of S and N, it can be impossible or highly improbable for the 4000 samples

180   from the feasible set to all be unique, but for large communities, all 4000 are usually unique. We refer to

181   this as the sampled feasible set.

182   *Comparing observed SADs to their baselines*

183   If all SADs in a feasible set are equally likely to occur, then an SAD with a particular S and N is likely to

184   have a shape similar to the shape that is most common among the SADs in the feasible set for the same S

185   and N; in contrast, strong processes may cause observed SADs to have shapes that deviate from this

186   statistical baseline (Locey and White 2013). We focus on two metrics to describe the shape of the SAD,

187   skewness and Simpson's evenness. Skewness measures the asymmetry of a distribution around its mean,

188   and Simpson's evenness is a commonly used metric in ecology for assessing how equitably abundance is

189   distributed across species. By calculating these metrics for each of the samples in the community's

190   sampled feasible set (see *Generating the statistical baseline*, above), we generated a distribution

191   describing the general shape (i.e. evenness or skewness) expected from the randomly sampled SADs.

192   Note that skewness, as implemented in the R package "e1071" (Meyer et al. 2019), always evaluates to 0

193   for distributions with only two species, and we therefore excluded those cases from analyses of skewness

194   (but included those communities for analyses using Simpson's evenness).

195   To assess whether the shape of an observed SAD was statistically unlikely, we calculated Simpson's

196   evenness and skewness for the observed SAD and compared these observed values to the distributions of

197   evenness and skewness obtained from that community's sampled feasible set. An observed SAD's

198   deviation from its feasible set was determined by computing the percentile rank of its skewness and

199   evenness relative to the sampled distributions for skewness and evenness, respectively. These percentile

200   ranks are then comparable across different community sizes, allowing broad-scale assessment across

201   wide ranges of S and N. After aggregating across communities, if observed SADs reflect random draws

202   from their feasible sets, their percentile rank values should be uniformly distributed from 0 to 100.

203   However, if observed SADs are consistently more skewed or even than their feasible sets, the percentile

204   values will be disproportionately concentrated towards the extremes. Because an earlier survey in this

205   space (Locey and White 2013) found that the tendency is for empirical SADs to be more skewed and less

206   even than their feasible sets, we used one-tailed 95% confidence intervals and tested for unusually *high*

207   values for skewness and *low* values for evenness. This comparison is not meaningful if there are very few

208   unique values in the distributions of skewness and evenness, which can occur for small feasible sets. We

209   therefore excluded communities for which the distribution of skewness or evenness values from the

210   sampled feasible set had fewer than 20 unique values (in these cases, it is impossible for an observation

211   to fall above or below the 95th or 5th percentile, respectively). Our final aggregated analyses included

212   22,142 communities for evenness and 22,325 communities for skewness.

213   *The narrowness of the expectation*

214   We also used the distributions of skewness and evenness from the sampled feasible set to describe the

215   relative specificity of the statistical baseline, in order to assess in what situations there could be

216   challenges in determining whether observed communities differ from their statistical baselines. We

217   quantified the narrowness of a distribution as the ratio of the range of values encompassed within a 95%

218   density interval relative to the full range of values in the distribution (Figure 1). This breadth index for the

219   statistical baseline ranges from 0 (a very narrow distribution and well-resolved baseline) to 1 (a very

220   broad distribution), and is comparable across feasible sets for varying combinations of S and N. This

221   metric corresponds qualitatively to more computationally-intensive approaches to measuring the self-

222   similarity of the elements of feasible sets (see Appendix S4).

223   **Results**

224   *Observed SADs compared to their feasible sets*

225   For four of the five datasets we analyzed – BBS, Gentry, Mammal Communities, and Misc. Abund –

226   empirical SADs are highly skewed and highly uneven relative to their feasible sets, much more frequently

227   than would be expected by chance (Figure 2, Table 1). Combined across these four datasets, 16% of

228   observed SADs are more skewed than 95% of their feasible sets, and 31% are less even than 95% of their

229   feasible sets. For SADs randomly sampled from the feasible set, we would expect only 5% of observed

230   distributions to fall in these extremes. In contrast to the other datasets, the SADs from the FIA dataset

231   exhibit percentile scores that are more uniformly distributed: 5% of observations are more skewed than

232   95% of their feasible sets, and 9% of observations are less even than 95% of their feasible sets.

233   *The narrowness of the expectation*

234   The ability to detect deviations from the statistical baseline depends on the distribution of SADs in the

235   feasible set. Here, the statistical baseline for both skewness and evenness becomes more narrowly

236   defined as the size of the feasible set increases (Figure 1; Figure S5), making even small deviations in

237   skewness or evenness statistically meaningful and readily detectable. However, for communities with

238   relatively small feasible sets – fewer than approximately 1000 elements for skewness, and approximately

239   200 elements for evenness – the breadth index approaches 1, meaning that a 95% density interval of the

240   values in the distribution spans nearly the entire range of values (Figure S5). In particular, the FIA dataset

241   is dominated by small communities for which the breadth index is very high, reflecting relatively broad

242   and nonspecific statistical expectations for the shape of the SAD derived from the feasible set (Figure 3).

243   **Discussion**

244   We found widespread evidence that SADs for a range of real ecological communities are more skewed

245   and less even than expected given the distribution of shapes within their feasible sets. These deviations

246   may signal that ecological processes operate on top of statistical constraints, thereby driving the SAD

247   away from common shapes that would be observed in the absence of a dominating non-statistical

248   process. Our results suggest that the prevailing processes structuring these communities tend to be those

249   that cause abundance distributions to be more uneven – rather than those that produce more even

250   abundances across species. Ecological processes may lengthen the rare tail of the SAD, for example by

251   promoting the persistence of rare species at very low abundances (e.g. Yenni et al. 2012). Or, they could

252   drive abundant species to have larger populations that would be statistically expected, without driving

253   other species entirely to extinction (Chesson 2000). Although a disproportionate number of communities

254   deviated statistically from their feasible sets, there were also many communities for which we did not

255   detect deviations. In such cases, multiple ecological processes may operate simultaneously and with

256   countervailing impacts on abundance distributions, resulting in no dominating net effect on the shape of

257   the distribution beyond that imposed by fundamental constraints (Harte 2011; Harte and Newman 2014).

258   Going forward, testing whether ecological theories or common functional approximations (e.g. the log-

259   normal distribution) accurately predict this range of variation in deviations between observed SADs and

260   their statistical baselines may be much more fruitful than focusing only on the general form of the SAD

261   that emerges from statistical constraints (McGill et al. 2007; Locey and White 2013).

262   Unlike the other four datasets, communities in the FIA dataset showed weak or no evidence of deviations

263   from their feasible sets. These results may reflect statistical phenomena related to community size. The

264   FIA communities are by far the smallest across our datasets (Figure S1). Communities with small values of

265   S and N have smaller feasible sets; when there are relatively few possible SADs, the distributions of

266   evenness and skewness values from the feasible set are less narrowly peaked, meaning there is a weaker

267   statistical distinction between "common" and "extreme" shapes for the SAD (Figure 1). In fact, across the

268   range of community sizes present in our datasets, the feasible sets for small communities generally

269   generated broader distributions of evenness, and especially skewness, than those for large communities

270   (Figure S5). For such communities, the deviations – or lack thereof – that we perceive are less

271   informative than for larger communities with more strongly defined statistical baselines (Jaynes 1957).

272   If the lack of discernable deviations from the feasible set for the FIA communities is indeed a byproduct of

273   their generally small size, then we would expect similarly-sized communities from other datasets to have

274   similar results. We identified 371 communities from other datasets with S and N matching communities

275   from FIA. We found no difference in the distribution of percentile scores between communities from FIA

276   and communities from other datasets (Figure 4; Table S6), confirmed via Kolmogorov-Smirnov tests (for

277   evenness, $D = 0.04$ and $p = 0.87$; for skewness, $D = 0.07$ and $p = 0.37$). Although 371 communities

278   constitute a small sample relative to the 20,355 FIA communities we analyzed, these results point to

279   community size, and not attributes specific to FIA, as a likely explanation for the weak evidence for

280   deviations across the full FIA dataset.

281   If this is indeed the case, it means that small-community considerations may affect our capacity to

282   meaningfully distinguish signal from randomness using this approach. FIA communities, with their broad

283   distributions of shape metrics and overall lack of detectable signal, have on the order of 10 species and

284   50-100 individuals. While these values do not constitute hard thresholds, they may indicate a general

285   range of values below which we have relatively diminished power to detect deviations from the statistical

286   baseline represented by the feasible set. To meaningfully draw inferences from deviations in these small

287   communities, we will likely need more sensitive metrics (than skewness and evenness), and/or theories

288   that generate more specific predictions for the SAD. In the absence of such, we may stand to learn the

289   most by focusing on SADs from relatively large communities.

290   It is also important to recognize that there are multiple plausible approaches to defining a statistical

291   baseline for the SAD, of which we have taken only one (Haegeman and Loreau 2008, Locey and White

292   2013). Our approach follows Locey and White (2013) and reflects the random partitioning of individuals

293   into species, with the resulting distributions considered unique if the species' abundance values are

294   unique, regardless of the order in which the values occur. The philosophy behind the feasible set reflects

295   a longstanding approach in the study of abundance distributions: to focus on the shape of the distribution

296   without regard to species' identities (McGill 2007). To include differences in *order* in the statistical

297   baseline would imply that identifying *which* species contain the most or least individuals is important for

298   evaluating theory. Other formulations for the statistical baseline may be equally valid and generate

299   different statistical expectations, including forms that approximate exponential, Poisson, or log-series

300   distributions (Harte et al. 2008, Favretti 2018). Comparing the results that emerge from different

301   baselines will be an important next step towards reinvigorating the use of the SAD as a diagnostic tool.

302   Our study demonstrates the utility, and the potential challenges, of applying tools from the study of

303   complex systems and statistical mechanics to the study of ecological communities (Haegeman and Loreau

304   2008, Harte 2011, White et al. 2012, Harte and Newman 2014). While concepts such as maximum

305   entropy and the feasible set are promising new horizons for macroecology, the small size of some

306   ecological communities may present difficulties that do not occur as often in the domains for which these

307   tools were originally developed (Jaynes 1957, Haegeman and Loreau 2008). When the observed numbers

308   of species and individuals are too small to generate highly resolved statistical baselines, these approaches

309   will be less informative than we might hope – as appears to be the case for the smallest communities in

310   our analysis. In larger communities, where mathematical constraints have stronger effects on the general

311   form of the SAD, our results show that these constraints alone do not fully account for the extremely

312   uneven SADs we often observe in nature – leaving an important role for ecological processes. This ability

313   to detect and diagnose the specific ways in which empirical SADs deviate from randomness can generate

314   new avenues for understanding how and when biological drivers affect the SAD. There are, of course, still

315   many elements to be improved in our ability to distinguish biological signal from randomness, including

316   assessing alternative statistical baselines and calibrating our expected power to detect deviations,

317   especially for small communities. Indeed, more sensitive metrics could also enable identification of

318   processes that operate through time (note that, in this analysis, we sampled only one time point for each

319   community). Continuing to explore and account for the interplay between statistical constraint and

320   biological process constitutes a promising and profound new approach to our understanding of this

321   familiar, yet surprisingly mysterious, ecological pattern.

322

328

329

330   **References**

331   Baldridge, E. (2015). Miscellaneous Abundance Database. figshare. Available at:

332   https://doi.org/10.6084/m9.figshare.95843.v4

333   Baldridge, E., Harris, D.J., Xiao, X. & White, E.P. (2016). An extensive comparison of species-abundance

334   distribution models. *PeerJ*, 4, e2823.

335   Baldridge, E., Harris, D.J., Xiao, X. & White, E.P. (2016). Data from *An extensive comparison of species-*

336   *abundance distribution models*. Zenodo. Available at: https://zenodo.org/record/166725.

337   Chesson, P. (2000). Mechanisms of Maintenance of Species Diversity. *Annual Review of Ecology and*

338   *Systematics*, 31, 343–366.

339   Favretti, M. (2018). Remarks on the Maximum Entropy Principle with Application to the Maximum

340   Entropy Theory of Ecology. *Entropy*, 20, 11.

341   Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943). The Relation Between the Number of Species and the

342   Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology*, 12,

343   42–58.

344   Frank, S.A. (2009). The common patterns of nature. *Journal of Evolutionary Biology*, 22, 1563–1585.

345   Frank, S.A. (2019). The common patterns of abundance: the log series and Zipf's law. *F1000Res*, 8, 334.

346   Haegeman, B. & Loreau, M. (2008). Limitations of entropy maximization in ecology. *Oikos*, 117, 1700–

347   1710.

348   Harte, J. (2011). *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*.

349   Oxford University Press.

350   Harte, J. & Newman, E.A. (2014). Maximum information entropy: a foundation for ecological theory.

351   *Trends in Ecology & Evolution*, 29, 384–389.

352   Harte, J., Zillio, T., Conlisk, E. & Smith, A.B. (2008). Maximum Entropy and the State-Variable Approach to

353   Macroecology. *Ecology*, 89, 2700–2711.

354   Jaynes, E.T. (1957). Information Theory and Statistical Mechanics. *Phys. Rev.*, 106, 620–630.

355   Lawton, J.H. (1999). Are There General Laws in Ecology? *Oikos*, 84, 177.

356   Locey, K.J. & White, E.P. (2013). How species richness and total abundance constrain the distribution of

357   abundance. *Ecology Letters*, 16, 1177–1185.

358   McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., *et al.* (2007). Species

359   abundance distributions: moving beyond single prediction theories to integration within an

360   ecological framework. *Ecol Letters*, 10, 995–1015.

361   Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. (2019). *e1071: Misc Functions of the*

362   *Department of Statistics, Probability Theory Group (Formerly: E1071),* TU Wien. R package version

363   1.7-4. https://CRAN.R-project.org/package=e1071

364   Phillips, O. & Miller, J.S. (2002). *Global patterns of plant diversity: Alwyn H. Gentry's forest transect data*

365   *set*. Missouri Botanical Press.

366   Sauer, J.R., Link, W.A., Fallon, J.E., Pardieck, K.L. & Ziolkowski, D.J. (2013). The North American Breeding

367   Bird Survey 1966–2011: Summary Analysis and Species Accounts. *North American Fauna*, 1–32.

368   Supp, S.R. & Ernest, S.K.M. (2014). Species-level and community-level responses to disturbance: a cross-

369   community analysis. *Ecology*, 95, 1717–1723.

370   Thibault, K.M., Supp, S.R., Giffin, M., White, E.P. & Ernest, S.K.M. (2011). Species composition and

371   abundance of mammalian communities. *Ecology*, 92, 2316–2316.

372   White, E.P., Thibault, K.M. & Xiao, X. (2012). Characterizing species abundance distributions across taxa

373   and ecosystems using a simple maximum entropy model. *Ecology*, 93, 1772–1778.

374   Woudenberg, S.W., Conkling, B.L., O'Connell, B.M., LaPoint, E.B., Turner, J.A. & Waddell, K.L. (2010). The

375   Forest Inventory and Analysis Database: Database description and users manual version 4.0 for

376   Phase 2. *Gen. Tech. Rep. RMRS-GTR-245. Fort Collins, CO: U.S. Department of Agriculture, Forest*

377   *Service, Rocky Mountain Research Station. 336 p.*, 245.

378   Xiao, X., O'Dwyer, J.P. & White, E.P. (2016). Comparing process-based and constraint-based approaches

379   for modeling macroecological patterns. *Ecology*, 97, 1228–1238.

380   Yenni, G., Adler, P.B. & Ernest, S.K.M. (2012). Strong self-limitation promotes the persistence of rare

381   species. *Ecology*, 93, 456–461.

382

383   **Figure legends**

384   Figure 1. Large feasible sets may allow better detection of deviations from the statistical baseline by

385   generating more specific, narrowly-defined baselines. We illustrate this phenomenon using 3 hypothetical

386   communities: a small community ($S = 4$, $N = 34$; top row), an intermediate community ($S = 7$, $N = 71$;

387   middle row), and a large community ($S= 44$, $N = 13360$; bottom row). The large communiity has

388   approximately 6.59e+70 possible SADs in its feasible set, while the intermediate community has 60,289

389   and the small community has only 297. For every SAD sampled from the feasible set (left column), we

390   calculate the skewness (color scale) and evenness (not shown). The distributions of these values (right

391   column) constitute the statistical baseline. We define a "breadth index" as the ratio of the range

392   encompassed in the one-tailed 95% density interval (distance between red lines, right), compared to the

393   full range of values for the statistic (distance between the maximum and minimum values). As S and N

394   increase, the size of the feasible set increases, resulting in a narrower statistical baseline (smaller breadth

395   index) – thus enabling easier detection of deviations that may be the result of ecological processes

396   affecting the SAD.

397

398   Figure 2. Many ecological communities are more skewed (left) or uneven (right) than their statistical

399   baselines. Percentile ranks are calculated by comparing each community to its sampled feasible set, with

400   very high or very low percentile ranks reflecting extreme values relative to statistical baselines. The

401   vertical red line marks the 95th percentile for skewness and the 5th percentile for evenness. Species

402   abundance distributions that are sampled at random from the feasible set will produce percentile ranks

403   that are uniformly distributed from 0 to 100, with approximately 5% of values above or below the 95th

404   and 5th percentiles, respectively. In contrast, most datasets have more communities that are highly

405   skewed or uneven than would be expected by chance.

406

407   Figure 3. Breadth indices of skewness (left) and evenness (right) indicate varying ability to detect the

408   deviations between observations and the statistical baseline. The breadth index (see Figure 1) quantifies

409   how narrowly-defined the statistical baseline is; high values indicate broad, poorly-defined statistical

410   baselines that may impede our ability to confidently detect deviations between observations and what is

411   expected given the baseline. Most datasets contain a mixture of communities with broad and narrow

412   statistical baselines, but some – particularly the skewness baseline for the Forest Inventory and Analysis –

413   have consistently high breadth indices across all of their communities, suggesting that skewness is not an

414   effective metric for distinguishing empirical observations from the feasible set. In general, the breadth

415   index for evenness (right panels) indicates more narrow statistical baselines than those for skewness.

416

417   Figure 4. Small communities exhibit consistently broad statistical baselines (top), and consistently weak

418   evidence of deviations for observed SADs (bottom), regardless of the originating dataset. For a subset of

419   371 communities from the Forest Inventory and Analysis with communities in other datasets with

420   matching S and N, we generate distributions of breadth indices for skewness and evenness (top) and

421   compute corresponding percentile ranks for the observed SADs (bottom). Visually, there is no difference

422   between FIA (left panels) and other datasets (right panels), when they are matched in S and N. This is

423   confirmed by Kolmogorov-Smirnov tests for the breadth indices (for evenness, $D = 0.04$ and $p = 0.91$; for

424   skewness, $D = 0.03$ and $p > 0.99$) and percentile ranks (for evenness, $D = 0.04$ and $p = 0.87$; for skewness,
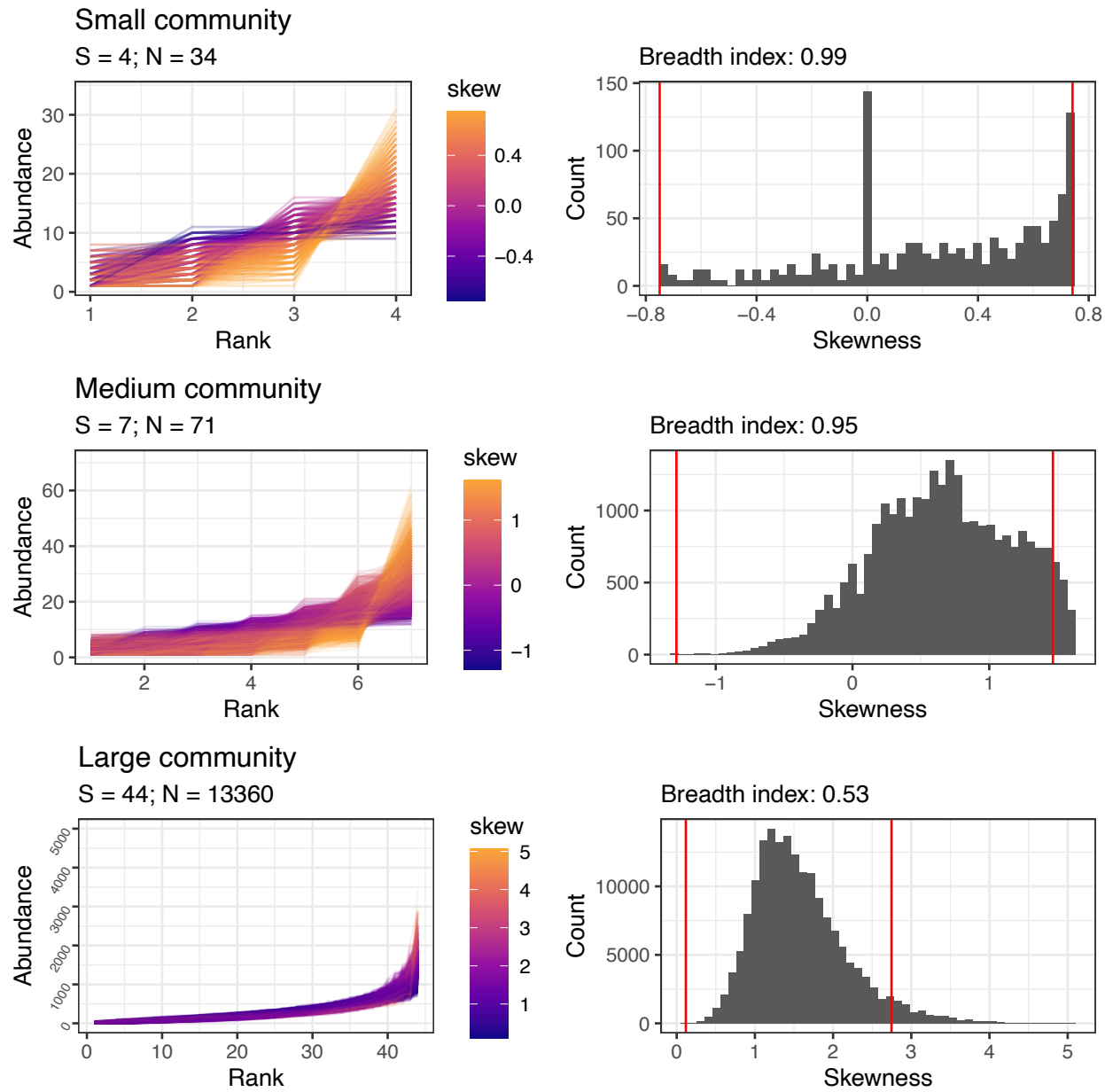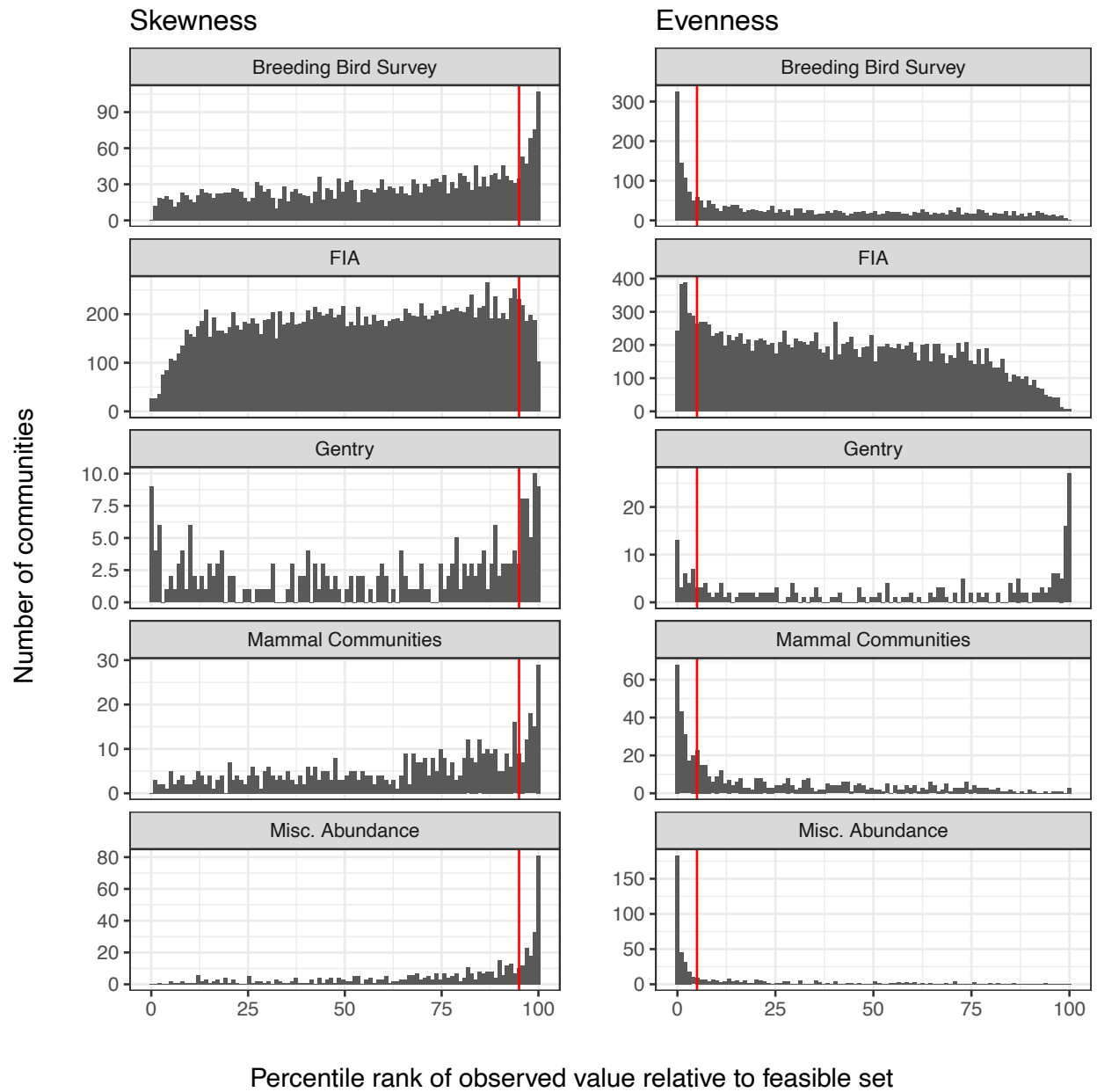
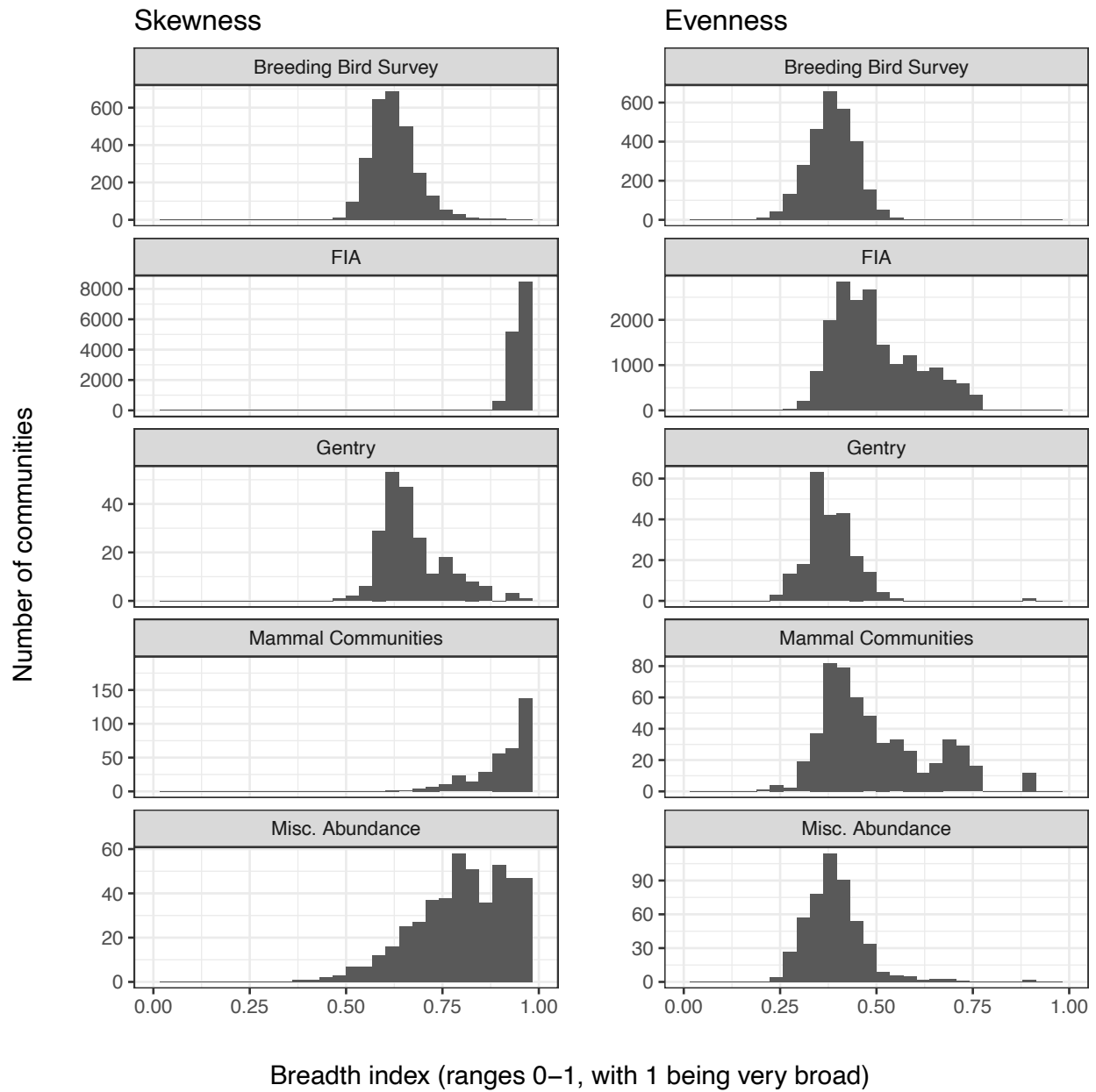425 $D = 0.07$ and $p = 0.37$).

426
427
428

429   **Figures and Tables**

430   **Figure 1.**

433   Figure 2.

434   Percentile rank of observed value relative to feasible set

435

436   Figure 3.

437   Breadth index (ranges 0–1, with 1 being very broad)

438

439   Figure 4.

Breadth index

Percentile rank
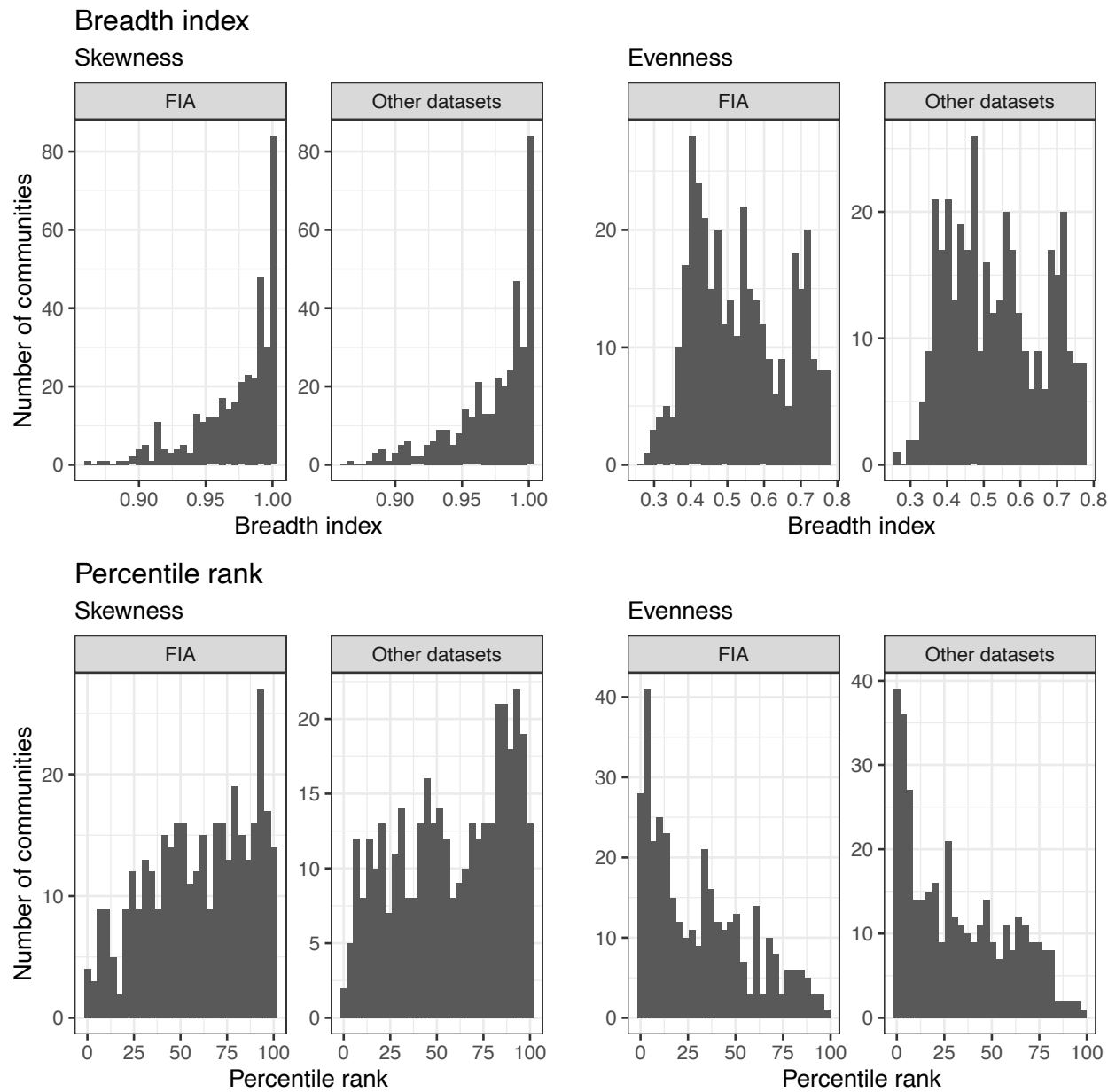

440

441

442   Table 1.

Dataset	Proportion of communities with skewness above 95th percentile	Proportion of communities with evenness below 5th percentile
Breeding Bird Survey	0.13 (N = 2773)	0.26 (N = 2773)
FIA	0.05 (N = 18300)	0.09 (N = 18113)
Gentry	0.19 (N = 223)	0.15 (N = 224)
Mammal Communities	0.16 (N = 537)	0.35 (N = 542)
Misc. Abundance	0.35 (N = 492)	0.6 (N = 490)

443   Table 1 Proportion of communities in each dataset with highly skewed or uneven SADs compared to their

444   sampled feasible sets. N, in parentheses, refers to the number of communities analyzed for each metric

445   for each dataset.

446