

Model averaging fails to improve the extrapolation capability of the island species–area relationship

Matthews, Thomas; Aspin, Thomas

DOI:

[10.1111/jbi.13598](https://doi.org/10.1111/jbi.13598)

License:

Other (please specify with Rights Statement)

Document Version

Peer reviewed version

Citation for published version (Harvard):

Matthews, T & Aspin, T 2019, 'Model averaging fails to improve the extrapolation capability of the island species–area relationship', *Journal of Biogeography*, vol. 46, no. 7, pp. 1558-1568.
<https://doi.org/10.1111/jbi.13598>

[Link to publication on Research at Birmingham portal](#)

Publisher Rights Statement:

Checked for eligibility: 26/06/2019

This is the peer reviewed version of the following article: Matthews, TJ, Aspin, TWH. Model averaging fails to improve the extrapolation capability of the island species–area relationship. *J Biogeogr.* 2019., which has been published in final form at: <https://doi.org/10.1111/jbi.13598>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1 Submission to: Journal of Biogeography

2

3 **Article Type: Research Paper**

4

5 **Model averaging fails to improve the extrapolation capability of the island species–area**
6 **relationship**

7

8 Thomas J. Matthews^{1,2} & Thomas W. H. Aspin¹

9

10 ¹GEES (School of Geography, Earth and Environmental Sciences) and the Birmingham
11 Institute of Forest Research, the University of Birmingham, Birmingham, B15 2TT

12

13 ²CE3C – Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity
14 Group and Univ. dos Açores – Depto de Ciências e Engenharia do Ambiente, PT-9700-042,
15 Angra do Heroísmo, Açores, Portugal.

16

17 *Correspondence: Thomas J. Matthews, School of Geography, Earth and Environmental
18 Sciences, University of Birmingham, Birmingham, B15 2TT, UK

19

20 Email: t.j.matthews@bham.ac.uk

21

22 Running header: ISAR extrapolation

23

24 Word count: abstract: 352 words; main text = 5742 words; 2 Tables; 2 Figures; 39 references;
25 2 appendices

26

27 **Keywords** extrapolation, habitat islands, multi-model inference, model averaging, power
28 model, species–area relationship, species richness

29

30 **ABSTRACT**

31

32 **Aim:** One of the main applications of the island species–area relationship (SAR) is to predict
33 species richness in areas of habitat too large to be sampled, but there are few clear guidelines
34 for choosing an appropriate model for this purpose. We therefore aimed to test whether a
35 multi-model averaging approach could improve the accuracy of predictions made by
36 extrapolating the ISAR. Specifically, we compared the performance of multi-model
37 averaging with that of the default ISAR model of choice, the power model, in predicting
38 species richness in large habitat islands.

39

38 **Location:** Global

39

39 **Taxa:** Vertebrates, invertebrates and plants

40

40 **Methods:** We removed the largest islands from 120 habitat island datasets, and fitted both
41 the power model and a multi-model average curve (averaging the predictions of up to 20
42 ISAR models) to this filtered dataset. We then assessed the accuracy of both approaches in
43 predicting the species richness of the largest island in the original dataset using the log error
44 of extrapolation (LEE) metric. A generalized additive regression modelling framework was

45 used to determine whether any dataset characteristics could explain variation in the LEE
46 values for the power model (LEE-POW).

47 **Results:** The power model gave the more accurate richness predictions for 58% of the
48 analysed datasets and the multi-model averaged curve gave the more accurate predictions for
49 the remaining 42%. Both the power models (61% of LEE-POW values were positive) and the
50 multi-model averaged curve (60% were positive) had a slightly greater tendency to over
51 predict the observed richness. The confidence intervals were also on average narrower for the
52 power model predictions (median 95% confidence interval width = 18 species) than for the
53 multi-model averaged curve predictions (median 95% confidence interval width = 78). The
54 range in island areas and richness values explained a small amount of the variation in LEE-
55 POW.

56 **Main conclusions:** Contrary to expectation, multi-model averaging was less accurate than
57 the power model in the majority of cases, and thus does not appear to be a panacea for
58 uncertainty in model choice when extrapolating the ISAR. However, further research is
59 urgently needed to evaluate the performance of a multi-model averaging approach at larger
60 spatial scales.

61 INTRODUCTION

62 The species–area relationship (SAR) describes the near-universally observed pattern whereby
63 the number of species increases with the area sampled (Rosenzweig, 1995; Tjørve & Tjørve,
64 2017). A number of different types of SARs have been described (Scheiner, 2003; Whittaker
65 & Fernández-Palacios, 2007), and these can be broadly split into island species–area
66 relationships (ISARs), whereby the number of species occurring within each of a set of
67 islands is analysed as a function of the area of each island, and species accumulation curves,
68 which describe the relationship between increasing cumulative species number with
69 increasing sampling area (see Matthews, Triantis, Rigal, Borregaard, Guilhaumon &
70 Whittaker, 2016). This paper is focused on ISARs (Type IV SARs in Scheiner’s 2003
71 typology). Although over twenty ISAR models have been proposed (Tjørve, 2003; Triantis,
72 Guilhaumon & Whittaker, 2012), the most widely used is the power model, $S = c * A^z$, where
73 S is the number of species on an island, A is the area of an island, and c and z are fitted
74 constants (Arrhenius, 1921). In comparative analyses, the power model has been found to
75 provide the best fit to a number of true and habitat island datasets, but it is not universally the
76 best model (Dengler, 2009; Triantis et al., 2012; Matthews, Guilhaumon, Triantis, Borregaard
77 & Whittaker, 2016), and the ISAR has been found to exhibit forms that the predominantly
78 convex power model cannot provide a good fit to, such as sigmoidal shaped relationships
79 (Lomolino, 2000; Triantis et al., 2012). For example, in an analysis of 182 habitat island
80 datasets, the power model provided the best fit, out of twenty candidate ISAR models, in only
81 24% of cases (Matthews, Guilhaumon et al., 2016). Put another way, there is considerable
82 model uncertainty in regards to the form of the ISAR, and a number of studies have argued
83 that ISAR analyses should incorporate a wider set of models rather than simply the power
84 model (Guilhaumon, Gimenez, Gaston & Mouillot, 2008; Guilhaumon, Mouillot & Gimenez,
85 2010; Triantis et al., 2012; Benchimol & Peres, 2013).

86 The SAR is a key tool in conservation biogeography and, amongst other things, has been
87 used to predict the number of extinctions resulting from habitat loss (e.g. Brooks, Pimm &
88 Collar, 1997; Martins & Pereira, 2017), improve protected area design (e.g. Diamond, 1975),
89 and predict the number of species occurring in large areas of natural habitat, such as a large
90 expanse of tropical forest (Palmer, 1990; Rosenzweig, 1995; Plotkin et al., 2000; Desmet &
91 Cowling, 2004; Santos et al., 2010; Smith, 2010; Basset et al., 2012; Gerstner, Dormann,
92 Václavík, Kreft & Seppelt, 2014; Kunin et al., 2018). In regards to the latter, the ability to
93 extrapolate the SAR to accurately predict the number of species occurring in large areas is of
94 significant importance given the logistical and financial constraints involved in sampling over
95 large spatial scales (Basset et al., 2012; Kunin et al., 2018). Typically, predicting richness at
96 large spatial scales using the SAR is achieved by using the power model to predict the
97 richness of an area (e.g. a large island, biome or region), either by using a set z value
98 (generally around 0.25; Rosenzweig, 1995) or by estimating z from empirical data. However,
99 as previously outlined, the power model may not always provide the best characterisation of
100 the ISAR in empirical systems, and thus previous extrapolation studies based solely on the
101 power model may have generated inaccurate predictions (this is true for any individual ISAR
102 model). For example, Dengler (2009) compared the extrapolation ability of 12 ISAR models
103 (in fact 25 models were compared as the same model was fitted using log-transformed and
104 untransformed data; one model was applied using three different transformations) to
105 accurately predict richness on large islands using six island archipelago datasets, and found
106 that the mean rank of the power model was only 11th out of 25. Figure 1 provides a further
107 illustration of this issue. Here, we have simulated eight islands of varying size (1, 3, 7, 14, 17,
108 22, 26, and 30; undefined units) that support reasonable numbers of species (3, 7, 14, 18, 20,
109 23, 24, and 25). We then fit five ISAR models (linear, logistic, negative exponential, power
110 and Weibull3; see Table 1 for more details on these models) to these eight data points. Using
111 these model fits, we estimated the number of species on an island of size 80 (grey dotted line
112 in Fig.1) for each model and extrapolated each curve to its respective predicted value. It can
113 be seen that the different models provide a range of predicted richness values for the
114 hypothetical largest island.

115 An alternative extrapolation approach to simply using the power model is to use multi-model
116 inference (MMI; Burnham & Anderson, 2002) and model averaging, whereby a larger
117 number of n models is fitted to a set of islands, the models ranked according to some criterion
118 (e.g. Akaike's information criterion, AIC; Burnham & Anderson, 2002) and the criterion
119 values converted into model weights (i.e. the conditional probabilities for each of the n
120 models; Wagenmakers & Farrell, 2004). The n models are then each used to predict the
121 richness of a larger area and these predictions are multiplied by the respective model weights
122 and summed to provide a multi-model averaged prediction (Burnham & Anderson, 2002; see
123 Guilhaumon et al., 2008 for a SAR example).

124 A MMI approach is arguably much more robust as it provides a framework to deal with the
125 model uncertainty observed in many SAR studies, and as Burnham & Anderson (2002, p.
126 198) note, such uncertainty can be much greater outside the range of the observed data.
127 However, the effectiveness of the MMI framework in ISAR extrapolation is unknown, and

128 with the exception of the Dengler (2009) study that only analysed six island datasets, the
129 question of model uncertainty in ISAR extrapolation has not been explored. As Dengler
130 (2009, p.733) states, “although extrapolation of species richness beyond the largest plot size
131 is one of the most frequent applications of SARs, there are only few and unsystematic
132 approaches to testing which model function types are most suitable for this purpose.”

133 It should be noted that using the ISAR is only one method for predicting the species richness
134 of larger areas. For example, species accumulation curves, rarefaction methods and various
135 extrapolation methods based on Hill numbers (Colwell & Coddington, 1994; Hsieh, Ma &
136 Chao, 2016) are also widely used. However, many of these approaches require abundance
137 data rather than incidence (i.e. presence-absence) data, although alternative methods are
138 available for incidence data (see Hsieh et al., 2016). Incidence data are commonly available
139 from biogeographical studies (e.g. Triantis et al., 2012; Matthews, Guilhaumon et al., 2016),
140 which likely explains why the ISAR (which only requires incidence data) has often been used
141 in extrapolation exercises (Dengler, 2009).

142 In this study, we use a set of 120 habitat island datasets to compare the accuracy of species
143 richness extrapolation predictions using the power model with predictions using a model
144 averaging approach based on twenty ISAR models. As such, our study goes beyond previous
145 ISAR meta-analyses (e.g. Triantis et al., 2012; Matthews, Guilhaumon et al., 2016), which
146 were focused on ISAR model goodness-of-fit evaluation, to explore ISAR model
147 extrapolation capability. We focus on habitat islands rather than true islands (see Whittaker &
148 Fernández-Palacios, 2007) as many applied SAR studies are focused on fragmented and
149 forested terrestrial landscapes (e.g. Hubbell et al., 2008; Hanski, Zurita, Bellocq & Rybicki,
150 2013; Matthews, Cottee-Jones & Whittaker, 2014). We hypothesise that, due to the high
151 degree of model uncertainty observed in many ISAR studies, the MMI framework will
152 generate more accurate extrapolation predictions than the use of the power model on its own.
153 The results of this analysis will provide useful information to guide future applications of
154 ISAR extrapolation in conservation biogeography studies.

155

156 **MATERIALS AND METHODS**

157 **Data collection**

158 We took a subset of the habitat island datasets collected by Matthews, Cottee-Jones &
159 Whittaker (2015) and Matthews, Guilhaumon et al. (2016). Habitat islands are defined as
160 discrete habitat patches surrounded by contrasting matrix habitat. However, as in Matthews,
161 Guilhaumon et al. (2016), we also included a small number of datasets consisting of protected
162 areas for which the contrast between the matrix and the island was not so pronounced, and we
163 included a few datasets of fragments within an aquatic matrix (e.g. rain forest fragment
164 systems created by the construction of a reservoir), as the dominant assembly processes are
165 considered to be more similar to those in habitat islands *sensu stricto* than oceanic islands (cf.
166 Matthews et al., 2015). The original criteria for dataset collection (see Matthews,
167 Guilhaumon et al., 2016) were: 1) the area and richness of each island were provided; 2) there

168 was no overlap between accepted datasets (data for different taxa within the same study
169 system were accepted); and 3) there were at least four habitat islands. For the present study,
170 we used datasets with at least eight islands and for which we could both successfully fit the
171 power model (i.e. the model fit converged) and construct a multi-model averaged ISAR curve
172 (i.e. at least two ISAR models could be successfully fitted to the dataset). We also manually
173 (i.e. no explicit scale threshold was applied) filtered out datasets that were focused at very
174 small spatial scales (e.g. insects on rose bushes or small experimental grassland plots) as
175 these are not the spatial scale at which ISAR extrapolation is typically undertaken.

176 A total of 120 habitat island datasets were used, comprising 80 vertebrate, 21 plant, and 19
177 invertebrate datasets (Table S1 in Appendix S1 provides a summary of the datasets, and the
178 source paper references are provided in Appendix S1).

179 **Extrapolating the ISAR**

180 To test the extrapolation ability of the various methods, we used the approach of Dengler
181 (2009) whereby, for each dataset, we removed the largest island and all islands within a
182 certain size threshold (*th*) relative to the largest island. For example, if the largest island was
183 100 ha and *th* was 0.5, we removed all islands larger than 50 ha. The new version of the
184 dataset with the largest islands removed is referred to herein as the ‘filtered dataset’.
185 Removing the largest islands from each dataset allowed us to use the model fits to the filtered
186 subset of islands to extrapolate and predict richness on larger islands for which we know the
187 number of species. The value of *th* used in the main analyses was 0.5, although we
188 experimented with different values as a sensitivity analysis (discussed below). For each
189 filtered dataset, we then fitted the power (non-linear) ISAR model using non-linear regression
190 and the ‘sars’ R package (version 1.1.1; Matthews, Triantis, Whittaker & Guilhaumon, 2019).
191 With the exception of a model convergence check, the power model was fitted to a dataset
192 regardless of the results of any model validation checks (the validity of this was tested as part
193 of a sensitivity test, outlined below). A multi-model averaged ISAR curve was then fitted to
194 the filtered dataset using the ‘sar_average’ function in the ‘sars’ R package. We attempted to
195 fit twenty ISAR models (Table 1). A model was excluded if: 1) the model fitting process did
196 not converge, 2) the model fit generated negative predicted values, 3) the residuals of the
197 model fit were not normally distributed (using a Shapiro-Wilks test for normality), or 4) the
198 residuals of the model fit were not homogeneous (assessed by correlating the residuals with
199 the fitted values). All of these checks were undertaken using the ‘sar_average’ function (see
200 Matthews et al., 2019). The remaining model fits were used to generate a multi-model
201 averaged ISAR curve using AIC corrected for small sample size (AIC_c; Burnham &
202 Anderson, 2002).

203 For each dataset, we followed the extrapolation procedure outlined in the introduction where
204 we used the power model fit and the multi-model averaged curve to predict the species
205 richness of the largest island in the original dataset (i.e. the largest of the islands that had
206 been removed; see Dengler, 2009). In regards to the multi-model averaged curve, this worked
207 by taking the multi-model fit object, using each of the individual model fits to predict the
208 richness of the largest island, and multiplying these predictions by the respective AIC_c

209 weights. As AIC_c was used, for datasets where the filtered dataset had only six islands (7
210 cases when $th = 0.5$) it was not possible to calculate AIC_c for the 4 parameter ISAR models.
211 Thus, the model weight was set to zero and the model fit had no bearing on the extrapolation
212 prediction. As there was no functionality to undertake these extrapolations in the ‘sars’ R
213 package, we wrote a new function to achieve this. The new function, ‘sar_pred’, takes two
214 arguments (fit and area) and extrapolates the ‘fit’ object to predict the richness on an island of
215 size ‘area’. The ‘fit’ argument can be an individual SAR model fit (e.g. the power model) or a
216 multi-model SAR curve. The new function is available in version 1.1.2 of the ‘sars’ package
217 which is currently on GitHub (txm676/sars) and will be uploaded to CRAN shortly.

218 To compare the predictions of the power model and the multi-model averaged curve for a
219 given dataset, we used the log error of extrapolation (LEE) metric of Dengler (2009) that
220 addresses extrapolation capability. LEE is simply the log of the model’s predicted richness
221 minus the log of the observed richness (following Dengler, 2009, log to the base 10 was
222 used); thus, the closer the LEE value is to zero the more accurate the prediction, and a
223 positive LEE value means the model has over predicted the observed richness and *vice versa*.
224 LEE was calculated for both the power model prediction and the multi-model averaged curve
225 prediction.

226 As an important part of model prediction is to generate an estimate of the error of a prediction
227 (Burnham & Anderson, 2002), the confidence intervals around the predictions were
228 calculated using bootstrapping (Davison & Hinkley, 1997). For each of the filtered datasets,
229 the data points (i.e. an individual island area and richness value) were sampled with
230 replacement until the bootstrap sample was the same size as the original filtered dataset. The
231 power model and multi-model curve prediction process was then undertaken using this
232 bootstrap sample and the predictions stored. For the multi-model curve, the same models that
233 were successfully fitted in the construction of the multi-model curve fit to the filtered dataset
234 were selected. We did not undertake residual checks (e.g. normality) here to ensure bootstrap
235 samples could be created, but we did still exclude model fits with negative predicted values.
236 This process was repeated 100 times for each dataset and a 95% confidence interval
237 constructed. Occasionally it was not possible to fit some of the relevant models to a bootstrap
238 sample, or the predicted value was negative; in these cases, the bootstrap sample was
239 discarded.

240 The main comparison of interest was the power model with the multi-model averaged curve.
241 However, we also re-ran the above analysis including the extrapolation predictions of the
242 additional 19 individual ISAR models. For each dataset, an individual model extrapolation
243 prediction was included in the comparison only if the fit of the model to the filtered dataset
244 passed all of the model validation checks.

245 **Modelling variation in prediction accuracy**

246 To determine whether any dataset characteristics could explain variation in the LEE values
247 for the power model predictions (LEE-POW), we used generalized additive models (GAMs;
248 Gaussian family) within a model selection framework. GAMs were used as there was evident

249 non-linear relationships between the predictors and the response. We used LEE-POW as the
250 response variable. It was not possible to use the LEE values from the multi-model averaged
251 curve (LEE-MMI) as the values were highly skewed and the residuals of the resultant models
252 did approximate a normal distribution. For predictor variables, for each dataset (here the
253 filtered dataset was used) we calculated the area of the smallest and largest islands and the
254 ratio between them (A_{\min} , A_{\max} and A_{scale}), the richness of the most species poor and species
255 rich islands and the ratio between them (S_{\min} , S_{\max} and S_{scale}), and the number of islands (N_i).
256 For each dataset, we also took the latitude (Lat.) of the dataset and the sampled taxon (i.e.
257 vertebrate, invertebrate or plant) from Matthews, Guilhaumon et al. (2016). Multicollinearity
258 between predictors was tested using variance inflation factors: A_{\max} and S_{\max} were removed
259 due to high multicollinearity and the remaining variance inflation factors were all below
260 three. All of the continuous predictors (with the exception of latitude) were log-transformed
261 to induce normality. The continuous predictors were modelled as penalized regression splines
262 and the GAMs were fitted using the ‘mgcv’ R package (Wood, 2011). Smoothing parameter
263 estimation was calculated using the Generalized Cross Validation (GCV) criterion.

264 A full set of models given all possible combinations of predictors were fitted using the
265 MuMIn R package (Bartoń, 2012), and models were compared using AIC_c . The model with
266 the lowest AIC_c value was considered the best model, and all models with $\Delta AIC_c \leq 2$
267 units of the best model were considered as having a similar degree of support (Burnham &
268 Anderson, 2002). Model fits were validated using histograms of the residuals and plots of the
269 residuals vs. the fitted values; the residuals of the full and best model roughly approximated a
270 normal distribution and there were no evident patterns in the residuals. The relative
271 importance of each predictor was calculated by summing the AIC_c weights for all models in
272 which a predictor was included (Giam & Olden, 2016).

273 To determine whether the relative fit of a model to the filtered dataset explained its
274 extrapolation performance, for each of the twenty models we calculated the LEE values
275 across all datasets. For each ISAR model separately, we then fitted a simple generalized
276 additive regression model (Gaussian family) whereby the absolute LEE values were the
277 response variable and the AIC_c weights were the predictor variable, modelled as a penalized
278 regression spline. Due to multiple testing, the critical P-value used was Bonferroni corrected
279 (i.e. $0.05 / 20 = 0.0025$).

280 **Sensitivity analyses**

281 To ensure our results were robust to the assumptions made during the analyses, we undertook
282 three sensitivity tests. First, we re-ran the extrapolation analysis using th values of 0.3 and 0.7
283 (i.e. removing all islands that were 30% or 70% the size of the largest island in the original
284 dataset). Second, in the main analyses, to ensure we could always compare the prediction of
285 the power model with the prediction of the multi-model averaged curve we fitted the power
286 model to all datasets regardless of the results of any model validation checks (with the
287 exception of model convergence; e.g. no normality of residuals check was undertaken). Thus,
288 we re-ran the prediction analysis after filtering out all datasets where the power model fit
289 failed any of the following validation checks: 1) the model fit generated negative predicted

290 values, 2) the residuals of the model fit were not normally distributed, 3) the residuals of the
291 model fit were not homogeneous, or 4) the z parameter was not significant. Third, we re-ran
292 the prediction analysis after removing the linear model from the multi-model averaged curve
293 fitting process (i.e. fitting of only 19 models was attempted; see Table 1). The reason for this
294 third check is that previous studies have found that the linear model tends to provide a better
295 relative fit to datasets with smaller numbers of islands, whereas in larger datasets its relative
296 performance declines (e.g. Matthews, Guilhaumon et al., 2016). As the removal of larger
297 islands necessarily generates datasets with fewer numbers of islands, it is possible that the
298 linear model might provide better fits to the filtered datasets which then leads to inaccurate
299 predictions if the ISAR of the full dataset is not linear. All analyses were undertaken using R
300 (version 3.5.2; R Core Team, 2017). Unless stated otherwise, an alpha level of 0.05 was used
301 in all significance tests.

302 **RESULTS**

303 When a th value of 0.5 was used, the power model provided the best fit to the most (filtered)
304 datasets ($n = 29$), followed by the linear model ($n = 21$), and then the Monod ($n = 19$) and
305 logarithmic models ($n = 16$) (see Table 1), according to AIC_c .

306 The full results of the main extrapolation and prediction analysis are provided in Table S2 in
307 Appendix S2. In contrast to our hypothesis, the power model provided the most accurate
308 prediction of the richness of the largest island (i.e. the lowest absolute LEE value) in 69 cases
309 (58%), with the multi-model averaged curve providing the more accurate prediction in the
310 remaining 51 cases (42%). The median LEE value of the power model was 0.04 (95%
311 quantiles = -0.32 and 0.31), whilst the median LEE value of the multi-model curve (LEE-
312 MMI) was 0.03 (95% quantiles = -0.35 and 0.74). However, as LEE values could be both
313 positive and negative, the median of the absolute LEE values provides a better summary of
314 the extrapolation capability: the median of absolute LEE-POW values was 0.08 (95%
315 quantiles = 0.01 and 0.34), whilst the median of absolute LEE-MMI values was 0.10 (95%
316 quantiles = 0.01 and 0.74). Both the power model (61% of LEE-POW values were positive)
317 and the multi-model averaged curve (60% of LEE-MMI values were positive) had a slightly
318 greater tendency to over predict the observed richness. The confidence intervals were on
319 average narrower for the power model predictions (median 95% confidence interval width =
320 18) than for the multi-model averaged curve predictions (median 95% confidence interval
321 width = 78) (Table S3 in Appendix S2). The confidence intervals around the multi-model
322 averaged curve predictions were sometimes very large (i.e. spanning multiple orders of
323 magnitude; see Table S3).

324 When the extrapolation predictions from all 20 ISAR models were considered, in addition to
325 the multi-model averaged curve, the power model provided the most accurate prediction of
326 the richness of the largest island in 11 cases, with the multi-model averaged curve providing
327 the most accurate prediction in five cases. The Extended Power 2 (see Table 1) model
328 provided the best prediction the most times, with 12 cases (the results for all models are
329 provided in Table 1).

330 The full GAM (i.e. the GAM with all predictors) had a lower AIC score (-121.9) than an
331 equivalent standard linear regression model (-115.4); this provides additional justification for
332 our use of GAMs. When LEE-POW was used as the response variable in a GAM model
333 selection analysis, the best model contained A_{scale} , S_{scale} , Lat. and S_{min} (Table 2). A plot of the
334 smoothers for these four variables is provided as Figure 2. The effective degrees of freedom
335 of the smoothers for A_{scale} and Lat. were one, indicating that these smoothers were straight
336 lines; increasing A_{scale} resulted in decreasing LEE-POW, while the opposite pattern was true
337 for Lat (Fig. 2). The S_{scale} and S_{min} relationships were more complex (Fig. 2), but increasing
338 S_{scale} resulted in an approximate increase in LEE-POW. However, there was a reasonable
339 degree of model uncertainty as the best model had an AIC_c weight of only 0.20, and there
340 were two additional models within 2 delta AIC_c units of the best model (Table 2). In addition,
341 the adjusted R^2 value of the best model was low (0.20). A_{scale} (0.98), S_{scale} (0.95) and S_{min}
342 (0.81) had quite high relative importance values, whilst the values for the remaining
343 predictors were all lower (Table 2).

344 For 18 of the ISAR models, the relative fit of a model to the filtered dataset (i.e. the model's
345 AIC_c weight) was a poor predictor of a model's extrapolation accuracy (measured using the
346 LEE metric). In only two cases (for the Power Rosenzweig and Extended Power 1 models;
347 see Table 1 for model descriptions) was the AIC_c weight a significant predictor of a model's
348 absolute LEE value (Table S4 in Appendix S2).

349 The choice of th value did not change the overall qualitative results. The power model
350 provided the more accurate prediction in 65 (54%) and 72 (61%; when a th value of 0.7 was
351 used there was one dataset for which no models could be successfully fitted) cases when th
352 values of 0.3 and 0.7 were used, respectively (see Table S5 & S6 in Appendix S2). In regards
353 to the power model validation sensitivity test, there were 23 datasets for which the power
354 model failed one of the validation checks. However, removing these 23 datasets and re-
355 running the prediction analysis using the remaining 97 datasets did not change the overall
356 qualitative results: the power model provided the most accurate prediction in 55 cases (57%).
357 Finally, re-running the prediction analysis after excluding the linear model from the multi-
358 model averaged curve resulted in a slight increase in the number of cases where the multi-
359 model averaged curve provided the more accurate prediction (60 out of 120 cases), but the
360 general picture remained the same.

361 **DISCUSSION**

362 Using 120 habitat island datasets, we compared the extrapolation capability of the power
363 ISAR model with that of a multi-model averaged ISAR curve constructed using up to twenty
364 ISAR models. In contrast to our hypothesis that the multi-model curve would produce more
365 accurate species richness predictions, we found that the power model provided the more
366 accurate prediction in a majority of cases.

367 **Model averaging is not a panacea for ISAR extrapolation**

368 It is rarely feasible to produce complete inventories of all species of a given taxonomic group
369 at large spatial scales (e.g. in a large expanse of tropical forest or on very large islands;

370 Colwell & Coddington, 1994). The question of how to extrapolate from samples collected at
371 relatively small scales to accurately predict richness over larger areas is therefore the subject
372 of considerable research effort (Hsieh et al., 2016). There has been particular focus on the
373 ISAR (in addition to SARs constructed using continuous habitat data) as it only requires
374 incidence data; yet a statistically rigorous ISAR extrapolation method, required for accurate
375 richness predictions, has proven elusive. The present study represents a formative step in the
376 development of such a method.

377 Based on the results and arguments presented in many recent SAR studies and other model
378 prediction exercises (Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004;
379 Guilhaumon et al., 2008, 2010; Triantis et al., 2012; Benchimol & Peres, 2013; Matthews,
380 Guilhaumon et al., 2016), we hypothesised that a model averaging framework would result in
381 more accurate extrapolation predictions than simply using the power model in isolation.
382 Model averaging based on a set of competing candidate models has been proposed for
383 improving predictions in cases where model uncertainty is prevalent (as seems to be the case
384 with the ISAR; e.g. Triantis et al., 2012). For example, Burnham & Anderson (2002, p.150)
385 state that “prediction is an ideal way to view model averaging, because each model in a set,
386 regardless of its parameterisation, can be used to make a predicted value.” However, in
387 contrast to our hypothesis we found that the power model provided the best prediction in the
388 most cases, regardless of which *th* value was used (0.3, 0.5 or 0.7). Although our study is the
389 first comprehensive evaluation of model averaging in ISAR extrapolation, Mazel et al. (2014)
390 found that the power model on its own provided similar results to a multi-model average
391 curve when using SARs, and functional diversity- and phylogenetic diversity-area
392 relationships, to select biodiversity hotspots. Thus, it may be that the power model is
393 generally a more appropriate tool than multi-model averaged curves in many applied SAR
394 contexts. More research is needed to examine the performance of multi-model averaging in
395 other areas of applied SAR research, such as predicting the number of extinctions resulting
396 from habitat loss.

397 Analysis of the raw LEE-POW and LEE-MMI values indicates that both the multi-model
398 averaged curve and the power model had a slightly greater tendency for over-prediction of
399 species richness. The tendency of the power model to overpredict richness has been
400 previously documented (Palmer, 1990; Smith, 2010), but the performance of multi-model
401 averaged ISAR curves when extrapolating richness has not been previously documented. A
402 tendency for over prediction is arguably preferable as, in a conservation context,
403 underprediction bias is likely to carry greater risk (for instance when forecasting the impacts
404 of habitat loss).

405 In general, the multi-model curve predictions also exhibited wider confidence intervals
406 (Table S3). In certain cases, these were very wide, reflecting the bootstrap procedure that we
407 employed, which works by sampling islands (paired area and richness values) with
408 replacement; this process can result in the same island being chosen multiple times,
409 particularly in smaller datasets, resulting in some bootstrapped samples having unusually-
410 shaped ISARs which do not bear much resemblance to the ISAR of the original
411 sample/dataset. As the multi-model curve combines multiple ISAR models it is necessarily

412 more flexible, allowing it to more accurately characterise the form of the unusually-shaped
413 bootstrapped ISARs, but which can then result in wayward extrapolation predictions (i.e.
414 predictions of the largest island in the original dataset).

415 Interestingly, when the extrapolation predictions from all individual twenty ISAR models
416 were compared the Extended Power 2 model (EPM2; Table 1) provided the most accurate
417 predictions the most times (12 times vs. 11 times for the standard power model). The EPM2
418 model, which is a sigmoidal model, is from within the same family as the standard power
419 model (i.e. it is defined by adding a single additional parameter to the standard power model;
420 Tjørve, 2009). The greater flexibility that arises from an additional parameter necessarily
421 means that the EPM2 model should explain more variation in richness than the standard
422 power model (i.e. have a larger R^2); however, this does not mean the model should produce
423 more accurate extrapolation predictions. For example, the Extended Power 1 model, which is
424 also in the same model family as the power model, only provided the most accurate
425 prediction 5 times. In addition, the other sigmoidal models generally performed poorly (Table
426 1). In contrast to Tjørve (2009), who postulated that extended power models may provide
427 poor extrapolation predictions, these results call for greater assessment of extended power
428 models in applied ISAR applications.

429 **Why does the power model provide better predictions on average?**

430 The rationale for the smaller confidence intervals around the extrapolation predictions of the
431 power model described in the preceding paragraph also provides an explanation for why the
432 multi-model curve provided less accurate predictions in a majority of cases more generally:
433 the greater flexibility of the multi-model curve is also its downfall. Regardless of the shape of
434 the ISAR of the full dataset, unless that shape is characterised by a linear model, the form of
435 the filtered dataset will differ, often considerably, from that of the full dataset. One of the
436 advantages of the MMI approach, if the model set contains a range of sensible models given
437 the situation, is that it often provides a better fit to a set of data than any one model on its own
438 (Burnham & Anderson, 2002). However, if the shape of the filtered dataset is not
439 representative of that of the full dataset, this greater flexibility may be a negative feature. For
440 example, the linear model has been shown to provide a better fit relatively speaking to
441 datasets with few, relatively smaller, islands (Matthews, Guilhaumon et al., 2016). Thus, it
442 can be assumed that the relative performance of the linear model is better for the filtered
443 datasets than for the full datasets; this better performance means it will have a larger
444 information criterion weight and thus a stronger influence on the multi-model curve.
445 However, if the full dataset is actually even just somewhat convex the multi-model curve
446 (with its linear element) will not provide an accurate extrapolation prediction. In addition, it
447 may be that habitat island datasets contain substantial amounts of noise due to the role of
448 factors other than area (e.g. human disturbance; Benchimol & Peres, 2013). These factors,
449 which may have a greater relative effect in small fragments (Matthews et al., 2014), may
450 result in “messy” ISAR datasets. The more complex models have greater flexibility to fit this
451 noise, resulting in poor extrapolation behaviour. For example, in a small number of cases, the
452 largest fragment in the filtered dataset had lower richness than some of the smaller fragments,

453 resulting in some of the more complex models predicting decreasing richness with increasing
454 area and thus predicting negative richness when extrapolated!

455 **Explaining variation in extrapolation capability of the power model across datasets**

456 Our generalized additive model selection analysis indicated that the most important variables
457 in driving variation in LEE-POW across datasets were A_{scale} , S_{scale} , S_{min} and Lat. (Table 2),
458 with A_{scale} , S_{scale} and S_{min} in particular having relative importance values greater than 0.80. It
459 should be noted that the amount of variation in LEE-POW explained by the best model was
460 relatively low (adjusted $R^2 = 0.20$). In the best model, the effect of A_{scale} on LEE-POW was
461 linear and negative, whilst the effect of S_{scale} was non-linear but broadly positive and convex
462 (Fig. 2). These results indicate that increasing A_{scale} results in lower LEE-POW values while,
463 in contrast, increasing S_{scale} results in larger LEE-POW values, although there is a flattening
464 out of this latter relationship at larger values of S_{scale} (Fig. 2). The negative effect of A_{scale} on
465 LEE-POW values is logical because the full convex shape of the empirical ISAR may only
466 become apparent when a large range of island sizes is studied (Martin, 1981; Matthews,
467 Guilhaumon et al., 2016); for a smaller range of island areas the relative performance of the
468 linear model is conversely greater. Thus, if A_{scale} is small and, in particular, there are no
469 relatively large fragments within the dataset, the ISAR is less likely to be characterised by a
470 power model (and more likely by a linear model) and attendant extrapolation predictions are
471 likely to over-predict the true richness value. The positive effect of S_{scale} is more surprising,
472 as one would expect the range in species richness in a dataset to scale positively with the
473 range in island area. Indeed, A_{scale} and S_{scale} were significantly, albeit weakly, positively
474 correlated (Spearman's $\rho = 0.38$; $P < 0.001$). We speculate that S_{scale} co-varies with another
475 variable that was not included in our analysis, such as sample completeness (Hsieh et al.,
476 2016). For example, if S_{scale} is related to the number of species across all fragments
477 (information that is not available from ISAR datasets) and more species-rich taxa are more
478 likely to have been under-sampled, particularly in the larger fragments, then the effect of S_{scale}
479 may in fact be evidence of a sampling artefact. Further research is needed to explore this
480 possibility.

481 We also found that, generally speaking, a model's relative fit to the filtered dataset provided a
482 poor predictor of that model's extrapolation accuracy. This further complicates providing
483 general guidelines for extrapolation as it rules out simply selecting the best fitting model
484 when undertaking ISAR extrapolation.

485 **Conclusions**

486 Our findings show that multi-model averaging is unlikely to provide a universally suitable
487 method for ISAR extrapolation, even though there is a large amount of model uncertainty
488 (e.g. see the mean AIC_c weights of each model in Table 1). Taking the specific characteristics
489 of the studied dataset into account (e.g. island size range, species richness range) could lead
490 to more informed ISAR model selection, though this requires further investigation. However,
491 the relevance of our results is likely to be restricted to the spatial scale of the analysed
492 datasets. Although some of our datasets contain very large islands (largest island across all

493 datasets = 19,604 km²), the median island size is much smaller (0.09 km²), and our results
494 may thus not be transferable to i) scenarios requiring the ISAR to be extrapolated to very
495 large areas (e.g. biotic regions or provinces; Rosenzweig, 1995; Gerstner et al., 2014), or ii)
496 other types of SARs (e.g. species accumulation curves; Bassett et al., 2012; Kunin et al.,
497 2018). It is also possible that habitat island datasets are particularly noisy and that we may
498 find different results when looking at true islands, for example.

499 Although the power model provided more accurate predictions in a majority of cases, it is
500 hard to advocate blanket use of the power model in future ISAR extrapolation analyses, as in
501 approximately 40% of cases the multi-model averaged curve provided a better prediction.
502 Depending on the aim of the study, a comparative selection of techniques (e.g. multiple
503 individual ISAR models and the multi-model averaged curve) may be useful, yielding a range
504 of predictions with confidence intervals that can be assessed together. In situations where a
505 single point estimate is required, our results would support judicious use of the power model.
506 However, further research at larger spatial scales is urgently needed to validate these
507 recommendations for ISAR extrapolation in a wider context.

508

509 **ACKNOWLEDGEMENTS**

510 The recently published ‘sars’ R package, which was used to run the analyses in the paper,
511 was written in collaboration with François Guilhaumon. François Rigal provided modelling
512 advice. Two anonymous reviewers provided comments that improved the paper.

513

514

515 **DATA ACCESSIBILITY**

516 All datasets are publicly available and the full source citations are provided in the Supporting
517 Information.

518

519 **REFERENCES**

520 Arrhenius, O. (1921). Species and area. *Journal of Ecology*, 9, 95-99.

521 Bartoń, K. (2012). MuMIn: multi-model inference (R package version 1.40.4). Retrieved
522 from <https://cran.r-project.org/web/packages/MuMIn/index.html>

523 Basset, Y., Cizek, L., Cuénoud, P., Didham, R. K., Guilhaumon, F., Missa, O., . . . Leponce,
524 M. (2012). Arthropod diversity in a tropical forest. *Science*, 338, 1481-1484.

525 Benchimol, M., & Peres, C. A. (2013). Anthropogenic modulators of species–area
526 relationships in Neotropical primates: a continental-scale analysis of fragmented
527 forest landscapes. *Diversity and Distributions*, 19, 1339-1352.

- 528 Brooks, T.M., Pimm, S.L. & Collar, N.J. (1997). Deforestation predicts the number of
529 threatened birds in insular Southeast Asia. *Conservation Biology*, 11, 382-394.
- 530 Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: a*
531 *practical information-theoretic approach* (2nd ed.). New-York: Springer.
- 532 Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through
533 extrapolation. *Philosophical Transactions: Biological Sciences*, 345, 101-118.
- 534 Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*.
535 Cambridge: Cambridge University Press.
- 536 Dengler, J. (2009). Which function describes the species–area relationship best? A review
537 and empirical evaluation. *Journal of Biogeography*, 36, 728-744.
- 538 Desmet, P., & Cowling, R. (2004). Using the species-area relationship to set baseline targets
539 for conservation. *Ecology and Society*, 9, 11-33.
- 540 Diamond, J.M. (1975) The island dilemma: lessons of modern biogeographic studies for the
541 design of natural reserves. *Biological Conservation*, 7, 129-146.
- 542 Gerstner, K., Dormann, C. F., Václavík, T., Kreft, H., & Seppelt, R. (2014). Accounting for
543 geographical variation in species–area relationships improves the prediction of plant
544 species richness at the global scale. *Journal of Biogeography*, 41, 261-273.
- 545 Giam, X., & Olden, J. D. (2016). Quantifying variable importance in a multimodel inference
546 framework. *Methods in Ecology and Evolution*, 7, 388-397.
- 547 Guilhaumon, F., Gimenez, O., Gaston, K. J., & Mouillot, D. (2008). Taxonomic and regional
548 uncertainty in species-area relationships and the identification of richness hotspots.
549 *Proceedings of the National Academy of Sciences USA*, 105, 15458-15463.
- 550 Guilhaumon, F., Mouillot, D., & Gimenez, O. (2010). mmSAR: an R-package for multimodel
551 species–area relationship inference. *Ecography*, 33, 420-424.
- 552 Hanski, I., Zurita, G. A., Bellocq, M. I., & Rybicki, J. (2013). Species–fragmented area
553 relationship. *Proceedings of the National Academy of Sciences USA*, 110, 12715-
554 12720.
- 555 He, F. & Legendre, P. (1996). On species-area relations. *The American Naturalist*, 148, 719-
556 737.
- 557 Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: an R package for rarefaction and
558 extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution*,
559 7, 1451-1456.
- 560 Hubbell, S. P., He, F., Condit, R., Borda-de-Água, L., Kellner, J., & ter Steege, H. (2008).
561 How many tree species are there in the Amazon and how many of them will go
562 extinct? *Proceedings of the National Academy of Sciences USA*, 105, 11498-11504.

- 563 Kunin, W. E., Harte, J., He, F., Hui, C., Jobe, R. T., Ostling, A., . . . Varma, V. (2018).
564 Upscaling biodiversity: estimating the species–area relationship from small samples.
565 *Ecological Monographs*, 88, 170-187.
- 566 Lomolino, M. V. (2000). Ecology's most general, yet protean pattern: the species-area
567 relationship. *Journal of Biogeography*, 27, 17-26.
- 568 Martin, T.E. (1981). Species-area slopes and coefficients: a caution on their interpretation.
569 *The American Naturalist*, 118, 823-837.
- 570 Martins, I. S., & Pereira, H. M. (2017). Improving extinction projections across scales and
571 habitats using the countryside species-area relationship. *Scientific Reports*, 7, 12899.
- 572 Matthews, T. J., Cottee-Jones, H. E., & Whittaker, R. J. (2014). Habitat fragmentation and
573 the species–area relationship: a focus on total species richness obscures the impact of
574 habitat loss on habitat specialists. *Diversity and Distributions*, 20, 1136-1146.
- 575 Matthews, T.J., Cottee-Jones, H.E.W. & Whittaker, R.J. (2015). Quantifying and interpreting
576 nestedness in habitat islands: a synthetic analysis of multiple datasets. *Diversity and
577 Distributions*, 21, 392-404.
- 578 Matthews, T. J., Guilhaumon, F., Triantis, K. A., Borregaard, M. K., & Whittaker, R. J.
579 (2016). On the form of species–area relationships in habitat islands and true islands.
580 *Global Ecology and Biogeography*, 25, 847–858.
- 581 Matthews, T. J., Triantis, K. A., Rigal, F., Borregaard, M. K., Guilhaumon, F., & Whittaker,
582 R. J. (2016). Island species–area relationships and species accumulation curves are
583 not equivalent: an analysis of habitat island datasets. *Global Ecology and
584 Biogeography*, 25, 607-618.
- 585 Matthews, T. J., Triantis, K. A., Whittaker, R. J., & Guilhaumon, F. (2019). sars: an R
586 package for fitting, evaluating and comparing species–area relationship models.
587 *Ecography*, *In press*.
- 588 Mazel, F., Guilhaumon, F., Mouquet, N., Devictor, V., Gravel, D., Renaud, J., . . . Thuiller,
589 W. (2014). Multifaceted diversity–area relationships reveal global hotspots of
590 mammalian species, trait and lineage diversity. *Global Ecology and Biogeography*,
591 23, 836-847.
- 592 Palmer, M. W. (1990). The estimation of species richness by extrapolation. *Ecology*, 71,
593 1195-1198.
- 594 Plotkin, J. B., Potts, M. D., Yu, D. W., Bunyavejchewin, S., Condit, R., Foster, R., . . .
595 Ashton, P. S. (2000). Predicting species diversity in tropical forests. *Proceedings of
596 the National Academy of Sciences USA*, 97, 10850-10854.

- 597 R Core Team. (2017). R: a language and environment for statistical computing (Version
598 3.5.1). Vienna, Austria: R foundation for statistical computing. Retrieved from
599 <https://www.R-project.org/>
- 600 Rosenzweig, M. L. (1995). *Species diversity in space and time*. Cambridge: Cambridge
601 University Press.
- 602 Santos, A.M.C., Whittaker, R.J., Triantis, K.A., Borges, P.A.V., Jones, O.R., Quicke, D.L.J.
603 & Hortal, J. (2010). Are species–area relationships from entire archipelagos congruent
604 with those of their constituent islands? *Global Ecology and Biogeography*, 19, 527-
605 540.
- 606 Scheiner, S. M. (2003). Six types of species-area curves. *Global Ecology and Biogeography*,
607 12, 441-447.
- 608 Smith, A. B. (2010). Caution with curves: caveats for using the species-area relationship in
609 conservation. *Biological Conservation*, 143, 555-564.
- 610 Tjørve, E. (2003). Shapes and functions of species–area curves: a review of possible models.
611 *Journal of Biogeography*, 30, 827-835.
- 612 Tjørve, E. (2009). Shapes and functions of species–area curves (II): a review of new models
613 and parameterizations. *Journal of Biogeography*, 36, 1435-1445.
- 614 Tjørve, E., & Tjørve, K. M. C. (2017). Species-area relationship. *eLS (Encyclopedia of Life
615 Sciences Online)*. Chichester: John Wiley & Sons.
- 616 Triantis, K. A., Guilhaumon, F., & Whittaker, R. J. (2012). The island species–area
617 relationship: biology and statistics. *Journal of Biogeography*, 39, 215-231.
- 618 Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights.
619 *Psychonomic Bulletin & Review*, 11, 192-196.
- 620 Whittaker, R. J., & Fernández-Palacios, J. M. (2007). *Island biogeography: ecology,
621 evolution, and conservation* (2nd ed.). Oxford: Oxford University Press.
- 622 Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood
623 estimation of semiparametric generalized linear models. *Journal of the Royal
624 Statistical Society: Series B (Statistical Methodology)*, 73, 3-36.

625

626 BIOSKETCH

627 **Tom Matthews** is a macroecologist and biogeographer at the University of Birmingham,
628 UK. He is interested in the application of macroecological methods to global environmental
629 change questions, and his previous work has focused on the impacts of habitat fragmentation
630 and the form of the species–area relationship in fragmented landscapes.

631 **Thomas Aspin** is a disturbance ecologist affiliated with the University of Birmingham. His
632 research broadly centres on the interface of disturbance ecology, macroecology and
633 conservation ecology, with particular emphasis on climate change and habitat loss.

634 Author Contributions: TJM designed the study and collected the data; TJM ran the analyses
635 with input from TWHA; TJM and TWHA wrote the paper.

636

637 **SUPPORTING INFORMATION**

638 Additional Supporting Information may be found online in the supporting information tab for
639 this article.

640

641

642

643 **TABLES**

644

645 **Table 1** The twenty models that were fitted to generate the multi-model averaged ISAR
 646 curve. The model shape is the general model shape, as in Triantis et al. (2012); the observed
 647 shape can deviate from the general model shape in cases when fitting certain models. For the
 648 model equation, A = sample area, and d, c, z and f are free parameters. Each equation is
 649 calculating the number of species. Mean weight is the mean AIC_c weight for a given model
 650 across all fits to the filtered datasets (excluding non-satisfactory fits). Best fit corresponds the
 651 number of times a model provided the best fit to a filtered dataset (i.e. had the lowest AIC_c
 652 value). Best prediction corresponds to the number of times a model provided the best
 653 extrapolated prediction in the all model comparison; these values do not sum to 120 (the
 654 number of datasets) as the multi-model averaged curve provided the best extrapolation
 655 prediction in five cases.

656

Model	No. parameters	Model shape	Equation	Mean weight	Best fit	Best Prediction
Asymptotic	3	Convex	$d - c \cdot z^A$	0.04	0	6
Beta-P	4	Sigmoid	$d \cdot (1 - (1 + (A/c)^z)^{-f})$	<0.01	0	4
Chapman–Richards	3	Sigmoid	$d \cdot (1 - \exp(-z \cdot A)^c)$	0.01	0	6
Logarithmic	2	Convex	$c + z \cdot \log(A)$	0.14	16	10
Extended Power 1	3	Convex/Sigmoid	$c \cdot A^z \cdot A^{-d}$	0.04	0	5
Extended Power 2	3	Sigmoid	$c \cdot A^{z - (d/A)}$	0.03	1	12
Gompertz	3	Sigmoid	$d \cdot \exp(-\exp(-z \cdot (A - c)))$	0.04	2	4
Kobayashi	2	Convex	$c \cdot \log(1 + A/z)$	0.15	13	5
Linear	2	Linear	$c + z \cdot A$	0.12	21	9
Logistic	3	Sigmoid	$c / (f + A^{(-z)})$	0.03	0	7
Monod	2	Convex	$d / (1 + c \cdot A^{(-1)})$	0.10	19	7
Morgan–Mercer–Flodin	3	Sigmoid	$d / (1 + c \cdot A^{(-z)})$	0.03	0	1
Negative Exponential	2	Convex	$d \cdot (1 - \exp(-z \cdot A))$	0.10	11	4
Persistence Function 1	3	Convex	$c \cdot A^z \cdot \exp(-d \cdot A)$	0.03	2	2
Persistence Function 2	3	Sigmoid	$c \cdot A^z \cdot \exp(-d/A)$	0.04	2	6
Power	2	Convex	$c \cdot A^z$	0.16	29	11
Power Rosenzweig	3	Convex	$f + c \cdot A^z$	0.03	1	6
Rational	3	Convex	$(c + z \cdot A) / (1 + d \cdot A)$	0.03	1	3
Weibull-3	3	Sigmoid	$d \cdot (1 - \exp(-c \cdot A^z))$	0.04	1	4
Weibull-4	4	Sigmoid	$d \cdot (1 - \exp(-c \cdot A^z))^f$	0.01	1	3

657

658

659

660

661

662

663

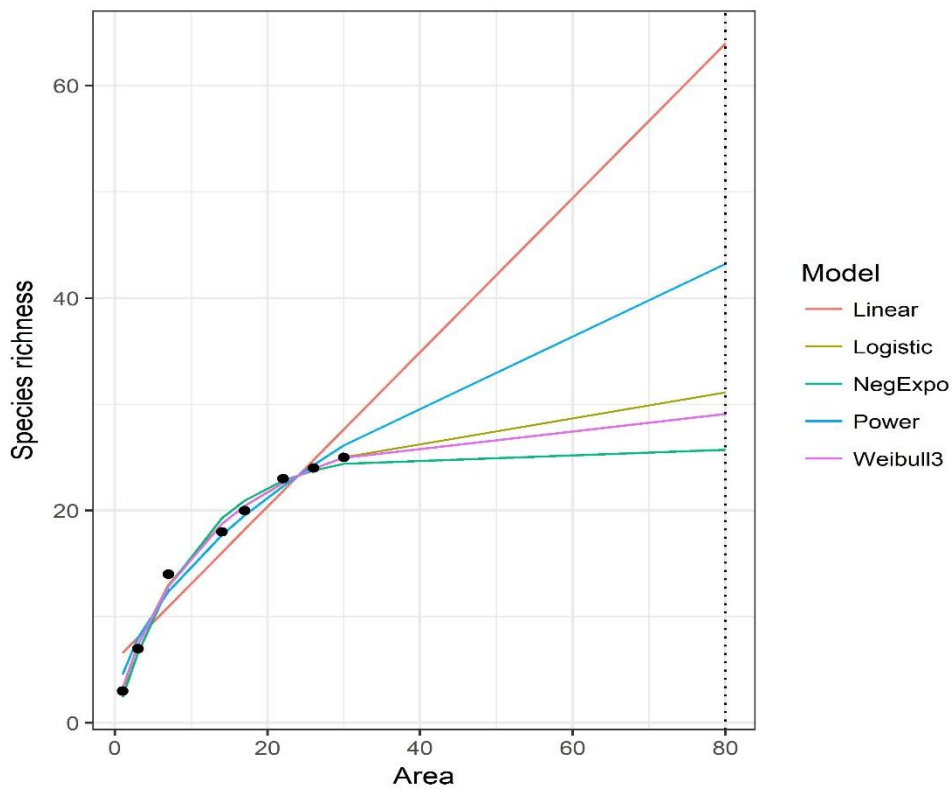
664 **Table 2** The results of the generalized additive model selection. The response variable was
665 the LEE values from 120 habitat island datasets for the power model curve (see the main
666 text), which provides an assessment of the extrapolation accuracy of the power ISAR model.
667 The predictor variables were the smallest island area in a dataset (A_{\min}) and the ratio between
668 the largest and the smallest island area (A_{scale}), the same two variables but for species richness
669 (S_{\min} and S_{scale}), the number of islands in a dataset (Ni), the latitude of the dataset (Lat.) and
670 the taxon sampled (Taxon). A_{\min} , A_{scale} , Lat, Ni, S_{\min} and S_{scale} were all modelled as ‘penalized
671 regression splines’, while taxon was modelled as a standard linear variable (as it was
672 categorical). A ‘+’ indicates that a variable was included within a model. Models were ranked
673 using AIC_c and all models with delta AIC_c values less than two are shown. The AIC_c weight
674 of each model is also provided. The relative importance (RI) of each predictor is shown on
675 the bottom row.

Model	A_{\min}	A_{scale}	Lat.	Ni	S_{\min}	S_{scale}	Taxon	Delta	Weight
1	-	+	+	-	+	+	-	0.00	0.20
2	-	+	-	-	+	+	-	0.48	0.16
3	+	+	+	-	+	+	-	1.89	0.08
RI	0.27	0.98	0.55	0.25	0.81	0.95	0.15		

676

677

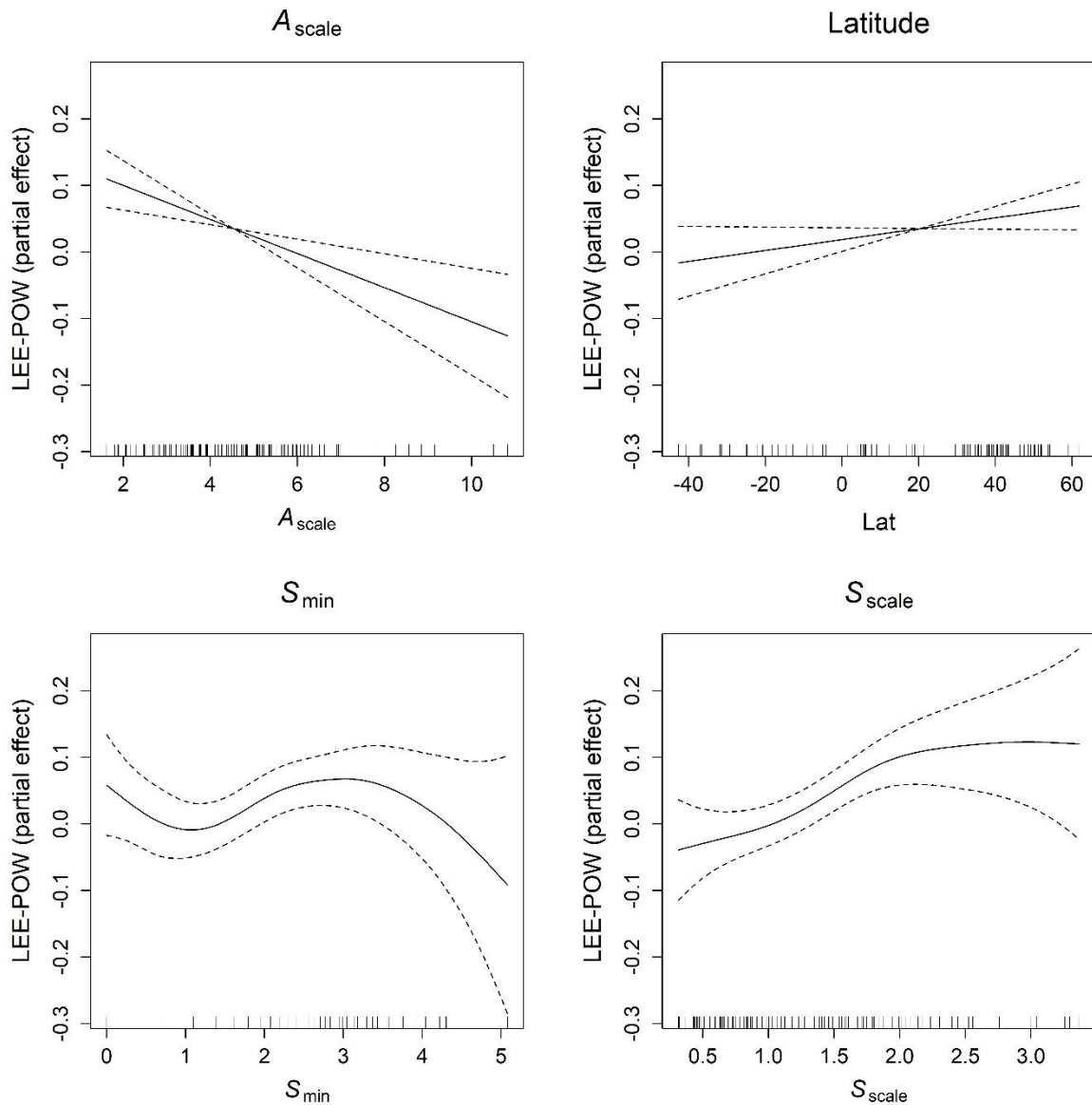
678



680

681 **Figure 1** The varying species richness predictions of five ISAR models. Each of the five
 682 models (see Table 1) was fitted to a simulated archipelago consisting of eight islands of
 683 varying size (1, 3, 7, 14, 17, 22, 26, and 30; undefined units) and richness (3, 7, 14, 18, 20,
 684 23, 24, and 25). These model fits were then used to predict the richness of an island of size 80
 685 (grey dotted line).

686



687

688 **Figure 2** Fitted smoothers from the best fit generalized additive model showing the partial
 689 effects of A_{scale} , Latitude, S_{min} and S_{scale} on the LEE-POW values. The fitted values have been
 690 shifted in each plot by adding the model intercept (0.04) value (using the shift argument in
 691 the plot.gam R function). The effective degrees of freedom for each smoother are: A_{scale}
 692 (1.00), Latitude (1.00), S_{min} (3.53) and S_{scale} (2.70). The dashed lines represent the standard
 693 error curves (two SE above and below). Each LEE-POW value relates to the accuracy of a
 694 prediction of the number of species on a habitat island using the power model. For each of
 695 120 habitat island datasets, the largest island and all islands larger than half the size of the
 696 largest island were removed and the power model fitted to the filtered dataset and
 697 extrapolated.

698

699