PNAS
www.pnas.org

Supplementary Information for

A global model of island species–area relationships

Thomas J. Matthews, François Rigal, Kostas A. Triantis, Robert J. Whittaker

Correspondence to: Robert J. Whittaker

Email: robert.whittaker@ouce.ox.ac.uk

This PDF file includes:

SI Materials and Methods
Tables S1 to S6
Figures S1 to S3
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Dataset S1

www.pnas.org/cgi/doi/10.1073/pnas.1818190116

SI Materials and Methods

Data collection. The datasets were originally sourced by (1) and the full dataset collection methodology is presented there. For present purposes, each possible dataset was checked to ensure the following conditions applied:

1. The datasets each pertained to archipelagos or geographically coherent groups of archipelagos of true geographical islands, i.e. areas of land surrounded by water.
2. Of these datasets, we retained only those where the source paper provided a full list of species per island, or at least the number of species present on each island and in the system (i.e. archipelago) as a whole.
3. Each extracted dataset was distinct from the other datasets already collected. We did include some cases where, for example, data were available from adjacent island groups and also were collated as a regional data set.
4. The slope of the power (log-log) model (details below) was significantly different from zero.

Each dataset provides species richness for a particular taxonomic group, e.g. beetles, spiders, land snails, etc, rather than for all invertebrates: reflecting that thorough sampling has typically been carried out only for particular groups. In testing for taxon effects, we grouped the data into three higher taxa (invertebrates, vertebrates, higher plants) in order to maintain large enough sample sizes, recognizing that the variability in response within vertebrate taxa and within invertebrate taxa may have added some noise to the analysis. In compiling and reviewing the database to select the eventual 151 data sets, we followed the original source papers as guidance on the validity of the datasets.

Variables of interest were extracted from the source papers. These included the taxon sampled (‘Taxon’, classified as vertebrates, invertebrates or plants), archipelago richness (‘Gamma’), the number of islands (NumIsl), the total area of the archipelago (ArchArea), the areas of the smallest (MinArea) and largest islands (MaxArea) in the archipelago, all areas being expressed in km^2, and the ratio of the largest island area to the smallest island area (AreaScale). We also measured the geographical isolation of the archipelago (Isolation) in metres from the nearest mainland. The variable Taxon was turned into dummy binary variables. Vertebrates was taken to be the base category, and then two dummy variables were created that classified a dataset as an ‘Invertebrate’ or ‘Plant’ (in all cases vascular plants) dataset (*SI, Data S1*). Prior to our analysis, we graphically examined the distributions of all variables for outliers and severe departures from normality. All variables, with the exception of logC (which is already logged) and the two dummy variables, were log-transformed (natural logarithms) to approximate normality. To derive comparable estimates in the following analysis (see below), all variables were standardized to a mean of zero and standard deviation of one. Multicollinearity was assessed among all the predictors using Pearson’s correlation with a threshold $|r<0.7|$, following (2). MinArea and MaxArea were both removed due to their strong correlation with ArchArea (Table S1). ArchArea is more informative: (i) because it represents the total landmass of an archipelago and thus incorporates information from all islands, smallest and largest included, and (ii) because of its established importance in explaining variation in Gamma (3).

Theoretical causal structural equation model(s). Our theoretical causal model is illustrated in Fig. 1 and the logic behind it is as follows. Based on the demonstration of power archipelago species–area relationships (ASARs; (3)), Gamma is hypothesized to be primarily a function of Taxon (diversity of Plants >Invertebrates >Vertebrates) and ArchArea, both of which are included as exogenous variables. Gamma is also hypothesized to be a function of geographical isolation given that island biogeography theory (4) predicts a reduction in richness with increasing distance from the source pool. Variation in both *z* and logC has been shown in previous ISAR meta-analyses to be linked to variation in AreaScale (4, 5; and see 6), but AreaScale is only one aspect of how archipelago area is distributed. Hence, we also included NumIsl, which captures additional information concerning the subdivision of total archipelago area, as an exogenous variable. As outlined above, the area of the smallest and largest islands in an archipelago were also each originally considered, but subsequently removed due to multicollinearity issues (above, Table S1). AreaScale was not considered as an exogenous variable since it is not causally independent from ArchArea and NumIsl. Thus, AreaScale was hypothesized to be a function of both ArchArea and NumIsl, and was included as an endogenous variable in the model (Fig. 1). Gamma was also included as an endogenous variable, and both Gamma and AreaScale were hypothesized to potentially explain variation in logC and *z* (6, 7). Based on previous work and theoretical considerations (e.g. 1, 4–9), we also permitted paths between the exogenous variables ArchArea, NumIsl, Taxon and Isolation, and logC and *z*. First, the species richness of both islands and archipelagos is known to be primarily a function of available resources and habitat diversity (10, 11), for which area provides a valuable proxy, and thus at the archipelago scale is measured by ArchArea. Second, it has also been shown that how the area of archipelagos is subdivided amongst islands and the range of variation in island area may also affect ISAR parameters (4–7), which is reflected through the inclusion in our analyses of ArchArea, NumIsl and AreaScale. Third, it is well known that taxa differ in multiple functional ways (e.g. dispersal ability, body size, lifespan) that can affect their carrying capacity in a given archipelago, and the rate at which their diversity scales with area (4, 6, 12). Fourth, the role of isolation on island diversity and thus on ISAR parameters has long been central to island theory (4–7, 13), although previous work has shown that the ISAR parameter space occupied by distant, nearshore, inland and habitat islands shows a great deal of overlap (1, 5, 14), with the hypothesized effect of steepening ISARs with increasing distance having been repeatedly questioned in the past (e.g. 6, 9, 15).

Whereas initial analysis demonstrated no significant bivariate correlation between logC and *z* (Pearson's correlation: -0.07, P = 0.42, Table S1), we hypothesized a trade-off between the two ISAR parameters, conditioned by the foregoing causal network (cf. 6). This causal hypothesis posits that, taking account of variation between taxa and in the location of archipelagos, increases in Gamma reflecting larger archipelagos and richer species pools should drive a trade-off between logC and *z* values, further modified by the distribution of total archipelago area across variable numbers of islands. Another way of putting this is to say that ISARs are predicted to steepen (and logC values to decrease) as ecological process regimes give way to increasing evolutionary regimes (3, 13, 16). However, our general working hypothesis is that the expected tendency for the slope (*z*) of the fitted power model to increase with isolation is modulated or canalized by variation in the disposition of area within the archipelago and by taxonomic differences in responses to area, isolation and archipelago configuration (e.g. pp. 25–31 in (4), and 6, 14).

The rationale for a link from logC to *z* rather than vice versa is based on the notion that given the same biological process regime, *z* values should be equivalent whilst logC may vary in relation to the biotic richness of the available species pool, reflected at the archipelago level by Gamma. We tested whether this rationale (i.e. whether logC was the causal agent affecting z, or *vice versa*) could be supported analytically using the method of Vinod Causality (17). We implemented the Vinod method with the generalCorr R package (18). Briefly, for a given pair of variables X and Y, the method calculates an unanimity index (UI) that quantifies the likelihood that either X or Y is causal. This index will always lie in the range [−100, 100]. Three decision rules based on the value of the UI determine the direction of the causal path. If the UI lies in the interval [−100, −15], then Y causes X, and if UI is in the interval [15, 100] then X causes Y. If the UI lies within the range [−15, 15], the causal direction is indeterminate. More details about the approach can be found in (17). We conducted our analysis separately using the raw variables logC and *z* and with the residuals of their respective models (Logc ~ AreaSum + NoIsl + Iso + AScale + Plants + Inverts; and *z* ~ AreaSum + NoIsl + Iso + AScale + Plants + Inverts). In both cases, we found that the causal path Log C→ Z was supported by Vinod's criteria for causality with a UI of 31.5 and 37.01 respectively.

Among the ISARs retrieved for our analysis, some belong to the same archipelago. For instance, for the Galapagos, ISAR data were obtained for land-snails, ants, mites, and plants. In total, our all-ISARs dataset includes 151 ISARs from 89 archipelagos. The role of the archipelagic context in driving island biogeographical patterns has been well documented over the last decade (e.g. 19, 20). Species diversity patterns of different taxa within the same archipelago may potentially be constrained by similar climatic conditions, distance from the potential species pool, intraarchipelagic isolation and geological and mainland connectivity history, thus generating ISARs that might not be considered as completely independent of each other, and arguably violating the assumption of independence of data points. Therefore, to account for the non-independence of our data within archipelagos, we included Archipelago identity (i.e. the archipelago name) as a random effect in the SEM analysis by using linear mixed models (LMM), fitted using restricted maximum likelihood.

Different types of archipelagos were considered in our model, namely oceanic (39 ISARs), continental (64 ISARs), atoll (8 ISARs), inland (22 ISARs) and mixed archipelagos (18 ISARs), the latter category including at least two (oceanic and one other marine form) of the aforementioned types. The process regime of volcanic oceanic islands is arguably distinct from other categories of island, as diversity patterns are shaped mainly by the geological dynamics of the archipelagoes and by their perpetual and considerable isolation from mainland species pools (e.g. 21–22). The resulting dominance of evolutionary dynamics results in high proportions of endemism and the expectation of steeper ISARs for these archipelagos than either low lying atoll systems, or the rather heterogeneous continental island types (e.g. involving land-bridge islands) or the much less isolated inland islands (13, 16). Our previous analyses support the expectation of steeper ISARs for oceanic archipelagos, although less clearly so than originally anticipated in MacArthur and Wilson’s equilibrium theory of island biogeography (4) (see 4, 5). For these reasons and to assess the generality of our model for specific archipelago types, we re-ran our analyses using the subset of oceanic-ISARs datasets (N = 39) and the subset of continental-ISARs datasets (N = 64): the two largest groups of ISARs in our dataset. The remaining subsets (e.g. atolls) contained too few datasets for analysis. This additional step should be viewed with slight caution due to the smaller number of datasets involved, relative to our main all-ISARs analysis. However, these analyses serve the purpose of indicating whether patterns in the all-ISARs results are caused by patterns in one specific archipelago type, or whether the results are consistent across different archipelago types.

Evaluation of the predictive power of the best path models. To assess the generality of our results, we adopted a repeated k-fold cross validation approach whereby we randomly partitioned the datasets into ten equal components (k = 10). We then put aside one component as the test data and fitted the model to the remaining nine components (the training data) and used the resultant path coefficients to predict the values of the four endogenous variables (*z*, logC, Gamma and AreaScale) in the training data. The next component was then selected as the test data and the process repeated, and so on, across all ten components. We assessed the predictive power of each model on the basis of the Pearson's correlation calculated between the predicted and observed values and subsequently averaged across the 10-folds. This ten-fold cross-validation process was then repeated 100 times and the mean correlation with its associated 95% confidence interval value taken. The above procedure was undertaken separately for the all-ISARs, the oceanic-ISARs and the continental-ISARs datasets (Table S5). In each case, the model used was the best SEM selected by the backward procedure (described above). The average correlation between the observed and the predicted endogenous variables values was > 0.5 in all cases, with the exception of AreaScale, for the three sets of datasets (Table S5). These findings support the generality of our best SEMs and indicate that the biological relationships described have predictive power, i.e. they may extend to other archipelagoes and island systems.

Testing for interaction effects. As a further test of the sensitivity of our modelling approach we explored the possibility that taxon effects may not have been fully captured in our structural equation modelling approach. To do this, we examined a total of six potential interactions involving the variable Taxon as the moderator. These were taxon moderating the relationship Gamma←Isolation and Gamma←ArchArea; LogC←Isolation and LogC ←ArchArea and Z←Isolation and Z←ArchArea. Thus, all the interactions including the key biogeographical variables and the three main endogenous variables were tested (Fig. S2). As taxon is represented by two dummy variables, this adds a total of four additional variables in the models for Gamma, logC and *z*. We then applied the same statistical model selection procedure to this new theoretical causal model. A summary of our backward stepwise selection procedure is presented in Table S6 and the best path model is presented in Fig. S3. We found that, with the exception of the added interactions, all remaining paths did not change (compare with Fig. 2 in the main text). As a consequence, the R^2_m values also did not change apart from a 3% increase for Gamma. Indeed, only two weak but significant interactions were detected, which were that Plants interacted with Isolation in explaining Gamma (negative effect) and Plants interacted with ArchArea in explaining Gamma (positive effect) (see Fig. S3). Therefore, these additional results are consistent with and lend support to the preferred model reported for the all-ISARs dataset (Fig. 2) in the main text.

Table S1. Pearson's pairwise correlations between all the variables used in the study for all-ISARs (n=151 datasets). Pairwise correlation > |0.7| are in bold. NumIsl = Number of islands; Inverts = invertebrates; MinArea = Area of the smallest island and MaxArea = Area of the largest island; AreaScale is the ratio between the largest and the smallest islands within each archipelago; and ArchArea is the total area of the archipelago. All variables with the exception of logC (which is already logged) and the two dummy variables were log-transformed (natural logarithms) prior to analysis. Pairwise correlations between logC and *z* and with all the predictors are given to illustrate the strength of the relationship between all the pair of variables prior to the path analysis.

	z	logC	Gamma	Inverts	Plants	NumIsl	ArchArea	MinArea	MaxArea	AreaScale
logC	-0.07									
Gamma	0.30	0.46								
Inverts	-0.07	-0.33	-0.09							
Plants	0.11	0.66	0.55	-0.46						
NumIsl	-0.23	0.31	0.18	-0.17	0.19					
ArchArea	-0.07	-0.60	0.20	0.21	-0.30	0.00				
MinArea	0.15	-0.59	0.10	0.16	-0.28	-0.31	**0.80**			
MaxArea	-0.07	-0.61	0.19	0.21	-0.30	-0.06	**0.99**	**0.77**		
AreaScale	-0.33	-0.01	0.13	0.07	-0.02	0.38	0.26	-0.36	0.32	
Isolation	0.11	-0.21	0.17	0.16	*0.04*	-0.26	0.39	0.40	0.39	-0.03

Table S2. Summary of the backward stepwise selection procedure for the theoretical causal model for the all-ISARs dataset (n = 151), the oceanic-ISARs dataset (n = 39) and the continental ISARs dataset (n = 64). Models were fitted using piecewise structural equation modelling (piecewiseSEM) and linear mixed effect models (LMM), with Archipelago identity as a random effect. After validating our hypothesized causal model (Fisher's C statistic, P < 0.05), we excluded non-significant paths with the highest P-values in a backward procedure until all remaining paths were statistically significant (P < 0.05). At each step of the backward procedure, the non-significant path with the highest *P*-value was dropped sequentially from the model and, at each step, the reduced model fit was evaluated using the Fisher's C statistic and the AIC_c value of the model stored. At each step, a reduced model was accepted as providing a good fit to the data if the Fisher's C statistic test was non-significant (P > 0.05). Finally, the best model was chosen by selecting the model, across all accepted models (i.e. the full model and any of the reduced models with a non-significant Fisher's C statistic), with the lowest AIC_c value. In the table, for each step of the procedure, the dropped path is given with arrows indicating the direction of the relationship. The values of Fisher's C statistic (*C*), the associated degree of freedom (*df*) and *P*-values (*P*), values of the marginal R^2 (R^2_m) for LMM for the endogenous variables (*z*, logC, Gamma, AreaScale), as well as values of AIC_c are reported. Results are given for our hypothesized causal model (row *full model*) and for each step of the backward procedure with the corresponding dropped path. Our best model, i.e. the one with the lowest AIC_c, is marked in bold and corresponds to the step 5, 10 and 8 of the backward procedure for all-ISARs, oceanic-ISARs and continental-ISARs subsets, respectively. In all steps, models had satisfactory fits. NumIsl = Number of islands; Inverts = invertebrates and AreaScale is the ratio between the largest and the smallest islands within each archipelago.

all-ISARs dataset	Dropped paths	C	*df*	*P*	R^2_m *z*	R^2_m logC	R^2_m Gamma	R^2_m AreaScale	AICc
Full model		13.283	10	0.208	0.483	0.768	0.448	0.336	102.323
Step 1	z ← NumIsl	13.316	12	0.346	0.485	0.768	0.448	0.336	98.938
Step 2	logC← Isolation	13.891	14	0.458	0.485	0.769	0.448	0.336	96.299
Step 3	Gamma ← Isolation	15.690	16	0.475	0.485	0.769	0.450	0.336	95.243
Step 4	z ← Isolation	13.637	10	0.190	0.481	0.769	0.450	0.336	89.398
Step 5	z ←Plants	**16.020**	**12**	**0.190**	**0.479**	**0.769**	**0.450**	**0.336**	**89.140**
Step 6	logC←AreaScale	19.330	14	0.153	0.479	0.768	0.450	0.336	90.023
Step 7	LogC ← Inverts	21.900	16	0.146	0.479	0.766	0.450	0.336	89.991

Oceanic-ISARs dataset	Dropped paths	C	*df*	*P*	R^2_m *z*	R^2_m logC	R^2_m Gamma	R^2_m AreaScale	AICc
Full model		11.446	10	0.324	0.812	0.556	0.415	0.235	604.079
Step 1	z ← NumIsl	11.788	12	0.463	0.818	0.556	0.415	0.235	492.622
Step 2	logC← Isolation	12.216	14	0.589	0.818	0.563	0.415	0.235	413.489
Step 3	*z*←Inverts	14.088	16	0.592	0.821	0.563	0.415	0.235	361.179
Step 4	*z*←Plants	14.343	18	0.706	0.822	0.563	0.415	0.235	313.486
Step 5	Gamma←Inverts	16.810	20	0.665	0.822	0.563	0.449	0.235	283.959
Step 6	logC←Inverts	12.310	14	0.581	0.822	0.556	0.449	0.235	235.099
Step 7	z ← Isolation	15.633	16	0.479	0.820	0.556	0.449	0.235	219.807
Step 8	logC←AreaScale	19.302	18	0.373	0.820	0.549	0.449	0.235	207.906
Step 9	AreaScale←ArchArea	21.668	20	0.359	0.820	0.549	0.449	0.196	194.075
Step 10	Gamma←Isolation	**14.359**	**14**	**0.423**	**0.820**	**0.549**	**0.356**	**0.196**	**156.933**

Continental-ISARs dataset	Dropped paths	C	*df*	*P*	R^2_m *z*	R^2_m logC	R^2_m Gamma	R^2_m AreaScale	AICc
Full model		9.516	10	0.484	0.446	0.798	0.721	0.231	161.101
Step 1	*z* ←NumIsl	9.584	12	0.652	0.451	0.798	0.721	0.231	151.915
Step 2	logC ←Inverts	10.138	14	0.752	0.451	0.801	0.721	0.231	144.276
Step 3	Gamma←Isolation	10.771	16	0.823	0.451	0.801	0.722	0.231	137.253
Step 4	logC ←Isolation	12.369	18	0.828	0.451	0.801	0.722	0.231	132.459
Step 5	*z* ← Isolation	7.746	12	0.805	0.450	0.801	0.722	0.231	116.564
Step 6	logC← NumIsl	11.090	14	0.679	0.450	0.791	0.722	0.231	115.716
Step 7	*z*←Plants	14.935	16	0.529	0.412	0.791	0.722	0.231	115.779
Step 8	*z* ←Inverts	**16.060**	**18**	**0.588**	**0.383**	**0.791**	**0.722**	**0.231**	**111.259**
Step 9	logC ← AreaScale	22.375	20	0.321	0.383	0.778	0.722	0.231	115.487

Table S3. Standardized path coefficients from the best models for the all-ISARs, oceanic-ISARs and continental-ISARs datasets. Models were fitted using piecewise structural equation modelling (piecewiseSEM) and linear mixed effect models with archipelago identity as a random effect. For each path, the arrow indicates the direction of the relationship. For each path, the estimated standardized path coefficients (Est.), the associated standard error (SE) and *P*-values (*P*) are reported. NumIsl = Number of islands; Inverts = invertebrates; AreaScale is the ratio between the largest and the smallest islands within each archipelago; and ArchArea is the total area of the archipelago.

	All-ISARs dataset			Oceanic-ISARs dataset			Continental-ISARs dataset		
Paths	*Est.*	*SE*	*P*	*Est.*	*SE*	*P*	*Est.*	*SE*	*P*
z←logC	-0.997	0.117	<0.001	-0.906	0.084	<0.001	-0.891	0.216	0.002
z←AreaScale	-0.256	0.063	<0.001	-0.172	0.082	0.049	-0.307	0.113	0.021
z←Gamma	0.930	0.091	<0.001	1.148	0.094	<0.001	0.776	0.163	0.001
z←Inverts	-0.146	0.064	0.025	-	-	-	-	-	-
z←ArchArea	-0.751	0.105	<0.001	-0.464	0.089	<0.001	-0.595	0.178	0.008
logC←NumIsl	0.157	0.046	0.001	0.385	0.113	0.003	-	-	-
logC←AreaScale	0.064	0.048	0.191	-	-	-	0.134	0.062	0.055
logC←Gamma	0.483	0.055	<0.001	0.416	0.145	0.01	0.387	0.094	0.002
logC←ArchArea	-0.650	0.050	<0.001	-0.398	0.125	0.005	-0.539	0.073	<0.001
logC← Inverts	-0.074	0.048	0.131	-	-	-	-	-	-
logC←Plants	0.145	0.061	0.021	0.377	0.133	0.01	0.293	0.100	0.015
Gamma←ArchArea	0.360	0.067	<0.001	0.304	0.136	0.035	0.323	0.075	0.001
Gamma←Inverts	0.190	0.070	0.009	-	-	-	0.385	0.081	0.001
Gamma←Plants	0.730	0.069	<0.001	0.467	0.124	0.001	1.071	0.085	<0.001
AreaScale←ArchArea	0.485	0.068	<0.001	-	-	-	0.391	0.118	0.006
AreaScale←NumIsl	0.382	0.079	<0.001	0.411	0.123	0.003	0.318	0.117	0.019

Table S4. Estimations of the direct and indirect effect of the predictors on the exogenous variables *z* and logC. Estimations are given for the best model obtained for the all-ISARs dataset, the oceanic-ISARs and the continental-ISARs dataset. Direct effects are standardized path coefficients while indirect effects are calculated by multiplying the direct path coefficients along the path mediated by associated variables. The total effect is calculated by summing the direct and indirect effect where both routes of influence apply. Gamma and AreaScale are not included because of the absence of indirect paths with the exogenous variables. Therefore, effects of the predictors to Gamma and AreaScale are only direct effects and correspond to the standardized path coefficients reported in Table S3. NumIsl = Number of islands; Inverts = invertebrates; AreaScale is the ratio between the largest and the smallest islands within each archipelago; and ArchArea is the total area of the archipelago.

		All-ISARs dataset			Oceanic-ISARs dataset			Continental-ISARs dataset		
Endogenous	Exogenous	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total
z	logC	-0.997			-0.906			-0.891		
	Gamma	0.930	-0.481	0.448	1.148	-0.377	0.771	0.776	-0.344	0.432
	ArchArea	-0.750	0.809	0.058	-0.464	0.595	0.131	-0.595	0.620	0.025
	AreaScale	-0.256	-0.064	-0.320	-0.172			-0.307		
	NumIsl		-0.281			-0.419			-0.098	
	Isolation									
	Plants		0.182			0.019			0.202	
	Invertebrates	-0.146	0.159	0.012					0.166	
logC	Gamma	0.483			0.416			0.386		
	ArchArea	-0.649	0.174	-0.475	-0.398	0.127	-0.271	-0.539	0.125	-0.414
	AreaScale	0.064						0.134		
	NumIsl		-0.281		0.385					
	Isolation									
	Plants	0.145	0.352	0.498	0.377	0.194	0.571	0.293	0.413	0.706
	Invertebrates	-0.074	0.092	0.018					0.149	

Table S5. Results of the repeated *k*-fold cross validation sensitivity analysis using, first, the best path model from the all-ISARs dataset analysis, second, the best path model from the oceanic-ISARs dataset analysis, and third, the continental-ISARs dataset analysis. The mean Pearson's correlation *r* between predicted and observed values and the associated 95% confidence interval values are given for the four endogenous variables *z*, logC, Gamma and AreaScale. These results illustrate the generality of our best models and indicate that the relationships we have reported may extend to other archipelagoes and island systems. AreaScale is the ratio between the largest and the smallest islands within each archipelago.

	Endogenous variables			
	Z	logC	Gamma	AreaScale
All-ISARs dataset	0.670 [0.636;0.700]	0.862 [0.827;0.88]	0.648 [0.600;0.690]	0.434 [0.390;0.485]
Oceanic-ISARs dataset	0.877 [0.746;0.944]	0.654 [0.445;0.794]	0.525 [0.266;0.726]	0.367 [0.143;0.623]
Continental-ISARs dataset	0.562 [0.450;0.669]	0.850 [0.757;0.906]	0.800 [0.705;0.86]	0.265 [0.137;0.432]

Table S6. Summary of the backward stepwise selection procedure for the theoretical causal model for the all-ISARs dataset (n = 151) including interaction effects for the key biogeographical variables (Isolation and ArchArea) and the three main endogenous variables (Gamma, LogC and *z*) involving Taxon as a moderator. Models were fitted using piecewise structural equation modelling (piecewiseSEM) and linear mixed effect models (LMM), with Archipelago identity as a random effect. After validating our hypothesized causal model (Fisher's C statistic, P < 0.05), we excluded non-significant paths with the highest P-values in a backward procedure until all remaining paths were statistically significant (P < 0.05). At each step of the backward procedure, the non-significant path with the highest *P*-value was dropped sequentially from the model and, at each step, the reduced model fit was evaluated using the Fisher's C statistic and the AIC$_c$ value of the model stored. At each step, a reduced model was accepted as providing a good fit to the data if the Fisher's C statistic test was non-significant (P > 0.05). Finally, the best model was chosen by selecting the model, across all accepted models (i.e. the full model and any of the reduced models with a non-significant Fisher's C statistic), with the lowest AIC$_c$ value. In the table, for each step of the procedure, the dropped path is given with arrows indicating the direction of the relationship. The values of Fisher's C statistic (*C*), the associated degree of freedom (*df*) and *P*-values (*P*), values of the marginal R^2 (R^2_m) for LMM for the endogenous variables (*z*, logC, Gamma, AreaScale), as well as values of AIC$_c$ are reported. Results are given for our hypothesized causal model (row *full model*) and for each step of the backward procedure with the corresponding dropped path. Our best model, i.e. the one with the lowest AIC$_c$, is marked in bold and corresponds to step 17. In all steps, models had satisfactory fits. NumIsl = Number of islands; Inverts = invertebrates and AreaScale is the ratio between the largest and the smallest islands within each archipelago.

all-ISARs dataset	Dropped paths	C	*df*	*P*	R^2_m *z*	R^2_m logC	R^2_m Gamma	R^2_m AreaScale	AICc
Full model		16.362	10	0.090	0.490	0.774	0.481	0.336	152.959
Step 1	*z* ←NumIsl	16.436	12	0.172	0.491	0.774	0.481	0.336	148.772
Step 2	*z* ← Isolation x Plants	16.428	12	0.172	0.493	0.774	0.481	0.336	144.548
Step 3	Gamma ←Inverts x Isolation	16.351	12	0.176	0.493	0.774	0.483	0.336	140.306
Step 4	*z* ←Plants x ArchArea	16.440	12	0.172	0.495	0.774	0.483	0.336	136.371
Step 5	Gamma ←Isolation	17.143	14	0.249	0.495	0.774	0.486	0.336	133.351
Step 6	logC ← ArchArea x Inverts	17.143	14	0.249	0.495	0.775	0.486	0.336	129.429
Step 7	logC ← Isolation	17.991	16	0.324	0.495	0.776	0.486	0.336	126.720
Step 8	*z* ← Inverts x ArchArea	17.938	16	0.328	0.495	0.776	0.486	0.336	122.855
Step 9	logC ← AreaScale	20.429	18	0.309	0.495	0.775	0.486	0.336	122.428
Step 10	*z* ← Plants	23.060	20	0.286	0.492	0.775	0.486	0.336	122.192

Step 11	Gamma ← ArchArea x Inverts	23.477	20	0.266	0.492	0.775	0.481	0.336	119.078
Step 12	*z* ← Isolation	20.988	14	0.102	0.488	0.775	0.481	0.336	112.267
Step 13	logC ← ArchArea x Plants	21.466	14	0.090	0.488	0.774	0.481	0.336	109.368
Step 14	logC ← Inverts	25.049	16	0.069	0.488	0.770	0.481	0.336	110.457
Step 15	logC ←Inverts x Isolation	24.360	16	0.082	0.488	0.767	0.481	0.336	106.153
Step 16	logC ← Plants x Isolation	24.400	16	0.081	0.488	0.766	0.481	0.336	102.830
Step 17	***z* ← Inverts x Isolation**	**25.087**	**16**	**0.068**	**0.479**	**0.766**	**0.481**	**0.336**	**100.362**

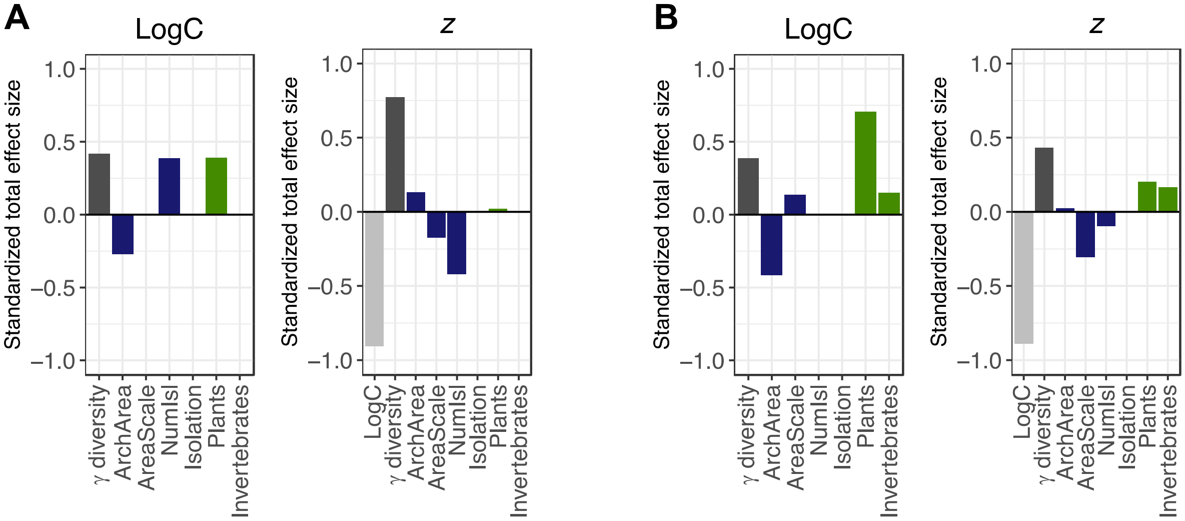

Fig. S1. Standardized total effect size of each variable on *z* and logC calculated by summing the direct and indirect effects derived from the best oceanic-ISARs (A) and continental-ISARs path models (*Materials and Methods/ SI Appendix,* Tables S3, S4).

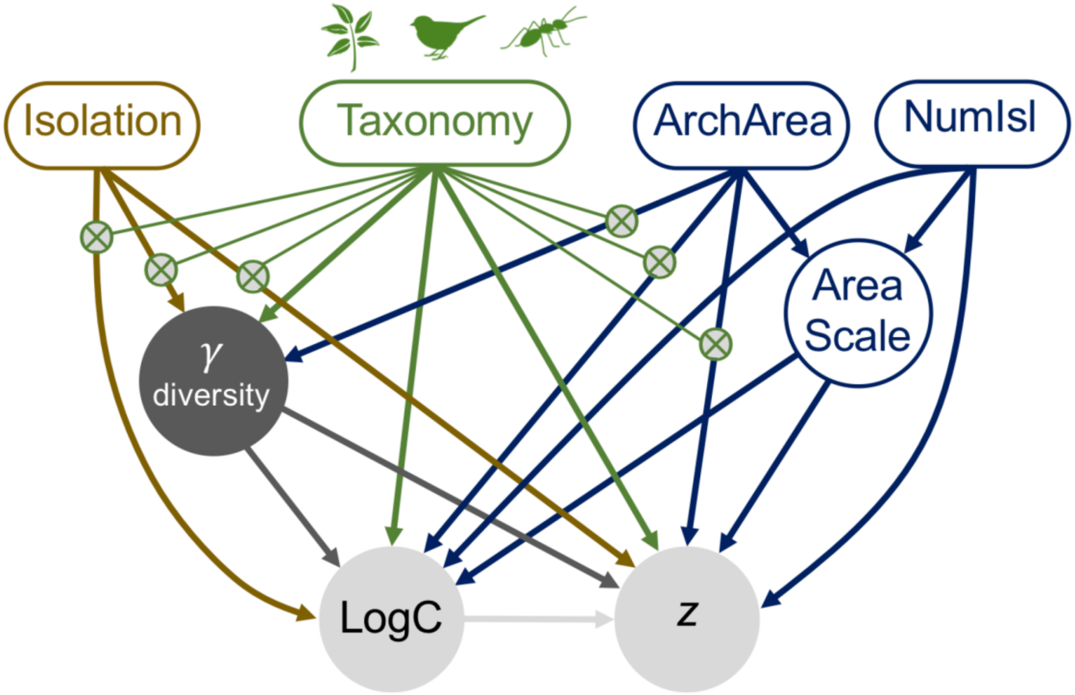

Fig. S2. Model structure of the analysis including interaction effects for the key biogeographical variables and the three main endogenous variables involving Taxon as a moderator. Interactions are represented by circles containing a cross. All other details of the figure are exactly as for Fig. 1.

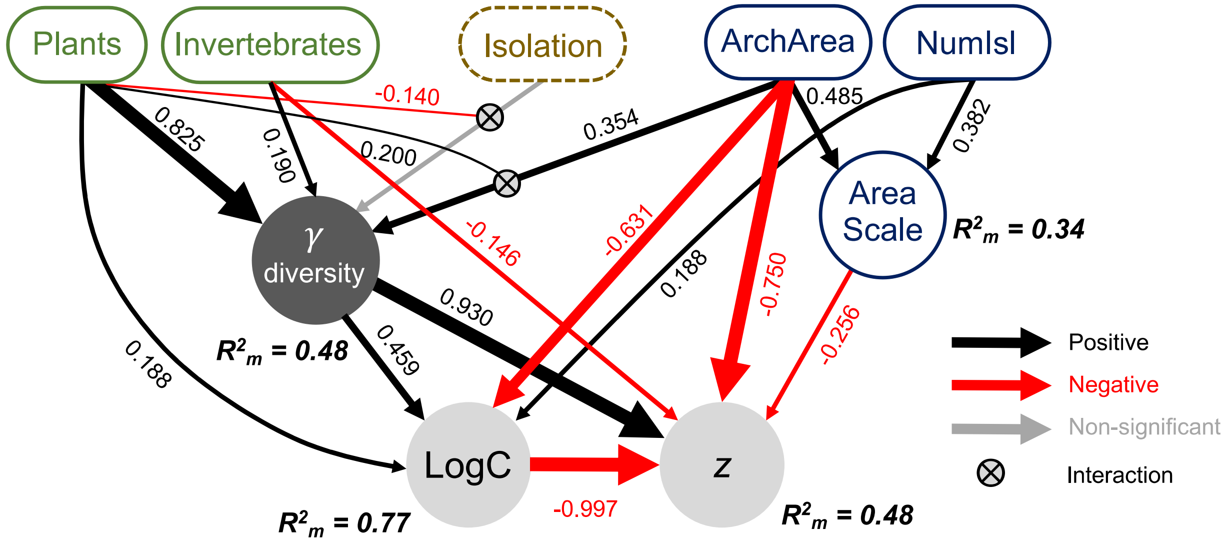

Fig. S3. Best path model for the all-ISARs dataset (n=151) testing for the interactions postulated in Fig. S2. Best path models were obtained using a backward stepwise selection procedure and AIC_C. Pathways show how taxon (Plant and Invertebrates with Vertebrates the base level), isolation, archipelago configuration (ArchArea, NumIsl and AreaScale) and Gamma influence logC and, together, *z*. Piecewise structural equation models were fitted using linear mixed models with Archipelago identity as a random effect. Arrow widths are proportional to standardized path coefficients (values are also given) and marginal R^2_m values (fixed effect) are given for each endogenous variable. Interactions are represented by circles containing a cross. The non-significant path between Isolation and Gamma diversity is not included in the best path model but is shown in the figure for graphical convenience because it is involved in a significant interaction. These models were supported by the data (see Table S6).

Dataset S1. The database of archipelagos (or groups of archipelagos) of island species–area relationships. The file comprises the data for all exogeneous and endogenous variables considered within our analyses, following initial filtering to remove duplicate or strongly overlapping datasets (e.g. alternative versions of ISAR data for the same taxon within a particular archipelago where different numbers of islands were included). The file includes the original source references of the ISAR datasets.

This dataset is to be found in a separate file.

References

1. Triantis KA, Guilhaumon F, Whittaker RJ (2012) The island species–area relationship: biology and statistics. *J Biogeogr* 39:215–231.
2. Dormann CF, et al. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36:27–46.
3. Triantis KA, Economo EP, Guilhaumon F, Ricklefs RE (2015) Diversity regulation at macro-scales: species richness on oceanic archipelagos. *Global Ecol Biogeogr* 24:594–605.
4. MacArthur RH, Wilson EO (1967) *The theory of island biogeography* (Princeton Univ. Press, Princeton).
5. Matthews TJ, Guilhaumon F, Triantis KA, Borregaard MK, Whittaker RJ (2016) On the form of species–area relationships in habitat islands and true islands. *Global Ecol Biogeogr* 25:847–858.
6. Martin TE (1981) Species–area slopes and coefficients: a caution on their interpretation. *Am Nat* 118:823–837.
7. Schoener TW (1976) The species–area relation within archipelagos: models and evidence from island land birds. *16th international ornithological congress*, pp 629–642.
8. Rosenzweig ML (1995) *Species diversity in space and time* (Cambridge Univ. Press, Cambridge).
9. Connor EF, McCoy ED (1979) Statistics and biology of the species–area relationship. *Am Nat* 113:791–833.
10. Triantis KA, Mylonas M, Lika K, Vardinoyannis K (2003) A model for the species–area–habitat relationship. *J Biogeogr* 30:19–27.
11. Kalmar A, Currie DJ (2006) A global model of island biogeography. *Global Ecol Biogeogr* 15:72–81.
12. Holt RD, Lawton JH, Polis GA, Martinez ND (1999) Trophic rank and the species–area relationship. *Ecology* 80:1495–1504.
13. Whittaker RJ, Fernández-Palacios JM, Matthews TJ, Borregaard MK, Triantis KA (2017)

Island biogeography: taking the long view of nature’s laboratories. *Science* 357: eaam8326.

14. Fattorini S, Borges PAV, Dapporto L, Strona G (2017) What can the parameters of the species–area relationship (SAR) tell us? Insights from Mediterranean islands. *J Biogeogr* 44:1018–1028.
15. Williamson M (1981) Relationship of species number to area, distance and other variables. *Analytical Biogeography: an integrated approach to the study of animal and plant distributions*, eds Myers AA, Giller PS (Chapman and Hall, London), pp 91–115.
16. Triantis KA, Mylonas M, Whittaker RJ (2008) Evolutionary species–area curves as revealed by single-island endemics: insights for the inter-provincial species–area relationship. *Ecography* 31:401–407.
17. Vinod HD (2017) Generalized correlation and kernel causality with applications in development economics. *Commun Stat Simul Comput 46*:4513–4534.
18. Vinod HD (2017) *generalCorr: generalized correlations and initial causal path* (R package).
19. Borregaard MK, et al. (2017) Oceanic island biogeography through the lens of the general dynamic model: assessment and prospect. *Biol Rev Camb Philos Soc* 92:830–853.
20. Bunnefeld N, Phillimore AB (2012) Island, archipelago and taxon effects: mixed models as a means of dealing with the imperfect design of nature's experiments. *Ecography* 35:15–22.
21. Borregaard MK, Matthews TJ, Whittaker RJ (2016) The general dynamic model: towards a unified theory of island biogeography? *Global Ecol Biogeogr* 25:805–816.
22. Whittaker RJ, Triantis KA, Ladle RJ (2008) A general dynamic theory of oceanic island biogeography. *J Biogeogr* 35:977–994.