

Title: A global model of island species–area relationships

Authors: Thomas J. Matthews^{1,2}, François Rigal^{2,3}, Kostas A. Triantis⁴, Robert J. Whittaker^{5,6*}

Affiliations:

¹School of Geography, Earth and Environmental Sciences, and Birmingham Institute of Forest Research, University of Birmingham, Birmingham B15 2TT, UK.

²Centre for Ecology, Evolution and Environmental Changes (CE3C)–Azorean Biodiversity Group and Universidade dos Açores–Depto de Ciências Agrárias e Engenharia do Ambiente, PT-9700-042, Angra do Heroísmo, Açores, Portugal.

³CNRS - Université de Pau et des Pays de l'Adour – E2S UPPA, Institut des Sciences Analytiques et de Physico-Chimie pour l'Environnement et les Matériaux, MIRA, UMR5254, 64000, PAU, France.

⁴Department of Ecology and Taxonomy, Faculty of Biology, National and Kapodistrian University of Athens, Athens GR-15784, Greece.

⁵School of Geography and the Environment, University of Oxford, Oxford OX1 3QY, UK.

⁶Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark.

*Corresponding author. Email: robert.whittaker@ouce.ox.ac.uk

Thomas J. Matthews: <https://orcid.org/0000-0002-7624-244X>

Kostas A. Triantis <https://orcid.org/0000-0003-2737-8890>

François Rigal <https://orcid.org/0000-0001-6882-1591>

Robert J. Whittaker <https://orcid.org/0000-0001-7775-3383>

Abstract:

The increase in species richness with island area (the ISAR) is a well-established global pattern, commonly described by the power model, the parameters of which are hypothesized to vary with system isolation and to be indicative of ecological process regimes. We tested a structural equation model of ISAR parameter variation as a function of taxon, isolation, and archipelago configuration, using a globally distributed dataset of 151 archipelagos encompassing a range of taxa and archipelago types. The resulting models revealed a negative relationship between ISAR intercept and slope as a function of archipelago species richness, in turn shaped by taxon differences and by the amount and disposition of archipelago area. These results suggest that local-scale (intra-archipelago) processes have a substantial role in determining ISAR form, obscuring the diversity patterns predicted by island theory as a function of archipelago isolation. These findings have implications for the use and interpretation of ISARs as a tool within biogeography, ecology and conservation.

Keywords:

archipelago effects | diversity model | island biogeography | macroecology | species–area relationship

Significance

The island species–area relationship (ISAR) is a fundamental diversity pattern, best described by the power model. Biogeographic theory assumes predictable variation in power model parameters in relation principally to system isolation, but these assumptions are only weakly supported by previous work, which has been limited in considering the two parameters

separately and over-simplistically. By developing and testing a hierarchical (structural equation) model of factors influencing ISAR form, we show that island species diversity patterns are shaped by intra-archipelago processes more strongly than by isolation from mainland source pools. These findings point to a need to quantify the role of differing scales of isolation in influencing propagule exchange among insular systems in order to develop improved predictive diversity models.

Main Text:

The island species–area relationship (ISAR) is a fundamental macroecological pattern, the study of which has been closely connected with the investigation of processes responsible for the generation, maintenance and loss of biological diversity (1–5). The ISAR is commonly described by the power model, which in its logarithmic form is given by $\log S = \log C + z \log A$ (where S = island species richness, A = island area, and z and $\log C$ are fitted parameters representing the slope and intercept of the model, respectively). Meta-analyses of 612 island datasets have shown the power model to provide: (i) significant fits in 75% of cases, and (ii) the best general model from 20 SAR models tested (6, 7). Island biogeographical theory posits that $\log C$ and z should increase, respectively, as a function of the biotic richness of the source pool region and the isolation of the archipelagos: the latter reflecting increased importance of evolutionary process with distance from the mainland (1, 2, 7, 8). These trends are evident in the aforementioned meta-analyses and wider literature, but only weakly so, reflecting the many confounding factors involved and undermining confidence in the biological interpretation of $\log C$ and z . Intriguingly, while both parameters are necessary to specify any given ISAR, variation in each has

traditionally been examined separately (1, 6–11). We therefore set out to explore whether departure from expected patterns of variation in z might in part be coupled to variation in $\log C$. Through exploratory analyses of our database of significant power model ISARs (*Materials and Methods*), we observed that despite $\log C$ and z values being statistically unrelated in bivariate correlation (below), negative co-variation between them appeared to emerge with increasing archipelago species richness (herein Γ). This led us, in turn, to question how the characteristics of each archipelago may modify the ISAR variation otherwise predicted to occur as a function of increasing isolation from mainland sources (above, and see 1, 12, 13). Drawing on previous work, we selected a set of key archipelagic properties shown or hypothesized to influence ISAR form and we developed a statement of plausible hierarchical causal influences (e.g. archipelago area and isolation may both influence Γ but not vice versa), constituting a general hypothesis of ISAR parameter variation (Fig. 1; *Materials and Methods/SI*). The model encapsulates classic hypothesized roles for differences between major taxa, geographical isolation and archipelago configuration (Fig. 1) (1, 11, 12).

We then applied piecewise structural equation modelling (SEM (14)) using backward stepwise selection and AICc, to a database of ISARs ($N = 151$), in order to test our general model (*Materials and Methods*). The parent–child relationships (paths) of variables in the full model were specified but not the path sign. Within the SEM, AreaScale (the ratio between the area of the largest and the smallest island within each archipelago), Γ , $\log C$ and z are all classed as endogenous variables (those influenced by one or more other variables), whilst the remaining variables are exogenous (Fig. 1). The model structure ultimately focuses attention on how richness scales with area and thus on z , which together with variation in $\log C$, is crucial to biological interpretations and applications of ISARs, for example, in island biogeography and

conservation science (3–5, 15–17). Also of interest is the extent to which archipelagos of differing long term geo-environmental dynamics (e.g. hotspot oceanic vs land-bridge) generate different ISAR form (6–8, 13, 16, 18–20). Accordingly, we also report analyses for our two largest sub-sets, oceanic ($n = 39$) and continental ($n = 64$) archipelagos.

Results

Our full dataset (hereafter ‘all-ISARs’) encompasses globally representative variation in: Gamma diversity from 5 to 3394 species; total archipelago area (ArchArea) from $<1\text{km}^2$ to 1,594,760 km^2 ; system type from islands within lakes to continental and oceanic archipelagos; and Number of islands (NumIsl) from 6 to 86 (*SI Appendix, Dataset S1*). The best all-ISARs model was identified using a backward stepwise procedure and AICc (Fig. 2, *SI Appendix, Table S2*) and shows that Gamma increases from Vertebrates, to Invertebrates, to Plants (the largest effect), which reflects trophic and general ecological differences (cf. 21) and that our plant data are in each case for all higher plants, while the animal datasets are for limited sub-taxa (e.g. spiders, snails, birds, lizards). Gamma is also (predictably (19)) a positive function of ArchArea (Fig. 2A). AreaScale increases with NumIsl and with ArchArea. The model has high explanatory power for both logC and z (Fig. 2A). The marginal $R^2\text{m}$ values (fixed factors only) for z , logC and Gamma (respectively 0.48, 0.77 and 0.45) were almost identical to the conditional $R^2\text{c}$ values (all factors including the random effect, respectively 0.48, 0.8 and 0.53), indicating that the random factor ‘Archipelago identity’ accounted for very little additional variance in the analysis, except for AreaScale for which ‘Archipelago identity’ captures a substantial amount of variance ($R^2\text{m} = 0.34$, $R^2\text{c} = 0.81$). At the core of the path diagram there is a strong negative relationship between logC and z , driven by increasing values of Gamma (Figs 2A, 3A). In

declining order of the size of the model coefficients (i.e. direct effects of predictors), logC values decrease in response to ArchArea, and increase in response to Gamma, and NumIsl, with a further taxon effect of Plants (Fig. 2A, [SI Appendix, Table S3](#)). There is also a small positive net effect of Invertebrates on logC through Gamma ([SI Appendix, Table S4](#)). ISAR slope (z) values decrease in response to logC, ArchArea, AreaScale, and Invertebrates, and increase in response to Gamma (Fig. 2, [SI Appendix, Table S3](#)). There is also a negative, indirect effect of NumIsl, and a positive indirect effect of Plants, on z ([SI Appendix, Table S4](#)).

Simplifying the contributions of each variable to the variation in ISAR parameters for the all-ISARs dataset (Fig 2B): (i) for logC, it is apparent that taxon, archipelago configuration and Gamma are each important, with plants and increasing Gamma driving higher values and archipelago configuration having more complex effects; (ii) for z , the interplay between logC and Gamma is of central importance, as are the effects of archipelago configuration, while there are relatively limited net effects deriving from taxon differences. Effects of archipelago configuration on z are also complex, involving a combination of direct and indirect effects, such that there is, for instance, a negligible net (positive) role for ArchArea on z (standardized path coefficient = 0.058), despite its large contribution to the overall model structure and its direct (negative) linkage to z ([SI Appendix, Tables S3 and S4](#)). It is also notable that: (i) based on standard bivariate Pearson's correlation coefficients there is no relationship between logC and z ($r = -0.07$; $P = 0.42$) ([SI Appendix, Table S2](#)) and it is only through analysis of the interactions with Gamma within the SEM that the relationship emerges; and (ii) the theoretical expectation of steeper slopes and lower intercepts with increased system isolation is not supported.

Theoretically, we may expect differences in ISAR parameters to arise between different types of archipelagos (e.g. oceanic volcanic, atolls, continental, mixed groups, inland) that have distinct and contrasting geodynamics. Visual scrutiny of the Gamma, logC, z interrelationship, appears to support this proposition (Fig. 3A). However, re-analysis of the two best represented subsets, oceanic and continental archipelagos, generated remarkably similar overall model structures (Fig. 2, *SI Appendix, Fig. S1A*). We found that the oceanic-ISARs model (Fig. 2C) has a much higher explanatory power for z ($R^2_m = 0.82$), explains slightly less variation in logC ($R^2 = 0.55$) and differs in paths in just two respects from the all-ISARs model (Fig. 2A, C, *SI Appendix, Tables S2–S4*). First, the taxon signal is reduced to a plant versus animal effect (links from invertebrates to Gamma and z being absent), perhaps simply reflecting the smaller number of datasets by which to demonstrate clear differences (Fig 3B). The second is the disappearance of the link from ArchArea to AreaScale. The model for continental-ISARs is comparable in power for logC ($R^2_m = 0.79$) and poorer for z ($R^2_m = 0.38$) than the all-ISARs analysis. These differences are evident also in the Gamma, logC, and z interrelationships shown in Fig. 3, wherein the relationship is not so evident for continental ISARs, especially for the plant datasets: most of which are from land-bridge archipelagos, ranging from the Mediterranean to the Baltic and Canada (*SI Appendix, Dataset S1*). Nonetheless, the respective importance of Gamma and archipelago configuration effects are little changed while the influence of taxon differences on logC and z are more pronounced than for the (smaller) oceanic-ISARs subset (Fig. 2B and *SI Appendix, Fig. S1B, Tables S2–S4*).

Further sensitivity tests were run to evaluate the predictive power of the best path models for all three datasets, using a repeated k-fold cross-validation approach. Outputs produced an average Pearson's correlation between the observed and the predicted endogenous variables values > 0.5

for the four endogenous variables, with the exception of AreaScale (*SI Appendix, Table S5*). Finally, we also tested for interactions between taxon and the three other main endogenous variables (*SI Appendix, SI Materials and Methods*), which once again showed the model structure to be stable (*SI Appendix, Table S6, Figs S2, S3*, and compare with Fig. 2). These findings support the generality of our best path models and indicate that the biological relationships described have predictive power, i.e. they may extend to other archipelagoes and island systems.

In summary, the SEMs reveal that, for a diverse range of taxa, from archipelagos ranging from inland lakes to remote atolls: (i) increases in Gamma are found for larger archipelagos and taxa that are known to have higher densities at community level (e.g. plants versus vertebrate groups such as reptiles), (ii) increases in Gamma drive a trade-off between logC and z values, (iii) further modified by archipelago configuration (the distribution of archipelago area across variable numbers of islands); and (iv) the role of system isolation in relation to Gamma and ISAR parameters anticipated in classic island theory (1) is not evident regardless of the dataset analyzed (Fig. 2, *SI Appendix, Tables S1–S5*).

Discussion

Notwithstanding recent advances, we lack a general consensus as to how individual factors and mechanisms contribute to ISAR form across different taxa, environmental conditions and spatial and temporal scales (7–11, 16–17, 21, 22–24). Such an understanding is essential if the ISAR is to be used effectively as a predictive tool in applied biodiversity science (3–5, 15). Previous work has, however, established a number of broad generalizations. First, species richness variation across islands and archipelagos is primarily a function of area, which not only provides more space for individual plant and animal populations, but is also frequently indicative of

available resources, energy supply, elevational range, or habitat diversity (11, 22, 23). The inclusion of e.g. climatic, habitat, environmental change and island age data can often also raise the explanatory power of diversity models (19, 22–24). Second, while $\log C$ tends to decrease and z to increase with geographical isolation, consistent with reduced immigration rates and consequent increases of *in situ* speciation (1, 6, 8), the ISAR parameter space occupied by distant, nearshore, inland and even habitat islands shows a great deal of overlap (6, 7, 11): the prompt for the present work. Third, alongside differences in ISAR parameters with isolation and between taxa, previous work has shown scale effects on ISAR slope (z) linked to range in island area (6, 7, 10, 25). Fourth, considering whole archipelagos as islands, it has been shown elsewhere that strong archipelago species–area relationships (ASARs) can be obtained when fitting power models to a set of 14 oceanic archipelagos (18), and that steeper slopes are found when restricting those analyses to archipelago endemic species, reflecting enhanced diversification effects among larger (isolated) archipelagos.

In the present analyses, we drew on these prior observations in developing a general hierarchical model of factors that may explain departure from the variation in ISAR form predicted by theories of island biogeography that are framed in relation to the notion of single, dominant source pools. These classic island theories predict increased slope (z) with increasing isolation, as rates of immigration and accompanying gene flow (rescue effects) are reduced and in-situ anagenesis and cladogenesis lead to high incidence of local endemics. Similarly, within conservation science, habitat loss and increased isolation of remnants is hypothesized to cause adjustment (‘relaxation’) of ISAR form, generating time-lagged extinctions (2–4).

Here we have shown that: (i) archipelago area provides a first order control on archipelago diversity (Gamma), (ii) the distribution of that diversity within archipelagos responds to the

internal configuration of the archipelago, such that (iii) regionally structured variation in ISAR parameters (decreasing $\log C$, increasing z with isolation) is modulated by local properties of archipelagos (cf. 1, 10, 12, 13). Thus, whereas ISAR slope (z) may indeed be responding to a changing balance in the role of ecological vs evolutionary processes (2, 8, 20), these effects are only evident in the all-ISARs model via variations in archipelago richness and configuration, and neither the expected negative relationship between system isolation and Gamma, nor the importance of system isolation for ISAR parameters, is evident for our full all-ISARs dataset, or for either the continental or oceanic subsets (Fig. 1, *SI Appendix*, Tables S1–S4).

It is remarkable how much of the global variation in ISAR parameters can be captured by models including taxon effects, archipelago area and two simple metrics of how that area is subdivided, without the need to consider other archipelagic features that are known to have explanatory power for island diversity patterns (e.g. elevational range, island age, climate, energy flow (20–24, 26)). Further work could usefully explore how incorporation of such variables might improve model fits, although the development of more complex models is analytically challenging with the available data if overfitting is to be avoided.

While previous work has shown that switching attention from generalist to specialist species groups, and from wide ranging natives to restricted endemics is accompanied by increasing ISAR slope (8, 27), as predicted with a shift from ‘ecological’ towards ‘evolutionary’ process regimes with geographical isolation, the expected combination of increasing slope and decreasing intercepts with distance from mainland sources (1, 2) fails to feature in our path models (Fig. 1), as combinations of taxon and archipelagic features drive local adjustment of ISAR fits.

We interpret these findings as supporting the importance of archipelago configuration (area distribution, number and spacing of islands), in turn influencing intra-archipelago dynamics, such as metapopulation-like ‘rescue effects’ between islands (12, 28), for adjustment of ISAR form. Interestingly, when considering solely volcanic oceanic islands, typically systems of high local endemism (13, 20), as Gamma increases, an increasingly constrained range of logC vs z combinations is apparent: a pattern which does not appear to have previously been documented empirically. The trade-off in ISAR slope and intercept as a function of archipelago richness (Figs 2, 3) is consistent with the importance of dynamic processes of propagule exchange, population extinction and in situ diversification within archipelagos (1, 2, 20, 26), *but* demonstrates that these processes need not result in a single or narrow-band of canonical slope values determined primarily by distance from major source pools. Future ISAR studies should (i) consider logC and z values in tandem and approach system description in ways that better capture scale-dependency in ISAR form and (ii) extend the approach developed herein to other forms of insular system (e.g. forest fragments, sky islands). Finally, to test these ideas and in order to improve understanding of the biological mechanisms at play, further work is needed to determine how the configuration of the constituent islands within archipelagos and the multiple scales of geographical isolation involved (within archipelagos, between archipelagos and with mainland areas; 12, 13, 26) influence propagule exchange.

Materials and Methods

Data compilation and ISAR models. The datasets were sourced from the literature according to protocols detailed elsewhere (6, 7). In brief, we used two main abstracting/indexing systems (ISI Web of Knowledge and Scopus) with a wide range of search strings. More than 800 journal

papers, books, doctoral theses, online databases, reports and unpublished resources were screened. For present purposes, all datasets were checked (as explained more fully in *SI, Materials and Methods*) to ensure that each dataset retained for analysis provided: (i) a significant power model; (ii) a unique taxon /archipelago combination; (iii) a system of real islands (land surrounded by water). Each of the resulting datasets represents, for that archipelago, a well-sampled taxonomic group (e.g. higher plants, birds, mammals, land snails, Homoptera: Cicadoidea), which we coded for analysis as belonging to: plants, invertebrates, or vertebrates.

Where possible we extracted the binary presence-absence matrix and used this to calculate archipelago species richness ourselves. However, certain studies did not provide the presence-absence matrix; instead simply reporting the number of species on each island and separately for the archipelago (Gamma). All island area values were converted to km² prior to analysis to ensure comparability of logC values. We fitted the power (log-log) ISAR model using linear regression in R (29), extracting the slope (z) and intercept (logC) values. The power model was chosen as it has been shown to be the best performing general model, significantly fitting and thus adequately describing high proportions of ISAR datasets (6, 7); and because the parameters permit comparison between studies and are used in further biogeographical analyses (e.g. 4, 5, 15). Natural logarithms were used. Following (6), for datasets that contained islands with zero species we added 1 to all island richness values prior to analysis. In total, we retrieved 151 datasets (= ‘all-ISARs’) meeting the above criteria from 113 separate sources (*SI Appendix, Dataset SI*).

Other 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 in km²), the areas of the smallest (MinArea) and largest islands (MaxArea) in the archipelago, 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. Measuring archipelagic isolation is not straightforward (26) but for these purposes and given the wide variety of island types/locations, we used: the minimum distance of each archipelago to the closest mainland (or lake edge for islands within lakes), where mainlands were taken to include the world's continents plus Madagascar (the world's largest continental fragment island at 587,040km²) and the largest (>130,000km²) of the world's land bridge islands that were relevant to our study systems (New Guinea, Java, Sumatra, Borneo, Great Britain). This provided an objective, conventional metric, but one that we recognize for some archipelagos may not be the best possible indicator of their isolation from their key source pools (26). The type of archipelago was classified using information in the source papers and the wider literature (e.g. 30) as: inland water body (i.e. islands within lakes), continental (includes land-bridge and other continental shelf islands), oceanic (islands of volcanic origin over an oceanic plate and with no history of connection to continental land masses), atolls, or mixed (a mix of oceanic and other island types found within oceans) (see *SI Appendix, Materials and Methods*).

Analytical strategy. Structural equation models (SEMs) are well suited for evaluating multivariate hypotheses because the direct and indirect effects of predictor variables can be tested simultaneously (31), and, if the SEM is set up correctly, it allows the user to infer causality

from observational data (32). Whereas in multiple linear regressions, the test is for whether a response variable is a linear function of a set of predictor variables, in SEMs we are testing whether the endogenous variables are caused by a set of other variables (these can be a mix of endogenous and exogenous variables) (32).

The first stage of an SEM is to develop a theoretical causal model that outlines the specific hypothesized causal structure between variables. Our model is illustrated in Fig. 1 and described fully in *SI Appendix, SI Materials and Methods*. Based on the previous demonstration of power archipelago species–area relationships (ASARs; 19), Gamma is hypothesized to be primarily a function of Taxon and ArchArea, both of which are included as exogenous variables. Gamma is also hypothesized to be a function of geographical isolation (e.g. 1, 12, 13). NumIsl is included as an exogenous variable as it captures additional information concerning the subdivision of total archipelago area. 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 (10, 25). Based on previous work and theoretical considerations (e.g. 1, 2, 6–9), we also permitted paths between the exogenous variables ArchArea, NumIsl, Taxon and Isolation, and z and logC.

Whereas initial analysis demonstrated no significant bivariate correlation between logC and z (*SI Appendix, Table S1*), we hypothesized a trade-off between the two ISAR parameters, conditioned by the foregoing causal network (cf. 10). 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 $\log C$ may vary in relation to the biotic richness of the available species pool, reflected at the archipelago level by taxon and Gamma. We tested whether this rationale could be supported analytically using the method of Vinod Causality (33, see *SI Appendix, SI Materials and Methods*). The model posits that, taking account of variation between taxa and in the location of archipelagos, increases in Gamma reflecting larger archipelagos and richer species pools may drive a trade-off between $\log C$ and z values, further modified by the distribution of total archipelago area across variable numbers of islands (cf. 10, 12, 13, 25, 26).

The processes driving community assembly differ in balance between island types. In particular, the process regime of volcanic oceanic islands is distinct from other categories of island, as diversity patterns are strongly shaped by the geological dynamics of the archipelagoes and by the opportunities for diversification presented by their permanent and significant isolation from mainland species pools (e.g. 19, 20, 30, 34, 35). 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 also for the subset of continental-ISARs datasets ($N = 64$), a high proportion of which comprise land-bridge islands (likely connected to larger landmasses during Pleistocene sea-level minima), since these represent the two largest groups of ISARs in our dataset. The remaining subsets of island types contained too few datasets for analysis.

Model fitting and validation. The SEMs were fitted using piecewise structural equation modelling (piecewiseSEM (14)). This SEM method enables the overall fit of a multifaceted hierarchical network to be tested, including the estimation of indirect effects, whilst allowing for

the consideration of random effects (36). We assessed overall model fit using direct separation tests (d-sep) based on Fisher's C statistic, with the model being accepted where the associated $P > 0.05$ (36). We first assessed the fit of our hypothetical causal model, before simplifying the model using a backward stepwise selection procedure (37–39). At each step of the backward procedure, the non-significant path with the highest P-value was dropped from the model and the fit of the resultant reduced model was evaluated using Fisher's C statistic (a model was accepted if Fisher's C was non-significant, i.e. $P > 0.05$). The AIC_c value of the resultant reduced model was also stored. This backward process carried on until there were no non-significant paths remaining in the model. 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 (38, 39).

As different archipelagos feature contrasting environmental and biological dynamics (6–8, 11, 13, 20), likely contributing both noise and signal, archipelago identity (i.e. the archipelago name; *SI Appendix, Dataset S1*) was accounted for in the analysis as a random effect within linear mixed models (LMM) fitted using restricted maximum likelihood (14).

For all endogenous variables (z, logC, Gamma and AreaScale), both the conditional R^2_c (all factors including the random effect) and marginal R^2_m (fixed factors only) were computed, following (40). In all best models, the effect of each predictor on the endogenous variables was evaluated through their standardized path coefficients. An overview of the backward procedure for the all datasets and the two subsets (oceanic, continental), is provided in *SI Appendix, Table S2*. Our hypothesized causal models (the full model including all hypothesized paths), as well as our best models, all had satisfactory fits (*SI Appendix, Tables S2*).

The direct and indirect effects of the predictors on z and $\log C$ were calculated using the standardized path coefficients (41). For a given predictor A, the strength of its indirect effect on C through B is obtained by multiplying the direct standardized path coefficients of A on B and B on C. The total effect of A on C is then calculated by summing the standardized path coefficient of its direct effect and the sum of its indirect effects (Fig. 2B, D *SI Appendix, Tables S3, S4*).

Finally, we evaluated the predictive power of the best path models using a repeated k -fold cross validation approach; this procedure was undertaken separately for the all-ISARs and the two island type subsets (see *SI Appendix, Materials and Methods*). As a further test of the sensitivity of our modelling approach we also explored an alternative model with several potential interactions involving the variable Taxon as the moderator (see *SI Appendix, Materials and Methods*). All the analyses were undertaken in R (29). Piecewise SEM models were fitted using the *piecewiseSEM* package (14). The LMM models were implemented with the *nlme* R package (42).

ACKNOWLEDGMENTS. We thank Andrés Baselga, Luís Borda-de-Água, Michael Borregaard, François Guilhaumon, Jamie Owens and Mike Rosenzweig for discussions during the execution of this study and three anonymous reviewers for their constructive criticism. Kostas Proios helped with obtaining isolation measurements. **Funding:** no specific project funding was involved. **Author Contributions:** RJW, TJM and KAT conceived the project, KAT and TJM collected the data, all authors were involved in analysis and writing, coding was led by TJM and FR, writing by RJW and TJM. The authors declare no competing interests.

References

1. MacArthur RH, Wilson EO (1967) *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton).
2. Rosenzweig M (1995) *Species diversity in space and time* (Cambridge Univ. Press, Cambridge).
3. Lewis OT (2006) Climate change, species–area curves and the extinction crisis. *Phil Trans R Soc B* 361:163–171.
4. Ladle RJ (2009) Forecasting extinctions: uncertainties and limitations. *Diversity* 1:133–150.
5. Gerstner K, Dormann, CF, Václavík T, Kreft H, Seppelt R (2014) Accounting for geographical variation in species–area relationships improves the prediction of plant species richness at the global scale. *J Biogeogr* 41:261–273.
6. Triantis, KA, Guilhaumon, F, Whittaker, RJ (2012) The island species–area relationship: biology and statistics. *J Biogeogr* 39:215–231.
7. 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.
8. 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.
9. Connor EF, McCoy ED (1979) Statistics and biology of the species-area relationship. *Am Nat* 113:791–833.

10. Martin TE (1981) Species-area slopes and coefficients: a caution on their interpretation. *Am Nat* 118:823–837.
11. 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.
12. Gascuel F, Laroche F, Bonnet-Lebrun A-S, Rodrigues ASL (2016) The effects of archipelago spatial structure on island diversity and endemism: predictions from a spatially-structured neutral model. *Evolution* 70: 2657–2666.
13. Price JP, et al. (2018) Colonization and diversification shape species–area relationships in three Macaronesian archipelagos. *J Biogeogr* 45:2027–2039.
14. Lefcheck JS (2016) piecewiseSEM: piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods Ecol Evol* 7:573–579.
15. Halley JM, Sgardeli V, Monokrousos N (2013) Species–area relationships and extinction forecasts. *Ann N Y Acad Sci* 1286:50–61.
16. Lomolino MV (2001) The species–area relationship: new challenges for an old pattern. *Progr Phys Geogr* 25:1–21.
17. Tjørve E, Tjørve KMC (2017) Species–area relationship. *eLS* (John Wiley & Sons, Ltd, Chichester). DOI: 10.1002/9780470015902.a0026330.
18. 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.
19. Borregaard MK, Matthews TJ, Whittaker RJ (2016) The general dynamic model: towards a unified theory of island biogeography? *Global Ecol Biogeogr* 25:805–816.

20. 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.
21. Holt RD, Lawton JH, Polis GA, Martinez ND (1999) Trophic rank and the species–area relationship. *Ecology* 80:1495–1504.
22. Triantis KA, Mylonas M, Lika K, Vardinoyannis K (2003) A model for the species-area-habitat relationship. *J Biogeogr* 30:19–27.
23. Kalmar A, Currie DJ (2006) A global model of island biogeography. *Global Ecol Biogeogr* 15:72–81.
24. Weigelt P, Steinbauer MJ, Cabral JS, Kreft H (2016) Late Quaternary climate change shapes island biodiversity. *Nature* 532:99–102.
25. Schoener TW (1976) The species-area relation within archipelagos: models and evidence from island land birds. *Proc 16th Int Ornithol Congr* 1976:629–642.
26. Weigelt P, Kreft H (2013) Quantifying island isolation – insights from global patterns of insular plant species richness. *Ecography* 36:417–429.
27. Matthews TJ, Cottee-Jones HEW, Whittaker RJ (2014) Habitat fragmentation and the species–area relationship: a focus on total species richness obscures the impact of habitat loss on habitat specialists. *Divers Distrib* 20:1136–1146.
28. Brown JH, Kodric-Brown A (1977) Turnover rates in insular biogeography: effect of immigration on extinction. *Ecology* 58:445–449.
29. R Core Team (2017) *R: A Language and Environment for Statistical Computing*. (R foundation for statistical computing, Vienna, Austria) <https://www.R-project.org/>.

30. Gillespie RG, Clague DA (eds) (2009) *Encyclopedia of islands* (Univ. of California Press, California).
31. Grace JB, et al. (2012) Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere* 3:1–44.
32. Shipley B (2016) *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. (Cambridge Univ. Press, Cambridge).
33. Vinod HD (2017) Generalized correlation and kernel causality with applications in development economics. *Commun Stat Simul and Comput* 46:4513–4534.
34. Whittaker RJ, Triantis KA, Ladle RJ (2008) A general dynamic theory of oceanic island biogeography. *J Biogeogr* 35:977–994.
35. 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.
36. Shipley B (2009) Confirmatory path analysis in a generalized multilevel context. *Ecology* 90:363–368.
37. Grace JB (2006) *Structural equation modeling and natural systems* (Cambridge Univ. Press, Cambridge).
38. Kim TN, Holt RD (2012) The direct and indirect effects of fire on the assembly of insect herbivore communities: examples from the Florida scrub habitat. *Oecologia* 168:997–1012.
39. Ando Y, Utsumi S, Ohgushi T (2017) Aphid as a network creator for the plant-associated arthropod community and its consequence for plant reproductive success. *Funct Ecol* 31:632–641.

40. Nakagawa S, Schielzeth H (2013) A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol Evol* 4:133–142.
41. Grace GB, Bollen KA (2005) Interpreting the results from multiple regression and structural equation models. *Bull Ecol Soc Am* 86:283–295.
42. Pinheiro P, Bates D, DebRoy S, Sarkar, D, R Core Team (2015) *nlme: linear and nonlinear mixed effects models* (R Package).

List of Figures

Fig. 1. Model structure. A priori pathways were hypothesized between exogenous (rectangles) and endogenous variables (circles) based on arguments from the literature (see text). Three categories of taxa (green) were considered: Plants (i.e. vascular plants), Vertebrates (e.g. birds, mammals) and Invertebrates (e.g. land snails, beetles). Archipelago configuration (blue) was represented by: NumIsl = Number of islands, AreaScale = the ratio between the largest and the smallest islands within each archipelago, and ArchArea = the total land area of the archipelago. Isolation (brown) = archipelago distance from mainland. Diversity properties (grey) are represented by: Gamma diversity = the total species richness of an archipelago; logC and z, i.e. the parameters of the ISAR of each archipelago. Archipelago identity (not shown) is included as a random factor as some archipelagos are represented by separate datasets for different taxa. For the full set of variables initially considered and their pairwise correlations see [SI Appendix, Supplementary text, Table S1](#).

Fig. 2. Best path models for all-ISARs (n=151) (A, B), oceanic-ISARs (n=39) (C) and continental-ISARs (n=64) (D). Best path models were obtained using a backward stepwise selection procedure and AIC_c. Pathways show how taxon (with vertebrates the base level), isolation, archipelago configuration (ArchArea, NumIsl and AreaScale) and Gamma influence logC and, together, z ([SI Appendix, Table S2](#)). Piecewise structural equation models were fitted using linear mixed models with Archipelago identity as a random effect. These models were supported by the data (all-ISARs: Fisher's C = 16.02, df= 12, P =

0.190; oceanic-ISARs: respectively 14.36, 14, 0.423, continental-ISARs: 16.06, 18, 0.588). 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. Panel B shows the standardized total effect size of logC and of z, calculated by summing the direct and indirect effects derived from the best all-ISARs path model (*Materials and Methods/ SI Appendix, Tables S3, S4*).

Fig. 3. Archipelago richness (Gamma) in relation to the parameters of each ISAR. The five archipelago types are distinguished for the all-ISARs dataset (n = 151) (A), and vascular plants, vertebrates, and invertebrates (note that animal data sets are for sub-sets such as e.g. birds or spiders) for the oceanic-ISARs dataset (n=39) (B) and continental-ISARs dataset (n=64) (C) (*SI Appendix, Dataset S1*).

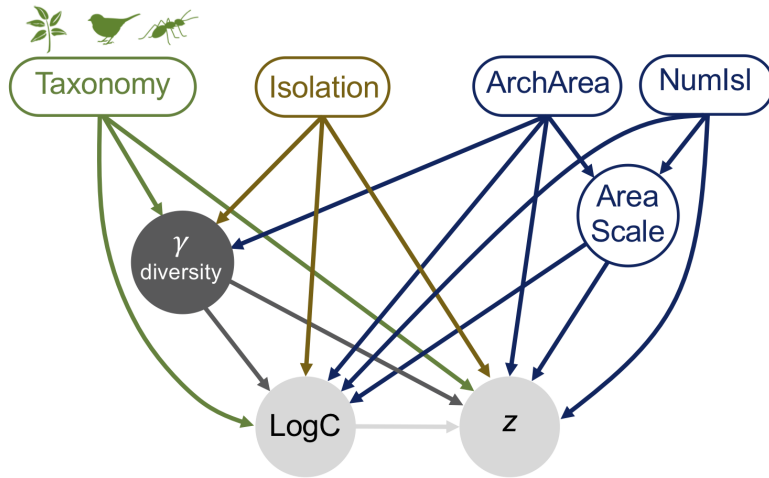


Figure 1

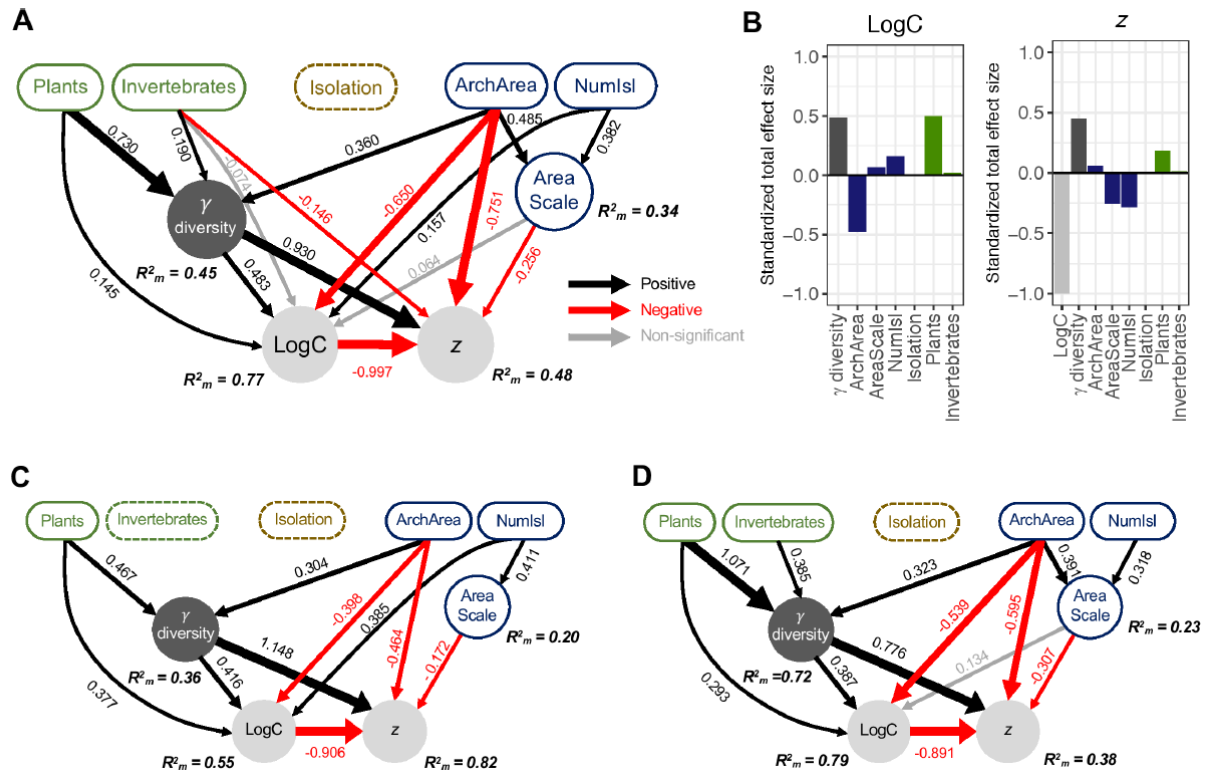


Figure 2

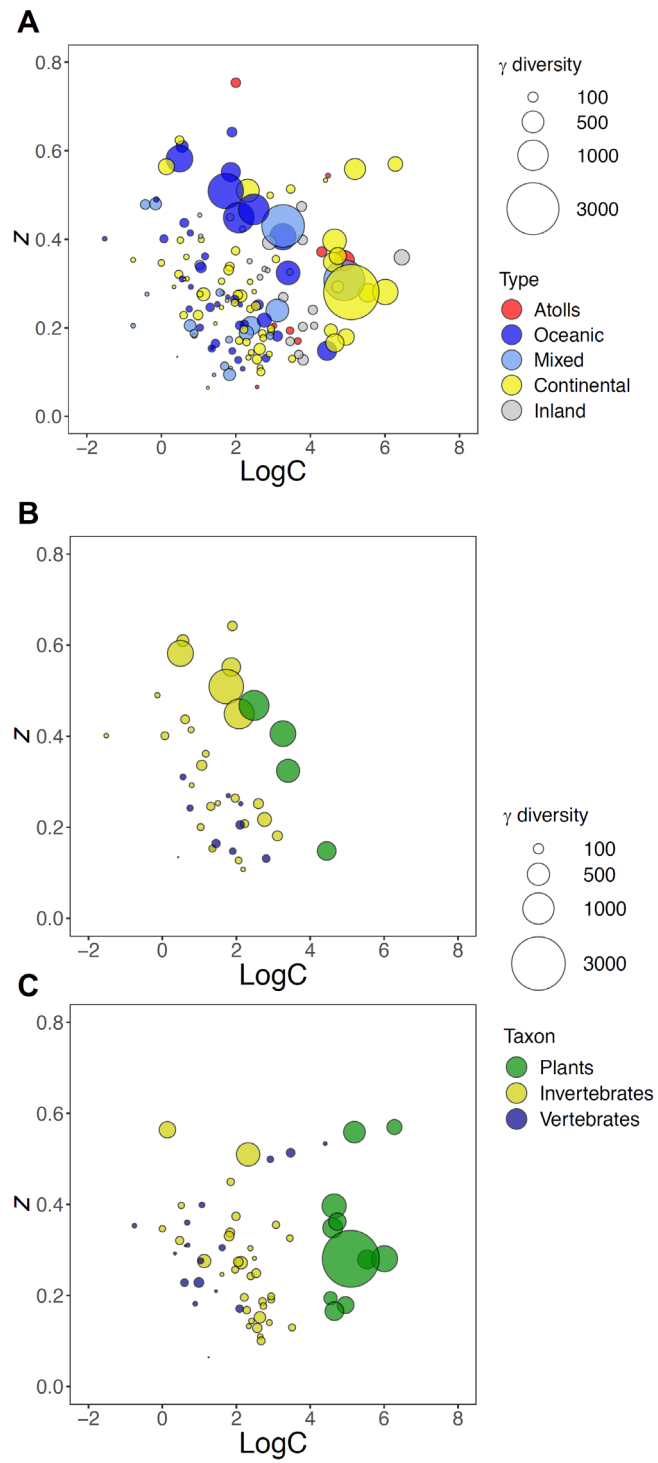


Figure 3

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

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 km², 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 $\log C$ 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 $\log C$ 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 $\log C$ 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 $\log C$ to z rather than vice versa is based on the notion that given the same biological process regime, z values should be equivalent whilst $\log C$ 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 $\log C$ 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 $\log C$

and z and with the residuals of their respective models ($\text{Logc} \sim \text{AreaSum} + \text{NoIsl} + \text{Iso} + \text{AScale} + \text{Plants} + \text{Inverts}$; and $z \sim \text{AreaSum} + \text{NoIsl} + \text{Iso} + \text{AScale} + \text{Plants} + \text{Inverts}$). In both cases, we found that the causal path $\text{Log C} \rightarrow Z$ 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, intra-archipelagic 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 , $\log C$, Γ 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 $\Gamma \leftarrow \text{Isolation}$ and $\Gamma \leftarrow \text{ArchArea}$; $\log C \leftarrow \text{Isolation}$ and $\log C \leftarrow \text{ArchArea}$ and $z \leftarrow \text{Isolation}$ and $z \leftarrow \text{ArchArea}$. 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 Γ , $\log C$ 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 , $\log C$, Γ , 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	$R^2_m z$	$R^2_m \log C$	$R^2_m \Gamma$	$R^2_m \text{AreaScale}$	AIC_c
<i>Full model</i>		13.283	10	0.208	0.483	0.768	0.448	0.336	102.323
Step 1	$z \leftarrow \text{NumIsl}$	13.316	12	0.346	0.485	0.768	0.448	0.336	98.938
Step 2	$\log C \leftarrow \text{Isolation}$	13.891	14	0.458	0.485	0.769	0.448	0.336	96.299
Step 3	$\Gamma \leftarrow \text{Isolation}$	15.690	16	0.475	0.485	0.769	0.450	0.336	95.243
Step 4	$z \leftarrow \text{Isolation}$	13.637	10	0.190	0.481	0.769	0.450	0.336	89.398
Step 5	$z \leftarrow \text{Plants}$	16.020	12	0.190	0.479	0.769	0.450	0.336	89.140
Step 6	$\log C \leftarrow \text{AreaScale}$	19.330	14	0.153	0.479	0.768	0.450	0.336	90.023
Step 7	$\log C \leftarrow \text{Inverts}$	21.900	16	0.146	0.479	0.766	0.450	0.336	89.991

Oceanic- ISARs dataset	Dropped paths	C	df	P	$R^2_m z$	$R^2_m \log C$	$R^2_m \text{Gamma}$	$R^2_m \text{AreaScale}$	AICc
<i>Full model</i>		11.446	10	0.324	0.812	0.556	0.415	0.235	604.079
Step 1	$z \leftarrow \text{NumIsl}$	11.788	12	0.463	0.818	0.556	0.415	0.235	492.622
Step 2	$\log C \leftarrow \text{Isolation}$	12.216	14	0.589	0.818	0.563	0.415	0.235	413.489
Step 3	$z \leftarrow \text{Inverts}$	14.088	16	0.592	0.821	0.563	0.415	0.235	361.179
Step 4	$z \leftarrow \text{Plants}$	14.343	18	0.706	0.822	0.563	0.415	0.235	313.486
Step 5	$\text{Gamma} \leftarrow \text{Inverts}$	16.810	20	0.665	0.822	0.563	0.449	0.235	283.959
Step 6	$\log C \leftarrow \text{Inverts}$	12.310	14	0.581	0.822	0.556	0.449	0.235	235.099
Step 7	$z \leftarrow \text{Isolation}$	15.633	16	0.479	0.820	0.556	0.449	0.235	219.807
Step 8	$\log C \leftarrow \text{AreaScale}$	19.302	18	0.373	0.820	0.549	0.449	0.235	207.906
Step 9	$\text{AreaScale} \leftarrow \text{ArchArea}$	21.668	20	0.359	0.820	0.549	0.449	0.196	194.075
Step 10	$\text{Gamma} \leftarrow \text{Isolation}$	14.359	14	0.423	0.820	0.549	0.356	0.196	156.933
Continental- ISARs dataset	Dropped paths	C	df	P	$R^2_m z$	$R^2_m \log C$	$R^2_m \text{Gamma}$	$R^2_m \text{AreaScale}$	AICc
<i>Full model</i>		9.516	10	0.484	0.446	0.798	0.721	0.231	161.101
Step 1	$z \leftarrow \text{NumIsl}$	9.584	12	0.652	0.451	0.798	0.721	0.231	151.915
Step 2	$\log C \leftarrow \text{Inverts}$	10.138	14	0.752	0.451	0.801	0.721	0.231	144.276
Step 3	$\text{Gamma} \leftarrow \text{Isolation}$	10.771	16	0.823	0.451	0.801	0.722	0.231	137.253
Step 4	$\log C \leftarrow \text{Isolation}$	12.369	18	0.828	0.451	0.801	0.722	0.231	132.459
Step 5	$z \leftarrow \text{Isolation}$	7.746	12	0.805	0.450	0.801	0.722	0.231	116.564
Step 6	$\log C \leftarrow \text{NumIsl}$	11.090	14	0.679	0.450	0.791	0.722	0.231	115.716
Step 7	$z \leftarrow \text{Plants}$	14.935	16	0.529	0.412	0.791	0.722	0.231	115.779
Step 8	$z \leftarrow \text{Inverts}$	16.060	18	0.588	0.383	0.791	0.722	0.231	111.259
Step 9	$\log C \leftarrow \text{AreaScale}$	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.

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

Endogenous	Exogenous	All-ISARs dataset			Oceanic-ISARs dataset			Continental-ISARs dataset		
		Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total
z	$\log C$	-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	
$\log C$	Invertebrates	-0.146	0.159	0.012					0.166	
	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 , $\log C$, Γ 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	$\log C$	Γ	AreaScale
All-ISARs dataset	0.670	0.862	0.648	0.434
	[0.636;0.700]	[0.827;0.88]	[0.600;0.690]	[0.390;0.485]
Oceanic-ISARs dataset	0.877	0.654	0.525	0.367
	[0.746;0.944]	[0.445;0.794]	[0.266;0.726]	[0.143;0.623]
Continental-ISARs dataset	0.562	0.850	0.800	0.265
	[0.450;0.669]	[0.757;0.906]	[0.705;0.86]	[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	$R^2_m z$	$R^2_m \log C$	R^2_m Gamma	R^2_m AreaScale	AIC _c
<i>Full model</i>		16.362	10	0.090	0.490	0.774	0.481	0.336	152.959
Step 1	$z \leftarrow \text{NumIsl}$	16.436	12	0.172	0.491	0.774	0.481	0.336	148.772
Step 2	$z \leftarrow \text{Isolation} \times \text{Plants}$	16.428	12	0.172	0.493	0.774	0.481	0.336	144.548
Step 3	$\text{Gamma} \leftarrow \text{Inverts} \times \text{Isolation}$	16.351	12	0.176	0.493	0.774	0.483	0.336	140.306
Step 4	$z \leftarrow \text{Plants} \times \text{ArchArea}$	16.440	12	0.172	0.495	0.774	0.483	0.336	136.371
Step 5	$\text{Gamma} \leftarrow \text{Isolation}$	17.143	14	0.249	0.495	0.774	0.486	0.336	133.351
Step 6	$\log C \leftarrow \text{ArchArea} \times \text{Inverts}$	17.143	14	0.249	0.495	0.775	0.486	0.336	129.429
Step 7	$\log C \leftarrow \text{Isolation}$	17.991	16	0.324	0.495	0.776	0.486	0.336	126.720
Step 8	$z \leftarrow \text{Inverts} \times \text{ArchArea}$	17.938	16	0.328	0.495	0.776	0.486	0.336	122.855
Step 9	$\log C \leftarrow \text{AreaScale}$	20.429	18	0.309	0.495	0.775	0.486	0.336	122.428
Step 10	$z \leftarrow \text{Plants}$	23.060	20	0.286	0.492	0.775	0.486	0.336	122.192
Step 11	$\text{Gamma} \leftarrow \text{ArchArea} \times \text{Inverts}$	23.477	20	0.266	0.492	0.775	0.481	0.336	119.078
Step 12	$z \leftarrow \text{Isolation}$	20.988	14	0.102	0.488	0.775	0.481	0.336	112.267
Step 13	$\log C \leftarrow \text{ArchArea} \times \text{Plants}$	21.466	14	0.090	0.488	0.774	0.481	0.336	109.368
Step 14	$\log C \leftarrow \text{Inverts}$	25.049	16	0.069	0.488	0.770	0.481	0.336	110.457
Step 15	$\log C \leftarrow \text{Inverts} \times \text{Isolation}$	24.360	16	0.082	0.488	0.767	0.481	0.336	106.153

Step 16	<u>logC \leftarrow Plants x Isolation</u>	24.400	16	0.081	0.488	0.766	0.481	0.336	102.830
Step 17	<u>z \leftarrow Inverts x Isolation</u>	25.087	16	0.068	0.479	0.766	0.481	0.336	100.362

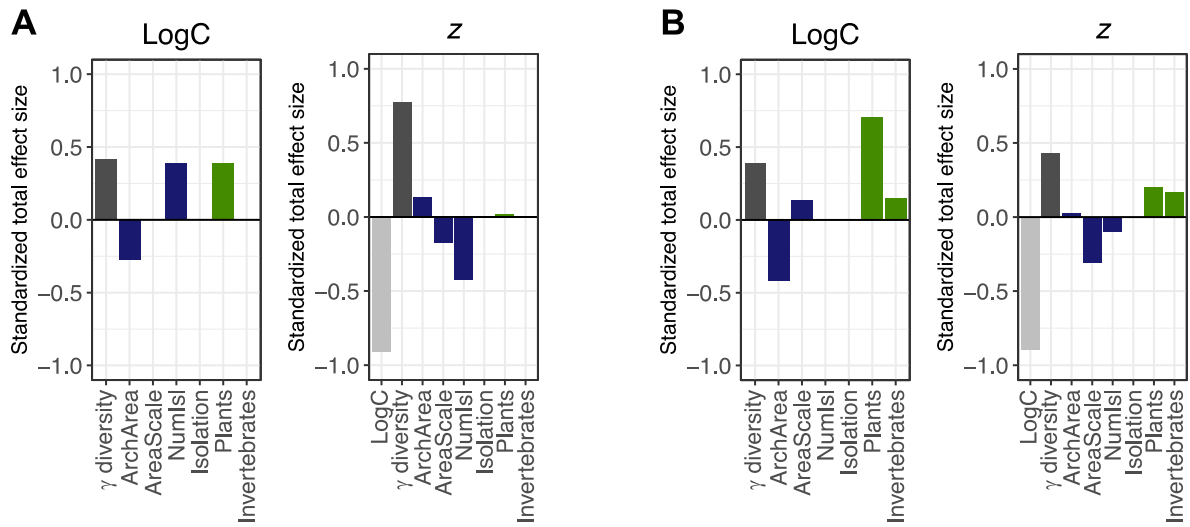


Fig. S1. Standardized total effect size of each variable on z and $\log C$ calculated by summing the direct and indirect effects derived from the best oceanic-ISARs (A) and continental-ISARs path models (*Materials and Methods/ SI Appendix, Tables S3, S4*).

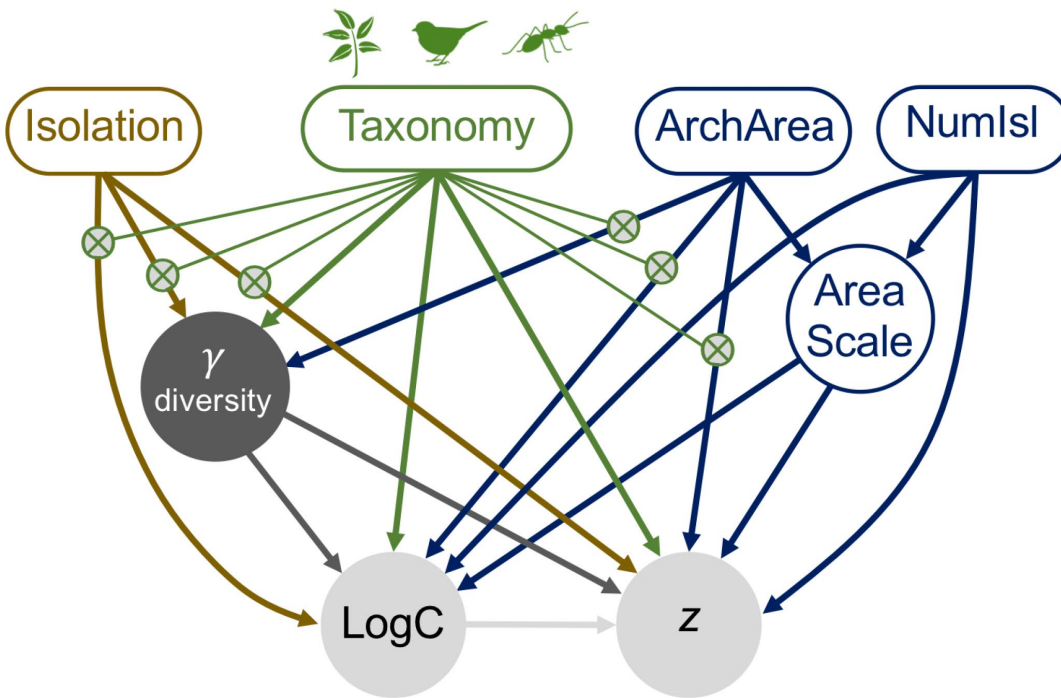


Fig. S2. Model structure of the analysis including interaction effects for the key biogeographical variables and the three main endogenous variables involving Taxon as a moderator. Interactions are represented by circles containing a cross. All other details of the figure are exactly as for Fig. 1.

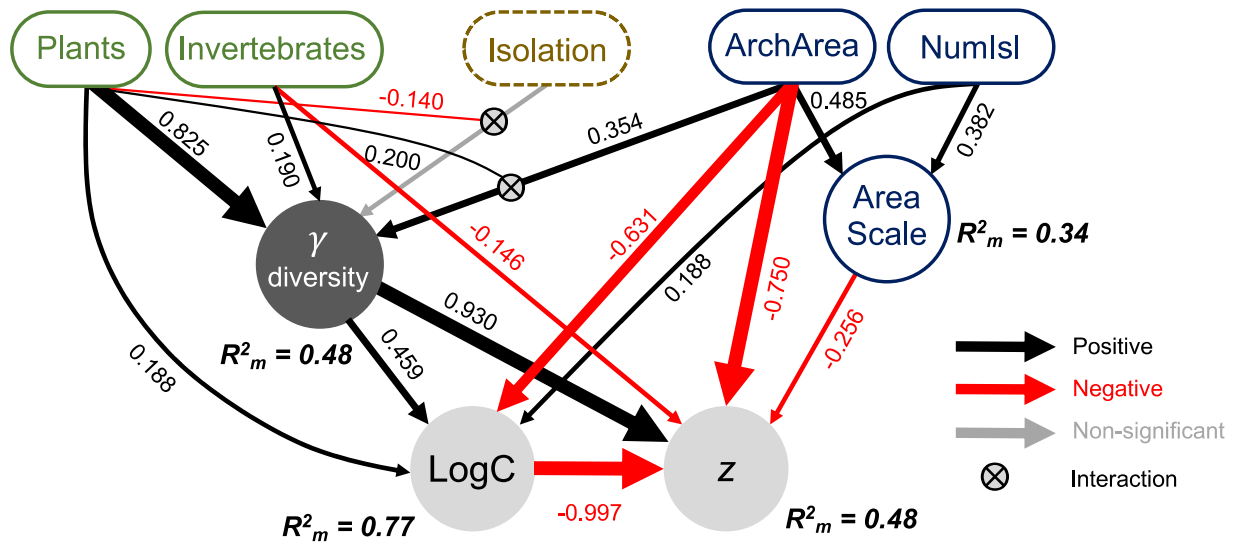


Fig. S3. Best path model for the all-ISARs dataset (n=151) testing for the interactions postulated in Fig. S2. Best path models were obtained using a backward stepwise selection procedure and AIC_C. Pathways show how taxon (Plant and Invertebrates with Vertebrates the base level), isolation, archipelago 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 exogenous 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.