Probabilistic models of species discovery and biodiversity comparisons

Stewart M. Edie[a,1], **Peter D. Smits**[b], and **David Jablonski**[a,b,1]

[a]Department of the Geophysical Sciences, University of Chicago, Chicago, IL 60637; and [b]Committee on Evolutionary Biology, University of Chicago, Chicago, IL 60637

Contributed by David Jablonski, February 16, 2017 (sent for review October 3, 2016; reviewed by Gene Hunt, Mark McPeek, and Andy Purvis)

Inferring large-scale processes that drive biodiversity hinges on understanding the phylogenetic and spatial pattern of species richness. However, clades and geographic regions are accumulating newly described species at an uneven rate, potentially affecting the stability of currently observed diversity patterns. Here, we present a probabilistic model of species discovery to assess the uncertainty in diversity levels among clades and regions. We use a Bayesian time series regression to estimate the long-term trend in the rate of species description for marine bivalves and find a distinct spatial bias in the accumulation of new species. Despite these biases, probabilistic estimates of future species richness show considerable stability in the currently observed rank order of regional diversity. However, absolute differences in richness are still likely to change, potentially modifying the correlation between species numbers and geographic, environmental, and biological factors thought to promote biodiversity. Applied to scallops and related clades, we find that accumulating knowledge of deep-sea species will likely shift the relative richness of these three families, emphasizing the need to consider the incomplete nature of bivalve taxonomy in quantitative studies of its diversity. Along with estimating expected changes to observed patterns of diversity, the model described in this paper pinpoints geographic areas and clades most urgently requiring additional systematic study—an important practice for building more complete and accurate models of biodiversity dynamics that can inform ecological and evolutionary theory and improve conservation practice.

species discovery | Bayesian time series model | species richness | taxonomic effort | marine bivalves

The number of biological species on Earth is notoriously uncertain, but such estimates are critical for a broad range of issues, from the environmental and biological limits of diversity to the design of conservation strategies in dwindling habitats (1–6). Geographic and phylogenetic differences in the discovery and description of species can change the patterns of species richness that are used, for example, to pinpoint biodiversity hotspots (7). A frequent approach to either anticipating or evaluating these taxonomically driven shifts is to estimate the "true," unknown species richness from a cumulative taxonomic description curve (8–10).

In theory, the cumulative count of newly described species should approach an asymptote as knowledge of the species pool nears the true value (Fig. 1*A*). However, many curves fail to "level off" or "saturate" because new species are being named at a steady or even accelerating rate (Fig. 1 *B* and *C*) (11, 12). These "unsaturated" curves lack a stable asymptote and therefore cannot provide robust estimates of the true species richness (12)—a result reflected in the many incongruent estimates of global diversity (13). Even with a robust estimate, a single value for the global number of species, or for high-level taxa such as Aves or Mammalia, is of limited utility in comparative diversity analyses across space, phylogeny, and time. Here, we develop a Bayesian model that can both accommodate nonasymptotic trends in species description to forecast species richness and operate at higher spatial and phylogenetic resolution. We use this model to assess the stability of observed differences in regional and among-clade diversity for a major animal group that has accrued newly described species at an unabated rate for the past 165 years: the marine bivalves.

In our Bayesian time series model [available from Zenodo (doi.org/10.5281/zenodo.159033)], the number of species described in a given year is a function of the long- and short-term trends in description rate. We first model the trajectory of species accumulation using only the history of currently valid species description beginning with Linnaeus (14), the starting point of formal taxonomy. We then add a simple estimate of taxonomic effort (TE), another factor relevant to estimates of taxonomic knowledge (15–19). For both approaches, we find strong regional differences in the long-term trend of species description, suggesting a spatial bias in the saturation of taxonomic knowledge. We also identify potential instability in the relative richness of closely related clades but find that, overall, the major geographic and phylogenetic diversity patterns in our example are robust to the spatial and taxonomic heterogeneity of description rates. Thus, these probabilistic estimates can be useful measures of data stability in comparative analyses of diversity when focal regions or clades have not reached taxonomic saturation.

Modeling Taxonomic Description

Model Design. Our model most closely resembles that of refs. 4, 10, and 11, with three key differences. (*i*) We balance our prediction of species description events by modeling the short-term

Significance

Estimates of species numbers are central to many analyses in fields ranging from conservation biology to macroecology and macroevolution. However, new species continue to be discovered and described at an uneven rate among regions and taxonomic groups, raising questions about the robustness of currently observed biodiversity patterns. We present a statistical approach to the rate of species description that incorporates uncertainty in species numbers across space and among clades. This approach identifies regions or clades where taxonomic knowledge is most complete, and provides estimates of stability in large-scale patterns given continued species discoveries through probabilistic forecasts of diversity levels.

Author contributions: S.M.E., P.D.S., and D.J. designed research; D.J. collected and standardized data; S.M.E. analyzed data; P.D.S. developed the description model; and S.M.E., P.D.S., and D.J. wrote the paper.

Reviewers: G.H., Smithsonian Institution; M.M., Dartmouth College; and A.P., National History Museum, London.

The authors declare no conflict of interest.

Data deposition: Bivalve data and Bayesian time series model are available from Zenodo (doi.org/10.5281/zenodo.159033).

[1] To whom correspondence may be addressed. Email: sedie@uchicago.edu or djablons@uchicago.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1616355114/-/DCSupplemental.

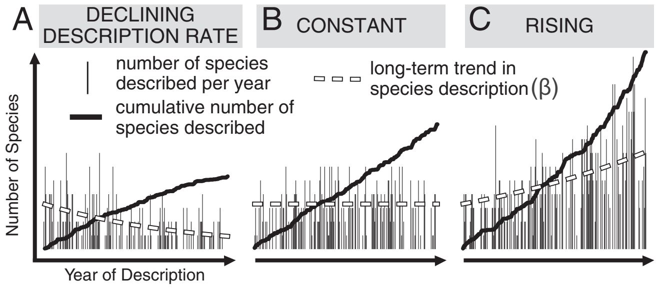

Fig. 1. Simulation of species description series under different long-term trends (β in *Modeling Taxonomic Description*). (*A*) A declining description rate produces the "saturated" asymptotic cumulative description curve. Constant and rising trends in description produce cumulative description curves with (*B*) linear and (*C*) exponential shapes.

volatility and the long-term trend in the description rate (including consecutive years with no description). (*ii*) We shift our analytical focus from attempting to calculate a single, unknown true species richness (as in ref. 11) to estimating the aforementioned long-term trend in the number of species described per year (β, Fig. 1). This approach can be applied to any species description curve regardless of its asymptotic shape. For example, we can directly compare the degree of taxonomic saturation for two regions with dramatically different description trajectories—the North (N) Temperate East Atlantic and the Tropical West Pacific margin (Fig. 2). (*iii*) We simultaneously estimate model parameters for all groups (i.e., regions and clades) in a hierarchical Bayesian framework so that diversity estimates can be compared among groups (estimates are relative to each other and the overall "average" regional pattern) (20). Thus, parameter estimates for groups with low statistical power (low species counts and/or erratic description events) are drawn toward the average regional pattern, whereas parameter estimates from regions with high statistical power vary more freely. This approach makes group estimates appropriately conservative when statistical power is highly uneven. Altogether, these three model features improve the characterization of taxonomic description at regional scales and clade levels where description events can be irregular in time and number.

Incorporating TE. In theory, an approach toward true taxonomic knowledge should be reflected by a decline in species description rate and an increase in TE—a broad concept largely distilled into the time, energy, and funds required to discover and describe a new species. Trends in TE and species description are often studied in parallel (15–19), but are difficult to bring into the same model framework (12, 16). When modeled simultaneously, trends in TE and species description mutually inform estimates of taxonomic saturation. Here, we follow the logic of "catch per unit effort" (12) and model the number of publications as an exposure term in our Poisson regression, where the long-term trend becomes the number of species described per publication per year (Fig. S2). Thus, we might infer an approach toward taxonomic saturation from a decline in the number of species described per publication—lower catch

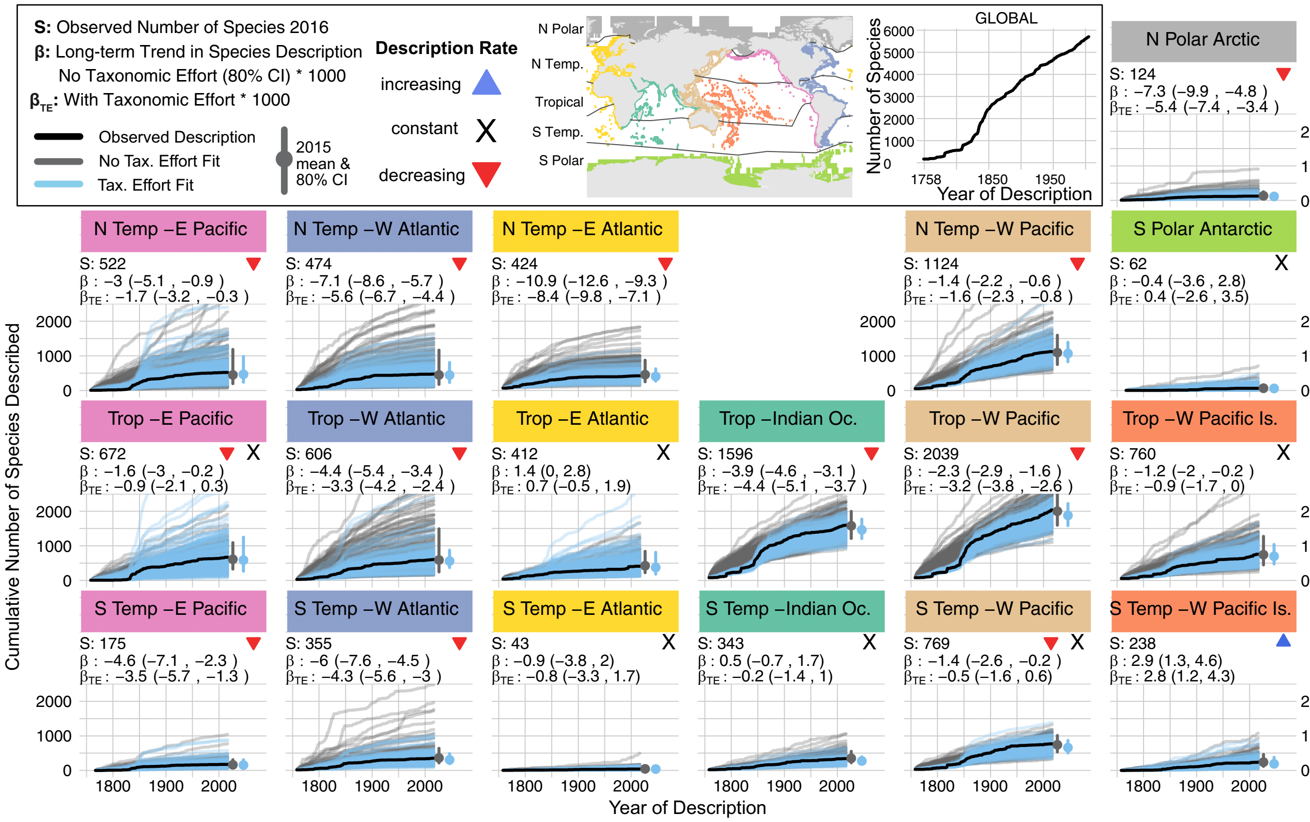

Fig. 2. Description of marine bivalves species recognized at a water depth of <200 m across geographic regions (climate zones and coastlines marked in *Inset* map). Each regional panel plots the cumulative species description curve (dark black line) and 200 modeled description curves for TE and noTE fits (blue and gray lines respectively). Summary boxes report the total number of species recognized in the region today (S), the long-term trend in species described per year β with its 80% credible interval, the long-term trend in species described per publication per year β_{TE}, and a colored symbol marking the sign of the long-term description trend (red down arrow, credible decline; blue up arrow, credible increase; and black X, no credible trend, i.e., constant). Model fits are plotted separately in Fig. S1. Is, islands; Oc, ocean.

per unit effort. This metric must be used cautiously because the steady attrition of professional taxonomists and the rise of nonprofessional publications (21, 22) drives a tendency for publications devoted to describing a single species. Thus, we cannot differentiate an increase in TE, i.e., an approach toward the true taxonomic knowledge, from a cultural shift to publishing in stand-alone journal articles rather than larger monographs.

We adopt this publications-per-year metric because alternative measures of TE are difficult to compile and apply across a variety of biological groups and are subject to their own biases in taxonomic culture [e.g., the number of authors per species, the number of junior synonyms, the number of journal or book pages dedicated to a species, and the average time required to describe a new species (19)]. We emphasize that the simple metric used here is only a first step toward evaluating the role of TE in developing a more robust and complete probabilistic model of species discovery.

Results and Discussion

Our primary goal is to shift the use of species description histories away from estimating global richness toward comparing differences among regions and clades. To that end, we estimate the long-term species description rate, examine the utility of one estimate of TE, and forecast the stability of ranked regional and clade richness.

Comparing Model Performance. Posterior predictive simulations show that both model fits, with and without the addition of TE, accurately recover the observed species richness values in 2016, albeit by very different trajectories (compare median estimates and their credible intervals in Fig. 2). The model without TE (noTE) fails to follow the exponential, 1800–1860 spike in description, but does track the constant description from 1860 to today (e.g., Tropical West Pacific margin in Fig. 2). The TE model follows both the exponential spike in description and the transition to a more constant rate. Including the number of publications in the TE model smooths the expected description events through time by transforming the modeled value to a rate—the number of species described per publication per year. Thus, the short-term trends in description rate become more predictable because the year-to-year variance in the number of species per publication per year is considerably lower than the number of species described per year (Fig. S3). Therefore, the TE model has a tighter tracking of the description trajectory that leads to a more constrained estimate of present-day richness.

Long-term trends are not directly comparable between models because of the differences in their units ($\beta = \mathrm{N_{species}}$ described per year; $\beta_{\mathrm{TE}} = \mathrm{N_{species}}$ described per publication per year). However, the rank order of trend estimates remains consistent across both model fits (Figs. S4 and S5), likely reflecting a correlated decline in the number of publications and number of species described per year (Fig. S2).

Geographic Variation in Species Description. Globally, bivalve systematists have slowed in their description of new species over time. This overall decline is inevitable because of the shift near 1860 from a rising description rate to a remarkably constant description rate of 21 (20 to 22 $\mathrm{CI_{80\%}}$) new species per year (Global, Fig. 2). Regionally, we find striking heterogeneity, where 12 of 18 climate–coastline regions show a decline in the number of species named per year (negative long-term trends β; red arrows in Fig. 2), five show constant description rates (black Xs), and only one shows a rising rate (blue arrow). In general, N Polar/Temperate coastlines have the strongest declines in description rate, followed by Tropical and then South (S) Polar/Temperate coastlines. Across climate zones, coastlines in the West Atlantic show some of the strongest declines in description rates, and those in the East Atlantic and West Pacific Islands show the weakest declines (Fig. S6).

The variation in regional rates of species description highlights a distinct spatial bias in the history of bivalve systematics. As with many other groups, formal description of bivalve species began in 1758 (14) and was pursued with zest for another 100 years by several prolific European systematists (e.g., Gmelin, Lamarck, Reeve, and Deshayes). Consequently, the N Temperate East Atlantic exhibits the strongest decline in description rate, likely reflecting the most complete taxonomic knowledge of any region. However, proximity to the early European systematists does not impart a similar level of taxonomic saturation on Tropical and S Temperate East Atlantic coastlines. Our model identifies these regions as two of the least described (Fig. S4), even compared with coastlines in the Tropical Indian and West Pacific Oceans that are considered highly undersampled (23).

More than half of the climate–coastlines show a decline in the number of species described per publication per year, which implies a decline in the catch per unit effort under the assumption of constant taxonomic culture. Thus, these regions may be nearing taxonomic saturation, but this inference must be made cautiously, because, as noted above, decreases in scientific funding and political limitations on sampling might also drive the description declines. Regardless of the link between description rates and taxonomic completeness, the variation in description rates among geographic regions indicates spatial differences in taxonomic activity that must be accounted for in comparisons of their observed species richness.

Geographic Comparisons of Species Richness. The long-term trends in description rates across geographic regions vary in sign, magnitude, and credibility, which, together, provide a relative sense of taxonomic activity. For example, the long-term trend in description rate is steeper in the Tropical West Pacific Islands than on the Tropical West Pacific margin, implying the West Pacific Islands are a comparatively undersaturated region (Fig. 2 and Fig. S3). However, estimating differences in diversity depends not only on the long-term trend in description but also on the baseline description rate (Fig. S3) and the current differences in observed diversity. Forecasts of species richness capture the effects of all of the factors above and become a useful tool for generating probabilistic estimates of species richness that help prevent overinterpretation in macroecological and macroevolutionary analyses (3, 24–26).

Forecasts of species richness after infinite time and effort could provide estimates of the true, unknown species richness. However, such estimates from our model accumulate a large forecasting error under the assumption that current trends in description rates will continue indefinitely (Fig. S7). Within the bivalve description series, poorest forecasting performance occurs during periods of relatively rapid change in description rate (1820–1860). Even during the long period of approximately constant global description (post-1860), longer forecasts create larger forecasting error, demonstrating that even small changes in description rate can compound into high predictive error. Thus, the credibility of a particular forecasting window depends on the likelihood that description rates remain constant, and that the size of the forecasting error is not comparable to the currently observed differences in diversity. Given these limitations, we conservatively interpret regional stability using a 20-y forecast, but we also compare those conservative estimates to a 50-y forecast with much greater inherent forecasting error.

Despite the regional heterogeneity in description rate, we find an overall stability in the estimated rank order of regional diversity in 2035 and 2065 (Fig. 3, Fig. S8, and Table S1). Forecasts from both the TE and noTE models show that regions within the Indo-West Pacific are expected to gain the bulk of newly

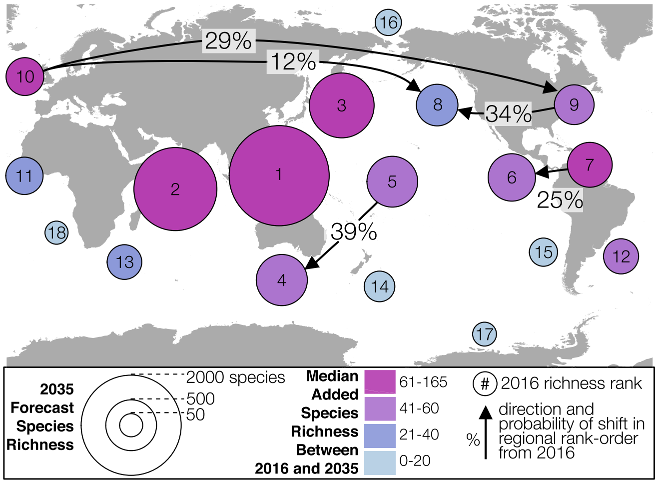

Fig. 3. The 20-y forecast of species richness (noTE model year 2035); forecast values with uncertainty provided in Table S1. Regional richness rank order is expected to remain stable to 2035, and most of the newly discovered species are expected to come from coastlines in the Indo-West Pacific.

described species and will remain the richest. A mixture of Tropical and Temperate coastlines will continue to occupy the middle richness ranks, with Polar regions toward the lowest ranks. A few regions show nonzero but low probabilities of diversity rank shift across the 20- and 50-y forecasts (Fig. 3 and Fig. S8). These unlikely shifts are mostly confined within climate zones, implying that the global latitudinal diversity gradient will persist in light of continued species discovery.

Forecasts are especially useful in targeted comparisons of species richness among regions. For example, an outstanding question in the geographic patterning of bivalve biodiversity has been the greater species richness in the Tropical East Pacific (TEP) than in the Tropical West Atlantic (TWA). Paleontological studies have proposed that differential extinction underlies this seemingly reversed diversity pattern given the larger continental shelf area and greater habitat heterogeneity in the reef-bearing TWA (27, 28). However, the difference is only 66 species, and we should consider the possibility that biases in taxonomic discovery may bias this interpretation. The TWA appears to be approaching taxonomic saturation faster than the TEP (joint probability $\beta_{\mathrm{TWA}} < \beta_{\mathrm{TEP}} = 1$; Fig. S4), but the TWA has a higher baseline rate of description and may still gain on the diversity of the TEP before reaching saturation (Fig. S3). Assuming trends in description rate remain constant for the next 20 and 50 y, we predict that the diversity of the TWA will get closer to that of the TEP, reducing the difference to 44 species [median forecast difference by 2035 and 20 species by 2065 (Table S1)]. The TEP has a 75% probability of remaining more diverse over the next 20 y and only a 58% probability over the next 50 y. This closing gap in estimated richness between regions should be considered when analyzing the oceanographic and biological factors that may underlie their diversity differences.

Clade Comparisons. The description model and its associated forecasts are also useful tools for comparisons of clade diversity. In the marine system, deep-sea exploration has dramatically elevated our estimates of species diversity in many groups (29), and we estimate that 43% of marine bivalve species described since 2005 were discovered in the deep sea (Fig. S9) (30). Thus, newly discovered species may be concentrated within particular clades, which may challenge the interpretation of many ecological and evolutionary patterns derived from strictly continental shelf occurrences (31).

Including newly discovered deep-sea species changes the relative richness of three well-studied, monophyletic bivalve families. When only considering continental shelf species (water depths of <200 m), true scallops (Pectinidae) are nearly 3 times as diverse as their closest relatives, the mainly tropical thorny oysters (Spondylidae) and the cold-water glass scallops (Propeamussiidae). However, recent deep-sea discoveries (e.g., ref. 32) have more than doubled the number of glass scallops, bringing their diversity much closer to that of their sister clade, the true scallops (Fig. 4). Still, even with their apparent taxonomic undersaturation, we do not predict the glass scallops to surpass or even match the diversity of the mainly continental shelf true scallops for the next 20 and 50 y (Table S2).

These probable estimates of clade diversity raise questions about the relationship between each clade's richness and biological or environmental factors. At least within these three families, bathymetric affinity alone appears to be a poor predictor of species richness. Instead, the greater ecological breadth of the true scallops may explain their higher diversity over the more restricted ecology of the mostly carnivorous glass scallops and sessile, filter-feeding thorny oysters. Estimating the probability of diversity shifts among clades with continued description of deep-sea species will be paramount for correctly interpreting evolutionary patterns.

Improving Estimates of Species Richness.

Alternative estimates of TE. Estimating true TE will require negative evidence, that is, the failure to recognize new species after repeated attempts. Combining recent region- and clade-specific faunal inventories can offer unparalleled insight into the taxonomic stability and saturation of the taxonomic record. In marine bivalves, recent rigorous molecular and morphological examination of a chemosybiotic group (Lucinidae) from Panglao, Philippines, in the Tropical West Pacific confirmed 50 existing species and discovered 26 new species (34); a similar treatment of lucinids from Guadeloupe in the TWA confirmed 25 existing and 1 new species (35). Despite all of the potential biases conflating the results of our model, these observed descriptions are precisely the dynamic that our model and other models (36) predict for the undiscovered diversity within these two regions.

Trends in biological characteristics. As the clade analysis shows, the biological properties of organisms can strongly affect the timing of the discovery and description of new species (8, 25). The earliest descriptions within many marine groups are commonly of species with larger body sizes, larger geographic ranges, and

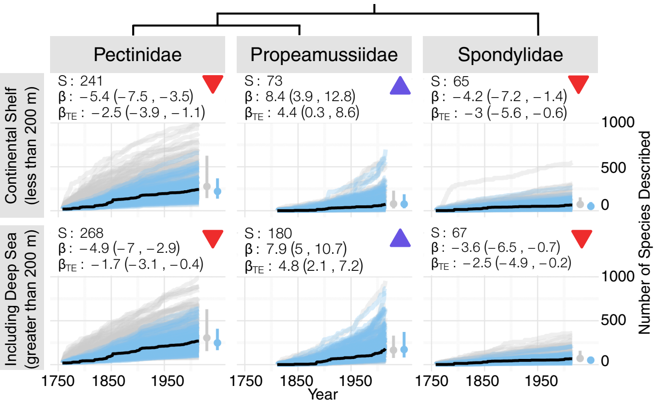

Fig. 4. Description of three closely related clades of marine bivalves excluding and including deep-sea species (phylogeny from ref. 33). Panels are organized as in Fig. 2. Despite the doubling of diversity in Propeamussiidae when including deep-sea species, these glass scallops are not forecast to overtake the diversity of their more speciose sister clade Pectinidae—the true scallops.

shallower bathymetric occurrences (37, 38). As the model stands here, we interpret a region or a clade with a relatively strong decline in species description rate as being closer to taxonomic saturation. However, if the body sizes and geographic range sizes of the species within that region show a temporally constant or increasing trend over time, we might conclude that the observed richness is unsaturated, because those species most easily encountered by systematists are still being described (39, 40). The challenge remains to directly incorporate these biological trends into a spatially and taxonomically explicit probabilistic model of species discovery.

Accounting for invalid descriptions. We modeled the description of currently accepted species and thus assumed the observed taxonomic record is completely stable. However, taxonomic revision on both morphological and molecular grounds can split (add) and synonymize (remove) species throughout the history of description. Reshaping the description curve changes the inferred rates of long-term description and the subsequent forecast of undiscovered species.

The history of taxonomic practice within a particular clade provides qualitative insight into the stability of an observed description curve. Most marine bivalve species have been defined by their morphology, and recent molecular work largely supports these lower-level taxonomic delimitations (22, 41). This general agreement between morphology and molecules reduces the likelihood of extensive synonymies or adding a large number of morphologically cryptic species. Cryptic species certainly exist, but their influence on the description curve is difficult to predict. Systematists reinstate an older, synonymized name (e.g., from year 1850) for a newly verified genetic unit or apply a new name entirely (e.g., in year 2017). Reinstatement of older names will produce stronger declines in the long-term trend of description, suggesting higher taxonomic saturation. Applying new names will contribute to a rise in long-term description rates, implying lower taxonomic saturation. Given the general congruency between molecules and morphology in bivalves, we expect most synonymized older names to remain synonymized and changes to the shape of the description curve to come primarily from new species descriptions.

Higher taxonomic groups such as birds, mammals, and bivalves are unlikely to exhibit similar histories of taxonomic revision, making the qualitative tactic above impractical for studying broad patterns in comparative biology (e.g., ref. 42). The net species description rate is the sum of the synonymization rate (reduce accepted species) and reinstatement rate (increase accepted species). Thus, in a given year, the probability of observing the currently accepted number of species is a function of the long- and short-term trends in description rate and the rate at which species are deemed invalid. Alternatively, the persistence of a species name could be modeled as a birth–death-type process in an extension of the "flux rate" method (43). Either proposed framework would provide the most probable "net taxonomic output" for a given year.

Comparing the idiosyncrasies of system-specific taxonomic records will be very important for designing and testing general models of species discovery. Removing the effects of taxonomic culture will always be difficult in comparative studies of higher taxonomic groups, but modeling the dynamics of description loss and reinstatement is likely the most promising method for future development.

Conclusions

Comparative macroecological and macroevolutionary studies often treat observed richness as known, but failing to account for the spatial and phylogenetic variation in taxonomic activity may mislead interpretations of biodiversity dynamics derived from currently observed species richness. Modeling the long-term species description rate provides a direct comparison of taxonomic knowledge among geographic regions or clades. Incorporating those trends and their associated uncertainties into short-term forecasts of species richness generates a set of probable values, which can be directly used in quantitative ecological and evolutionary models and in assessing the knowledge of diversity in and around biological reserves. Integrating description rates with forecasts of species richness not only improves our interpretations of current biodiversity patterns but also highlights areas where continued systematic research and discovery is necessary for building more rigorous quantitative analyses at higher spatial and phylogenetic resolution.

Materials and Methods

Marine Bivalve Database. Our marine bivalve database includes 5,744 currently valid species with 62,059 georeferenced occurrences (44) (Dataset S1). For the regional richness study, we focus on intertidal to continental shelf bivalves (living at depths from 0 m to 200 m), as deep-sea bivalves are widely acknowledged to be an independent and undersampled system (45). We also exclude two clades of exceptionally small body size (<1 cm) that have poorly understood taxonomy [Cyamioidea and Galeommatoidea (21, 30, 46)]. For the clade study, we include taxonomically standardized deep-sea occurrences from a low-resolution taxonomic dataset of 136 deep-sea species (largely from ref. 30; Dataset S2).

We define 18 geographic regions termed "climate–coastlines" using a combination of coastline geography, climate zones, and major biogeographic turnover (map in Fig. 2) (47). Our climate–coastlines resemble the 12 "realms" in the Marine Ecoregions of the World (48), but we split the realms by coastline to reflect the biogeographic structure of shelf biotas. We assigned species to one or more climate–coastlines by intersecting the individual occurrences for each species with the climate–coastline boundaries. Approximately 48% of species are endemic to one climate–coastline, and $\sim$40% of species occur across two and three climate–coastlines (Fig. S10). Allowing species to occur across more than one climate–coastline makes the regional description histories more similar, which biases against the test for differences in description history.

Modeling Species Description. We generate the number of species described in a given year following a zero-inflated Poisson distribution (49). Zero inflation accommodates an excess of individual years having zero description events above that expected under a Poisson distribution—a common characteristic of regional and clade description curves. We modified the zero-inflation component to allow for long runs of consecutive years with zero species described by modeling the occurrence probability of a description event as a two-state Markov chain. We characterize the long-term temporal trend in species description series using an autoregressive conditional Poisson regression (50). Within this regression, the predicted number of species described per year is a function of time and the long-term and short-term autoregressive components of the description rate. Finally, we incorporated TE, defined here as the number of unique publications describing new bivalve species for a given year, as an offset term for the number of species named per year (51). Including TE in the model transforms the interpretation of the estimated parameters from the expected number of species described per year to the rate of species described per publication per year. Full model description, formulation, and choice of priors is in *Supporting Information*, and model code is available from Zenodo (doi.org/10.5281/zenodo.159033).

The joint posterior of our model parameters was estimated using a variant of Hamiltonian Monte Carlo called the No U-Turn Sampler, as implemented in the probabilistic programing language Stan (52). Four independent chains were run for 15,000 steps each (5,000 warm-up) and were well mixed [$\hat{R} = 1$ (52)]. Model adequacy was assessed using posterior predictive simulations to determine whether patterns generated from the parameter estimates resemble the empirically observed patterns—the fundamental determinant of model fit. We made 1,000 independent draws from the marginal posterior distributions of each parameter and compared these posterior estimates to the observed patterns of taxonomic discovery through graphical comparisons (Fig. 2).

Forecasting Species Richness. We forecast species richness across groups (regions and clades) by simulating forward in time from the posterior predictive distribution. We examined the forecasting error using a variant of leave-p-out time series cross-validation ["rolling forecast origin" (53), recommended in ref. 4], where we fit the model to incremental time series from k blocks of p years each starting with 1758–1765. For $p = 5$, the series is 1758–1765, 1758–1770,..., 1758–2010. We estimated the species richness p years into the future for each block by drawing parameter estimates from the model posterior (for the TE model, we used random samples of publication counts p years before the end of the time series). We estimated the forecasting error as the difference between the observed and forecast counts within a forecast window (Fig. S7).

ACKNOWLEDGMENTS. We thank G. Hunt, A. Purvis, and M. McPeek for valuable reviews that expanded the breadth of this paper. We thank K. Roy, M. Foote, S. Huang, and the combined D.J.-Price Lab group at University of Chicago for discussions, and K. S. Collins and M. Ingalls for revision edits. We thank the following for taxonomic advice, assistance, and/or access to collections in their care: M. Aberhan, L. C. Anderson, K. Amano, A. G. Beu, R. Bieler, D. C. Campbell, J. G. Carter, R. von Cosel, J. S. Crampton, E. V. Coan, T. A. Darragh, H. H. Dijkstra, E. M. Harper, C. S. Hickman, M. Huber, S. Kiel, K. Lam, K. Lamprell, K. A. Lutaenko, N. Malchus, T. Matsubara, P. A. Maxwell, P. M. Mikkelsen, P. Middelfart, N. J. Morris, J. Nagel-Myers, G. Paulay, A. F. Sartori, F. Scarabino, J. A. Schneider, P. Valentich-Scott, J. T. Smith, J. D. Taylor, J. J. ter Poorten, J. D. Todd, T. R. Waller, A. Warén, and F. P. Wesselingh. This work was supported by National Science Foundation (NSF) and NASA (D.J.) and NSF Graduate Research Fellowship Program and Doctoral Dissertation Improvement Grant (S.M.E.).

1. Hey J, Waples RS, Arnold ML, Butlin RK, Harrison RG (2003) Understanding and confronting species uncertainty in biology and conservation. *Trends Ecol Evol* 18(11): 597–603.
2. McPeek MA (2008) The ecological dynamics of clade diversification and community assembly. *Am Nat* 172(6):270–284.
3. Isaac NJB, Purvis A (2004) The 'species problem' and testing macroevolutionary hypotheses. *Divers Distrib* 10(4):275–281.
4. Wilson SP, Costello MJ (2005) Predicting future discoveries of European marine species by using a non-homogeneous renewal process. *J R Stat Soc Ser C Appl Stat* 54(5): 897–918.
5. May RM (2010) Tropical arthropod species, more or less? *Science* 329(5987):41–42.
6. Price T (2015) The debate on determinants of species richness. *Am Nat* 185(5):571.
7. Meijaard E, Nijman V (2003) Primate hotspots on Borneo: Predictive value for general biodiversity and the effects of taxonomy. *Conserv Biol* 17(3):725–732.
8. May RM (1988) How many species are there on Earth? *Science* 241:1441–1449.
9. Patterson BD (1994) Accumulating knowledge on the dimensions of biodiversity: Systematic perspectives on Neotropical mammals. *Biodiv Lett* 2(3):79–86.
10. Solow AR, Smith WK (2005) On estimating the number of species from the discovery record. *Proc Bio Sci* 272(1560):285–287.
11. Bebber DP, Marriott FHC, Gaston KJ, Harriss SA, Scotland RW (2007) Predicting unknown species numbers using discovery curves. *Proc Bio Sci* 274(1618):1651–1658.
12. Joppa LN, Roberts DL, Pimm SL (2011) How many species of flowering plants are there? *Proc Bio Sci* 278(1705):554–559.
13. Caley MJ, Fisher R, Mengersen K (2014) Global species richness estimates have not converged. *Trends Ecol Evol* 29(4):187–188.
14. Linnaeus C (1758) *Systema Naturæ per Regna Tria Naturæ, Secundum Classes, Ordines, Genera, Species, cum Characteribus, Differentiis, Synonymis, Locis* (Impensis Direct. Laurentii Salvii, Stockholm), 10th Ed. Latin.
15. Appeltans W, et al. (2012) The magnitude of global marine species diversity. *Curr Biol* 22(23):2189–2202.
16. Joppa LN, Roberts DL, Myers N, Pimm SL (2011) Biodiversity hotspots house most undiscovered plant species. *Proc Natl Acad Sci USA* 108(32):13171–13176.
17. Costello MJ, Wilson S, Houlding B (2012) Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Syst Biol* 61(5):871–883.
18. Costello MJ, Wilson S, Houlding B (2013) More taxonomists describing significantly fewer species per unit effort may indicate that most species have been discovered. *Syst Biol* 62(4):616–624.
19. Sangster G, Luksenburg JA (2015) Declining rates of species described per taxonomist: Slowdown of progress or a side-effect of improved quality in taxonomy? *Syst Biol* 64(1):144–151.
20. Gelman A, et al. (2013) *Bayesian Data Analysis* (Chapman & Hall, Boca Raton, FL).
21. Bieler R, Mikkelsen PM, Giribet G (2013) Bivalvia–A discussion of known unknowns. *Am Malacol Bull* 31(1):123–133.
22. Mikkelsen PM (2011) Speciation in modern marine Bivalves (Mollusca: Bivalvia): Insights from the published record. *Am Malacol Bull* 29(1-2):217–245.
23. Bouchet P, Lozouet P, Maestrati P, Heros V (2002) Assessing the magnitude of species richness in tropical marine environments: Exceptionally high numbers of molluscs at a New Caledonia site. *Biol J Linn Soc* 75(1996):421–436.
24. Isaac NJB, Mallet J, Mace GM (2004) Taxonomic inflation: Its influence on macroecology and conservation. *Trends Ecol Evol* 19(9):464–469.
25. Scheffers BR, Joppa LN, Pimm SL, Laurance WF (2012) What we know and don't know about Earth's missing biodiversity. *Trends Ecol Evol* 27(9):501–510.
26. Gray A, Cavers S (2014) Island biogeography, the effects of taxonomic effort and the importance of island niche diversity to single-island endemic species. *Syst Biol* 63(1):55–65.
27. Jackson JBC, Jung P, Coates AG, Collins LS (1993) Diversity and extinction of tropical American mollusks and emergence of the Isthmus of Panama. *Science* 260(5114):1624–1626.
28. Allmon WD, Rosenberg G, Portell RW, Schindler KS (1993) Diversity of Atlantic coastal-plain mollusks since the Pliocene. *Science* 260(5114):1626–1629.
29. Rex MA, Etter RJ (2010) *Deep-Sea Biodiversity* (Harvard Univ Press, Cambridge, MA).
30. Huber M (2015) *Compendium of Bivalves II* (ConchBooks, Hackenheim, Germany).
31. Woolley SNC, et al. (2016) Deep-sea diversity patterns are shaped by energy availability. *Nature* 533(7603):393–396.
32. Dijkstra HA, Maestrati P (2008) New species and new records of deep-water Pectinoidea (Bivalvia: Propeamussiidae, Entoliidae and Pectinidae) from the South Pacific. *Mem Mus Natl Hist Nat* 196:77–113.
33. Bieler R, et al. (2014) Investigating the bivalve tree of life – an exemplar-based approach combining molecular and novel morphological characters. *Invertebr Syst* 28(1):32–115.
34. Glover EA, Taylor JD (2016) Lucinidae of the Philippines: Highest known diversity and ubiquity of chemosymbiotic bivalves from intertidal to bathyal depths (Mollusca: Bivalvia). *Mem Mus Natl Hist Nat* 208:65–234.
35. Taylor JD, Glover EA (2016) Lucinid bivalves of Guadeloupe: Diversity & systematics in the context of the tropical Western Atlantic (Mollusca: Bivalvia: Lucinidae). *Zootaxa* 4196(3):301–380.
36. Bouchet P, Bary S, Héros V, Marani G (2016) How many species of molluscs are there in the world's oceans, and who is going to describe them? *Tropical Deep-Sea Benthos Mémoires du Muséum National d'Histoire Naturelle*, eds Héros V, Strong E, Bouchet P (Mus Natl Hist Nat, Paris), Vol 29, pp 9–24.
37. Gibbons MJ, et al. (2005) What determines the likelihood of species discovery in marine holozooplankton: Is size, range or depth important? *Oikos* 109(3):567–576.
38. Costello MJ, Lane M, Wilson S, Houlding B (2015) Factors influencing when species are first named and estimating global species richness. *Glob Ecol Conserv* 4:243–254.
39. Adamowicz SJ, Purvis A (2005) How many branchiopod crustacean species are there? Quantifying the components of underestimation. *Glob Ecol Biogeogr* 14(5):455–468.
40. Stork NE, McBroom J, Gely C, Hamilton AJ (2015) New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc Natl Acad Sci USA* 112(24):7519–7523.
41. Jablonski D, Finarelli JA (2009) Congruence of morphologically-defined genera with molecular phylogenies. *Proc Natl Acad Sci USA* 106(20):8262–8266.
42. Tomašových A, et al. (2016) Unifying latitudinal gradients in range size and richness across marine and terrestrial systems. *Proc Bio Sci* 283(1830):20153027.
43. Alroy J (2002) How many named species are valid? *Proc Natl Acad Sci USA* 99(6): 3706–3711.
44. Jablonski D, et al. (2013) Out of the tropics, but how? Fossils, bridge species, and thermal ranges in the dynamics of the marine latitudinal diversity gradient. *Proc Natl Acad Sci USA* 110(26):10467–10469.
45. Valentine JW, Jablonski D (2015) A twofold role for global energy gradients in marine biodiversity trends. *J Biogeogr* 42(6):997–1005.
46. Valentine JW, Jablonski D, Kidwell S, Roy K (2006) Assessing the fidelity of the fossil record by using marine bivalves. *Proc Natl Acad Sci USA* 103(17):6599–6604.
47. Belanger CL, et al. (2012) Global environmental predictors of benthic marine biogeographic structure. *Proc Natl Acad Sci USA* 109(35):14046–14051.
48. Spalding MD, et al. (2007) Marine ecoregions of the world: A bioregionalization of coastal and shelf areas. *Bioscience* 57(7):573–583.
49. Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34(1):1–14.
50. Fokianos K, Rahbek A, Tjøstheim D (2009) Poisson autoregression. *J Am Stat Assoc* 104(488):1430–1439.
51. Gelman A, Hill J (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge Univ Press, Cambridge, UK).
52. Carpenter B, et al. (2017) Stan: A probabilistic programming language. *J Stat Softw* 76(1):1–32.
53. Hyndman RJ, Athanasopoulos G (2013) *Forecasting: Principles & Practice*. Available at https://www.otexts.org/fpp. Accessed September 27, 2016.
54. Stan Development Team (2016) *Stan Modeling Language Users Guide and Reference Manual, Version 2.9.0*. Available at mc-stan.org. Accessed September 21, 2016.