Supporting Information

Edie et al. 10.1073/pnas.1616355114

Formal Model Definition

We analyze the record of species named per analytical group (e.g., biogeographic region) as a hierarchical Bayesian model (20). There are J groups where $j=1,2,\ldots,J$. Each group has a specific time series of species named per year y^j (minimum 0) with length T^j, which is the number of years between the first year a species was described in a group j and the last year.

We model y^j as a zero-inflated Poisson regression where the Poisson distribution follows an autoregressive conditional Poisson model (ACP) and the Bernoulli distribution follows a two-state Markov model (Eq. **S1** and refs. 50 and 51, following formulation in ref. 55):

$$p(y_t^j|\theta^j,\lambda_t^j) = \begin{cases} \theta_t^j + (1-\theta_t^j)\times \mathrm{Pois}(0|(u_t^j+1)\lambda_t^j) & \text{if } y_t^j = 0\text{, and} \\ (1-\theta_t^j)\times \mathrm{Pois}(y_t^j|\lambda_t^j) & \text{if } y_t^j > 0, \end{cases} \tag{S1}$$

where θ_t^j is the probability for region j at time t that y_t^j equals 0, and $1-\theta_t^j$ is the probability that y_t^j is a draw from a Poisson distribution with rate λ_t^j (Eq. **S3**); u_t^j is the number of publications necessary to name y_t^j species plus 1, because the Poisson rate is undefined at 0.

Here θ^j is a length T^j vector whose elements are modeled as a two-state Markov process (Eq. **S2**; this accounts for possible long runs of no species being named. The probability for region j that $y_t^j=0|y_{t-1}^j=0$ is γ^j, and η^j is the probability for region j that $y_t^j \neq 0|y_{t-1}^j=0$; z^j is then a length T indicator vector where $z_t^j=0$ if $y_t^j=0$ and $z_t^j=1$ if $y_t^j\neq 0$:

$$\begin{aligned} \theta_t^j &= z_t^j\gamma^j + (1-z_t)\eta^j \\ \gamma^j &\sim \mathrm{U}(0,1) \\ \eta^j &\sim \mathrm{U}(0,1). \end{aligned} \tag{S2}$$

We use an ACP regression to account for the temporal autocorrelation present in the time series, where λ^j is a length T^j vector of Poisson rates (Eq. **S3**). The short-term autocorrelation in number named α_j is the number of species named at time $t-1$, and the long-term autocorrelation in rate β_j is the value of λ^j at time $t-1$. Both α_j and β_j are constrained to be between 0 and 1 so that $\alpha_j+\beta_j \leq 1$. These parameters are given independent uniform priors; α_j and β_j are not defined for the first time point, so, when $t=1$, these terms are not included (Eq. **S3**). Thus, the ACP part of the mixture model is defined as:

$$\begin{aligned} \lambda_t^j &= \begin{cases} \phi_j & \text{if } t=1\text{, and} \\ \omega_t^j + \alpha_j y_{t-1}^j + \beta_j\lambda_{t-1} & \text{if } t>1 \end{cases} \\ \phi_j &\sim \log N(\mu_\phi,\sigma_\phi) \\ \mu_\phi &\sim N(0,1) \\ \sigma_\phi &\sim \mathrm{C}^+(1) \\ \alpha_j &\sim \mathrm{U}(0,1) \\ \beta_j &\sim \mathrm{U}(0,1-\alpha_j). \end{aligned} \tag{S3}$$

Increased temporal autocorrelation can be modeled by adding lags. In that case, λ_t^j would be defined as $\lambda_t^j=\omega_t^j + (\sum_{k=1}^{K}\alpha_{j,k}y_{t-1}^j) + (\sum_{k=1}^{K}\beta_{j,k}\lambda_{t-1,k})$, where K = number of lags and $\alpha_{j,k}+\beta_{j,k}\leq 1$.

The expected number of species in group j named in the first year (t = 1) is ϕ_j; ϕ_j is considered a realization from the shared distribution of species named in the first year and was assumed to be log-normally distributed with log-mean μ_ϕ and log-scale σ_ϕ, which were both given weakly informative priors.

The intercept of the ACP regression ω^j is modeled as a hierarchical linear model, also known as a generalized linear mixed effect model (20) with a log-link function (Eq. **S4**). X is a $T\times 2$ matrix of predictors, where T is the number of years since the first species was identified. The first column of X corresponds to the intercept term (all set to 1), and the second column is the year corresponding to that row; δ is a $2\times J$ matrix of regression coefficients, where J is the number of regions in the study.

The matrix of correlation coefficients δ is given a multivariate normal prior with mean vector μ of length D and covariance matrix Σ, which is $D\times D$. The elements of μ correspond to the overall average intercept and slope terms, where the slopes correspond to the temporal trend and are given independent normally distributed priors. Σ is decomposed into a length two vector of scales τ and a 2×2 correlation matrix Ω. The elements of τ are given independent half-Cauchy priors, and Ω is given a prior from the Lewandowski, Kurowicka, and Joe (LKJ) Correlation Distribution, following the Stan manual (54). Here μ represents the average values of δ across all biogeographic regions, τ are the standard deviations of the region-specific estimates δ_j, and Ω is the correlation between the intercept and slope terms of δ as they vary by region.

$$\begin{aligned} \omega_t &= \exp(X_t\delta_j) \\ \delta &\sim \mathrm{MVN}(\mu,\Sigma) \\ \Sigma &= \mathrm{diag}(\tau)\Omega\mathrm{diag}(\tau) \\ \Omega &\sim \mathrm{LKJ}(2) \\ \tau_j &\sim \mathrm{C}^+(1). \end{aligned} \tag{S4}$$

Full model code is available from Zenodo (doi.org/10.5281/zenodo.159033).

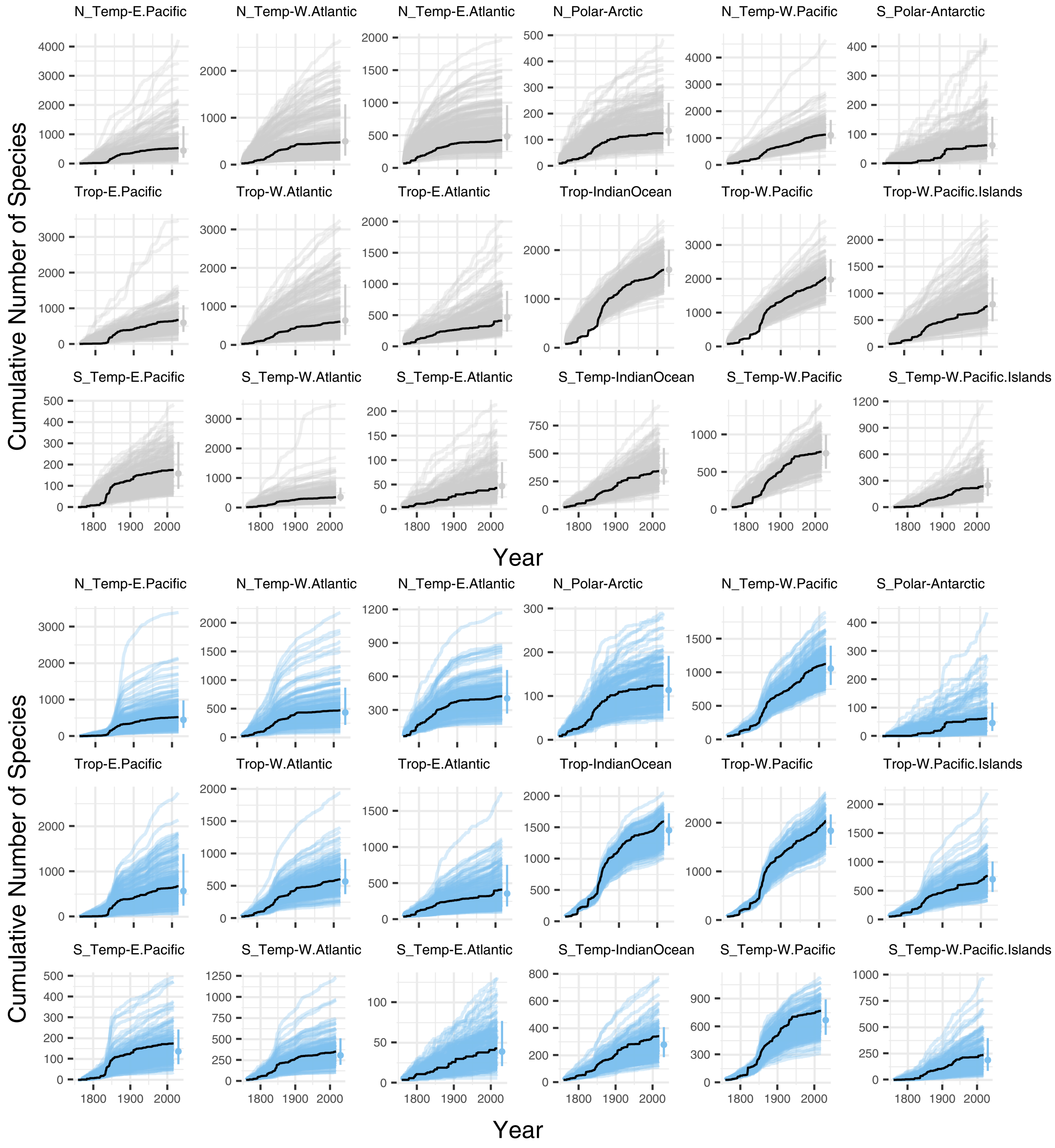

Fig. S1. Cumulative count of described species through time per geographic region. The range of the *y* axis is set by region-specific counts, to better visualize model fits within regions. Lines are the same as in Fig. 2, where gray lines are model fits without TE, blue lines are fits with TE, and black lines are the observed description series. Point ranges at the end of the description series show the predicted 2016 richness with 80% credible intervals. The range of model estimates for 2016 (median estimate and 80% CI to right of each time series) bracket the observed counts across all regions for models fit with and without TE, indicating that both models successfully predict the observed count of species today.

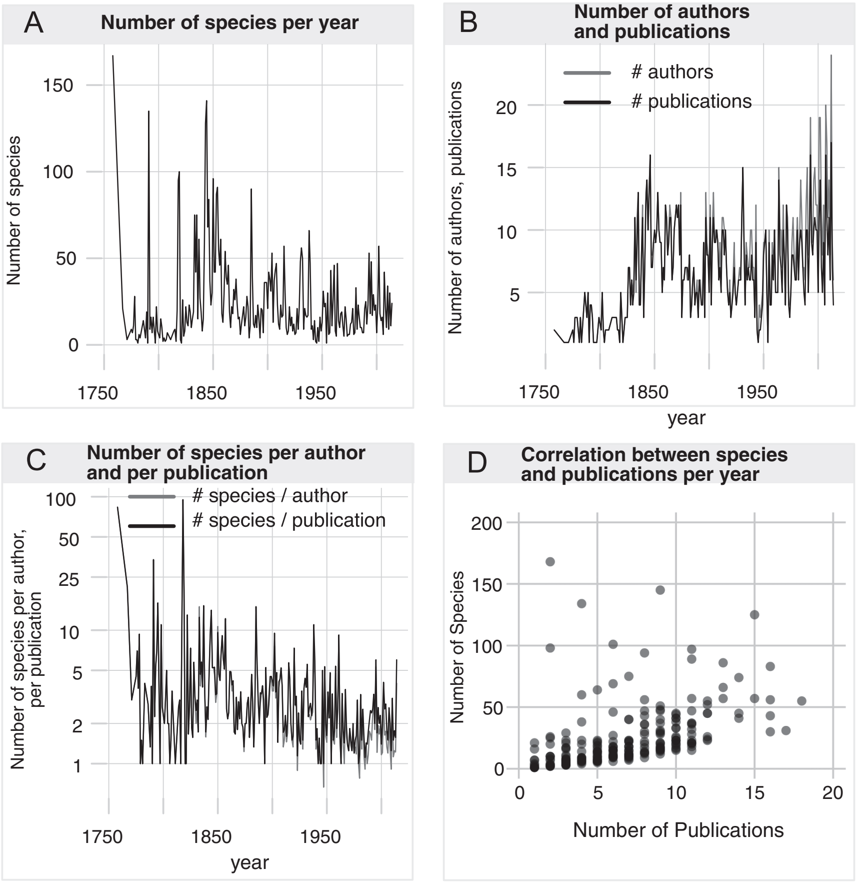

Fig. S2. Measure of TE through time. (*A*) The number of species described each year largely declines over time. (*B*) The number of publications and authors increases over time. (*C*) The number of species described per author and per publication declines over time. (*D*) The number of species described in a given year tends to be roughly proportional to the number of taxonomic publications in that year.

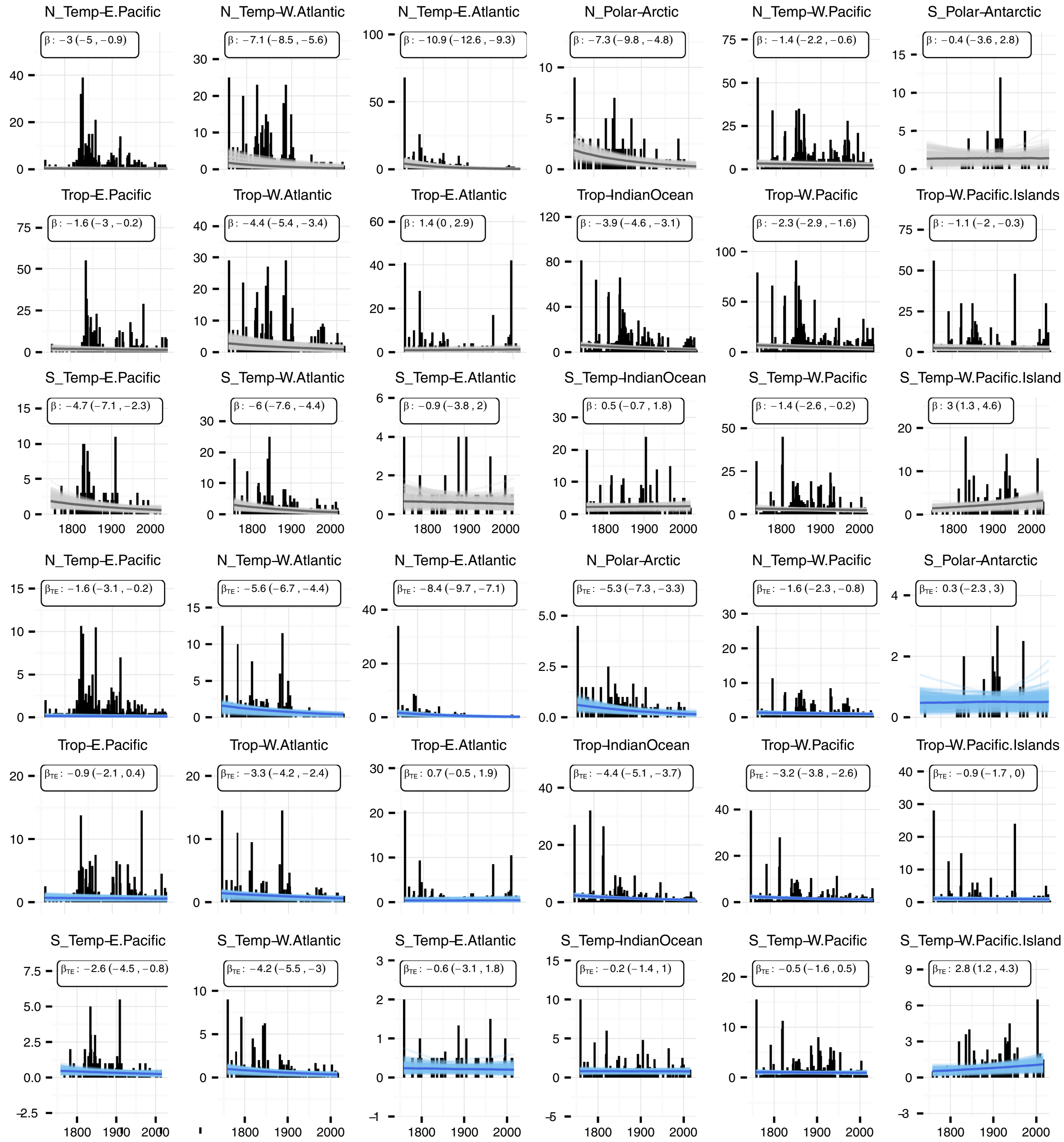

Fig. S3. Fit of the long-term trend in species description to the observed count of descriptions per year. Top three rows show the model fit without TE (i.e., the number of species described per year), and the bottom three rows show the fit with TE (i.e., the number of species described per publication per year). Note that the black vertical segments are the numbers of species described in a year. Single dark gray and blue lines are the median long-term trend estimates (β), for the model fits with and without TE, respectively. The light gray and blue lines are 200 posterior predictive simulations of the long-term trends for the models without and with TE, respectively. Panels show the median estimated long-term trend and 80% credible interval same as in Fig. 2. The overall rank order of long-term trend coefficients is similar between models fit with and without TE. Temp, temperate; Trop, tropical.

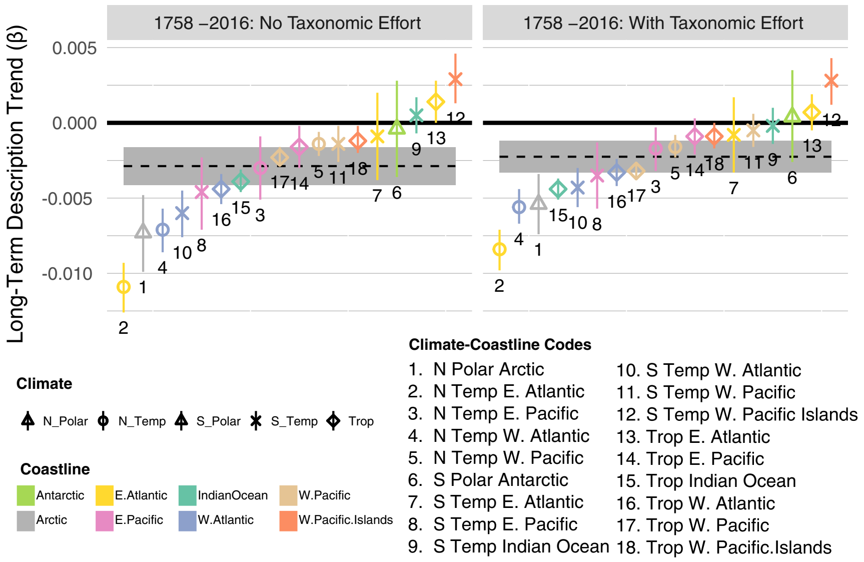

Fig. S4. Long-term trend coefficients for species description among regions. Symbols mark the median estimate, and bar represents the 80% credible interval. The horizontal dashed black lines mark the median estimated long-term trend of description rate among regions (80% credible interval in gray polygon). The overall rank order of long-term trend coefficients is similar between models fit with and without TE. Temp, temperate; Trop, tropical.

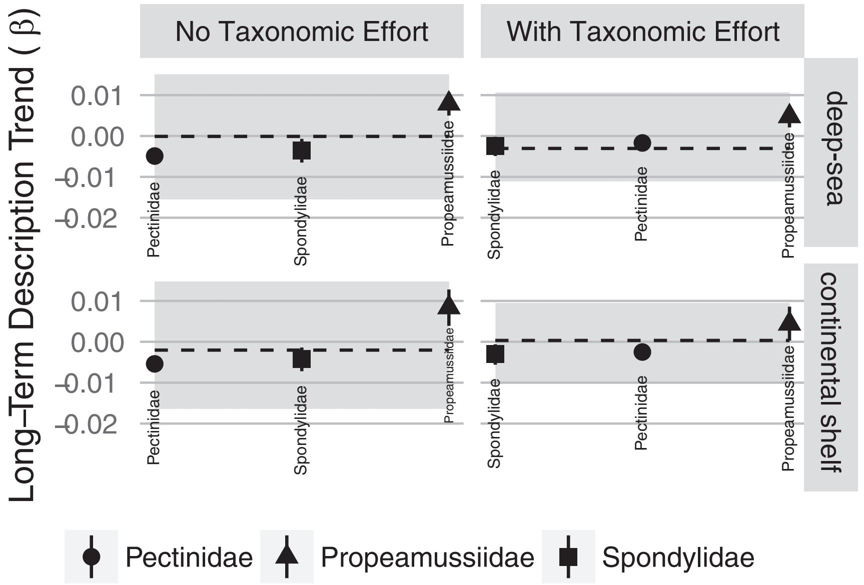

Fig. S5. Long-term trend coefficients for species description among clades. Symbols mark the median estimate, and bar represents the 80% credible interval. The horizontal dashed black lines mark the median estimated long-term trend of description rate (80% credible interval in gray polygon). As with regions (Fig. S3), the overall rank order of long-term trend coefficients is similar between models fit with and without TE, for both the deep-sea and continental shelf datasets.

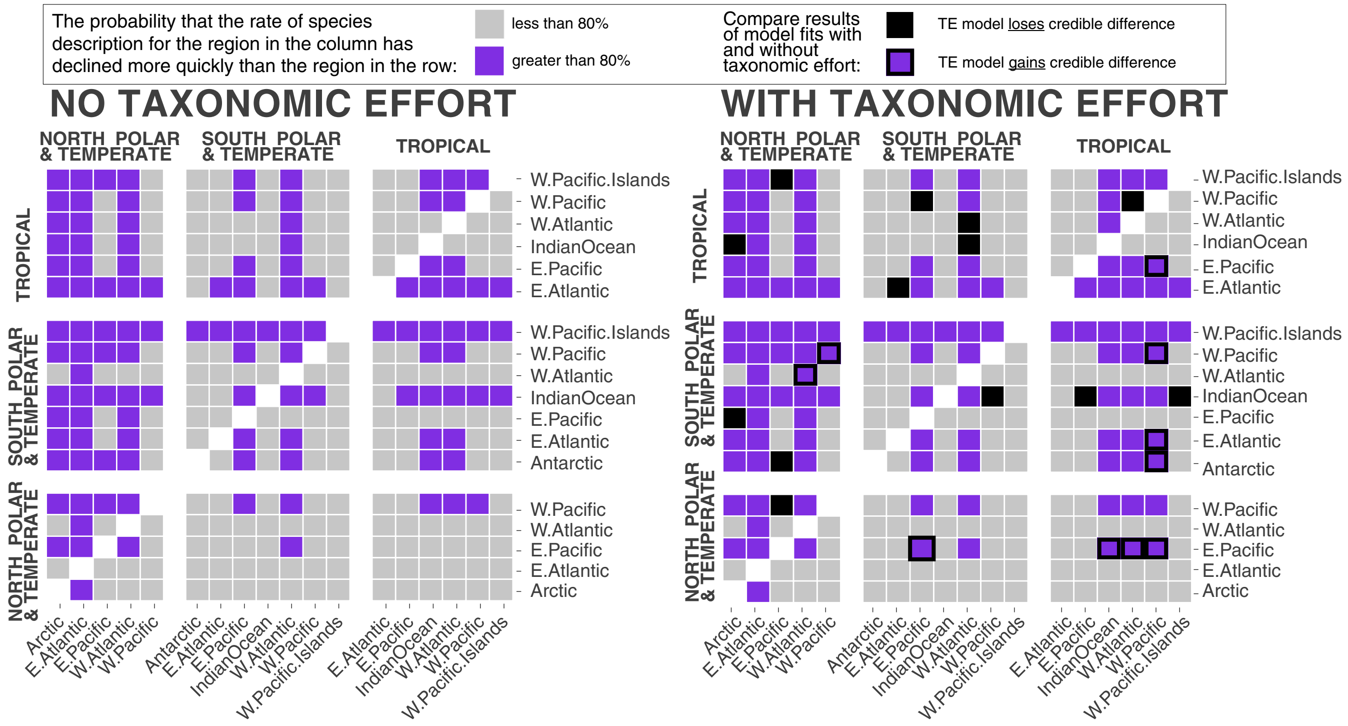

Fig. S6. Pairwise probabilities of differences in long-term description trends among climate–coastlines. For example, in the model without TE, there is at least an 80% probability that the long-term trend in description rate for the North Polar Arctic coastline (labeled column) has declined more rapidly than the trend for the North Temperate East Pacific (labeled row).

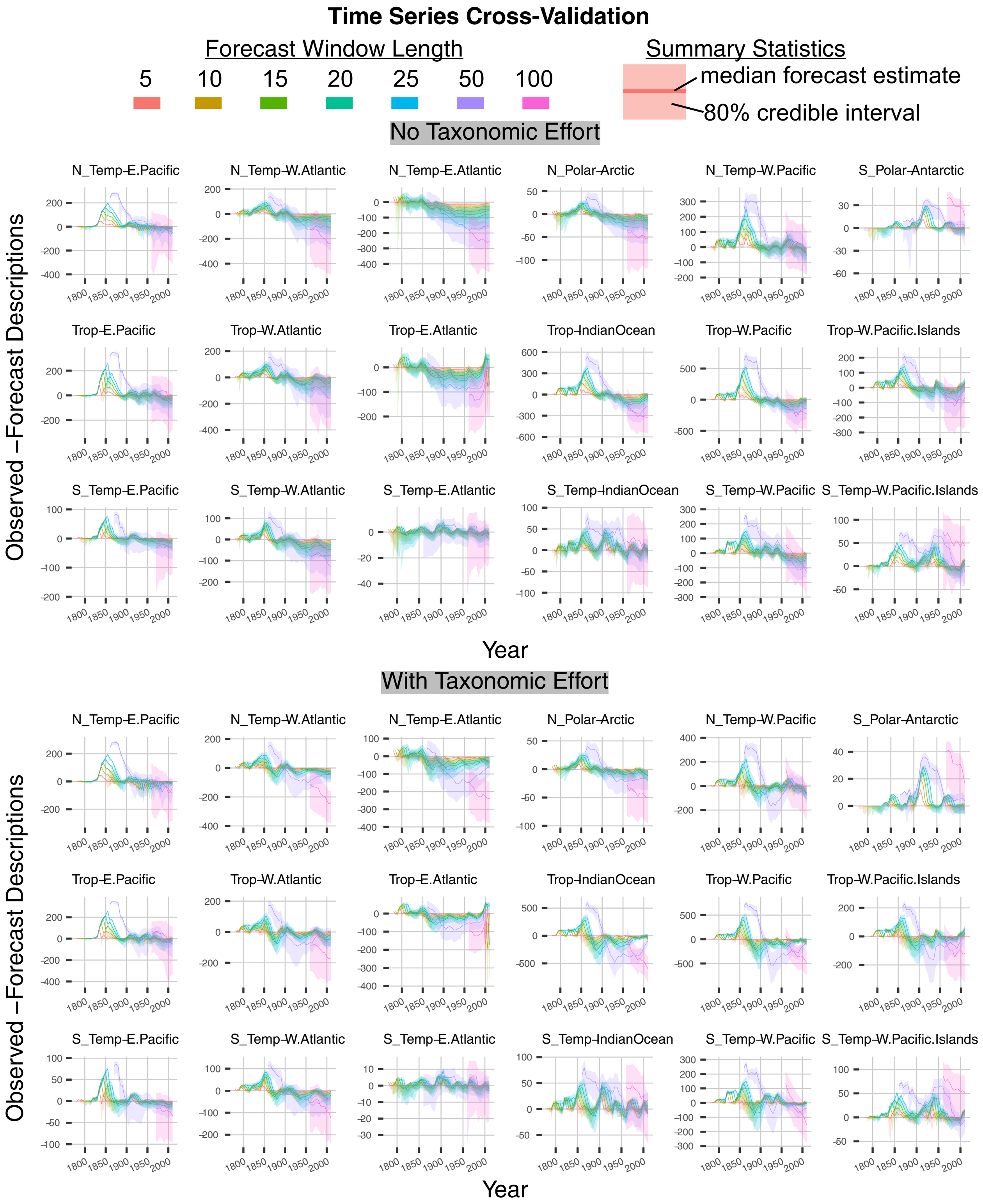

Fig. S7. The forecasting error of the taxonomic description model. The *y* axis represents the forecasting error: the difference between the number of observed and forecast descriptions within the forecasting window. A value of zero means a perfect match between the observed and forecast values. Values above zero represent underprediction of species descriptions, and values below zero represent overprediction. For example, in the North Temperate East Pacific, there are ∼200 more species observed in 1855 than expected under the 50-y forecast. Poorest forecasting error occurs early in the taxonomic time series, during the exponential phase of species description. The forecasting error increases as the regional description series approach a more constant description rate. Forecasts tend to underpredict species richness in regions with rising long-term trends in description (see Fig. S4 and the South Temperate West Pacific Islands). Forecasts tend to overpredict in regions with declining trends in description (e.g., the N Temperate East Atlantic). Temp, temperate; Trop, tropical.

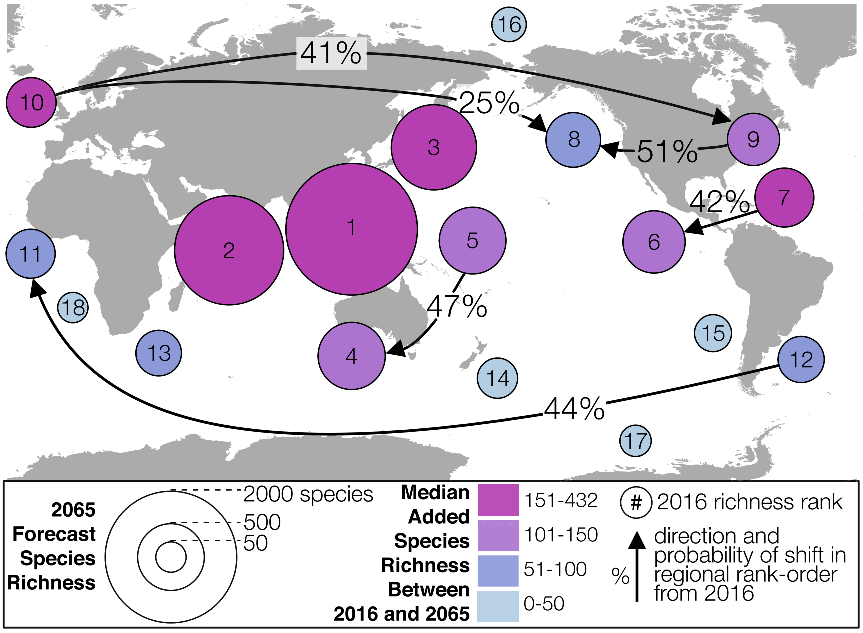

Fig. S8. The 50-y forecast of species richness (model without TE, year 2065; forecast values with uncertainty are given in Table S1). Many of the newly discovered species are expected to come from coastlines in the Indo-West Pacific. Despite this regional bias in the accumulation of new species, the regional richness rank order is expected to remain stable to 2065. For example, the Tropical West Pacific Islands (circle 5) have only a 47% chance of overtaking the South Temperate West Pacific (circle 4), suggesting that more time for discovery or changes in future description rate will be necessary to overcome the current (2016) observed difference in diversity.

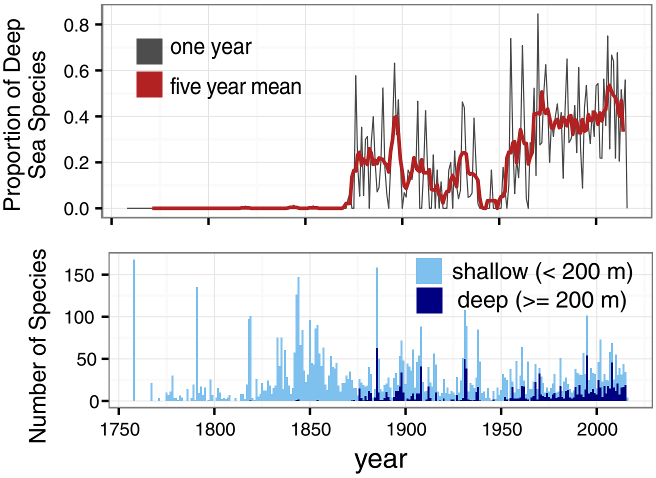

Fig. S9. (*Top*) Proportion of bivalve species described from the deep sea per year. (*Bottom*) Actual counts of species described in the deep sea and continental shelf per year. For the last 50 years, approximately 40% of newly described species were discovered in the deep sea.

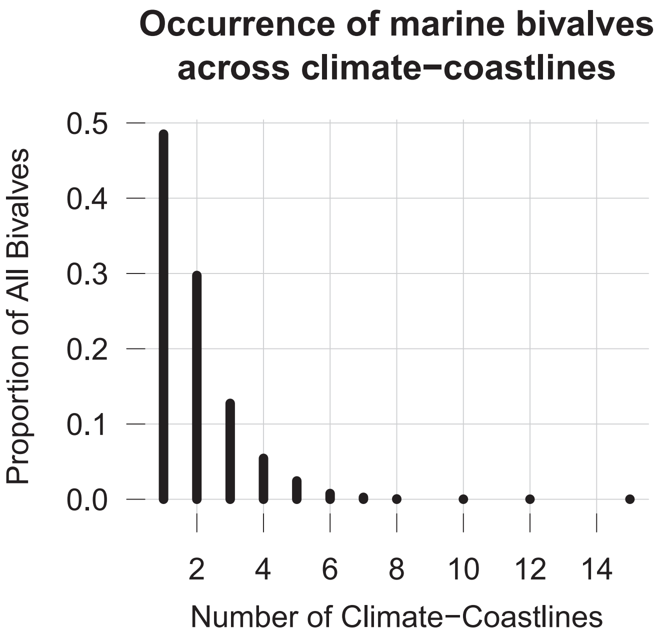

Fig. S10. Proportion of bivalves occurring in climate–coastlines. Approximately 50% of bivalve species occur in only one or two climate–coastlines. A combined 92% of bivalve species occur in three or fewer climate–coastlines. Only around 8%, or ∼500 species, of bivalves could be considered widespread in that they occur across four or more climate–coastlines.

Table S1. Twenty-year (2035) and 50-y (2065) forecasts of species richness from TE and noTE models

		2035		2065	
Climate–coastline	2016 richness	noTE median richness (80% CI)	TE median richness (80% CI)	noTE median richness (80% CI)	TE median richness (80% CI)
Trop-W.Pacific	2,039	2,204 (2,149 to 2,275)	2,289 (2,212 to 2,387)	2,474 (2,356 to 2,618)	2,585 (2,479 to 2,725)
Trop-IndianOcean	1,596	1,744 (1,688 to 1,817)	1,793 (1,737 to 1,866)	1,973 (1,871 to 2,120)	1992 (1,906 to 2,083)
N_Temp-W.Pacific	1,124	1,209 (1,175 to 1,260)	1,188 (1,167 to 1,215)	1,345 (1,267 to 1,466)	1,328 (1,275 to 1,397)
S_Temp-W.Pacific	769	824 (797 to 861)	797 (786 to 813)	915 (863 to 983)	841 (819 to 868)
Trop-W.Pacific.Islands	760	814 (787 to 859)	836 (806 to 879)	907 (841 to 1025)	907 (851 to 978)
Trop-E.Pacific	672	718 (689 to 764)	706 (684 to 747)	795 (735 to 897)	751 (706 to 842)
Trop-W.Atlantic	606	674 (632 to 776)	663 (640 to 697)	778 (673 to 1001)	749 (691 to 828)
N_Temp-E.Pacific	522	546 (530 to 574)	539 (530 to 555)	605 (556 to 712)	574 (549 to 620)
N_Temp-W.Atlantic	474	526 (492 to 615)	510 (490 to 543)	618 (526 to 860)	568 (518 to 661)
N_Temp-E.Atlantic	424	492 (454 to 569)	468 (447 to 498)	587 (502 to 758)	528 (482 to 599)
Trop-E.Atlantic	412	431 (418 to 458)	427 (417 to 451)	468 (436 to 536)	456 (430 to 514)
S_Temp-W.Atlantic	355	401 (378 to 443)	388 (374 to 410)	462 (414 to 555)	433 (402 to 480)
S_Temp-IndianOcean	343	364 (353 to 381)	358 (351 to 369)	400 (376 to 435)	386 (368 to 413)
S_Temp-W.Pacific.Islands	238	248 (241 to 263)	245 (240 to 254)	270 (252 to 298)	257 (246 to 277)
S_Temp-E.Pacific	175	192 (182 to 215)	187 (180 to 200)	222 (197 to 265)	202 (189 to 221)
N_Polar-Arctic	124	141 (131 to 162)	133 (128 to 141)	167 (145 to 207)	150 (138 to 168)
S_Polar-Antarctic	62	66 (62 to 76)	65 (62 to 73)	74 (65 to 95)	69 (64 to 82)
S_Temp-E.Atlantic	43	46 (43 to 51)	46 (43 to 50)	51 (46 to 62)	50 (45 to 58)

Temp, temperate; Trop, tropical.

Table S2. Twenty- and 50-y forecasts of species richness within clades across model fits with and without TE

Bivalve family	2016 richness	2035 noTE	2035 TE	2065 noTE	2065 TE
Continental shelf only					
Pectinidae	241	261 (249 to 291)	257 (249 to 272)	263 (263 to 331)	293 (268 to 313)
Propeamussiidae	69	73 (71 to 78)	75 (71 to 85)	83 (77 to 85)	77 (73 to 79)
Spondylidae	65	70 (65 to 80)	68 (66 to 74)	77 (73 to 103)	77 (75 to 85)
Including deep sea					
Pectinidae	268	289 (276 to 317)	285 (277 to 299)	320 (302 to 364)	337 (323 to 421)
Propeamussiidae	176	189 (185 to 197)	197 (188 to 223)	200 (195 to 208)	202 (200 to 252)
Spondylidae	67	72 (67 to 81)	70 (67 to 76)	82 (79 to 110)	73 (71 to 75)

Values in the 2035 or 2065 columns are the median forecast species richness for that year, with the 80% credible interval in parentheses.

Other Supporting Information Files

Dataset S1 (XLSX)

Dataset S2 (XLSX)