Received: 4 August 2016 | Revised: 20 February 2017 | Accepted: 3 March 2017

DOI: 10.1111/geb.12589

RESEARCH PAPER

WILEY Global Ecology and Biogeography A Journal of Macroecology

# Estimating regional species richness: The case of China's vascular plant species

Muyang Lu[1] | Fangliang He[1,2]

[1]SYSU–Alberta Joint Lab for Biodiversity Conservation, State Key Laboratory of Biocontrol and School of Life Sciences, Sun Yat-sen University, Guangzhou, China

[2]Department of Renewable Resources, University of Alberta, Edmonton, Alberta, Canada

**Correspondence**
Fangliang He, Department of Renewable Resources, University of Alberta, Edmonton, Alberta, Canada.
Email: fhe@ualberta.ca

**Present address**
Muyang Lu, Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT, USA.

Editor: Michael Borregaard

**Funding information**
Sun Yat-sen University; Natural Sciences and Engineering Research Council of Canada (NSERC)

## Abstract

**Aim:** The estimation of regional species richness has been a major challenge in ecology but is crucial for setting up conservation priorities. Discovery curves are a principal tool for estimating regional richness, but they have been criticized for being too sensitive to historical fluctuations in species discovery. In this study we propose a new discovery model that considers historical influences. We have applied the model to estimate the number of vascular plant species in China, a country of mega-biodiversity that has also suffered endless wars and social turmoil in its modern history.

**Location:** China.

**Time period:** 1755–2000.

**Major taxa studied:** Vascular plant species.

**Methods:** We compiled the discovery time for each vascular species from the complete volumes of *Flora of China*, leading to 31,220 valid species names. We applied our model to these data and compared the performance of our model with existing discovery models and two other methods (species–area relationships and a taxonomic rank curve model). We also tested our model with three other independent datasets and one simulation study.

**Results:** Our new method estimated there to be 36,554 (± 2,708) vascular plant species in China. Our model accounted very well for the effect of historical events and was robust to different periods of data that were used to estimate the total richness. Our model outperformed all other models that were compared. We found that 5,334 species remained to be discovered in China and it would take about 50 years to discover all the species should the current discovery rate of 110 species per year persist.

**Main conclusions:** Species discovery curves, with historical effects being properly accounted for, offer a promising tool for estimating regional species richness. Our model is robust to the effect of historical events and provides by far the most accurate and reliable estimates of species richness of test data, including vascular plant richness in China.

KEYWORDS
accumulation curve, *Flora of China*, historical effects, species-area relationships, species discovery, taxonomic effort

## 1 | INTRODUCTION

Many methods have been proposed to answer a seemingly simple question: "How many species are there?" (Costello & Wilson, 2011; Curtis, Sloan, & Scannell, 2002; Joppa, Roberts, & Pimm, 2011; May, 1988; Medellín & Soberón, 1999; Mora, Tittensor, Adl, Simpson, & Worm, 2011) However, efforts to answer this question have so far met with mixed success. We are more successful at answering this question at local scales than at regional and global scales. The success also depends on the taxa of interest, with some taxa (e.g., mammals, birds) being better estimated than others (e.g., arthropods, nematodes) due to the variation in the completeness of species discoveries (Bebber, Harris, Gaston, & Scotland, 2007; Essl, Rabitsch, Dullinger, Moser, & Milasowszky, 2013; Nabout, da Silva Rocha, Carneiro, & Sant'Anna, 2013; Randhawa, Poulin, & Krkošek, 2015). Given the uncertainties and difficulties in knowing the true regional or global species richness, biologists are usually contented at regional scales with pinning down the magnitude of life rather than looking for unbiased estimators (Hamilton et al., 2013; Joppa, Roberts, & Pimm, 2011; Locey & Lennon, 2016; Stork, 1993; Stork, McBroom, Gely, & Hamilton, 2015). However, knowledge on regional or global species diversity is critical for delineating biodiversity hotspots and prioritizing conservation efforts (Giam, Ng, Yap, & Tan, 2010; Joppa, Roberts, Myers, & Pimm, 2011; Myers, Mittermeier, Mittermeier, da Fonseca, & Kent, 2000; Tedesco et al., 2014). Reducing the uncertainties in estimating species richness at regional scales is of important theoretical and applied significance and is urgently needed.

The many methods (e.g., area-based, abundance-based, and non-parametric methods) that have been developed to estimate species richness are shown to perform well at estimating local richness (Chao, 1987; Colwell & Coddington, 1994; Palmer, 1990; Peterson & Slade, 1998) and their biases at local scales are well studied (Chao & Bunge, 2002; Colwell & Coddington, 1994; Soberón & Llorente, 1993). However, when applied to the regional or global scales, few of these methods work satisfactorily (Bebber, Marriott, Gaston, Harris, & Scotland, 2007; Gerstner, Dormann, Václavík, Kreft, & Seppelt, 2014; Ugland, Gray, & Ellingsen, 2003; Xu, Liu, Li, Zang, & He, 2012). The failure of these methods is primarily caused by the fact that in the majority of applications sample size is simply too small and too incomplete to extrapolate species richness to large areas. We know very little about the minimum number of samples or specimen records required for a reliable inference of regional diversity (Hubbell, 2015; Locey & Lennon, 2016; Slik et al., 2015). Extrapolation to large regions of different evolutionary histories, for example regions comprising tropical, subtropical and temperate biomes, is especially challenging (O'Dea, Whittaker, & Ugland, 2006; Ugland et al., 2003).

Faced with these difficulties, biologists have resorted to the use of species discovery curves to estimate the number of species in regions (Bebber, Marriott, et al., 2007; Costello & Wilson, 2011; Frank & Curtis, 1979; Joppa, Roberts, Myers, et al., 2011; Joppa, Roberts, & Pimm, 2011; Medellín & Soberón, 1999; Soberón & Llorente, 1993; Wilson & Costello, 2005; Zapata & Robertson, 2006). A species discovery curve is a kind of species accumulation curve that tallies the increase in species number as new species are discovered over time. The assumption underlying the use of discovery curves is that, with the continued discovery of species, the number of undiscovered species in a region is constantly depleted, leading to an asymptotic discovery curve. Although the application of species discovery curves in estimating regional diversity seems promising (Giam et al., 2010, 2012; Joppa, Roberts, Myers, et al., 2011; Joppa, Roberts, & Pimm, 2011), empirical up-to-date data have shown little sign of depletion but rather fluctuating discoveries for many taxa (Costello & Wilson, 2011; Costello et al., 2012; Essl et al., 2013; Nabout et al., 2013; Randhawa et al., 2015; Zapata & Robertson, 2006). Among the many factors that may contribute to uncertainties associated with regional richness estimation, the effect of historical events (e.g., wars, unbalanced economic development, or technological breakthroughs) has been considered to be particularly serious (Bebber, Harris, et al., 2007; Bebber, Marriott, et al., 2007; Gaston, 1995; Rosenberg, Johansson, Powers, & Miller, 2013) and has invoked much discussion and modelling effort, but little consensus has been reached on how to model historical effects. Some studies argue that data from initial discovery stages are largely uninformative and should be discarded so as to remove the effect of historical events (Bebber, Marriott, et al., 2007; Woolhouse et al., 2008), while others continue to fit the whole dataset (Costello, Wilson, & Houlding, 2012; Wilson & Costello, 2005). We also noted that most previous studies avoid assessing their models by comparing predictions made from subsets of data from different periods of time, because such predictions are highly variable and are strongly subject to which subset of data is used (Costello et al., 2012; Giam et al., 2010; Wilson & Costello, 2005). Having assessed the haphazard behaviour of historical influences, Bebber, Marriott, et al. (2007) suggest that discovery curves should not be used at all for estimating regional richness (also see Hammond, 1995).

In this study we took an 'adaptive learning' approach (Caley, Fisher, & Mengersen, 2014) to revisit discovery curves. After a thorough evaluation of the existing methods in relation to the 31,220 vascular plant species compiled from *Flora of China*, we have learned insufficiencies of those methods in modelling species discovery in regions such as China, which is very rich in biodiversity but has suffered endless social turmoil until very recently. We proposed a new formulation to take account of the effects of historical events on species discovery. We showed that our model very well described historical effects and significantly improved richness estimation for vascular plants in China, compared with other species discovery models. We also tested our model with three other empirical datasets (world monocots, European butterflies and European spiders) and simulated data. For vascular plants in China, we further compared the performance of our model with that of other two widely noted models for estimating regional richness, the species–relationship and a recently developed taxonomic rank curve (Mora et al., 2011), to appreciate our species discovery model. Our results show that species discovery curves, formulated appropriately, are the most (probably the only) reliable tool for estimating species richness at a regional scale.

## 2 | METHODS

### 2.1 | Data

Data for 31,220 species, including species names, higher-rank taxonomies (genus, family, order, class, phylum and kingdom), and publication dates were obtained from *Flora of China*, (see Appendix S1 in the Supporting Information). Note that the Introduction on http://foc.eflora.cn/ states that *Flora of China* contains 31,362 species, but we were only able to compile 31,220 of them from http://www.efloras.org/flora_page.aspx?flora_id=2 with 142 species missing. Our study will thus be based on the data we have compiled (i.e., 31,220 species). To construct species–area relationships we used the publication dates and province-level geographical distribution compiled from *Flora Republicae Popularis Sinicae* (http://frps.eflora.cn/), which includes 31,066 species. There were 28 provinces after the municipalities were merged into adjacent provinces. We used the publication dates of accepted names as the earliest publication dates. The numbers of the first four Linnaean hierarchies did not change over time in our data, while the numbers of families and genera increased slightly over the past 50 years. For species discovery models, we used data between 1755 and 2000 for analysis (which contain 29,821 of the 31,220 species) to avoid the influences of Linnaeus' publication of *Species Plantarum* in 1753 and the inclusion of basionyms published in recent volumes of *Flora of China* after 2000 (Appendix S2). As suggested by Bebber, Marriott, et al. (2007), a sudden decrease in taxonomic effort will cause a false plateau in cumulative discovery curves and thus invalidate predictions. We identified the major events that affected species discovery processes and marked them on cumulative curves to display their influences. To show the effects of historical events on predictions, we first estimated species richness using data for the range from 1755 to 2000, then sequentially excluded the data for the most recent 10 years and repeated the analysis until data of the most recent 70 years (i.e., data after 1930) were excluded. If data were further excluded, we found that most models failed to converge or give sensible predictions.

Three other datasets with relatively well known species richness were used to further test our model. They include world monocots (59,238 described species excluding grasses, provided by Dr Joppa; see Joppa, Roberts, Myers, et al., 2011), European butterflies (9,268 described species) and European spiders (3,913 described species) from Essl et al. (2013).

## 2.2 | Species discovery models

The basic form of species discovery model is proposed by Bebber, Marriott, et al. (2007), by which the cumulative number of discovered species ($S_{t+1}$) at time $t+1$ equals the cumulative number of discovered species ($S_t$) at time $t$ plus a proportion of the number of undiscovered species (we adopted a different but more consistent notation than the original nation of Bebber et al.): $S_{t+1}=S_t+k(S_{\text{tot}}-S_t)+\varepsilon_t$. Rearranging the equation, we have

$$\Delta S_t=S_{t+1}-S_t=k(S_{\text{tot}}-S_t)+\varepsilon_t, \tag{1}$$

where $\Delta S_t$ is the number of species discovered per time interval, $k$ is the efficiency of species discovery, which depends on factors such as the number of taxonomists, visibility of species and the number of unexplored habitats. $S_{\text{tot}}$ is the total number of species (discovered + undiscovered) in a region. $\varepsilon_t$ is the error term. Our goal is to estimate $S_{\text{tot}}$.

Most species discovery models can be reformulated as in Equation 1 and they differ from each other in three key aspects: (i) the functional forms of the cumulative curve ($S_t$), (ii) how the variation (dispersion) in discovery rate ($\Delta S_t$) is accounted for, and (iii) the assumed error ($\varepsilon_t$) distributions (Table 1). The logistic model and negative exponential model are the two most widely used models (see below). In the logistic model, the discovery rate first increases then decreases with time, while it decreases linearly with time in the negative exponential model. However, both models have been criticized for being sensitive to historical events (Bebber, Marriott, et al., 2007).

We noted that the logistic and negative exponential models can be derived by assuming different parameterizations of $k$ in Equation 1. Three types of $k$ are commonly used (in the following, $a$ and $b$ are two parameters):

1. $k=a$: Equation 1 becomes $\Delta S_t=a(S_{\text{tot}}-S_t)+\varepsilon_t$. Solving this equation (by ignoring the error term), we obtain $S_t$ as a negative

**TABLE 1** Species discovery models used in the literature

| | Cumulative curve | Dispersion accounted for? | Error distribution |
|---|---|---|---|
| Nabout et al. (2013) | Logistic/Gompertz | No | |
| Giam et al. (2012) | Nonlinear | No | Normal |
| Costello et al. (2012) | Logistic | Yes | Poisson |
| Joppa, Roberts, & Pimm (2011) | Nonlinear | Yes | Normal |
| Joppa, Roberts, Myers, et al. (2011) | Nonlinear | No | Normal |
| Giam et al. (2010) | Negative exponential | Yes | Negative binomial |
| Woolhouse et al. (2008) | Negative exponential | No | Poisson |
| Woodley et al. (2008) | Logistic/Michaelis–Menton | No | |
| Bebber, Marriott, et al. (2007) | Logistic/negative exponential | Yes | Poisson |
| Zapata & Robertson (2006) | Logistic/Von Bertalanffy | No | |
| Wilson & Costello (2005) | Logistic | Yes | Poisson |
| Solow & Smith (2005) | Negative exponential | No | Binomial |
| Medellín & Soberón (1999) | Logarithmic | No | |

1. Giam et al. and Joppa, Roberts, & Pimm, respectively, square-root transformed and log transformed their data before fitting the model.
2. Wilson & Costello applied a non-homogeneous renewal stochastic process to model dispersion.
3. Joppa et al. used $SD=z\sqrt{\text{mean}}$ to approximate a Poisson distribution with the parameter $z$ accounting for dispersion.
4. Bebber et al. used a quasi-Poisson model to estimate the dispersion parameter.

exponential function of $t$ (Bebber, Marriott, et al., 2007; Soberón & Llorente, 1993; Woolhouse et al., 2008).

2. $k=a+bS_t$: Equation 1 becomes $\Delta S_t=(a+bS_t)(S_{\text{tot}}-S_t)+\varepsilon_t$, Solving this equation, we obtain $S_t$ as a logistic function of $t$ (Bebber, Marriott, et al., 2007; Costello & Wilson, 2011; Nabout et al., 2013; Wilson & Costello, 2005; Zapata & Robertson, 2006).
3. $k=T_t(a+bt)$: Equation 1 becomes $\Delta S_t=T_t(a+bt)(S_{\text{tot}}-S_t)+\varepsilon_t$, where $T_t$ is the number of authors describing species which is used as a proxy for taxonomic effort during the time interval ($t$, $t+1$). This model is proposed by Joppa, Roberts, Myers, et al. (2011) and Joppa, Roberts, and Pimm (2011) to deal with temporal variation (and thus historical events) in species discoveries, leading to a nonlinear model of $S_t$ that does not have an analytical form. Although Joppa et al.'s formulation has improved the performance of Equation 1, the robustness of their model against historical variation has not been tested.

In this study, we propose a different formulation of $k$, as a function of $\Delta S_t$ (the number of species discovered per time interval) instead of a function of time: $k=T_t(a+b\Delta S_t)$. As will be clear later, this seemingly simple change substantially improves the performance of the method. This is not surprising because the relationship between the efficiency of species discovery ($k$) and time is deemed to be haphazard due to historical events: a high $k$ at time $t$ would not guarantee a high $k$ at time $t+1$ precisely because of the unpredictability of these events. In contrast, the historical effects on $k$ could be captured by the number of species discovered per time interval $\Delta S_t$: a high $k$ at time $t$ naturally leads to a high $\Delta S_t$ regardless of what happened.

Substituting $k=T_t(a+b\Delta S_t)$ into Equation 1, our model is

$$\Delta S_t=T_t(a+b\Delta S_t)(S_{\text{tot}}-S_t)+\varepsilon_t, \qquad (2)$$

as in Joppa et al.'s model, $S_t$ in Equation 2 cannot be explicitly solved. In this study, we used generalized nonlinear least square regression (the gnls function in the R package 'nlme') to fit Equation 2. A power function was used to link the variance function with the mean to account for overdispersion or underdispersion in the error term: var=mean$^{2z}$, where $z$ is a parameter. When $z$ is close to 0.5, the error approximates a Poisson distribution. $z$ values larger than 0.5 suggest overdispersion, while those smaller than 0.5 suggest underdispersion. Data lumped in 5-year intervals were used in the estimates (lumping data in 10-year interval did not change the behaviours of the models). The approximate 95% confidence intervals of $S_{\text{tot}}$ were constructed from the standard errors of the estimated parameters in the fitted generalized nonlinear least square model, which could be directly obtained by the intervals function in 'nlme'. R codes for fitting species discovery models are included in Appendix S3.

Model 2 and three other discovery models (negative exponential model, logistic model, and Joppa et al.'s model) were applied to estimate the richness of vascular plants in China using species data over different time ranges. The performances of the four models were further tested and compared using discovery data for world monocots, European butterflies and European spiders (see Section 2.1). We also implemented a simulation test of the models (Appendix S4).

### 2.3 | Species–area relationships and the taxonomic rank curve

To construct the species–area relationship, we randomly drew a certain number of provinces from the 28-province pool: one province, two provinces, and up to the whole country (28 provinces). Each draw (i.e., one set of provinces) was done without replacement and was repeated five times (10 or 20 times each draw did not change the results). The total area and total number of species were counted for each draw. In cases where different draws had the same set of provinces (e.g., the five repeated draws for the 'whole country' would have the same 28 provinces), the replicated data points were removed. We used the 'mmSAR' package to fit the species–area data (Guilhaumon, Mouillot, & Gimenez, 2010). Six asymptotic parametric models were used to extrapolate species richness with infinite area: the Lomolino model, the negative exponential model, the Monod model, the rational model, the logistic model and the cumulative Weibull model. Because the data points are not independent in our analysis, we used a resampling procedure to generate confidence intervals: we repeated the above-mentioned sampling procedure 1000 times, and obtained a point estimate for each sample. The 95% confidence intervals were given by the 0.025 and 0.975 quantiles of the 1000 point estimates.

We also used the taxonomic rank curve that was used by Mora et al. (2011) to estimate global eukaryotic species richness. The taxonomic rank curve describes a log-linear relationship between the number of taxa along the Linnaean hierarchies which are ranked as kingdom (1), phylum (2), class (3), order (4), family (5), genus (6), and species (7). The ranked values were log-transformed for plotting the taxonomic rank curve. Although the number of higher taxa discovered per year changed very little with taxonomic effort over time in our data, the number of families and genera increased slightly over the last 50 years. We estimated the 'true' number of families and genera from the family and genus discovery curves, as Mora et al. (2011) did. Because only the negative exponential and logistic discovery models could be used to estimate the asymptotic values of numbers of families and genera, we used the same logistic model of species discovery curves (i.e., Equation 1) by assuming $k=a+bS_t$ to estimate the asymptotic values for these two hierarchies and used those values to construct the taxonomic rank curve. The estimate of the total number of species and its confidence intervals was obtained using an ordinary linear regression to the log-linear taxonomic rank data, assuming data independence. The independence assumption could be relaxed by assuming autocorrelated errors if the confidence interval becomes unrealistically narrow, but such an exercise changed the result very little (Mora et al., 2011) and was not implemented here.

## 3 | RESULTS

### 3.1 | Species discovery curves

The number of vascular plant species discovered in China per year underwent large fluctuations over time (Figure 1a). In the early discovery stage, John Reeves (1774–1856), a prominent British naturalist,

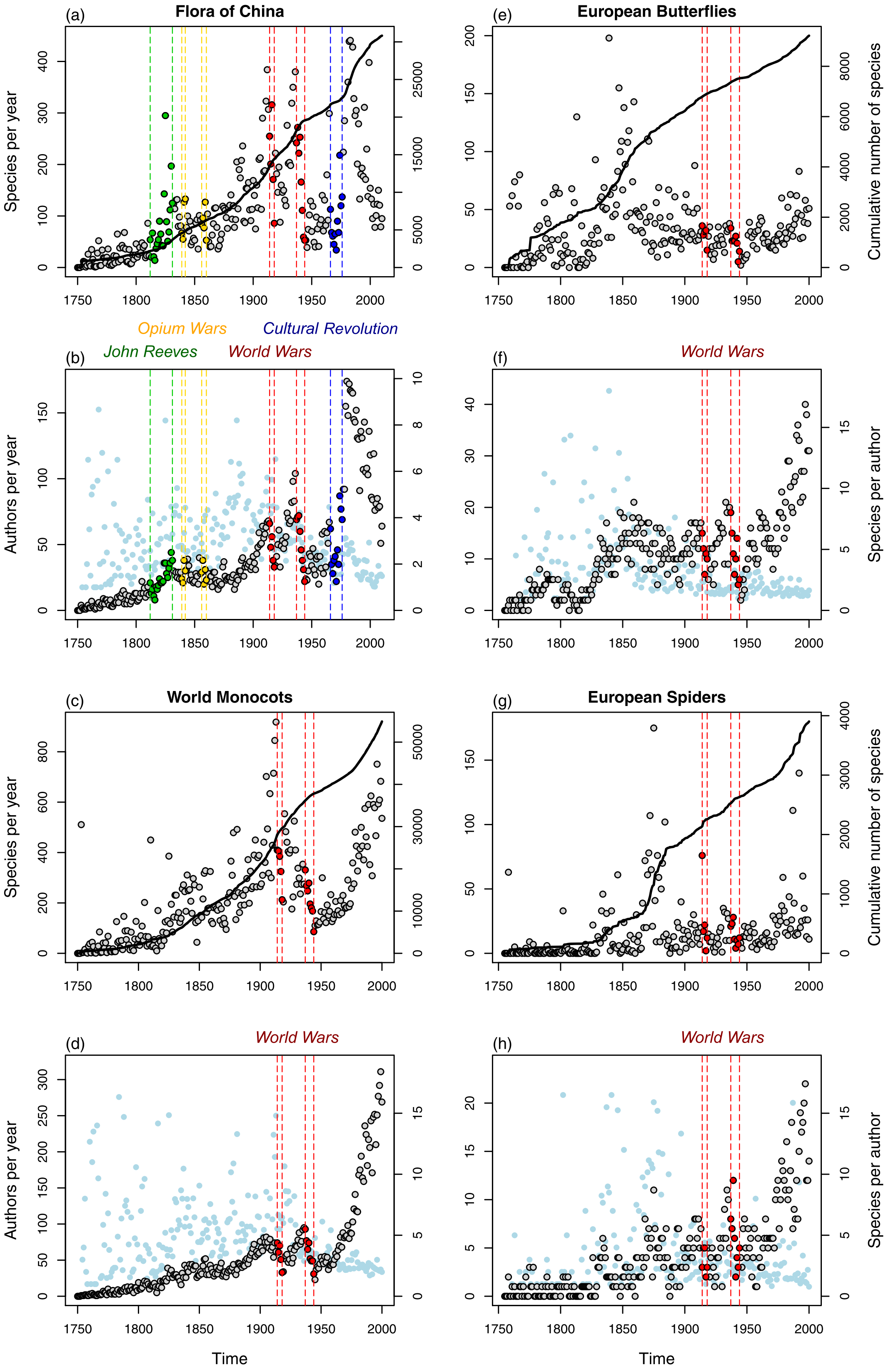

**FIGURE 1** Species discoveries over time. Panels (a), (c), (e) and (g) show the number of species discovered per year (grey points) and the cumulative number of discovered species (solid black lines) for each of the datasets. Panels (b), (d), (f) and (h) show the number of authors per year (grey points) and the number of species discovered per author (light blue points). Coloured points and dashed lines are used to highlight the periods influenced by major historical events

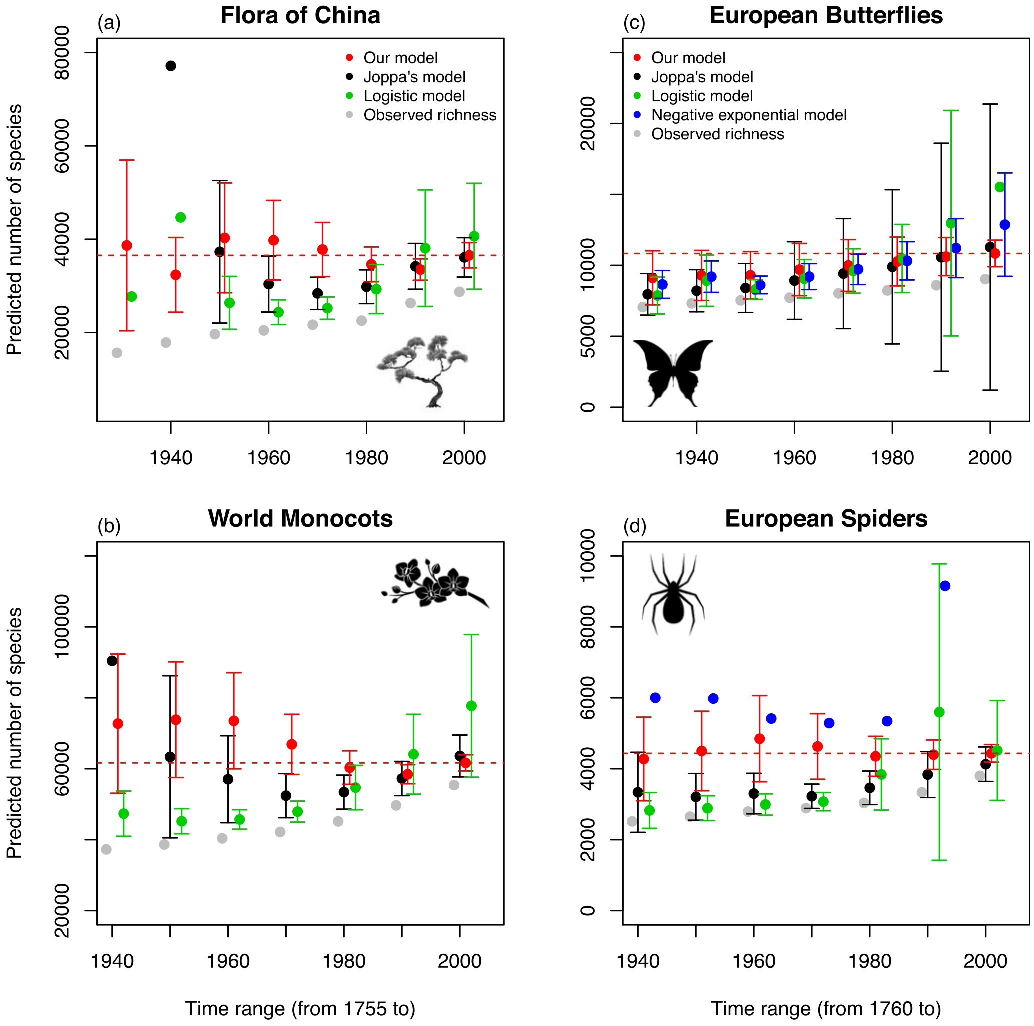

**FIGURE 2** Predictions of different species discovery models using data for different time ranges. The models are: our new model (Model 2), Joppa et al.'s model (2011), the logistic model, and the negative exponential model. For the negative exponential model, data starting from 1865 were used because the model can only be fitted if discovery rates are decreasing. The 95% confidence intervals are shown in bars. The dashed horizontal red lines are the final predictions of our new model using the full range of data: *Flora of China*, $S_{tot} = 36{,}554$; world monocots, $S_{tot} = 61{,}646$; European butterflies, $S_{tot} = 10{,}840$; European spiders, $S_{tot} = 4{,}436$ (see Table 2)

spent 19 years in China (1812–31) as a tea inspector working in the East India Company in Canton (now Guangzhou). It was due to his collections that many species in China were described by taxonomists in London or Paris, resulting in the first peak in discovery rate between 1812 and 1831 (98 species were discovered per year; Figure 1a). After defeated in the two Opium Wars (the first from 1839 to 1842 and the second from 1856 to 1860), China was forced to open to the West. The number of species discovered per year increased steadily in the following 50 years, and peaked in 1911 when the anti-Qing Revolution broke out (142 species were discovered per year between 1860 and 1911). The discovery rate decreased dramatically around the two World Wars (1914–19 and 1938–45) and the Cultural Revolution (1966–76) (Figure 1a). During the Cultural Revolution an average of 93 species were discovered per year. The following rapid increase in discovery in the 1980s was due to the increased investment in science resulting from China's 'reform and opening-up' policies and the advent of molecular techniques (371 species were discovered per year between 1978 and 1990). The discovery rate in the first decade of the 21st century was 110 species per year. The number of authors per year and the number of species discovered per author followed a similar historical pattern (Figure 1b). Discovery rates of world monocots, European butterflies and spiders also underwent fluctuations over time, especially during the two World Wars (Figure 1c–h).

The results in Figure 2 show the predictions of four discovery models including Model 2 which outperforms other models in two aspects. First, it is robust to the use of discovery data from different time ranges. For example, the richness of China's vascular plants estimated by our model varies from 32,376 to 40,284 by using species discovered over periods of 1755–1930, 1755–1940, and so on, while the richness estimated by the next closest model (that of Joppa et al.) varies from 28,397 to 77,151. This difference is even more noticeable for other test data (Figure 2b–d). Second, our model overall has narrower confidence intervals than the other models, particularly for data that include more recent discoveries (Figure 2). When the full range of

**TABLE 2** Comparison of different species discovery models using data from 1755 (1760 for European butterflies and spiders) to 2000

| Data | Model | $S_{tot}$ | Lower 95% CI | Upper 95% CI | $a$ | $b$ | $z$ |
|---|---|---|---|---|---|---|---|
| *Flora of China* | Our model | 36,554 | 33,847 | 39,261 | $1.12 \times 10^{-4}$ | $1.54 \times 10^{-7}$ | 0.77 |
| | Joppa's model | 36,113 | 31,892 | 40,333 | $-1.32 \times 10^{-4}$ | $8.12 \times 10^{-7}$ | 0.84 |
| | Logistic model | 40,643 | 29,294 | 51,993 | $-4.77 \times 10^{-4}$ | $2.63 \times 10^{-6}$ | 1.06 |
| World monocots | Our model | 61,646 | 59,372 | 63,922 | $1.11 \times 10^{-4}$ | $7.78 \times 10^{-8}$ | 0.64 |
| | Joppa's model | 63,605 | 57,700 | 69,509 | $-7.32 \times 10^{-4}$ | $4.94 \times 10^{-7}$ | 0.88 |
| | Logistic model | 77,750 | 57,648 | 97,853 | $1.09 \times 10^{-3}$ | $1.36 \times 10^{-6}$ | 0.98 |
| European butterflies | Our model | 10,840 | 9,902 | 11,779 | $5.17 \times 10^{-4}$ | $2.70 \times 10^{-6}$ | −0.12 |
| | Joppa's model | 11,286 | 1,202 | 21,370 | $6.63 \times 10^{-3}$ | $-2.94 \times 10^{-6}$ | −0.04 |
| | Logistic model | 15,527 | −1,967 | 33,021 | $1.22 \times 10^{-2}$ | $1.07 \times 10^{-6}$ | 3.34 |
| | Negative exponential model | 12,872 | 9,230 | 16,514 | $3.07 \times 10^{-2}$ | – | 0.59 |
| European spiders | Our model | 4,436 | 4,188 | 4,684 | $6.37 \times 10^{-4}$ | $2.11 \times 10^{-5}$ | 0.58 |
| | Joppa's model | 4,129 | 3,642 | 4,615 | $-3.71 \times 10^{-2}$ | $2.12 \times 10^{-5}$ | 1.03 |
| | Logistic model | 4,517 | 3,110 | 5,924 | $-2.32 \times 10^{-4}$ | $3.01 \times 10^{-5}$ | 0.89 |

The estimated total number of species $S_{tot}$ and its 95% CI are shown for each model. $a$ and $b$ are estimated parameters. $z$ is the estimated parameter of the variance function var=mean$^{2z}$.

discovery data was used, the richness estimated by Joppa et al.'s model was very close to that of our model (see the estimate for the last point in each panel of Figure 2; also see Table 2). However, the confidence intervals of Joppa et al.'s method are not reliable and can vary unexpectedly (e.g., the European butterflies data). Nevertheless, Joppa et al.'s model performs better than the logistic model and negative exponential model but is sensitive to the effect of 'false plateau' which is caused by decreasing taxonomic efforts due to historical effects rather than the depletion of undiscovered species (Figure 1a,c,e,g). The logistic model is very sensitive to historical events and provides unreliable estimates in most cases. The negative exponential model performs the worst of all; only when applied to the European butterfly data did it make sensible predictions (Figure 2c).

The estimated numbers for China's vascular plant species using the entire data range (1755–2000) are shown in Table 2. Our model estimated 36,554 species with a 95% confidence interval of 33,847–39,261. Joppa et al.'s estimate is very close to ours but has a wider confidence interval, while the estimate of the logistic model is unreasonably high (Table 2). Our estimate indicates that there are about 5,334 vascular plant species remaining to be discovered in China, representing 14.6% of total number of species.

Results of simulations (Appendix S4) are consistent with that of the empirical data (Figure 2), showing that our model has similar coverage probabilities but narrower confidence intervals than Joppa et al.'s model.

### 3.2 | Species–area relationships and the taxonomic rank curve

The province-level species–area relationship (SAR) and the six fitted species–area models are shown in Figure 3a. The Lomolino and cumulative Weibull SAR models estimated unrealistically high numbers of species, while the estimates of the logistic and the negative exponential models were unrealistically low (fewer than 31,220 species) (Figure 3b). The Monod and rational models estimated the total number of species more reasonably, but none of the models can be trusted because the performance of all the SAR models is highly sensitive to the data from different time periods that were used to fit the SAR models (Figure 3b).

Mora et al.'s taxonomic rank curve only estimated 10,127 vascular plant species in China, which is just about one-third of the 31,220 species discovered (Figure 3c). The method produced a very wide 95% confidence interval of 1,924–53,315 species.

## 4 | DISCUSSION

The estimation of regional richness has been a major challenge for the study of biodiversity. The problem is not that there is a shortage of methods but rather that there are too few data—only a tiny fraction of landscapes can be reasonably sampled in almost all applications. For example, in an exceptional study Xu et al. (2012) sampled 164 25 m × 25 m quadrats distributed on a 1 km × 1 km grid covering a 472-km$^2$ reserve in Hainan Island, China, but the study was still only able to sample about 0.022% of the reserve landscape. Few methods work with such a small fraction of samples. Faced with this challenge, species discovery curves have been used as a major tool for estimating regional richness (Bebber, Marriott, et al., 2007; Costello & Wilson, 2011; Joppa, Roberts, Myers, et al., 2011; Joppa, Roberts, & Pimm, 2011; Medellín & Soberón, 1999; Soberón & Llorente, 1993; Solow & Smith, 2005; Wilson & Costello, 2005; Zapata & Robertson, 2006). However, discovery curves are known to be notoriously affected by historical events such as wars, unbalanced economic developments and technological advances (Essl et al., 2013; Gaston, 1995; Randhawa et al., 2015; Rosenberg et al., 2013). This is especially true for botanical discoveries in China, which has been affected by endless wars and civil conflicts in the past two centuries. How to account for the effect of historical events on species discovery is key to the success of discovery curves in estimating regional richness. In this study, following an 'adaptive learning' approach

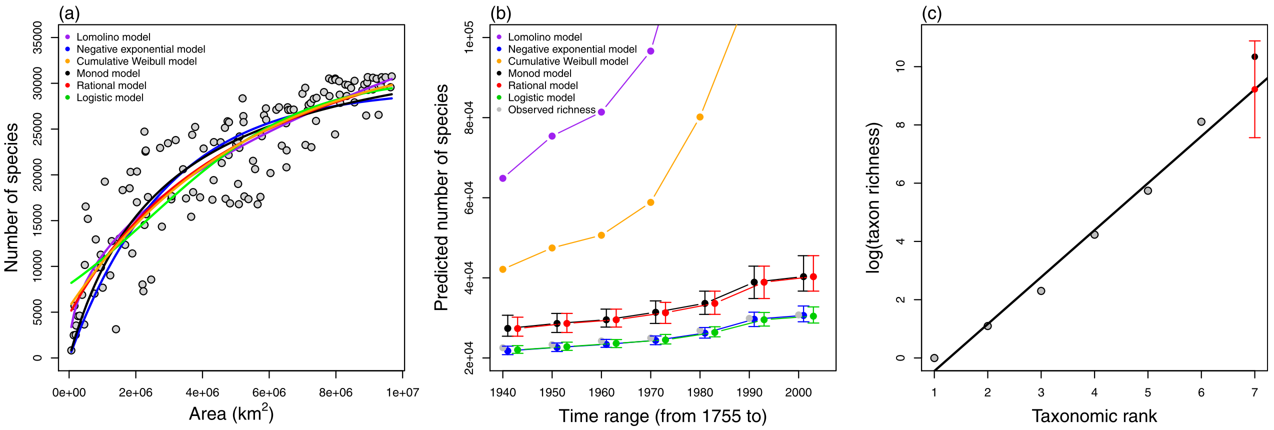

**FIGURE 3** (a) The province-level species–area relationship and the fitted SAR models. (b) Predictions of different species–area models using data for different time ranges. Confidence intervals are not shown for the Lomolino and cumulative Weibull models because they are too wide to be displayed. (c) Mora et al.'s taxonomic curve in which *y* axis is the log-transformed number of taxa and the *x* axis is the taxonomic rank of the Linnaean hierarchies: kingdom, phylum, order, class, family, genus and species. The red point shows the estimated number of species, with the 95% confidence interval bar. The black point shows the observed number of species

(Caley et al., 2014), we proposed a new formulation to account for historical effects. The new model (Model 2) estimated a total of 36,554 vascular plant species in China, which is 1.17 times the number of species documented in *Flora of China*. If the current rate of 110 species discovered per year persists, it would take about 50 years to discover all the vascular plant species in China. New discoveries might still come from lumping and splitting existing taxa, but this is unlikely to significantly change our result because the taxonomic status of all species in *Flora of China* has been thoroughly revised since the completion of the Chinese version, *Flora Republicae Popularis Sinicae*, in 2004.

What really makes our method remarkable is the robustness of the model to data of different periods and the narrower confidence intervals (Figure 2, Table 2). For example, we were able to make a sensible estimation for vascular plants in China based on data collected up to the year 1930—by then only 52% of today's species had been discovered (the total number of recorded species by 1930 was 16,330). The superior performance of our method arises from the way the effect of historical events on taxonomic efficiency is modelled. Taxonomic efficiency is defined as $\Delta S_t/[T_t(S_{\mathrm{tot}}-S_t)]$, which is the number of species discovered per author corrected by the number of undiscovered species (Joppa, Roberts, Myers, et al., 2011). (Note that *k* in Equation 1 is taxonomic efficiency multiplied by the number of authors $T_t$.) By this measure, the taxonomic efficiency against the number of species discovered per time interval ($\Delta S_t$) follows a linear relationship (rearranging Equation 2, we have $\Delta S_t/[T_t(S_{\mathrm{tot}}-S_t)]=a+b\Delta S_t$). We suggest that $\Delta S_t$ is a good indicator of historical effects because historical events influenced both the number of authors and their efficiency: a high $\Delta S_t$ naturally represents a high taxonomic efficiency and vice versa. This linear relationship between $\Delta S_t$ and taxonomic efficiency is well supported by the four independent datasets (Figure 4a–d). Joppa, Roberts, Myers, et al. (2011) and Joppa, Roberts, and Pimm (2011) assume that the taxonomic efficiency increases linearly with time. However, the linear relationship between taxonomic efficiency and time (Figure 4e–h) is generally weaker than that of taxonomic efficiency and $\Delta S_t$ (Figure 4a–d). In some cases, for example European butterflies (Figure 4g), the taxonomic efficiency even decreases with time, suggesting inconsistency of the relationship.

Historical effects are not explicitly accounted for in the two commonly used models, the logistic and negative exponential discovery models (Costello & Wilson, 2011; Giam et al., 2012; Medellín & Soberón, 1999; Nabout et al., 2013; Solow & Smith, 2005; Woodley, Naish, & Shanahan, 2008; Woolhouse et al., 2008; Zapata & Robertson, 2006). The effects of historical fluctuation in species discovery on these two models could be examined by plotting species discovery per year against the cumulative number of discovered species (Figure 4i–l). From Equation 1, species discovered per year, $\Delta S_t$, is expected to have a quadratic function with cumulative number of discovered species for the logistic model but a linear decreasing function for the negative exponential model. Data in Figure 4i–l show an approximate quadratic form but with huge fluctuations in $\Delta S_t$ that seem to increase with time. The decreasing function underlying the negative exponential model is only found in part of the European butterfly data (Figure 4k).

The choice of data to fit models becomes controversial because of historical effects. Due to the occurrence of a 'false plateau', all models except ours underestimated the total number of China's vascular plant species when the upper limit of the data range fell in the period between 1950 and 1980 (Figure 2a). This is in agreement with the observation that the estimates of discovery models are sensitive to the choice of data range (Alroy, 2002; Bebber, Marriott, et al., 2007). Unlike other studies (Bebber, Marriott, et al., 2007; Medellín & Soberón, 1999; Woolhouse et al., 2008; Zapata & Robertson, 2006), the application of our model does not require identification of which part of the cumulative curve represents the true decline of discovery rate in order to make reliable predictions. There is one possible caveat about our model: as our model relies on using discovery rates as a proxy for historical effects, it might not be robust if the total number of species is too small (all examples in this study have a few thousand species—the case for regional diversity).

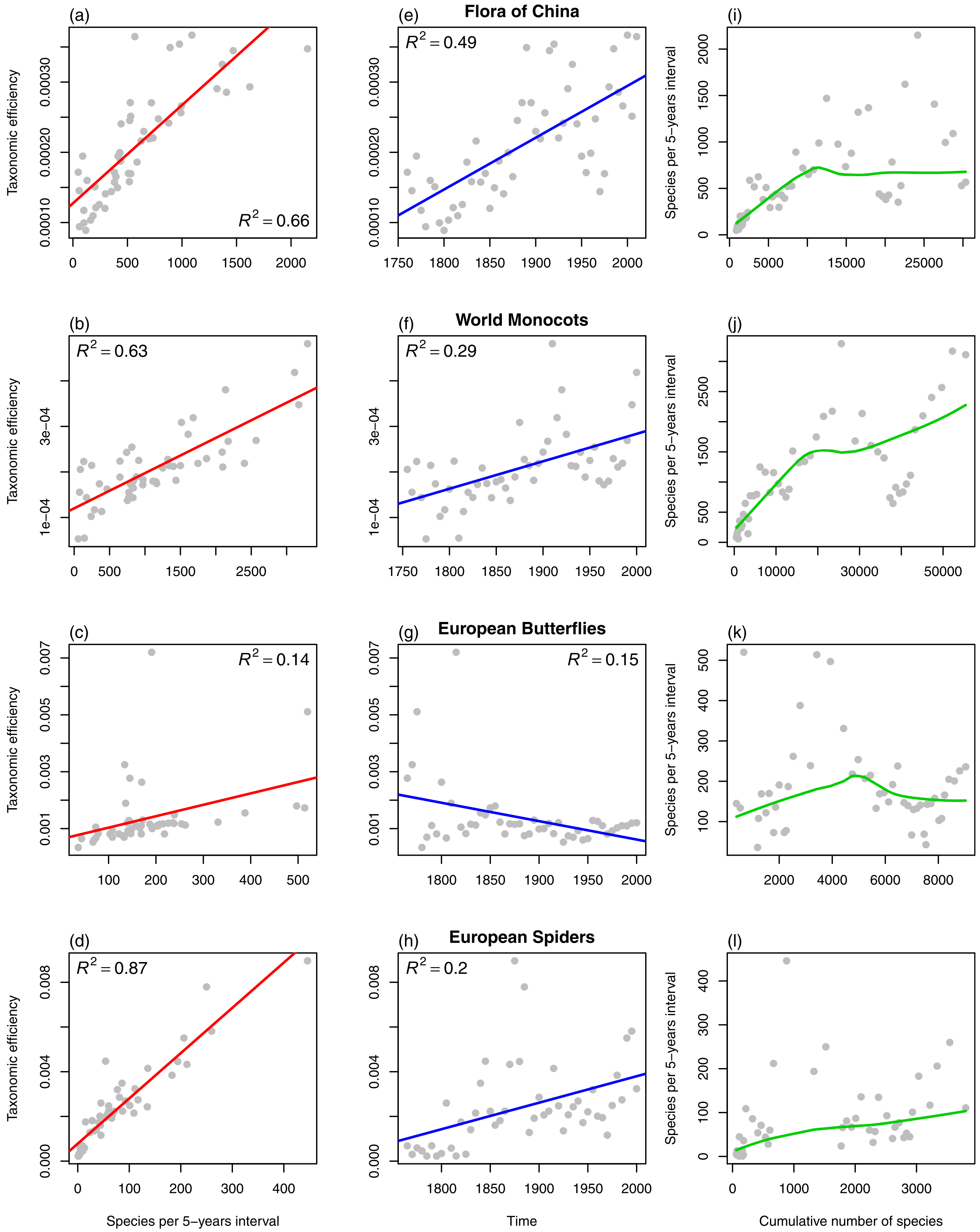

**FIGURE 4** (a)–(d) Taxonomic efficiency (defined as the number of species described per author corrected by the number of undiscovered species) against species discovered per 5-year time interval. (e)–(h) Taxonomic efficiency per 5-year interval against time. (i)–(l) Species discovered per 5-year time interval against cumulative number of discovered species. Solid lines show the locally weighted smoothing curves. For *Flora of China* $S_{\text{tot}} = 36{,}554$ was used; for world monocots $S_{\text{tot}} = 61{,}646$; for European butterflies $S_{\text{tot}} = 10{,}840$; and for European spiders $S_{\text{tot}} = 4{,}436$ (see Table 2)

In addition to the discovery curves, we also applied six SAR models and Mora et al.'s (2011) taxonomic rank curve to *Flora of China* data, but none of them performed satisfactorily. The SAR models either estimated unrealistically high or low richness or their estimates were too sensitive to the use of different subsets of data, meaning that the incompleteness of species inventories could have serious consequence for SAR predictions (Figure 3b). The sensitivity of SAR models to incomplete inventory data is also recognized by Gerstner et al. (2014). SAR models are not often used to estimate regional richness for the reason that at the regional scale areal samples are simply too small for most SAR models to reliably extrapolate richness (Hammond, 1992). More importantly, it is well known that area is not the only (or may not even the main) factor determining species richness at regional or global scales (Ricklefs & He, 2016). For successful applications of SAR models at the regional scale, species turnover patterns (beta diversity) and environmental heterogeneity must be considered (Beck & Kitching, 2007; Jobe, 2008).

Mora et al.'s (2011) taxonomic rank curve only estimated about one-third of the observed richness (Figure 3c). The failure suggests that this method is largely *ad hoc* and lacks the generality necessary for a reliable estimation of regional species richness. In our opinion, a major problem with Mora et al.'s method is the use of the rank of Linnaean hierarchies. There is no a priori reason to suggest that the distances between adjacent ranks should be equally spaced in the taxonomic hierarchies. It is difficult to interpret the associated confidence intervals without more theoretical justification.

Other methods for estimating regional richness commonly used in the literature also include nonparametric estimators (Cao, Larsen, & White, 2004) and the taxa ratio (e.g., insect:plant richness ratio) method (Gaston, 1992). But those methods are of little use in most cases because data on the number of singleton or doubleton species are often unknown at the regional scale (for nonparametric methods), while the application of the taxa ratio relies on the condition that richness for one of the taxa is already known, which is very unlikely at the regional scale. The results of this study together with others (e.g., Joppa, Roberts, Myers, et al., 2011; Joppa, Roberts, & Pimm, 2011) show that species discovery curves, with historical effects being properly accounted for, offer a promising tool for estimating regional species richness. We argue that no data on a regional diversity are more informative than the recorded time of species discoveries.

## ACKNOWLEDGMENTS

We thank Lianming Gao and Hong Qian for their valuable advice and assistance on the use of *Flora of China*. We thank Lucas Joppa for kindly providing the data on world monocots. We thank two anonymous referees for providing valuable comments which substantially improved the study. We also thank Yanqing Liu for helping with compiling the data, Xinghua Sui, and Junfeng Tang and other members in the lab for helpful discussions. This work was funded by Sun Yat-sen University and the Natural Sciences and Engineering Research Council of Canada (NSERC).

## AUTHOR CONTRIBUTION

F.H. and M.L. conceived the study; M.L. collected and analysed the data; M.L. and F.H. wrote the paper.

## REFERENCES

Alroy, J. (2002). How many named species are valid? *Proceedings of the National Academy of Sciences USA*, *99*, 3706–3711.

Bebber, D. P., Harris, S. A., Gaston, K. J., & Scotland, R. W. (2007). Ethnobotany and the first printed records of British flowering plants. *Global Ecology and Biogeography*, *16*, 103–108.

Bebber, D. P., Marriott, F. H. C., Gaston, K. J., Harris, S. A., & Scotland, R. W. (2007). Predicting unknown species numbers using discovery curves. *Proceedings of the Royal Society B: Biological Sciences*, *274*, 1651–1658.

Beck, J., & Kitching, I. J. (2007). Estimating regional species richness of tropical insects from museum data: A comparison of a geography-based and sample-based methods. *Journal of Applied Ecology*, *44*, 672–681.

Caley, M. J., Fisher, R., & Mengersen, K. (2014). Global species richness estimates have not converged. *Trends in Ecology and Evolution*, *29*, 187–188.

Cao, Y., Larsen, D. P., & White, D. (2004). Estimating regional species richness using a limited number of survey units. *ÉcoScience*, *11*, 23–35.

Chao, A. (1987). Estimating the population size for capture–recapture data with unequal catchability. *Biometrics*, *43*, 783–791.

Chao, A., & Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, *58*, 531–539.

Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *345*, 101–118.

Costello, M. J., & Wilson, S. P. (2011). Predicting the number of known and unknown species in European seas using rates of description. *Global Ecology and Biogeography*, *20*, 319–330.

Costello, M. J., Wilson, S., & Houlding, B. (2012). Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Systematic Biology*, *61*, 871–883.

Curtis, T. P., Sloan, W. T., & Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences USA*, *99*, 10494–10499.

Essl, F., Rabitsch, W., Dullinger, S., Moser, D., & Milasowszky, N. (2013). How well do we know species richness in a well-known continent? Temporal patterns of endemic and widespread species descriptions in the European fauna. *Global Ecology and Biogeography*, *22*, 29–39.

Frank, J. H., & Curtis, G. A. (1979). Trend lines and the number of species of staphylinidae. *The Coleopterists Bulletin*, *33*, 133–149.

Gaston, K. (1992). Regional numbers of insect and plant species. *Functional Ecology*, *6*, 243–247.

Gaston, K. (1995). Patterns in species description: A case study using the Geometridae (Lepidoptera). *Biological Journal of the Linnean Society*, *55*, 225–237.

Gerstner, K., Dormann, C. F., 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. *Journal of Biogeography*, *41*, 261–273.

Giam, X., Ng, T. H., Yap, V. B., & Tan, H. T. W. (2010). The extent of undiscovered species in Southeast Asia. *Biodiversity and Conservation*, *19*, 943–954.

Giam, X., Scheffers, B. R., Sodhi, N. S., Wilcove, D. S., Ceballos, G., & Ehrlich, P. R. (2012). Reservoirs of richness: Least disturbed tropical forests are centres of undescribed species diversity. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 67–76.

Guilhaumon, F., Mouillot, D., & Gimenez, O. (2010). mmSAR: An R-package for multimodel species–area relationship inference. *Ecography*, *33*, 420–424.

Hamilton, A. J., Novotný, V., Waters, E. K., Basset, Y., Benke, K. K., Grimbacher, P. S., . . . Stork, N. E. (2013). Estimating global arthropod species richness: Refining probabilistic models using probability bounds analysis. *Oecologia*, *171*, 357–365.

Hammond, P. M. (1992). Species inventory. In B. Groombridge (Ed.), *Global biodiversity: Status of the earth's living resources* (pp. 17–39). London: Chapman and Hall.

Hammond, P. M. (1995). Described and estimated species numbers: An objective assessment of current knowledge. In D. Allsopp, D. L. Hawksworth, & R. R. Colwell (Eds.), *Microbial diversity and ecosystem function* (pp. 29–71). Wallingford, UK: CAB International.

Hubbell, S. P. (2015). Estimating the global number of tropical tree species, and Fisher's paradox. *Proceedings of the National Academy of Sciences USA*, *112*, 7343–7344.

Jobe, R. T. (2008). Estimating landscape-scale species richness: Reconciling frequency- and turnover-based approaches. *Ecology*, *89*, 174–182.

Joppa, L. N., Roberts, D. L., Myers, N., & Pimm, S. L. (2011). Biodiversity hotspots house most undiscovered plant species. *Proceedings of the National Academy of Sciences USA*, *108*, 13171–13176.

Joppa, L. N., Roberts, D. L., & Pimm, S. L. (2011). How many species of flowering plants are there? *Proceedings of the Royal Society B: Biological Sciences*, *278*, 554–559.

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences USA*, *2016*, 201521291.

May, R. M. (1988). How many species are there on earth? *Science*, *241*, 1441–1449.

Medellín, R. A., & Soberón, J. (1999). Predictions of mammal diversity on four land masses. *Conservation Biology*, *13*, 143–149.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biology*, *9*, 1–8.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*, 853–858.

Nabout, J. C., da Silva Rocha, B., Carneiro, F. M., & Sant'Anna, C. L. (2013). How many species of cyanobacteria are there? Using a discovery curve to predict the species number. *Biodiversity and Conservation*, *22*, 2907–2918.

O'Dea, N., Whittaker, R. J., & Ugland, K. I. (2006). Using spatial heterogeneity to extrapolate species richness: A new method tested on Ecuadorian cloud forest birds. *Journal of Applied Ecology*, *43*, 189–198.

Palmer, M. W. (1990). The estimation of species richness by extrapolation. *Ecology*, *71*, 1195–1198.

Peterson, A. T., & Slade, N. A. (1998). Extrapolating inventory results into biodiversity estimates and the importance of stopping rules. *Diversity and Distributions*, *4*, 95–105.

Randhawa, H. S., Poulin, R., & Krkošek, M. (2015). Increasing rate of species discovery in sharks coincides with sharp population declines: Implications for biodiversity. *Ecography*, *38*, 96–107.

Ricklefs, R. E., & He, F. (2016). Region effects influence local tree species diversity. *Proceedings of the National Academy of Sciences USA*, *113*, 674–679.

Rosenberg, R., Johansson, M. A., Powers, A. M., & Miller, B. R. (2013). Search strategy has influenced the discovery rate of human viruses. *Proceedings of the National Academy of Sciences USA*, *110*, 13961–13964.

Slik, J. W. F., Arroyo-Rodríguez, V., Aiba, S., Alvarez-Loayza, P., Alves, L. F., Ashton, P., . . . Venticinque, E. M. (2015). An estimate of the number of tropical tree species. *Proceedings of the National Academy of Sciences USA*, *112*, 7472–7477.

Soberón, M. J., & Llorente, B. J. (1993). The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, *7*, 480–488.

Solow, A. R., & Smith, W. K. (2005). On estimating the number of species from the discovery record. *Proceedings of the Royal Society B: Biological Sciences*, *272*, 285–287.

Stork, N. E. (1993). How many species are there? *Biodiversity and Conservation*, *2*, 215–232.

Stork, N. E., McBroom, J., Gely, C., & Hamilton, A. J. (2015). New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proceedings of the National Academy of Sciences USA*, *112*, 7519–7523.

Tedesco, P. A., Bigorne, R., Bogan, A. E., Giam, X., Jézéquel, C., & Hugueny, B. (2014). Estimating how many undescribed species have gone extinct. *Conservation Biology*, *28*, 1360–1370.

Ugland, K. I., Gray, J. S., & Ellingsen, K. E. (2003). The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology*, *72*, 888–897.

Wilson, S. P., & Costello, M. J. (2005). Predicting future discoveries of European marine species by using a non-homogeneous renewal process. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *54*, 897–918.

Woodley, M. A., Naish, D., & Shanahan, H. P. (2008). How many extant pinniped species remain to be described? *Historical Biology*, *20*, 225–235.

Woolhouse, M. E. J., Howey, R., Gaunt, E., Reilly, L., Chase-Topping, M., & Savill, N. (2008). Temporal trends in the discovery of human viruses. *Proceedings of the Royal Society B: Biological Sciences*, *275*, 2111–2115.

Xu, H., Liu, S., Li, Y., Zang, R., & He, F. (2012). Assessing non-parametric and area-based methods for estimating regional species richness. *Journal of Vegetation Science*, *23*, 1006–1012.

Zapata, F. A., & Robertson, D. R. (2006). How many species of shore fishes are there in the tropical eastern Pacific? *Journal of Biogeography*, *34*, 38–51.

## BIOSKETCHES

**Muyang Lu** is a PhD student primarily interested in macroecology, biogeography and biodiversity conservation.

**Fangliang He** is a theoretical ecologist and conservation biologist whose primary research involves understanding maintenance, dynamics and conservation of biodiversity.

## SUPPORTING INFORMATION

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

**How to cite this article**: Lu M, He F. Estimating regional species richness: The case of China's vascular plant species. *Global Ecol Biogeogr*. 2017;00:1–11. https://doi.org/10.1111/geb.12589