Ecological Applications, 29(3), 2019, e01866

A new baseline for countrywide α-diversity and species distributions: illustration using >6,000 plant species in Panama

Brian Leung,[1,2,3,5] Emma J. Hudgins,[1] Anna Potapova,[1,2] and Maria C. Ruiz-Jaen[4]

[1]*Department of Biology, McGill University, Montreal, Quebec H3A1B1 Canada*
[2]*School of Environment, McGill University, Montreal, Quebec H3A2A7 Canada*
[3]*Smithsonian Tropical Research Institute, P.O. Box 0843-03092, Panama City, Panama*
[4]*Subregional Office for Mesoamerica, Food and Agriculture Organization of the United Nations, Panama City, Panama*

Citation: Leung, B., E. J. Hudgins, A. Potapova, and M. C. Ruiz-Jaen. 2019. A new baseline for countrywide α-diversity and species distributions: illustration using >6,000 plant species in Panama. Ecological Applications 29(3):e01866. 10.1002/eap.1866

Abstract. Estimating α-diversity and species distributions provides baseline information to understand factors such as biodiversity loss and erosion of ecosystem services. Yet, species surveys typically cover a small portion of any country's landmass. Public, global databases could help, but contain biases. Thus, the magnitude of bias should be identified and ameliorated, the value of integration determined, and application to current policy issues illustrated. The ideal integrative approach should be powerful, flexible, efficient, and conceptually straightforward. We estimated distributions for >6,000 species, integrating species sightings (S) from the Global Biodiversity Information Facility (GBIF), systematic survey data (S_2), and "bias-adjustment kernels" (BaK) using spatial and species trait databases (S_2BaK). We validated our approach using both locational and species holdout sets, and then applied our predictive model to Panama. Using sightings alone (the most common approach) discriminated relative probabilities of occurrences well (area under the curve [AUC] = 0.88), but underestimated actual probabilities by ~4,000%, while using survey data alone omitted over three-quarters of the >6,000 species. Comparatively, S_2BaK had no systematic underestimation, and substantially stronger discrimination (AUC = 0.96) and predictive power (deviance explained = 47%). Our model suggested high diversity (~200% countrywide mean) where urban development is projected to occur (the Panama Canal watershed) and also suggested this is not due to higher sampling intensity. However, portions of the Caribbean coast and eastern Panama (the Darién Gap) were even higher, both for total plant biodiversity (~250% countrywide mean), and CITES listed species. Finally, indigenous territories appeared half as diverse as other regions, based on survey observations. However, our model suggested this was largely due to site selection, and that richness in and out of indigenous territories was roughly equal. In brief, we provide arguably the best estimate of countrywide plant α-diversity and species distributions in the Neotropics, and make >6,000 species distributions available. We identify regions of overlap between development and high biodiversity, and improve interpretation of biodiversity patterns, including for policy-relevant CITES species, and locations with limited access (i.e., indigenous territories). We derive a powerful, flexible, efficient and simple estimation approach for biodiversity science.

Key words: bias; CITES; GBIF; global change; Global South; Neotropics; presence-only data; species distribution model.

Introduction

Estimating the spatial distribution of biodiversity provides critical baseline information for fundamental ecological questions on the drivers and patterns of living organisms on Earth, as well as for applied issues in conservation. For instance, such baseline information could help address questions on the distributions of ecosystem function and services in relation to local richness (Balvanera et al. 2014), or refine analyses of connectivity, where different plant species form a key part of the habitat (García-Feced et al. 2011). For more applied contexts, knowledge of the distribution of threatened species could help minimize impacts on these species through land-use planning or design of protected areas (Venter et al. 2014). Clearly, these considerations are central to the maintenance of our living planet, given projected socioecological changes (IPCC 2014).

Such baseline spatial estimates of biodiversity are rare and would benefit from further analyses. Existing estimates have been undeniably useful, but are based largely

Manuscript received 2 July 2018; revised 2 January 2019; accepted 7 January 2019. Corresponding Editor: Bradley J. Butterfield.
[5] E-mail: brian.leung2@mcgill.ca

on coarse range maps, which are often based on expert opinion, and have been critiqued in terms of overestimating occurrences (Hurlbert and Jetz 2007), or species distribution models (SDMs) based on species sightings (Hastie and Fithian 2013, Warton et al. 2013), which have well known biases, including incomplete detections (Lahoz-Monfort et al. 2014), causing underestimation bias, and unequal reporting across locations (Beck et al. 2014) and species (Troudet et al. 2017). For instance, areas near roads or population centers may be more accessible and more likely to be sampled; some species may possess visually conspicuous traits or may be more popular, leading to spatial and species-specific biases in the estimation of species distributions. Other methods that estimate unbiased probabilities rely on full presence–absence data (e.g., Wang et al. 2003, but see Fithian et al. 2015), time-series information for spreading species (Gertzen and Leung 2011), or show lack of bias only qualitatively, for single specific species (Gormley et al. 2011, Catterall et al. 2012). These methods are difficult to apply to entire suites of species, given that the necessary data are typically lacking, and/or they do not integrate easily with existing SDMs, and therefore cannot take advantage of continuing advances in the field. Given the increase of publicly available databases, disparate information can now be integrated, potentially increasing predictive power. Yet, given that each database contains imperfect information, the value of integration for estimating spatial biodiversity patterns needs to be clearly demonstrated. Further, given their importance, it is imperative to assess and improve the reliability of baseline biodiversity estimates.

In addition to high predictive power, an ideal estimation approach might contain the following characteristics: (1) *flexibility*, allowing researchers to choose amongst the wide array of SDMs available, to utilize continuing advances in the field; (2) *efficiency*, to estimate the thousands of species comprising countrywide biodiversity, computational complexity should scale near linearly with species number; and (3) *simplicity*, easily understandable approaches are more likely to be adopted and used correctly. If these elements could be achieved, they would yield a viable, general, biodiversity tool, facilitating applied and fundamental ecological inquiry.

Our objectives are twofold. (1) We derive an integrative estimation approach, which is simple, efficient and flexible, and we demonstrate its predictive power through validation. In so doing, we demonstrate the value of integration across global databases, despite inherent biases, for creating baseline spatial biodiversity estimates. (2) We estimate countrywide spatial plant biodiversity, using Panama as our model system, and illustrate how our novel baseline estimates improve our understanding of biodiversity patterns. Specifically, we identify regions of highest overall plant biodiversity and species of specific interest (CITES listed species), and estimates for regions with limited accessibility (indigenous territories). Further, we make available the distributions of >6,000 plant species across Panama.

Materials and Methods

Study system

Panama is an exemplar of issues facing the Global South. It has high biodiversity, but also is characterized by rapid economic (10.6% in 2011; Central Intelligence Agency 2016) and population growth (>40% by 2050; World Bank 2013), and is the focus of an integrative effort centered on sustainability science (Panama Research and Integrated Sustainability Model).

Data sources and processing

We used information from both Panama and Costa Rica, given their data availability and overlapping environmental conditions. We collated information from publicly available databases and government sources, including (1) occurrence information, species sightings and five sources of systematic plant surveys; (2) species information, species life history traits, correlates of species "popularity" via web-based analytics; (3) environmental variables, bioclimatic variables, elevation and slope, evapotranspiration, vegetation, tree cover; (4) human demography, country boundaries, distance to roads, distance to cities, and city population sizes (Table 1).

All data were standardized, cleaned, and discretized into grid cells to match the WorldClim data set (Table 1), which comprised the majority of our predictors (~0.9 × 0.9 km cells). This resulted in 331 grid cells comprising the survey data, and 19,778 grid cells containing all GBIF sightings (Table 1). To avoid double-counting, species names were standardized using the Taxonomic Name Resolution Service (TNRS: GCC [Flann 2009], ILDIS [International Legume Database and Information Service 2017], TPL [The Plant List 2013]), following the APGIII classification (iPlant Collaborative 2013). Survey data were converted to presences/absences for each species. The PNFI and Hilje et al. (Table 1) surveyed only tree species, and therefore these sites were excluded from analyses of non-tree species. For distance calculations, we generated a distance raster and Voronoi polygons, using the raster and dismo R packages. We corrected an error in the Global Roads Open Access Dataset (CIESIN and ITOS 2013), which had positioned a road through the Darién Gap, where no road access exists. We excluded species with <10 sightings, as a minimum threshold to build SDMs. This reduced the number of sightings from 430,336 to 404,354 (i.e., a loss of ~6%), allowing analyses of 7,360 species. For species trait data, we removed duplicated rows and values classified as "variable or conflicting reports," used the mean where multiple values existed, and set missing values to zero (corresponding to the normalized mean). All continuous predictor variables were normalized (mean of zero; standard deviation of unity), for comparability.

Table 1. Description of data sources used in the analyses.

Datum	Description	Source
Occurrence information		
Species sightings	records of species presence (1 or no data) across Panama and Costa Rica	Global Biodiversity Information Facility (GBIF 2017)
Species surveys	systematic surveys of species presence/absence (1/0) across Panama and Costa Rica	Panamanian National Forest Inventory (PNFI, Melgarejo et al. 2015), Tropical Ecology Assessment and Monitoring Network (TEAM, http://www.teamnetwork.org/), Alwyn H. Gentry Forest Transect Data Set and Missouri Botanical Gardens (Phillips and Miller 2002), Hilje et al. (2015), BioTreeNet (Cayuela et al. 2012), Center for Tropical Forest Science (CTFS, Condit et al. 2013)
Species variables		
Species life history traits	species life history predictors potentially related to bias: vegetation type (herbs, shrubs, trees, vines, lianas, non-woody epiphytes; deciduous, evergreen), size (plant height, diameter, leaf area, seed mass), and flowering duration	Botanical Information and Ecology Network database (BIEN; Maitner et al. (2018), http://bien.nceas.ucsb.edu/bien/)
Google hits	number of Google search results ("hits") for a species' Latin name	Google (google.ca)
Taxonomic key presence	presence of a species' taxonomic family in one or more online taxonomic keys	Kew Gardens Neotropikey (https://www.kew.org/science/tropamerica/neotropikey_generic.htm), Hansen and Rahn (1969) key (http://www.colby.edu/info.tech/BI211/)
Environmental variables		
Bioclimatic variables	original 19 bioclimatic variables reduced to 9 uncorrelated variables at 30" resolution or ~0.9 × 0.9 km grid cells (Bio1 = annual mean temperature, Bio2 = mean diurnal range, Bio3 = isothermality, bio4 = temperature seasonality, bio12 = annual precipitation, bio13 = precipitation of wettest period, bio15 = precipitation seasonality, bio18 = precipitation of warmest quarter, bio19 = precipitation of coldest quarter)	WorldClim (Fick and Hijmans 2017)
Elevation	estimated from digital elevation model (DEM)	Shuttle Radar Topography Mission (SRTM), Consultative Group for International Agriculture Research Consortium for Spatial Information (CGIAR-CSI, http://srtm.csi.cgiar.org)
Slope	calculated from DEM	As above, plus raster package in R (Jarvis et al. 2008)
Vegetation	Normalized Difference Vegetation Index, NDVI, at 10 km resolution	Moderate Resolution Imaging Spectroradiometer (MODIS) from the NASA Earth Observations (NEO)
Evapotranspiration	mean values across 2000–2013	Moderate Resolution Imaging Spectroradiometer (MODIS) from the NASA Earth Observations (NEO)
Tree cover	mean value in cell	Global Forest Cover database (Hansen et al. 2013)
Human demographic traits		
Country boundary	shape files for Panama and Costa Rica	The Global Administrative Boundaries Dataset (GADM [Global ADMinistrative Areas 2015])
Distance to road	value from centroid of cell	Global Roads Open Access Dataset (CIESIN and ITOS 2013)
Distance to urban center	value from centroid of cell	Global Roads Open Access Dataset (CIESIN and ITOS 2013)
Human population	population size of nearest town (~320 towns across Panama and Costa Rica)	Free World Cities Database (MaxMind 2008)

Models

We analyzed five models. The first two are basic models, presented for comparison. The third model builds upon these by introducing a simple bias adjustment, comparing survey to sightings data. The fourth model combines survey and sightings data (for species where both data exist) by means of a single dummy variable, and the fifth model is a combination of models 3 and 4.

Model 1.—We built SDMs using survey data (henceforth termed the "Survey Only" model), necessarily restricting analyses to 1,480 species found in the surveys. We based our SDM on the Generalized Additive Model (GAM; mcgv package in R, Wood 2011), given its ease of use, flexibility, and wide adoption. We used the generalized cross-validation score (GCV) to select variables, which penalized based on prediction error, as an alternative to stepwise approaches (Marra and Wood 2011). Smoothers were set to a maximum of 5 knots, providing high flexibility while balancing computational efficiency. Other default options within GAM were used. In notation

$$z_{i,j}^{\text{surv}} = b_0 + s(x_{1,i}) + s(x_{2,i}) + \cdots + s(x_{m,i}) \quad (1)$$

$$\hat{y}_{i,j}^{\text{surv}} = \frac{1}{1 + e^{-z_{i,j}^{\text{surv}}}} \quad (2)$$

in R notation

$$\text{gam}(y_{i,j} \sim s(x_{1,i}) + s(x_{2,i}) + \cdots + s(x_{m,i}), \text{family} = \text{binomial}, \text{select} = \text{TRUE}, \text{method} = \text{"GCV.Cp"}) \quad (3)$$

where $y_{i,j}$ was the binary presence/absence response for survey grid cell i, $\hat{y}_{i,j}^{\text{surv}}$ was the expected probability of occurrence, $x_{1\ldots m,i}$ were m continuous spatial predictors, and s were "smoothers" allowing non-linear relations. Additionally, we included an unsmoothed categorical variable for country (Panama vs. Costa Rica). SDMs were fit for each species j, and predict.gam was used to predict probability of occurrence across all grid cells, using fitted model coefficients.

Model 2.—We used sightings to estimate species distributions (henceforth the "Sightings Only" model). As absences were unmeasured, we generated 10,000 random reference (pseudoabsence) sites, which approximated the relative frequencies of different environmental conditions. We then proceeded as for Model 1, except that we could perform SDMs for 7,360 sighted species.

With sightings only, one does not expect all occurrences to be reported, and probability estimates are also affected by the (arbitrary) number of reference sites used in fitting. Thus, we calibrated the output using the survey data and a generalized linear model (GLM) to yield absolute probability estimates:

$$z_{i,j}^{\text{so}} = b_0' + s(x_{1,i}) + s(x_{2,i}) + \cdots + s(x_{m,i}) \quad (4)$$

$$\hat{y}_{i,j}^{\text{so}} = \frac{1}{1 + e^{-z_{i,j}^{\text{so}}}} \quad (5)$$

$$\hat{y}_{i,j}^{\text{soc}} = \frac{1}{1 + e^{-(b_0 + b_1 z_{i,j}^{\text{so}})}} \quad (6)$$

$$\text{glm}(y_{i,j} \sim z_{i,j}^{\text{so}}, \text{family} = \text{"binomial"}) \quad (7)$$

where $z_{i,j}^{\text{so}}$ was the output from GAMs fitted to each species j individually, using sightings and pseudoabsences, $\hat{y}_{i,j}^{\text{so}}$ and $\hat{y}_{i,j}^{\text{soc}}$ were the expected "Sightings Only" probabilities without (so) and with calibration (soc), respectively. The parameters b_0 and b_1 were coefficients in a GLM fitted to all species j in the same model. This corrected for systematic bias across species by statistically rescaling detection rates according to a constant bias relationship. The recalibration could be applied to all 7,360 species, including those without survey information, but did not consider spatial or species-specific factors influencing bias (compare with Model 3).

Model 3.—We derived a simple "bias kernel" to adjust "Sightings Only" predictions (henceforth termed the "Bias adjustment Kernel" or "BaK" model). Biases in reporting or detection may arise due to environmental or species-specific factors. If spatial factors causing bias have consistent effects, we would expect that bias would be higher in some locations across multiple species. Thus, we compared the cumulative discrepancy between observed occurrences in the surveyed cells and the unadjusted "Sightings Only" predictions. This metric of bias was then fit to spatial predictors using a 2nd order polynomial GLM, as the simplest non-monotonic option. More advanced models could be substituted (e.g., GAMs), however, second-order polynomials are more easily interpreted (important given our purpose to identify predictors of bias):

$$B_i = \log\left(\frac{1 + \sum_j y_{i,j}}{1 + \sum_j \hat{y}_{i,j}^{\text{so}}}\right) \quad (8)$$

$$\text{glm}(B_i \sim x_{1,i} + x_{1,i}^2 + x_{2,i} + x_{2,i}^2 + \cdots + x_{m,i} + x_{m,i}^2) \quad (9)$$

where B_i was the bias at surveyed location i, with the numerator being the occurrences observed in the surveys ($y_{i,j}$) summed across all species j, and the denominator being the predicted probability of occurring at location i, summed across all species j, using the unadjusted "sightings only" probabilities ($\hat{y}_{i,j}^{\text{so}}$; Eq. 5), and $x_{1\ldots m}$ were spatial predictors. Unity was added to the numerator and denominator to avoid zero values. Expected levels of bias ($\hat{B}_i$) were then calculated for all locations across Panama, using predict.glm and spatial predictors.

Likewise, some species could be generally more prone to bias. We estimated species-driven bias as

$$B_j = \log\left(\frac{1 + \sum_i y_{i,j}}{1 + \sum_i \hat{y}_{i,j}^{\text{so}}}\right) \quad (10)$$

$$\text{glm}(B_j \sim T_{1,j} + T_{1,j}^2 + T_{2,j} + T_{2,j}^2 + \cdots + T_{s,j} + T_{s,j}^2) \quad (11)$$

where B_j was the bias for a given species j, $T_{1\ldots m}$ were species-specific variables for species j, and $y_{i,j}$ and $\hat{y}^{\text{so}}_{i,j}$ were summed across all surveyed locations i. Again, predict.glm would be used to derive expected bias ($\hat{B}_j$), given species variables.

Finally, to generate occurrence probabilities, both spatial ($\hat{B}_i$) and species ($\hat{B}_j$) biases were combined with the Sightings Only predictions, using GLM.

$$\hat{y}^{\text{BaK}}_{i,j} = \frac{1}{1 + e^{-b_0 + b_1 z^{\text{so}}_{i,j} + b_2 \hat{B}_i + b_3 \hat{B}_j}} \tag{12}$$

$$\text{glm}(y_{i,j} \sim z^{\text{so}}_{i,j} + \hat{B}_i + \hat{B}_j, \text{family} = \text{“binomial”}) \tag{13}$$

where $\hat{y}^{\text{BaK}}_{i,j}$ was the final predicted probability from the BaK model, $z^{\text{so}}_{i,j}$ was the combined linear predictors from the Sightings Only model, and $b_{0\ldots3}$ were fitted parameters. Thus, the calibrated Sightings Only model differed from the BaK model only through spatial and species bias terms, yielding a direct comparison of the benefit of including those bias terms. Both Sightings Only and BaK models could be applied to all 7,360 species.

To fit Eqs. 12, 13, the data were combinations of each species j by each surveyed location i. We note that one could have fit spatial and species variables directly on presence/absences across species and locations (i.e., essentially skipping Eqs. 8–11). While we considered this option (and also analyzed other more sophisticated approaches, e.g., Fithian et al. 2015), the BaK model (Eqs. 8–13) yielded more accurate predictions, and therefore we did not consider these alternatives further. Additionally, calculating the bias terms using Eqs. 8–11 avoids pseudoreplication (each species or location contributes only a single datum to the analyses of bias, Eq. 9, 11), and is computationally light.

Model 4.—We combined survey and sightings data as a species-specific means of accounting for bias (henceforth Surveys and Sightings, S_2 model). We modified Eq. 1, adding a categorical (dummy) variable (d_i) to distinguish survey sites ($d_i = 0$) from locations of sighting (and pseudoabsences) ($d_i = 1$), fitting each species separately. In R notation

$$\begin{aligned}&\text{gam}(y_{i,j} \sim s(x_{1,i}, d_i) + s(x_{2,i}, d_i) + \cdots + s(x_{m,i}, d_i),\\ &\text{family} = \text{binomial}, \text{select} = \text{TRUE}, \text{method}\\ &= \text{“GCV.Cp”})\end{aligned} \tag{14}$$

This could only be applied to the 1,480 species observed in the survey records. To predict occurrences across Panama, we used predict.gam, forcing $d_i = 0$.

Model 5.—We considered a composite model, using the S_2 model for species found in surveys, and the BaK model for all other species (henceforth termed the S_2BaK model).

Analyses

Sightings-only underestimation bias.—While underestimation was expected, given incomplete sightings of species, the magnitude of discrepancy is of interest. We compared the unadjusted sightings only expectations ($\hat{O}$) (the most common approach in the literature) to the actual occurrences of species found in surveys (O), aggregating across species. Mathematically, the expected value from a binomial distribution is simply the number of trials multiplied by probability of occurrence. If probabilities vary, the expected value (i.e., number of occurrences) is the sum of individual probabilities ($p(y|X)$) across all the surveyed species (j) and sites (i):

$$\hat{O} = \sum_j \sum_i p(y_{ij}|X_i) \tag{15}$$

The probabilistic prediction (e.g., from a logistic function, Eq. 1) can be conceptualized as the number of sites containing the species compared to the total number of sites, given environmental conditions (X). However, the probabilities from an SDM would be in comparison to the number of reference sites used for the analysis, whereas the actual bias should be in comparison to all sites. For computational reasons, we had arbitrarily chosen 10,000 reference sites to approximate the relative frequency of each set of environmental conditions X. Thus, to calculate the magnitude of bias, we corrected for the total number of sites (161,161 grid cells in Panama and Costa Rica),

$$N(X) = N_r(X) \times \frac{N}{N_r} \tag{16}$$

$$p(y|X) = \frac{Y(X)}{N(X)} = \frac{Y(X)}{N_r(X)} \frac{N_r}{N} \tag{17}$$

where $p(y|X)$ was the probability of occurrence, given environment X, $Y(X)$ was the number of occurrences, given environment X. N and N_r were the total number of sites and number of sampled reference sites, respectively. $N(X)$ and $N_r(X)$ were the total number of sites and sampled reference sites, respectively, given environment X.

Factors contributing to bias.—We explored the factors contributing to bias using the GLM analysis (Eqs. 8–11). We reported the coefficient magnitude and significance for each term in the model.

Validation of predictive ability.—We withheld 100 randomly chosen survey sites (henceforth termed the “validation set”), and fit all models on the remaining 231 survey sites (henceforth termed the “fitting set”). We examined model predictions in validation sites, focusing on species found in the fitting set (henceforth termed “spatial validation”). Additionally, withholding sites naturally excluded some species in the fitting set. This provided a meaningful test of a model’s ability to predict

new (non-fitted) species occurring in "unsampled" validation sites (henceforth termed "species validation").

We considered two metrics, the area under the curve (AUC; Hanley and McNeil 1982) and the proportion of deviance explained. Area under the curve measured the ability to rank probabilities of occurrence from high to low, and was good for relative comparisons, but did not indicate whether probabilities were accurate on an absolute scale. Deviance measured the accuracy of absolute probabilities (Lawson et al. 2014). For ease of interpretation, we used the proportion of null deviance explained (D):

$$D = 1 - \frac{\sum_i \sum_j \log\left(1 - \left|y_{ij} - \hat{y}^{\mathrm{M}}_{i,j}\right|\right)}{\sum_i \sum_j \log\left(1 - \left|y_{ij} - \hat{y}^{\mathrm{NULL}}\right|\right)} \quad (18)$$

where $y_{i,j}$ was the observed presence/absence at surveyed site i for species j, $\hat{y}^{\mathrm{M}}_{i,j}$ was the predicted probability from a model (M), and $\hat{y}^{\mathrm{NULL}}$ was the mean expectation from a null model with no species or location specific predictors. The numerator and denominator were the log-likelihoods for the predictive model (M) and the null model, respectively. We truncated values of $\hat{y}^{\mathrm{M}}_{i,j}$ outside $0.0001 < \hat{y}^{\mathrm{M}}_{i,j} < 0.9999$, as we reasoned that we could not meaningfully distinguish values very close to zero or one, given the survey data set, yet extreme values could have disproportionately large consequences on likelihoods during validation. We also tested sensitivity across two orders of magnitude ($0.001 < \hat{y}^{\mathrm{M}}_{i,j} < 0.999$ and $0.00001 < \hat{y}^{\mathrm{M}}_{i,j} < 0.99999$). These did not change the conclusions, and therefore were not presented. The entire validation process was repeated 10 times.

Alpha diversity.—We examined predictions for α-diversity in the validation sites, using the mean squared error between predicted and observed local richness, standardized by the variation in observed local richness (r^2_{mse}). The parameter r^2_{mse} is analogous to regression, except that r^2_{mse} measures the mismatch between predicted and observed richness (i.e., intercept = 0 and slope = 1), and thus is always either less than or equal to r^2 values from regression.

We generated estimates of α-diversity by summing the probabilities across species in each location (i.e., stacked SDMs), for those models able to estimate richness beyond the surveyed species (i.e., could consider all 7,360 species), namely the Sightings Only, BaK and S_2BaK models. Additionally, we fit richness directly to spatial predictors (henceforth termed the "Environment Only" model), analogous to using habitat filtering macroecological predictors (Finch et al. 2008). While stacked SDM approaches have been criticized for ignoring species interactions (Pineda and Lobo 2009), this limitation could be outweighed if stacked SDMs incorporated more information than the Environment Only model.

Predicting plant distributions in Panama.—We used the best model, based on the above analyses, to estimate plant distributions across Panama. We refit the relevant models, using the entire survey data set (331 sites). We restricted environmental predictors to the ranges used for fitting. For projecting $z^{\mathrm{so}}_{i,j}$ or $z^{\mathrm{S}_2}_{i,j}$ values, we used the range of values occurring in the GBIF locations across Panama (and Costa Rica), and for the bias kernels, we used the range of values from the survey data. We examined general patterns and also focused on species listed under the Convention on International Trade in Endangered Species (CITES) for Panama (UNEP-WCMC 2017), and indigenous territories (INEC 2010), which have limited accessibility.

Code for S_2BaK is made available; see *Data Availability* statement.

Results

Bias estimates and spatial and trait-based predictors

As expected, SDMs based on unadjusted sightings underestimated probabilities of occurrences. In the surveys, 23,426 occurrences across 1,480 species were observed in 331 sites. In comparison, the unadjusted Sightings Only model predicted 572.61 occurrences for these species in these sites. This corresponded to an underestimation bias of 40.91 times (or ~4,000%).

Biases were correlated with both spatial (Fig. 1a) and species factors (Fig. 1b), explaining 57.4% and 34.3% of the deviance, respectively. For spatial factors, anthropogenic factors had the strongest effects, with distance to the nearest road being the largest in magnitude and most significant predictor of bias. Additionally, bias was substantially greater in Panama compared to Costa Rica (Costa Rica contained two-thirds of the GBIF sightings). Abiotic factors also affected bias. Significance largely occurred for second-order polynomial terms, with predominantly negative coefficients (Fig. 1a). Negative coefficients reflect smaller bias away from average environmental conditions. Given that bias should relate to the number of samples as a proportion of the number of sites, this finding could potentially reflect the smaller number of sites with extreme values. This likewise could account for the reduction of bias at higher elevations, which comprised less area than lower elevations. Additionally, tree cover, precipitation in the wettest quarter and annual precipitation (second-order term) were significantly related to bias (Fig. 1a).

For species variables, growth forms were the most significant predictors of bias, with trees showing higher occurrences for a given level of reporting, relative to other growth forms (Fig. 1b). Interestingly, when deciduous or evergreen statuses were recorded, both were positively related to bias. This was in comparison with species where these data were absent. While other variables had larger coefficients (e.g., dispersal syndrome), they were not nearly as significant. Finally, second order terms for size related variables (height and seed mass) and web-based analytics (number of Google hits for a

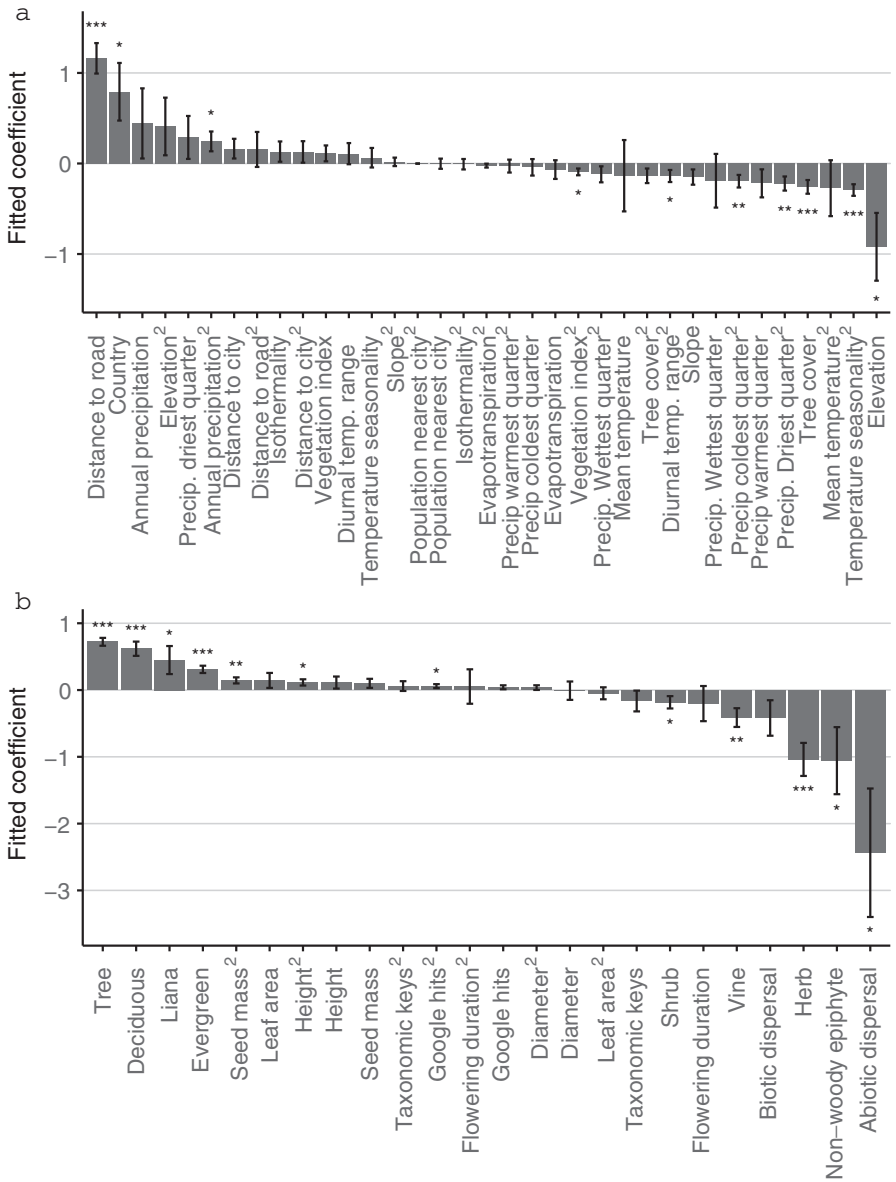

FIG. 1. Factors correlated with bias (Eqs. 8–11), for (a) environmental variables and (b) species variables. Variables were standardized, so that coefficients correspond to the relative magnitude of effect. Error bars represent standard errors. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$).

species name) showed small but significant effects. Unlike spatial factors, these second order coefficients were positive, where extreme values tended to result in larger underestimation bias.

Validation analysis

S_2BaK yielded the most predictive results, with $D = 0.47$ and AUC = 0.96 (Table 2a), improving predictions by 10% deviance explained compared to BaK alone. The Sightings Only SDM, performed considerably worse for absolute probabilities (Table 2a), but performed well for relative rankings (AUC = 0.88, Table 2b), and thus, this status quo approach remains useful in the right context. Nonetheless, using S_2BaK was clearly worthwhile.

For spatial validation, the S_2 model yielded the best results ($D = 0.44$, AUC = 0.88), with BaK performing second best (Table 2). Although Survey Only had the strongest result using the fitting set ($D = 0.81$, AUC = 0.97), it performed poorly for predicting absolute probabilities in validation (Table 2a), but retained good relative rankings (AUC = 0.82, Table 2b). Neither the Survey Only nor the S_2 models could be applied to species not found in the fitting set, and therefore were not considered in the remaining analyses.

TABLE 2. Comparative results from the different models (rows), showing (a) proportion of deviance explained as a measure of fit on an absolute scale, and (b) AUC as a measure of fit on a relative scale.

Model	Validated all	Species validation	Spatial validation	Fitted
(a) Proportion deviance explained (D)				
S_2BaK	0.47 (0.0052)	0.64 (0.005)	0.44 (0.0066)	NA
BaK	0.37 (0.004)	0.64 (0.005)	0.32 (0.0051)	0.38 (0.0017)
Sightings only	0.19 (0.0022)	0.16 (0.0075)	0.2 (0.0024)	0.19 (0.00098)
S_2	NA	NA	0.44 (0.0066)	0.51 (0.0021)
Survey only	NA	NA	−0.57 (0.022)	0.81 (0.0026)
(b) Area under the curve (AUC)				
S_2BaK	0.96 (0.00085)	0.9 (0.0061)	0.88 (0.0016)	NA
BaK	0.94 (0.001)	0.9 (0.0061)	0.82 (0.0018)	0.95 (0.00046)
Sightings only	0.87 (0.0014)	0.85 (0.0086)	0.79 (0.0016)	0.87 (0.00065)
S_2	NA	NA	0.88 (0.0016)	0.91 (0.00053)
Survey only	NA	NA	0.82 (0.0025)	0.97 (0.00051)

Notes: "Validated All" means predicting all species in validation sites, "Spatial Validation" means predicting only species found in the fitting set (but extrapolating to validation sites), "Species Validation" means predicting all species *not found* in the fitting set (i.e., ability to extrapolate to new species in new locations), and "fitted" means species and locations used in fitting set. Sightings only, Survey Only, S_2, BaK and S_2BaK refer to the models described in equations 1 to 14. A total of 100 of 331 randomly chosen survey sites were retained for validation. The process of random selection of sites and validation was repeated 10 times. Values shown represent the mean, with SD in parentheses.

On average, the validation set contained occurrences for 112 (SD = 21) "new" species not found in the fitting set, with 5,880 species absent from both fitting and validation sets. The BaK model consistently outperformed Sighting Only (Table 2a, b), and for the species validation BaK comparatively increased D by approximately fourfold (Table 2a). Thus, our results demonstrated the benefit of including bias adjustment kernels.

Alpha diversity

Stacked SDMs using the BaK model yielded the best predictions (mean $r^2_{mse} = 0.34$ [SD = 0.078]). This was followed by the Environment Only model (mean $r^2_{mse} = 0.28$ [SD = 0.07]). Interestingly, S_2BaK did not perform as well as BaK alone for estimating richness (mean $r^2_{mse} = 0.23$ [SD = 0.14]). Finally, Sightings Only performed the worst (mean $r^2_{mse} = 0.13$ [SD = 0.098]).

Predicting plant distributions in Panama

About 7,360 species across Panama and Costa Rica were used to build the models. Of these, 1,080 species were predicted by S_2BaK to have negligible probabilities in Panama (<0.0001 in any location, and also did not occur in Panama GBIF records), and therefore were not considered further. Thus, in total, we applied our model to 6,280 species across Panama.

While only 22% of these species were detected in surveys, they accounted for 67% of occurrences (based on summed probabilities of occurrence using S_2BaK). Thus, the surveys captured many prevalent species, but missed most of the rarer ones. The Panama Canal watershed, central Panama, had the most GBIF sightings and survey locations (Fig. 2). Our model confirmed high richness in this area (Fig. 3a), with α-diversity ~2× higher than average. However, our analyses suggested that the highest richness actually occurred in eastern Panama (Darién region) and along the Caribbean coast (Fig. 3a), with α-diversity ~2.5× higher than average. The BaK model separated high richness from observation biases, by calibrating sightings against occurrences *within* survey sites (i.e., BaK would be robust to unrepresentative *locational* selection of survey sites), assuming that the observations *within* survey sites were the "gold standard" and reflected species occurrences at the local level.

Only eight CITES species were found in surveys in Panama. However, combined with GBIF sightings, we could analyze 267 species. We note that we could not build distributions for 422 CITES species in Panama. For the 267 analyzable species, they were sighted a total of 2,178 times in Panama in the GBIF records (an average of eight sightings per species). However, S_2BaK predicted a combined total of 31,815 occurrences, suggesting over an order of magnitude greater frequency. These predictions were heavily skewed, with a median of 19.7 occurrences per species (0.016% of the cells), and only eight species occurring in more than 1% of the cells (*Cyathea multiflora*, *Cyathea petiolata*, *Swietenia macrophylla*, *Alsophila cuspidata*, *Cyathea bicrenata*, *Cyathea delgadii*, *Cyathea poeppigii*, and *Cyathea williamsii*). Comparatively, for the 6,280 species analyzed, CITES species occurred 42.4% as frequently as the other species, using GBIF records. Whereas S_2BaK suggested CITES species occurred only 20% as often as other species. Thus, despite the greater absolute numbers of CITES occurrences predicted, they were predicted to be relatively even rarer than apparent from the GBIF data set alone. Spatially, "hot spots" of CITES species were predicted in similar regions as for overall richness, with the highest CITES diversity in the Darién region and the Panama Canal watershed (Fig. 3b).

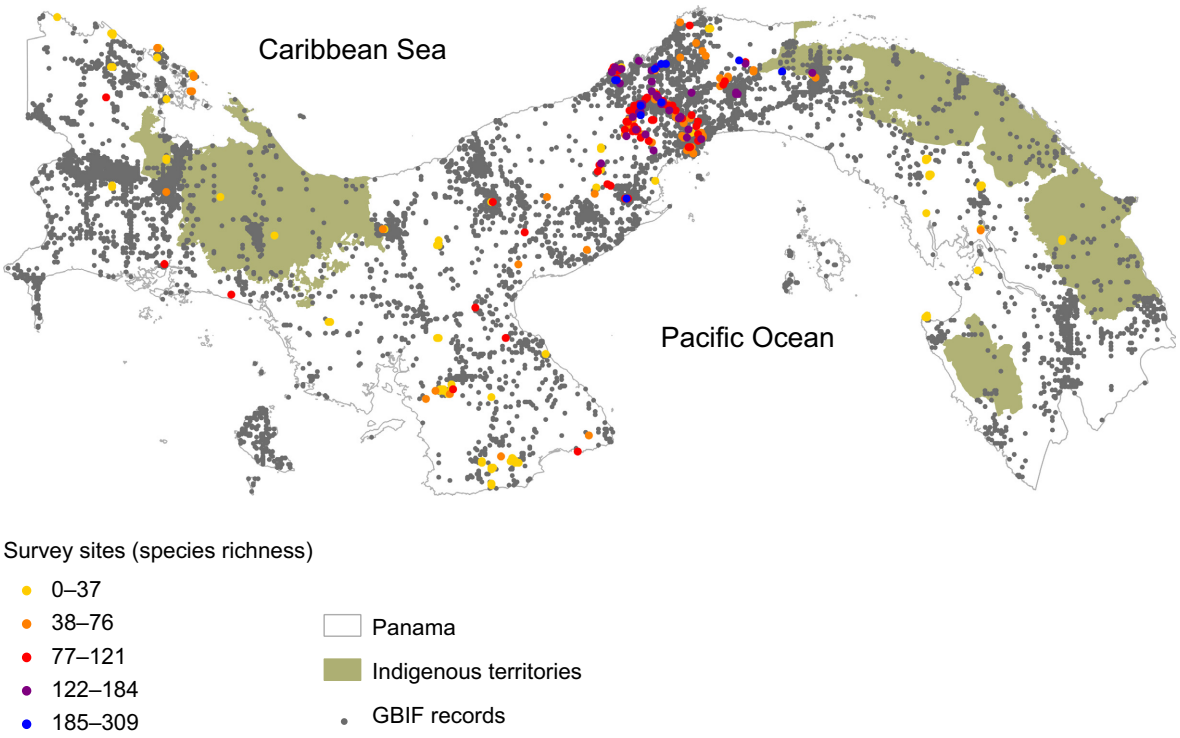

FIG. 2. Map of Panama, including official Indigenous territories (Comarcas), locations of Global Biodiversity Information Facility (GBIF) sightings, and locations and richness found in survey sites.

Finally, within Panama, 22% of the area is classified as indigenous territories, yet contains only 9.6% of GBIF sightings and 2.8% of survey sites (Fig. 2). In surveys, substantially lower α-diversity was found in indigenous territories, with the average richness of 33.4 species/cell vs. 78.1 species/cell outside the territories. While indigenous regions might appear to contain lower biodiversity, our results suggested otherwise. Applying the BaK model to those same survey sites, we found similar discrepancies (44.1 species/indigenous cell vs. 74.8 species/nonindigenous cell), indicating that the differences were largely environmentally driven. However, when BaK was applied to all cells, we found averages of 54.9 species/indigenous cell vs. 60.5 species/nonindigenous cell, respectively, indicating that α-diversity was approximately equal within and outside indigenous territories. Taken together, these results also suggest that the locations surveyed in indigenous territories had unrepresentatively low diversity, but there was a tendency to survey high diversity sites outside indigenous territories.

DISCUSSION

Estimating the spatial distribution of biodiversity is needed to understand what may be lost as anthropogenic pressures accelerate. Yet, it is unfeasible to sample local biodiversity at every location, and while impressive databases have become available, they invariably suffer from biases (Troudet et al. 2017). Particularly, global databases are necessarily coarse, often with imperfect information and uncertainties, and it is questionable whether their integration would improve predictive power. Our results built upon and demonstrated the value of the substantial previous scientific endeavors, integrating systematic surveys, species sightings and trait databases, environmental and demographic databases, as well as web-based analytics. Their combination, through S_2BaK, offers the best current distributional estimates of plant biodiversity in Panama, performing demonstrably better than models using survey or sightings data alone. These 6,280 species distributions provide the building blocks for further ecological analyses, for instance, predicting faunal occurrences (Hudgins et al. 2017), defining connectivity across landscapes (García-Feced et al. 2011), or yielding information on targeted species (e.g., *Anacardium excelsum*, predicted to cover 25% of Panama, and ranked in the top five species for carbon storage in Panama; Melgarejo et al. 2015).

S_2BaK generated excellent predictions, while being conceptually straightforward, computationally efficient, and easily transferrable to alternative SDM approaches. In a nutshell, S_2BaK relied on two relatively simple concepts: (1) S_2 required only a single dummy variable to differentiate sightings from survey data, and (2) BaK estimated bias by numerically comparing survey data to predictions from a sightings-only SDM model, and then used a general linear model (GLM) to combine sightings-only SDM with the bias estimates. As such, one

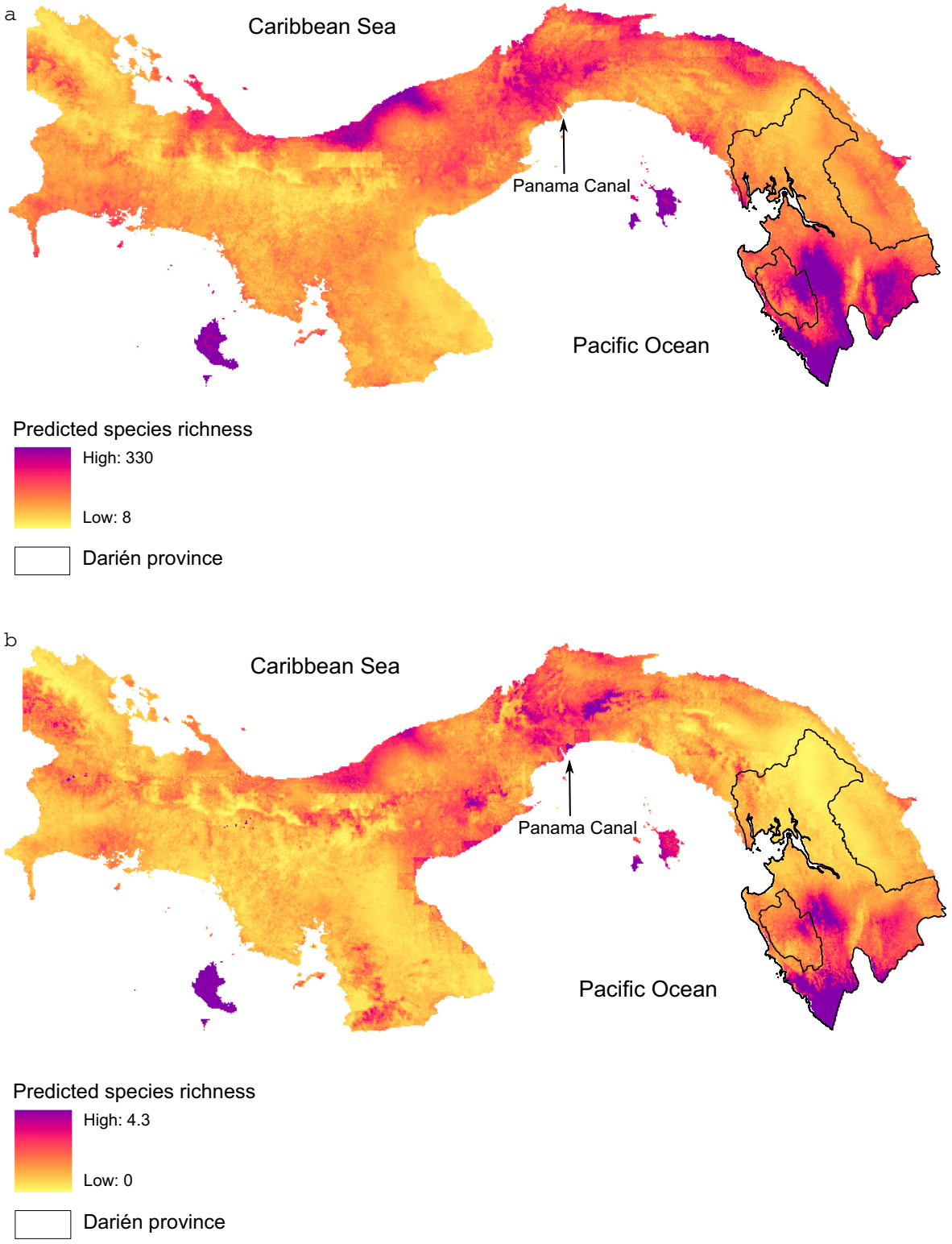

FIG. 3. Modeled α-diversity (richness), for (a) all species and (b) CITES-listed species across Panama.

could conceivably employ any sightings-only SDM in approaches S_2, BaK, and S_2BaK, providing simplicity and high flexibility. Moreover, since SDMs are fit separately for each species, computational complexity increases approximately linearly with the number of species, assuming that fitting SDMs (and not the GLM) is

the rate limiting step. Despite its simplicity, we found that S_2BaK was powerful, and that even for species not observed in the survey data sets, we could discriminate between those species that occurred in each location and those that did not.

While the performance of S_2BaK was exemplary for plant species in Panama (AUC = 0.96, D = 0.47), other approaches conceivably could work better, given different data characteristics. For instance, BaK should theoretically outperform the S_2 model when environmental biases are consistent across species. However, our empirical finding was that, when both survey and sightings data were available, S_2 performed better than BaK, arguably reflecting species idiosyncrasies. Nonetheless, sufficient consistency in biases existed, given that BaK substantially outperformed the Sightings Only model. Likewise, more complex approaches to distribution modeling (e.g., using spatial point processes; Dorazio 2014, Koshkina et al. 2017, and independently derived by Fithian et al. 2015) could be stronger given different data sets, but performed worse in this system: we conducted additional analyses using the multispeciesPP package (Fithian et al. 2015) and found weaker results (AUC = 0.77, D = 0.31) than either S_2 or BaK, based on locational validation of the 1,480 surveyed species (the existing package did not permit extrapolation to unsurveyed species). Nonetheless, for other systems, we suggest exploration of diverse models (e.g., reviewed in Guillera-Arroita 2017), given potential interactions with different data characteristics. However, in the case of plant distributions in Panama, the mixture of S_2 and BaK models was the logical choice.

Our spatial estimates offered baseline information important for applied issues. For instance, our results indicate high potential α-diversity in the Panama Canal watershed, the Darién region (Eastern Panama) and the Caribbean coast (between Colön and Veraguas Provinces). Importantly, the Canal watershed, including areas surrounding Panama City, face the greatest pressure population increases projected in the coming decades (World Bank 2013), and there are ongoing development plans along the Caribbean coast, overlapping the high diversity regions. For conservationists, our analyses of 276 CITES species will be of interest. The "frontier" Darién region in Eastern Panama is a CITES hotspot, but is expected to experience substantial human migration as farming land becomes scarcer (Heckadon-Moreno 2009, IUCN 2017). For government, S_2BaK complements existing efforts, leveraging the national forest inventory of Panama (Melgarejo et al. 2015), to produce the most accurate estimates available for plant distributions, and providing insight into regions with low accessibility. For instance, our results suggest indigenous and nonindigenous territories have similar average α-diversity, despite large differences found in surveys.

We view our distributional estimates as the next step in our understanding of biodiversity, rather than as the final answer. While predictive accuracy may be difficult to greatly improve, given the performance of S_2BaK, there are other worthwhile considerations. First, these are correlational models, and alternative, well-fitting formulations could suggest different driving factors. For instance, many anthropogenic activities were not explicitly considered, but only included via correlations with environmental or other factors (e.g., distance to roads). Thus, Panama City could have lower α-diversity than estimated here, as "built environments" were not included as predictors. Instead, predictions should be interpreted as "biodiversity potential" outside of anthropogenic activities. Second, we recognize that the burgeoning SDM literature offers many different procedures on constructing SDMs (e.g., alternative SDM approaches or pseudoabsence selection). We do not focus on these debates, but instead provide the flexibility to use diverse SDMs, and employed validation using systematic surveys as evidence of good predictive ability. While survey data themselves may be imperfect, they arguably are more robust to issues of bias and represent the most reliable (if limited) data available for validation. However, we acknowledge other analyses could provide insight. For instance, we did not consider biotic interactions. Nonetheless, predictions for individual species can remain strong (as occurred in this system), if species interactions correlate with the environment (Leung and Bradie 2017). In brief, additional analyses could be worthwhile, and should be considered in terms of (1) improved predictive ability or (2) reinterpretation of processes/predictors.

Here, we derived a powerful approach to spatial estimation of biodiversity. We illustrated its utility using current issues in Panama, and demonstrated the value of integrating information from publicly available databases. We make our distributional estimates of 6,280 species available, which will serve as baseline information for fundamental and applied scientific inquiry (http://prism.research.mcgill.ca).

Acknowledgments

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant to B. Leung and scholarships to E. J. Hudgins and A. Potapova. This product includes data created by MaxMind, available from http://www.maxmind.com/. The authors acknowledge Alwyn H. Gentry, the Missouri Botanical Garden, and collectors who assisted Gentry or contributed data for specific sites in the Gentry Forest Transect Data Set. Some data in this publication were provided by the Tropical Ecology Assessment and Monitoring (TEAM) Network, a collaboration between Conservation International, the Missouri Botanical Garden, the Smithsonian Institution, and the Wildlife Conservation Society, and partially funded by these institutions, the Gordon and Betty Moore Foundation, and other donors. The MODIS NDVI data were retrieved from https://neo.sci.gsfc.nasa.gov/, and the evapotranspiration data were retrieved from http://www.ntsg.umt.edu/project/modis/mod16.php, both courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS)

Center, Sioux Falls, South Dakota. The views expressed in this publication are those of the author(s) and do not necessarily reflect the views of FAO.

Literature Cited

Balvanera, P., I. Siddique, L. Dee, A. Paquette, F. Isbell, A. Gonzalez, J. Byrnes, M. I. O'Connor, B. A. Hungate, and J. N. Griffin. 2014. Linking biodiversity and ecosystem services: current uncertainties and the necessary next steps. BioScience 64:49–57.

Beck, J., M. Böller, A. Erhardt, and W. Schwanghart. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics 19:10–15.

Catterall, S., A. R. Cook, G. Marion, A. Butler, and P. E. Hulme. 2012. Accounting for uncertainty in colonisation times: a novel approach to modelling the spatio-temporal dynamics of alien invasions using distribution data. Ecography 35:901–911.

Cayuela, L., et al. 2012. The Tree Biodiversity Network (BIOTREE-NET): prospects for biodiversity research and conservation in the Neotropics. Biodiversity & Ecology 4:211–224.

Center for International Earth Science Information Network & Information Technology Outreach Services. 2013. Global Roads Open Access Data Set, Version 1 (gROADSv1). NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York, USA. https://doi.org/10.7927/h4vd6wct

Central Intelligence Agency. 2016. The world factbook. https://www.cia.gov/library/publications/the-world-factbook/

Condit, R., B. M. Engelbrecht, D. Pino, R. Pérez, and B. L. Turner. 2013. Species distributions in response to individual soil nutrients and seasonal drought across a community of tropical trees. Proceedings of the National Academy of Sciences USA 110:5064–5068.

Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography 23:1472–1484.

Fick, S. E., and R. J. Hijmans. 2017. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37:4302–4315.

Finch, O.-D., T. Blick, and A. Schuldt. 2008. Macroecological patterns of spider species richness across Europe. Biodiversity and Conservation 17:2849–2868.

Fithian, W., J. Elith, T. Hastie, and D. A. Keith. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6:424–438.

Flann, C. 2009. Global compositae checklist. www.compositae.org/checklist

García-Feced, C., S. Saura, and R. Elena-Rosselló. 2011. Improving landscape connectivity in forest districts: a two-stage process for prioritizing agricultural patches for reforestation. Forest Ecology and Management 261:154–161.

Gertzen, E. L., and B. Leung. 2011. Predicting the spread of invasive species in an uncertain world: accommodating multiple vectors and gaps in temporal and spatial data for *Bythotrephes longimanus*. Biological Invasions 13:2433–2444.

Global ADMinistrative Areas. 2015. http://gadm.org/country

Global Biodiversity Information Facility. 2017. Occurrence download. https://www.gbif.org/. https://doi.org/10.15468/dl.28brf5

Gormley, A. M., D. M. Forsyth, P. Griffioen, M. Lindeman, D. S. Ramsey, M. P. Scroggie, and L. Woodford. 2011. Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. Journal of Applied Ecology 48:25–34.

Guillera-Arroita, G. 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography 40:281–295.

Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36.

Hansen, M. C., et al. 2013. High-resolution global maps of 21st-century forest cover change. Science 342:850–853.

Hansen, B., and K. Rahn. 1969. Determination of angiosperm families by means of a punched-card system. Dansk Botanisk Arkiv 26:1–46.

Hastie, T., and W. Fithian. 2013. Inference from presence-only data; the ongoing controversy. Ecography 36:864–867.

Heckadon-Moreno, S. 2009. De Selvas a Potreros: La Colonización Santeña en Panamá: 1850–1980. Exedra Books, Panama City, Panama.

Hilje, B., J. Calvo-Alvarado, C. Jiménez-Rodríguez, and G. A. Sánchez-Azofeifa. 2015. Tree species composition, breeding systems, pollination and dispersal syndromes in three forest successional stages in a tropical dry forest in Mesoamerica. Tropical Conservation Science 8:76–94.

Hudgins, E. J., A. M. Liebhold, B. Leung, and R. Early. 2017. Predicting the spread of all invasive forest pests in the United States. Ecology Letters 20:426–435.

Hurlbert, A. H., and W. Jetz. 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proceedings of the National Academy of Sciences USA 104:13384–13389.

Instituto Nacional de Estadística y Censo. 2010. División Política de la República de Panamá, por provincias y comarcas. Contraloría General de la República, Panamá City, Panamá.

International Legume Database and Information Service. 2017. http://www.ildis.org/

IPCC. 2014. Climate change 2014: synthesis report. Page 151 *in* Core Writing Team, R. K. Pachauri, and L. A. Meyer, editors. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland.

IUCN. 2017. IUCN World Heritage Outlook 2: a conservation assessment of all natural World Heritage sites. IUCN, Gland, Switzerland.

Jarvis, A., H. I. Reuter, A. Nelson, and E. Guevara. 2008. Hole-filled seamless SRTM data Version 4. International Centre for Tropical Agriculture (CIAT). http://srtm.csi.cgiar.org

Koshkina, V., Y. Wang, A. Gordon, R. M. Dorazio, M. White, and L. Stone. 2017. Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. Methods in Ecology and Evolution 8:420–430.

Lahoz-Monfort, J. J., G. Guillera-Arroita, and B. A. Wintle. 2014. Imperfect detection impacts the performance of species distribution models. Global Ecology and Biogeography 23:504–515.

Lawson, C. R., J. A. Hodgson, R. J. Wilson, S. A. Richards, and R. Freckleton. 2014. Prevalence, thresholds and the performance of presence–absence models. Methods in Ecology and Evolution 5:54–64.

Leung, B., and J. Bradie. 2017. Estimating non-indigenous species establishment and their impact on biodiversity, using the relative suitability richness model. Journal of Applied Ecology 54:1978–1988.

Maitner, B. S., et al. 2018. The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. Methods in Ecology and Evolution 9:373–379.

Marra, G., and S. N. Wood. 2011. Practical variable selection for generalized additive models. Computational Statistics & Data Analysis 55:2372–2387.

MaxMind. 2008. Free world cities database. https://www.maxmind.com/en/free-world-cities-database

Melgarejo, C., A. Calderón, and M. C. Ruiz-Jaén. 2015. Inventario nacional forestal y de carbono de Panama—Diseno de la fase piloto 2013-2015 y propuesta para la fase final. Ministerio de Ambiente, Panama City, Panama.

Panama Research and Integrated Sustainability Model. *nd*. McGill University, Montreal, Canada. http://prism.research.mcgill.ca

Phillips, O. L., and J. S. Miller. 2002. Global patterns of plant diversity: Alwyn H. Gentry's Forest Transect Data Set. Missouri Botanical Garden Press, St. Louis, Missouri, USA.

Pineda, E., and J. M. Lobo. 2009. Assessing the accuracy of species distribution models to predict amphibian species richness patterns. Journal of Animal Ecology 78:182–190.

iPlant Collaborative. 2013. The Taxonomic Name Resolution Service. Version 4.0. http://tnrs.iplantcollaborative.org

The Plant List. 2013. Version 1.1. http://www.theplantlist.org/

Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre. 2017. Taxonomic bias in biodiversity data and societal preferences. Scientific Reports 7:9132.

UNEP-WCMC (Comps.). 2017. The checklist of CITES species website. CITES Secretariat, Geneva, Switzerland. Compiled by UNEP-WCMC, Cambridge, UK.

Venter, O., et al. 2014. Targeting global protected area expansion for imperiled biodiversity. PLoS Biology 12:e1001891.

Wang, H. G., R. D. Owen, C. Sánchez-Hernández, and M. L. Romero-Almaraz. 2003. Ecological characterization of bat species distributions in Michoacán, Mexico, using a geographic information system. Global Ecology and Biogeography 12:65–85.

Warton, D. I., I. W. Renner, and D. Ramp. 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. PLoS One 8:e79168.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73:3–36.

World Bank. 2013. World development indicators. http://databank.worldbank.org/data/download/WDI-2013-ebook.pdf

Data Availability

Code and data are available on GitHub: https://doi.org/10.5281/zenodo.2530382