Received: 8 January 2023 | Revised: 2 May 2023 | Accepted: 11 May 2023

DOI: 10.1002/ecs2.4589

ECOSPHERE
AN ESA OPEN ACCESS JOURNAL

ARTICLE

Methods, Tools, and Technologies

# The clustering of spatially associated species unravels patterns in tropical tree species distributions

**Sean E. H. Pang**[1] | **J. W. Ferry Slik**[2] | **Damaris Zurell**[3] | **Edward L. Webb**[1,4,5]

[1]Department of Biological Sciences, National University of Singapore, Singapore, Singapore

[2]Environmental and Life Sciences, Faculty of Science, Universiti Brunei Darussalam, Gadong, Brunei Darussalam

[3]Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany

[4]Viikki Tropical Resources Institute, Department of Forest Sciences, University of Helsinki, Helsinki, Finland

[5]Helsinki Institute of Sustainability Science (HELSUS), University of Helsinki, Helsinki, Finland

**Correspondence**
Sean E. H. Pang
Email: s.pang@u.nus.edu

**Handling Editor:** Charles D. Canham

## Abstract

Complex distribution data can be summarized by grouping species with similar or overlapping distributions to unravel spatial patterns and separate trends (e.g., of habitat loss) among spatially unique groups. However, such classifications are often heuristic, lacking the transparency, objectivity, and data-driven rigor of quantitative methods, which limits their interpretability and utility. Here, we develop and illustrate the clustering of spatially associated species, a methodological framework aimed at statistically classifying species using explicit measures of interspecific spatial association. We investigate several association indices and clustering algorithms and show how these methodological choices drive substantial variations in clustering outcomes and performance. To facilitate robust decision-making, we provide guidance on choosing methods appropriate to one's study objective(s). As a case study, we apply our framework to modeled tree distributions in Borneo and subsequently evaluate the impact of land-cover change on separate species groupings. Based on the modeled distribution of 390 tree species prior to anthropogenic land-cover changes, we identified 11 distinct clusters that unraveled ecologically meaningful patterns in Bornean tree distributions. These clusters then enabled us to quantify trends of habitat loss tied to each of those specific clusters, allowing us to discern particularly vulnerable species clusters and their distributions. This study demonstrates the advantages of adopting quantitatively derived clusters of spatially associated species and elucidates the potential of resultant clusters as a spatially explicit framework for investigating distribution-related questions in ecology, biogeography, and conservation. By adopting our methodological framework and publicly available codes, practitioners can leverage the ever-growing abundance of distribution data to better understand complex spatial patterns among species distributions and the disparate effects of global changes on biodiversity.

KEYWORDS
biogeography, cluster analysis, complex spatial data, habitat loss, land-cover change, multivariate analysis, ordination, R-mode analysis, spatial association, species classification, species distribution modeling, species grouping

# INTRODUCTION

With recent advancements in species distribution modeling (SDM) and the greater availability of biodiversity and environmental data, detailed predictions on species distributions are becoming increasingly accessible (Elith & Leathwick, 2009; GBIF, 2022; Norberg et al., 2019; Wüest et al., 2020). This new wealth of high-resolution data opens avenues for spatial analyses involving large species pools, which can serve to provide greater insights into community ecology and conservation science (Hannah et al., 2020; Pang et al., 2021; Santini et al., 2021; Wüest et al., 2020). However, as the number of species increases, so does the inherent complexity of the biogeographical data. Without a way to decompose species distributions, large stacks of distribution data may instead encumber analyses and obscure patterns or trends in the results (Kreft & Jetz, 2010; Marquet et al., 2004; Villalobos et al., 2013). To better summarize and interpret complex spatial datasets, there is a need for robust methods for classifying species based on their distribution (Jongman et al., 1995; Legendre & Legendre, 2012).

Species with highly similar or overlapping distributions indicate shared environmental requirements, biotic requirements, or dispersal barriers, or direct interactions between species (e.g., mutualism or predation) (Keddy, 1992; Peterson, 2011). Conversely, dissimilar distributions indicate differences in those processes instead (e.g., cold vs. warm temperature requirements or competitive exclusion). Classifying species based on their distribution, therefore, reflects a combination of processes—like the hierarchical filters of community assembly theory—that have led to recurrent patterns of associations or disassociations across geographic space (Calatayud et al., 2020; Keddy, 1992; Shipley & Keddy, 1987). Investigating these patterns and processes is key to understanding biodiversity patterns and species coexistence (Clements, 1936; HilleRisLambers et al., 2012; Roxburgh & Chesson, 1998; Shipley & Keddy, 1987). In other words, by decomposing species into relatively homogeneous subsets, with shared geographic distributions, we can make apparent the spatial structure of species communities and their driving processes.

Classifying species with similar distributions is the "sister analysis" of classifying sites with similar compositions, that is, the R-mode versus Q-mode analysis, respectively (Legendre & Legendre, 2012). Although classifying sites is the more prevalent method in biogeography and spatial ecology (Kreft & Jetz, 2010), classifying species offers an alternative view of spatial patterns. Site classifications are useful when the focus is on understanding compositional relationships among sites, for example, how differences in species composition among areas might reflect historical biogeography and evolution or events of vicariance and geodispersal (Hazzi et al., 2018; Holt et al., 2013; Kreft & Jetz, 2010; Leroy et al., 2019). Conversely, species classifications are useful when the focus is on understanding spatial relationships among species. For instance, consider the impact of deforestation and climate change on species distributions. As spatially heterogeneous threats, their impact varies substantially depending on the species' initial distribution (Bellard et al., 2014; Newbold, 2018; Pang et al., 2021; Trisos et al., 2020). Such variations are difficult to separate and investigate in across-species summaries that only reveal the most prevailing trend, whereas species-specific interpretations are impractical for studies involving hundreds or thousands of species (Marshall et al., 2018; Torres et al., 2014; Velazco et al., 2019). However, by grouping species and conducting group-specific summaries instead, we might uncover differing trends of loss and gain linked to each group's unique distributional pattern and better understand their vulnerability to a given threat (e.g., lowland vs. montane vulnerability to deforestation) (Manchego et al., 2017; Pang et al., 2021; Yanahan & Moore, 2019). The classification of species with similar distributions, therefore, functions as concise summaries of species distribution data, which can provide a spatially explicit framework for investigating distribution-related questions in ecology, biogeography, and conservation.

Despite the potential usefulness of grouping spatially associated species, few studies have adopted replicable quantitative methods for doing so. Instead, studies often adopt a heuristic approach toward grouping species, using classifiers based on a putative understanding or description of species associations (Baatar, 2019; Manchego et al., 2017; Pompe et al., 2010; Yanahan & Moore, 2019). Such qualitative approaches lack the transparency, objectivity, and data-driven rigor of more quantitative methods, which limits their interpretability and utility (Jongman et al., 1995; Kahneman & Tversky, 1972; Legendre & Legendre, 2012; Marquet et al., 2004). In this regard, multivariate methods—based on explicit measures of interspecific spatial association—hold immense potential as a quantitative approach for unraveling patterns in species distributions and delineating species groupings (Jongman et al., 1995; Keil et al., 2021; Legendre & Legendre, 2012; Roxburgh & Chesson, 1998).

Multivariate methods can reduce the inherent complexity of biogeographical data, and their strength lies in their ability to generate statistically derived species groupings, with within- and between-cluster variances that are quantifiable and testable. The reproducibility of such groupings, and therefore transparency, is especially relevant given the prevalent use of SDMs for evaluating species vulnerabilities and informing management decisions (Feng et al., 2019; Guisan et al., 2013; Titeux et al., 2017; Zurell et al., 2020).

The challenge with clustering spatially associated species, however, is that there is no definitive measure of spatial association; multiple interpretations and indices exist (Cramér, 1924; Hubálek, 1982; Keil et al., 2021; Roxburgh & Chesson, 1998). Likewise, there are multiple strategies and techniques for unsupervised clustering but no single universal approach (Erman et al., 2015; Guerra et al., 2012; Jongman et al., 1995; Legendre & Legendre, 2012; Seif, 2018). Furthermore, such a technique has not been applied to detailed distribution data before (i.e., modeled distributions) and a methodological framework for doing so yet exists. Thus, practitioners face a suite of methodological options that can lead to highly varied clustering outcomes but lack clear guidance on how to choose between outcomes or what the implications are. If clustering of spatially associated species is to be taken up more broadly, there is a need to understand how users' methodological choices affect variations in clustering results and their subsequent ecological interpretations.

In this study, we develop and illustrate a methodological framework for the clustering of spatially associated species (CSAS), which aims at helping practitioners navigate the steps involved with forming and applying species clusters to leverage the abundance of distribution data. To further guide user decision-making, we test several association indices and clustering algorithms and investigate resulting variations in clustering outcomes. As a case study, we apply the framework to the modeled distribution of 390 tree species in Borneo. We then demonstrate the use of the resulting clusters to separate trends of habitat loss due to land-cover change and further discuss other applications of spatially associated species clusters.

# METHODS

## Framework

We developed our framework based on those suggested by Kreft and Jetz (2010) and Dufrêne and Legendre (1997) for classifying sites—the "sister analysis" of classifying species. We also incorporated key considerations noted by Legendre and Legendre (2012) and Keil et al. (2021) for quantifying interspecific spatial associations and clustering species based on those associations. Our methodological framework for the CSAS involves six main steps (Figure 1).

1. Define the objective and purpose of the study: This sets the context and premise of the entire analysis and influences how subsequent steps are taken.
2. Obtain distribution data: The data type determines the spatial scale and extent at which associations are quantified (e.g., 50-m$^2$ plot data vs. 10-km$^2$ raster maps) and what further spatial analyses are possible (e.g., raster maps allow visualizations of each cluster's distributional pattern), whereas the species list determines what patterns of associations can be found (e.g., montane patterns if only montane species are included).
3. Select an association index: A relevant index is selected to quantify interspecific spatial associations and produce a pairwise dissimilarity matrix as required for clustering.
4. Clustering analysis: A chosen clustering algorithm is applied to the dissimilarity matrix to cluster species. A stopping rule can be implemented to determine the optimal number of clusters, which often uses information on within- and between-cluster variances.
5. Evaluate clusters: Clustering results can be evaluated quantitatively using a variety of metrics, each measuring a different aspect of clustering performance. Clusters and their underlying dissimilarity matrix can also be visualized using ordination techniques for more qualitative comparisons of data structure and cluster performance.
6. Further geographical, ecological, and conservation analyses: Clusters of spatially associated species can be directly analyzed (e.g., highlighting distinct patterns of spatial distributions) or used as a spatially explicit framework for further analyses (e.g., disentangling trends of habitat loss).

In the following, we elaborate on each of these steps and describe how we implemented them within the context of our case study. In steps (3) and (4), we also included analyses for investigating variations due to the choice of association index and clustering algorithm.

1. Define the objective and purpose of the study

The objective of the study directly shapes how each subsequent step is taken. In selecting clustering strategies, for instance, one may choose between hierarchical or nonhierarchical clustering algorithms. If the only requirement of the clustering analysis is to form a given number of clusters for comparison with existing classifications (e.g., morphological adaptations; Boyce & Wong, 2019), discrete nonhierarchical clustering algorithms like *k*-means may be a good choice. On the other hand, if the focus is on investigating patterns of interspecific spatial associations and their relatedness to other interspecific relationships (e.g., phylogenetic or functional dissimilarity; Rüger et al., 2020; Villalobos et al., 2017), a hierarchical clustering algorithm is likely more appropriate. In our case study of Bornean tree species, we are interested in uncovering spatial relationships and identifying discrete clusters of similarly distributed species for further spatial analyses (i.e., habitat loss). Therefore, we focus on clustering

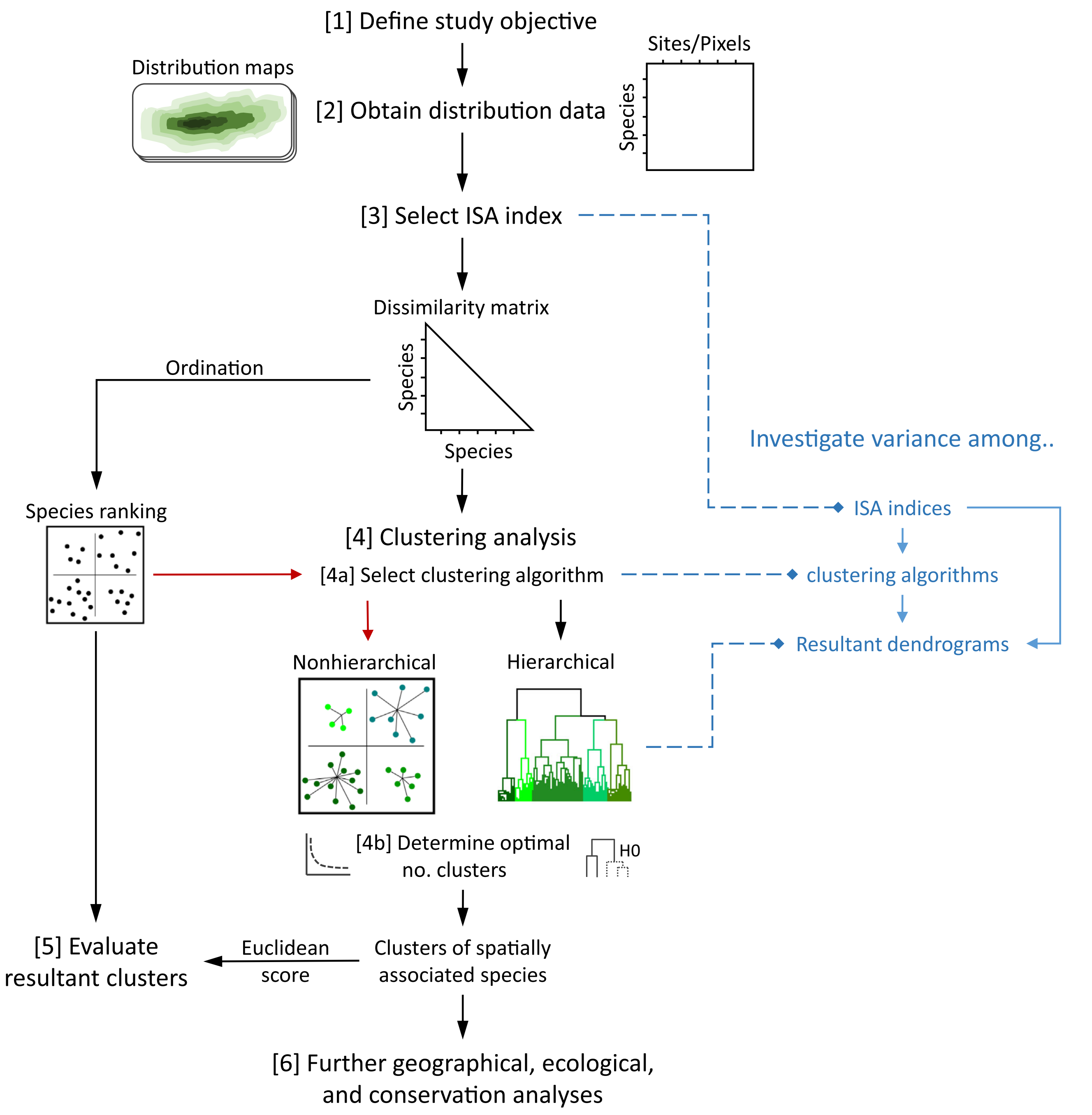
[1] Define study objective
Distribution maps
[2] Obtain distribution data
Sites/Pixels
Species
[3] Select ISA index
Dissimilarity matrix
Ordination
Species
Species
Investigate variance among..
Species ranking
ISA indices
[4] Clustering analysis
[4a] Select clustering algorithm
clustering algorithms
Resultant dendrograms
Nonhierarchical
Hierarchical
[4b] Determine optimal no. clusters
H0
[5] Evaluate resultant clusters
Euclidean score
Clusters of spatially associated species
[6] Further geographical, ecological, and conservation analyses

**FIGURE 1** The methodological framework for the clustering of spatially associated species (CSAS) that consists of six main steps. The analysis of variance at different steps are included in blue. The red lines indicate the alternate route of applying nonhierarchical clustering algorithms, where the dissimilarity matrix is first ordinated. ISA, interspecific spatial association.

algorithms that produce discrete hierarchical groupings of species (Kaufman & Rousseeuw, 2005).

2. Obtain distribution data

Species distribution data are the primary data required. The scale and extent of species distribution data, and the target species involved, all determine what can be inferred from the spatial analyses (Allen & Hoekstra, 1990; Hurlbert & Jetz, 2007; Kreft & Jetz, 2010; Owen-Smith et al., 2015). However, more fundamentally, the type of distribution data itself affects how we interpret resultant clusters. There are three general types of distribution data: sampled, extent-of-occurrence, and modeled.

I. Sampled presence or abundance community data (e.g., plot data) give empirical information on species distributions. Their advantage is that they indicate directly observed patterns of associations across varying spatial and temporal scales (Ledo, 2015; Owen-Smith et al., 2015). However, sampled data are often spatially biased and provide an incomplete representation of species distributions and their associations (Boakes et al., 2010; Stolar & Nielsen, 2015).

II. Extent-of-occurrence maps describe the maximum geographical extent of species and are typically hand-drawn by experts using occurrence records, a heuristic understanding of species' habitat requirements, or both (Gaston, 1996; Lomolino et al., 2006). Extent-of-occurrence maps thus represent scale-dependent abstractions of species' ranges (Gaston, 2003; Hurlbert & Jetz, 2007). Although extent-of-occurrence maps are highly qualitative and overestimate fine-scale occurrences (Graham & Hijmans, 2006; Jetz et al., 2008), they are useful for broadscale analyses where such data problems are less consequential (e.g., Kreft & Jetz, 2010; Trisos et al., 2020).

III. Modeled data typically represent statistical inferences of species distributions derived from modeling occurrence records against prevailing environmental conditions (Peterson & Soberón, 2012; Soberon & Peterson, 2005) and are the data type of interest for this study. Modeled data are advantageous in that they can be used to examine, measure, and predict changes in species distributions across space and time and are frequently used to support biodiversity assessments and conservation prioritizations (Guisan et al., 2013; Guisan & Thuiller, 2005; Peterson, 2011). One inevitable limitation of this method is that the processes modeled to produce the data dictate what can be inferred. For instance, SDMs based on climate predictors alone cannot reflect variations in species distributions due to varying soil requirements (Corlett & Tomlinson, 2020) and cannot be used to explicitly infer biotic interactions (i.e., modeled associations only reflect shared environmental requirements) (Blanchet et al., 2020; Peterson et al., 2020); however, this can potentially be addressed if data on such interactions are incorporated (Ovaskainen et al., 2017; Tikhonov et al., 2017).

Our case study is of a regional scale, encompassing the island of Borneo, and aims to measure associations among tree species' natural distributions (i.e., distributions before anthropogenic disturbances). These distributions will be used to identify clusters of similarly distributed species and evaluate habitat loss due to land-cover change for each cluster. Modeled distributions were therefore the most appropriate choice as they allow fine-scale, statistical, and empirically based estimates of species distribution before anthropogenic disturbances like deforestation.

As our study focuses primarily on the methods for clustering spatially associated species, we present only a summary of the SDM methods (for full details, see Appendix S1: Sections S1–S4):

a. We obtained 19 bioclimatic (30 arcsec; Karger et al., 2017), 5 soil-water (30 arcsec; Trabucco & Zomer, 2010, 2018), and 9 soil property (250 m$^2$; Hengl et al., 2017) variables, resampled to 30 arcsec (~1 km$^2$) and reduced using a principal components analysis (PCA) to their first 10 principal component (PC) axes (87% cumulative variance) via the "stats" and "raster" packages in R (Hijmans & Etten, 2012; R Core Team, 2013) (Appendix S1: Section S2 and Table S1). A land-cover map (300 m$^2$) from the year 1992 (earliest available) was obtained from the European Space Agency (ESA), reclassified as forested and nonforested following Intergovernmental Panel on Climate Change (IPCC) land categories and resampled to 30 arcsec (ESA, 2017) (Appendix S1: Table S1). A binary land-cover categorization was adopted to provide a conservative approach to identify intact habitats; nonforested pixels were considered unsuitable. This was a reasonable assumption given our focus on tree species and that deforestation by definition entails clearing the land of all or most trees.

b. Occurrence data of vascular (Tracheophyta) species in Borneo were obtained from the Global Biodiversity Information Facility (GBIF, 2019) (Appendix S1: Section S3). Spelling errors and synonyms were corrected using the "Taxonstand" package in R (Cayuela et al., 2012; R Core Team, 2013), and occurrences with low accuracy or precision were removed (Gueta & Carmel, 2016) (Appendix S1: Section S3). Tree species were identified using the GlobalTreeSearch database (Beech et al., 2017). To prevent mismatches between occurrences and contemporary land-cover data and account for anthropogenic niche truncations (Faurby & Araújo, 2018; Milanesi et al., 2020; Pang et al., 2022), occurrences within forested and nonforested areas were separated following recommendations in Pang et al. (2022). While occurrences within forested areas were used for model training and cross-validation, occurrences within nonforested areas were used exclusively to validate historical distributions (i.e., projections onto a manually calibrated zero anthropogenic disturbance land-cover scenario). Occurrences were then thinned using a 10 km buffer (Aiello-Lammens et al., 2015), separately for each split. Species with fewer than 10 or 5 occurrences within forested or nonforested areas, respectively, were excluded.

c. All species were individually modeled and tuned using the MaxEnt (3.4.1) algorithm via the "ENMeval" (2.0.0) package in R (Kass et al., 2021; Phillips et al., 2006, 2017; R Core Team, 2013). To contrast occurrence data, 10,000 background points were sampled from pixels of ≥10-km distance from species' occurrence points, where sampling probability was derived from a kernel density estimate representing the geographical sampling bias (Kramer-Schadt et al., 2013; VanDerWal et al., 2009; Vollering et al., 2019). For model tuning, 50 candidate models based on five combinations of feature classes and 10 regularization multipliers were considered (Boria et al., 2017; Morales et al., 2017), each trained using occurrences within forested areas and the 10 environmental PC axes and reclassified land-cover map (year 1992) as predictors (Appendix S1: Section S4 and Figure S1). Candidate models were evaluated using a nested checkerboard cross-validation technique for species with >25 occurrences (else, 10-fold cross-validation) and had their projections of historical distributions validated using occurrences within nonforested areas (testing for niche truncation; Pang et al., 2022). The best performing candidate model was determined using a combination of area under the curve (AUC), true skill statistics (TSS), and omission rates (OR) but accepted only if AUC > 0.7, TSS > 0.4, and OR < 0.2 (for cross-validation and previously excluded occurrences) (Appendix S1: Section S4). Model projections of historical distributions as continuous estimates of habitat suitability (or probability of occurrence) were converted into binary range maps using the maximizing the sum of sensitivity and specificity threshold (Liu et al., 2016).

3. Select interspecific spatial association index

Clustering analyses require a distance/dissimilarity matrix. Site clustering—the "sister analysis"—requires a site-by-site matrix of beta diversity (compositional dissimilarity between sites) (Kreft & Jetz, 2010), whereas species clustering here requires a species-by-species matrix containing interspecific measures of distributional dissimilarity. To calculate the required dissimilarity matrix, an appropriate association index must first be selected. From a theoretical or conceptual standpoint, the choice of association index is arguably the most influential step, as the index used reflects the study's mathematical and ecological definition of association (Box 1), and by extension, the clusters they inform (Hubálek, 1982; Keil et al., 2021; Legendre & Legendre, 2012).

Keil et al. (2021) recently evaluated several association indices and found substantial variation in their performance and sensitivity, which provides vital information for selecting well-performing indices. However, within the context of clustering, we also need to consider how different indices affect the clustering outcome and distributional patterns identified. Moreover, it is unclear whether indices perform differently when using modeled distributions; Keil et al. (2021) used community matrices and simulated positions of individuals in a bound space. Thus, we

**BOX 1 Significance of the choice of association index: the double-zero problem.**

A simple example of how association indices can differ, relevant specifically to binary indices, is in their treatment of co-absences: this is known as the double-zero problem (Legendre & Legendre, 2012). It is often difficult to sensibly define which sites of co-absence provide univocal or useful information for quantifying associations (Legendre & Legendre, 2012). In extreme cases, the answer is clear, for example, co-absences across Europe are not meaningful when measuring associations among Bornean species. But in many cases, the answer is less obvious. Binary indices are based on four quantities: the number of sites uniquely occupied by species 1 ($b$) or species 2 ($c$), and the number of sites occupied by both ($a$) or neither ($d$) species, where the total number of sites is $n = a + b + c + d$. As an example of the double-zero problem, compare the Jaccard (1901) index of association ($a/(a + b + c)$) against the Sokal and Michener (1958) matching coefficient ($(a + d)/n$). As co-absences ($d$) and the number of sites ($n$) by extension become very large, the Jaccard index remains unchanged while matching coefficients approach a value of one. On the other hand, without co-absences ($d$), an index essentially ignores differences in species prevalence. Consider a map of 1000 pixels, two widespread species occupying 500 pixels each that co-occur in 250, and two range-restricted species occupying 50 pixels each that co-occur in 25. If we adopt the Jaccard index, associations between the two widespread and two range-restricted species are identical (widespread: 250/(250 + 250 + 250) = 0.33; range-restricted: 25/(25 + 25 + 25) = 0.33). However, it is generally considered "harder" for range-restricted than widespread species to co-occur, which the matching coefficient reflects (widespread: (250 + 250)/1000 = 0.5; range-restricted: (25 + 925)/1000 = 0.95) (Legendre & Legendre, 2012).

selected and tested 12 indices from Keil et al. (2021)—six binary and six continuous—each representing various aspects of interspecific spatial association (Table 1). Broadly, the selected binary indices can be divided into two families: Jaccard index (jacc), Dice–Sorensen index (dice), and Alroy coefficient (alroy), which exclude co-absences; and matching coefficient (match), tetrachoric correlation (tetra), and scaled *C*-score (scalec), which include co-absences. Correspondingly, continuous indices also vary in their mathematical properties and the dissimilarities they measure: difference-based indices, Bray–Curtis (bray) and Ruzicka (ruz), which essentially measure the cumulative differences in occurrence probabilities across pixels; distance-based indices, Hellinger (hell) and chi-squared (chi), which normalize probabilities by their sum before calculating differences; and correlation-based indices, Pearson (pears) and Spearman (spear), which examine the scaled covariances in probabilities (occurrence probabilities because of our use of modeled distribution data). Clustering analyses require non-negative distance or dissimilarity values, and so we transformed indices violating this requirement, for example, [(1 − pears)/2] (see Table 1).

These indices were applied to the modeled historical distribution of our 390 tree species, which resulted in 12 dissimilarity matrices of dimension 390 × 390, each corresponding to one index. While continuous indices used continuous distribution maps, binary indices used their binary equivalents. To examine variations in interspecific spatial association due to the choice of index, we calculated the Pearson correlation of association values among the 12 dissimilarity matrices. The correlation matrix was then passed into a PCA and visualized using a variable plot.

## 4. Clustering analysis

Cluster analysis falls under the family of unsupervised learning methods in exploratory data analysis, and its primary aim is to classify similar objects while identifying boundaries between groups (Kaufman & Rousseeuw, 2005); the object as sites in the case of clustering sites—the "sister analysis"—but species in our case of clustering spatially associated species. Before clustering, one needs to justify why discontinuities might exist or explain a practical need to divide a continuous set of objects into groups (Legendre & Legendre, 2012). Justifying discontinuities among species distributions is complex. The community-unit model by Clements (1936)

**TABLE 1** The binary and continuous indices of interspecific spatial associations tested for this study.

| Binary indices | Acronym | Formula (as dissimilarity) | Brief notes |
|---|---|---|---|
| Jaccard index | jacc | $\frac{b+c}{a+b+c}$ | Proportional overlap; excludes *d* |
| Dice–Sorensen index | dice | $\frac{b+c}{2a+b+c}$ | Proportional overlap; excludes *d* |
| Alroy coefficient | alroy | $\frac{3bc}{2(a+b)(a+c)+2a\sqrt{z}+bc}$ | Variant of Forbes coefficient of association; excludes *d* |
| Matching coefficient | match | $\frac{b+c}{n}$ | Proportional overlap; includes *d* through *n* |
| Pearson tetrachoric correlation | tetra | $1-\frac{ad-bc}{[(a+b)(c+d)(a+c)(b+d)]^{0.5}}$ | Correlational; includes *d* |
| Scaled *C*-score | scalec | $\frac{bc}{n(n-1)/2}$ | Includes *d* through *n* |
| **Continuous indices** | **Acronym** | **Formula (as dissimilarity)** | **Brief notes** |
| Bray–Curtis dissimilarity (percentage difference) | bray | $\frac{\sum_{i=1}^{n}\lvert x_i-y_i\rvert}{x_+ + y_+}$ | Proportional difference |
| Ruzicka dissimilarity | ruz | $\frac{2\,C_{\text{bray}}}{1+C_{\text{bray}}}$ | Proportional difference |
| Hellinger distance | hell | $\sqrt{\sum_{i=1}^{n}\left(\sqrt{\frac{x_i}{x_+}}-\sqrt{\frac{y_i}{y_+}}\right)^2}$ | Distance metric; Euclidean distance after Hellinger transformation |
| Chi-squared distance | chi | $\sqrt{(x_+ + y_+)\sum_{i=1}^{n}\frac{1}{x_i+y_i}\left(\frac{x_i}{x_+}-\frac{y_i}{y_+}\right)^2}$ | Distance metric |
| Pearson correlation (scaled covariance) | pears | $\left(1-\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y(n-1)}\right)\Big/2$ | Correlational; parametric |
| Spearman correlation (Rho) | spear | Pears between rank values of *x* and *y* as dissimilarity | Correlational; nonparametric |

*Note*: Binary indices are based on four quantities: the number of sites uniquely occupied by species 1 ($b$) or species 2 ($c$), and the number of sites occupied by both ($a$) or neither ($d$) species, where the total number of sites is $n = a + b + c + d$ and the total number of occupied sites is $z = a + b + c$. Continuous indices are based on the vectors of continuous distribution data (occurrence probability or habitat suitability) of two species as represented by $x$ and $y$, their means as $\bar{x}$ and $\bar{y}$, their sums as $x_+$ and $y_+$, and their standard deviations as $\sigma_x$ and $\sigma_y$, where $x_i$ and $y_i$ are their values at site $i$ and $n$ equals the total number of sites.

states that communities result from nonoverlapping groups of species–response curves along an environmental gradient, which supports discrete groupings of species distributions. In contrast, the continuum model of Whittaker (1951, 1953) and Curtis (1959) contends that communities vary gradually along complex environmental gradients and that no distinct groupings exist. Other studies suggest that neither of these two views is correct, or some amalgamation of them (Roberts, 1987; Shipley & Keddy, 1987; Westman, 1985), or that it depends on the scale (Allen & Hoekstra, 1990; Collins et al., 1993; Hoekstra et al., 1991). However, there is evidence supporting recurrent patterns of associations along environmental gradients for Bornean flora (Raes et al., 2009; Slik et al., 2003, 2009), which clustering analyses may serve to uncover. More practically, clustering species distributions provides a spatially explicit framework for investigating distribution-related questions and applications in ecology, biogeography, and conservation. For this study, cluster-specific summaries of habitat loss would offer greater insights into trends of biodiversity loss in Borneo.

### (4a) Select clustering algorithm

Two main families of clustering algorithms exist: nonhierarchical and hierarchical (Jain et al., 1999; Kaufman & Rousseeuw, 2005). Nonhierarchical algorithms partition the data into a predetermined number of clusters (*k*). Algorithms from this family include *k*-means and partitioning around medoids (Kaufman & Rousseeuw, 2005). However, nonhierarchical algorithms are limited because they require the user to specify the number of clusters and they do not yield relationships among clusters (Legendre & Legendre, 2012). Thus, we do not consider nonhierarchical algorithms further. By contrast, hierarchical algorithms construct a hierarchy of clusters, where a predetermined number of clusters is not required and relationships among clusters are depicted through a dendrogram. Hierarchical relationships are especially relevant for spatially associated ecological communities (Clements, 1936; Collins et al., 1993; Keddy, 1992).

Hierarchical algorithms fall into two main categories: agglomerative and divisive (Jain et al., 1999; Kaufman & Rousseeuw, 2005). Although divisive algorithms are typically more efficient and accurate, they are also more complex and harder to interpret, whereas agglomerative algorithms rely on simpler merging steps and are also among the most popular (Erman et al., 2015; Rajalingam & Ranjini, 2011; Roux, 2018; Singh & Singh, 2012). Thus, we focus on seven easy-to-implement and frequently used agglomerative clustering algorithms (Table 2): unweighted pair-group method using arithmetic averages (UPGMA), weighted pair-group method using arithmetic averages (WPGMA), complete linkage (CL), single linkage (SL), unweighted pair-group method using centroids (UPGMC), weighted pair-group method using centroids (WPGMC),

**TABLE 2** The seven hierarchical agglomerative clustering algorithms tested for this study.

| Clustering algorithm (common synonyms) | Acronym | Type | Brief description of distance between clusters |
|---|---|---|---|
| Unweighted pair-group method using arithmetic averages (average linkage) | UPGMA | Proximity | Distance between clusters equal the mean of all distances between objects of each cluster |
| Weighted pair-group method using arithmetic averages (McQuitty's method) | WPGMA | Proximity | Distance between clusters equal the weighted mean of all distances between objects of each cluster, where the subclusters of the most recently merged cluster have equal influence on that distance |
| Complete linkage (furthest neighbor) | CL | Proximity | Distance between clusters equal the maximum distance between objects of each cluster |
| Single linkage (nearest neighbor) | SL | Proximity | Distance between clusters equal the minimum distance between objects of each cluster |
| Unweighted pair-group method using centroids (centroid linkage) | UPGMC | Geometric | Distance between clusters equal the Euclidean distance between their geometric centroids |
| Weighted pair-group method using centroids (median linkage) | WPGMC | Geometric | Distance between clusters equal the Euclidean distance between their weighted centroids, where the subclusters of the most recently merged cluster have equal influence on its centroid |
| Ward's method (minimum increase in sum-of-squares) | WARD | Geometric | Distance between clusters equal the magnitude by which the sum-of-squares in their joint cluster is greater than their combined sum-of-squares |

and Ward's method (WARD). Clustering algorithms were applied to each dissimilarity matrix using the "linkage" function from the mdendro package in R, generating 84 candidate clustering outcomes (12 dissimilarity indices by seven clustering algorithms) and their dendrograms (Fernández & Gómez, 2020; R Core Team, 2013).

To investigate variations among the 84 candidate dendrograms, we measured variations in dendrogram structure using Baker's gamma. Baker's gamma essentially compares the relative position of nodes between dendrograms as the Spearman correlation in lowest common branches, where the lowest common branch is the highest possible number of clusters for which two species belong to the same cluster (i.e., merging node in relation to other nodes) (Baker, 1974). The correlation values between dendrograms were then passed into a PCA and visualized using variable plots. We also compared dendrograms using co-phenetic correlation, common nodes, and Full-text index in Minute space (FM-index) (Appendix S1: Figures S5–S12). However, we focused on Baker's gamma as it assessed the general structure of each dendrogram and is unaffected by the height of each branch, because branch heights may be distorted for dendrograms resulting from space-dilating clustering algorithms like WARD (Fernández & Gómez, 2020).

### (4b) Determine optimal number of clusters

Determining the optimal number of clusters is an age-old challenge of clustering analysis (Chouikhi et al., 2015; Milligan & Cooper, 1985). While taxonomists often seek naturally formed clusters with small distances between member objects and large distances between objects from different clusters, ecologists seek to understand a world that often exists along a continuum and must usually contend with somewhat arbitrary clusters (Gauch & Whittaker, 1981; Legendre & Legendre, 2012). Clusters may be arbitrary when objects are evenly spread through dissimilarity space. In such cases, groupings are partially imposed by the clustering algorithm and are less intrinsic to the data, and a range of optimal number of clusters may be more appropriate than a single value (Gauch & Whittaker, 1981). This does not imply, however, that a stopping rule or an internal validation criterion for determining cluster boundaries is unnecessary. On the contrary, a more rigorous selection procedure is required to ensure transparency, objectivity, and replicability in determining and justifying the (range of) optimal number of clusters (Guerra et al., 2012; Legendre & Legendre, 2012).

Quantitative inspections of diagnostic graphs (i.e., an evaluation metric plotted against the number of clusters) offer a rigorous and data-driven procedure to determine a meaningful and useful number of clusters (Milligan & Cooper, 1985; Salvador & Chan, 2004; e.g., Kreft & Jetz, 2010). We assessed the diagnostic graph of three evaluation metrics: merging height, within-cluster variance, and between-cluster variance. Plotting these metrics against the number of clusters produced a scree-like evaluation plot (Appendix S1: Figure S2). The L-method of Salvador and Chan (2004) was used to quantitatively identify the knee or elbow in these evaluation plots, that is, the maximum curvature of the graph (for details, see Salvador & Chan, 2004). Although other advanced stopping rules certainly exist, many of them are difficult to implement for ecological data where clusters are largely arbitrary and filled with outliers and noise (Chouikhi et al., 2015; Guerra et al., 2012; Legendre & Legendre, 2012; Milligan & Cooper, 1985). The Caliński and Harabasz (1974) index, Duda and Hart (1973) ratio criteria, Hubert and Levin (1976) *C*-index, and Rousseeuw (1987) silhouettes identified optimal number of clusters that were less meaningful and useful (see Appendix S1: Table S3). Hence, to complement inspections of diagnostic graphs, we developed and employed a bifurcation paired *t*-test stopping rule. Moving down the dendrogram, we tested for a significant decrease in within-cluster variance using a paired *t*-test at each passing bifurcation (i.e., cluster partitioning). The bifurcation for which no significant decrease was observed determined the stopping point for partitioning the data and therefore the optimal number of clusters (for details, see Appendix S1: Figure S3). For stopping rules involving within- and between-cluster variance, the centers used to calculate variances were either the aggregated (centroid; mean for continuous data and mode for binary data) or indicator species distribution of each cluster (medoid; the object with the lowest sum of within-cluster dissimilarities).

### 5. Evaluate resultant clusters

We propose a set of quantitative and qualitative evaluations for assessing clustering performance. First, we quantitatively assessed the dendrogram of each candidate clustering outcome using the following three dendrogram performance metrics:

1. Co-phenetic correlation measures the faithfulness of the co-phenetic distances (dendrogram branch heights) to the original dissimilarity matrix (Sokal & Rohlf, 1962).
2. Agglomerative coefficient measures the strength of resulting clusters (Rousseeuw, 1986).
3. Tree balance measures the equality in the number of objects between clusters at each merger or partition (Fernández & Gómez, 2020).

A min–max scaling was then applied to each metric and combined using the Euclidean formula, where the Euclidean score quantifies the dendrogram's performance as a distance from an origin representing the worst performance possible. We used the Euclidean formula because of its flexibility, where metrics can be added, removed, or weighted, depending on the clustering characteristic(s) deemed most relevant to the study objective.

While the raw score of each metric covered specific aspects of a dendrogram's performance, the combined score provides an overview of its performance, aiding the selection of the most appropriate dendrogram. Although internal validation criteria could also provide useful information on cluster performances, we focused on dendrogram performance metrics because they evaluated the performance of the entire clustering result rather than a defined set of clusters (Fernández & Gómez, 2020; Legendre & Legendre, 2012; Milligan & Cooper, 1985). Dendrogram metrics were therefore consistent measures of clustering performance, indifferent to the number of clusters selected. This was vital because the relationship between clusters across the hierarchy was an important facet of the clustering result that needed to be assessed and because exact cluster boundaries are less crucial when dealing with arbitrary clusters (Gauch & Whittaker, 1981; Legendre & Legendre, 2012).

Second, we performed nonmetric multidimensional scaling (NMDS) on each dissimilarity matrix to visualize species distributions in ordination space. Ordination is a widely used tool for projecting multivariate data into low-dimensional space, where (in our case) species are arranged along reduced axes of geographic distributions (Legendre & Legendre, 2012). NMDS is regarded as the most robust unconstrained method and most effective at reducing complex data (Legendre & Legendre, 2012; Minchin, 1987). Additionally, NMDS requires no underlying assumption about linearity or normality, in that any distance or dissimilarity matrix can be used (Ludwig et al., 1988). Ordination via NMDS, therefore, represents a useful approach for visualizing distributional dissimilarities and investigating the spatial structure of community data. Paired alongside their respective dendrograms, NMDS ordinations also provide insight into the formation of cluster boundaries, thereby aiding interpretations of cluster memberships and hierarchical relationships. We performed the NMDS using the "metaMDS" function from the vegan package in R, with 100 random starts based on a fixed initial seed (Minchin, 1987; Oksanen et al., 2007; R Core Team, 2013).

6. Further geographical, ecological, and conservation analyses

The representative distribution of each cluster was generated by summing the binary data of member species at each pixel and dividing values by the total number of member species (i.e., the proportion of member species present). The distribution can be used to visually determine whether clusters are ecologically meaningful and justifiable (Di Febbraro et al., 2018; Mainali et al., 2020; Peterson, 2011), to identify consistencies among clustering outcomes, or as a spatially explicit framework for further spatial analyses (Currie, 2019; Keddy, 1992; Roxburgh & Chesson, 1998). Binary distributions were used because variable sampling biases, species prevalence or rarity, and assumptions of occurrence probability across species meant that continuous distributions are debatably noncomparable and cannot be combined (Elith et al., 2011; Elith & Leathwick, 2009; Merow et al., 2013; Phillips et al., 2006). Moreover, many applications of SDM require binary outputs, and binary distributions are easier to interpret than their continuous counterpart (Fithian & Hastie, 2013; Guisan & Thuiller, 2005; Liu et al., 2016; Royle et al., 2012).

As a demonstration of the application of distribution-based species clusters, representative distributions from the final clustering outcome were used to assess cluster-specific habitat loss due to land-cover change. Annual, 300-m$^2$, land-cover maps for the years 1992–2020 (ESA, 2017) were similarly reclassified to forested and nonforested and resampled to 30 arcsec as in step (2) and overlayed onto each representative distribution; habitat loss occurred when a pixel transitioned from forest to nonforest. For each cluster, habitat availability was quantified as the sum of representative distribution values (i.e., the proportion of member species present) within forested pixels, such that pixels with higher proportion values were weighted higher and nonforested/deforested pixels were valued at zero. We then calculated and presented habitat loss for each cluster in three ways: (1) the percentage of historically available habitat lost by 1992, lost between 1992 and 2020, and remaining in the year 2020; (2) annual percentages of 1992 habitats remaining from 1992 to 2020; and (3) annual rates of habitat loss from 1993 to 2020 as a percentage of available habitats in the year before. Historical baselines assumed all pixels were forested and reforested pixels were not considered.

# RESULTS

Of the 743 species with sufficient occurrence data, we accepted the SDM of 390. Accepted models had an average AUC of 0.76, TSS of 0.52, and OR of 0.10 and 0.04 for cross-validated occurrences and excluded occurrences (i.e., within nonforested pixels), respectively.

## Variance among association indices

We found association indices to generally capture one of three aspects of interspecific spatial association, as reflected by the three distinct groupings in our PCA of association values (Figure 2a). The first group consisted of indices that measure association as differences (bray and ruz) or distances (hell and chi) in site values (i.e., occurrence probabilities), which also formed the tightest and most distinct group. The second group consisted of binary indices that exclude co-absences (jacc, dice, and alroy). The last group consisted of two continuous indices that measure association as the correlation in site values (pears and spear) and three binary indices that include co-absences.

We also observed a comparatively higher correlation between pairs of indices with related mathematical properties, even within already tightly formed groups (Figure 2b). For example, within the first group, the two difference-based indices were highly correlated, as were the two distance-based indices. Combined with the tight grouping of continuous and binary correlation-based indices (pears, spear, and tetra) (Figure 2a,b), our results indicate the underlying mathematical property of the index, or its interpretation of association, as the main factor driving differences in measurements of association.

## Variance among clustering algorithms and dendrograms

Although it was difficult to separate variances in dendrogram outcomes because of the choice of clustering algorithm rather than association index, we observed some general trends. The most striking trend was the presence of reversals, or the upward branching of nodes, among dendrograms resulting from clustering algorithms UPGMC and WPGMC (Appendix S1: Figure S4). Reversals greatly hindered the interpretation of hierarchical relationships and delineation of discrete clusters, often also resulting in statistically incomprehensible dendrogram structures (Abe et al., 2017; Miyamoto, 2012; Wedley et al., 1993). Hence, we rejected clustering outcomes resulting from the UPGMC and WPGMC algorithms, regardless of their dendrogram performance (for variable plots with UPGMC and WPGMC, see Appendix S1: Figure S5).

Among the remaining five clustering algorithms, variances in dendrogram structure depended on the underlying association index (for variable plots of all 12 association indices, see Appendix S1: Figure S5). Dendrograms based on Bray–Curtis dissimilarity were generally less varied, as seen through closely grouped vectors (small between arrow angles) and high loading scores (long arrow lengths) in the PCA variable plot (Figure 3a). In contrast, dendrograms based on Jaccard index were more varied; their vectors were more dispersed (Figure 3b). Although vectors representing dendrograms based on Spearman correlation were also grouped, their loading scores were lower than those based on Bray–Curtis dissimilarity, which indicated weaker correlations and higher variances in dendrogram structure (Figure 3c; for full correlation plots, see Appendix S1: Figure S9).

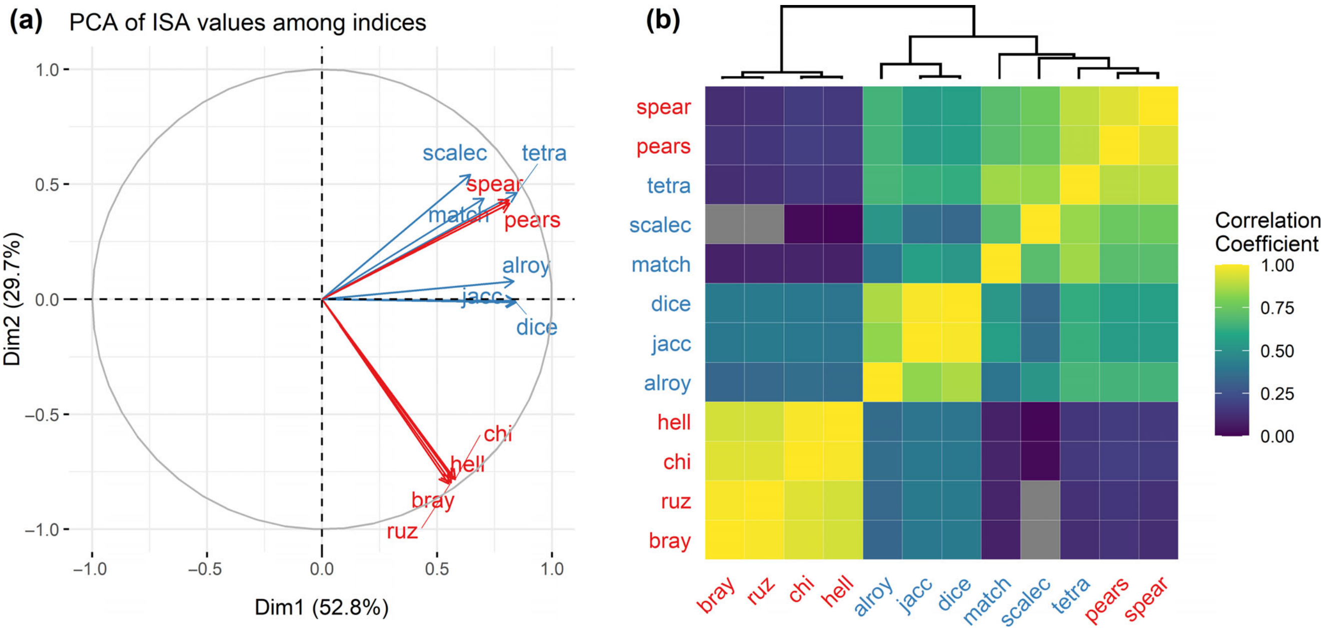

FIGURE 2 Comparison of interspecific spatial association (ISA) values among 12 indices, which were subjected to a principal components analysis (PCA). (a) A variable plot showing the first two PC axes and loadings of the 12 indices. (b) A correlation plot with a dendrogram showing the topological relationships between indices. For simplicity and ease of visualization, correlation values were colored from 0 to 1 only. Continuous indices were in red and binary indices were in blue.

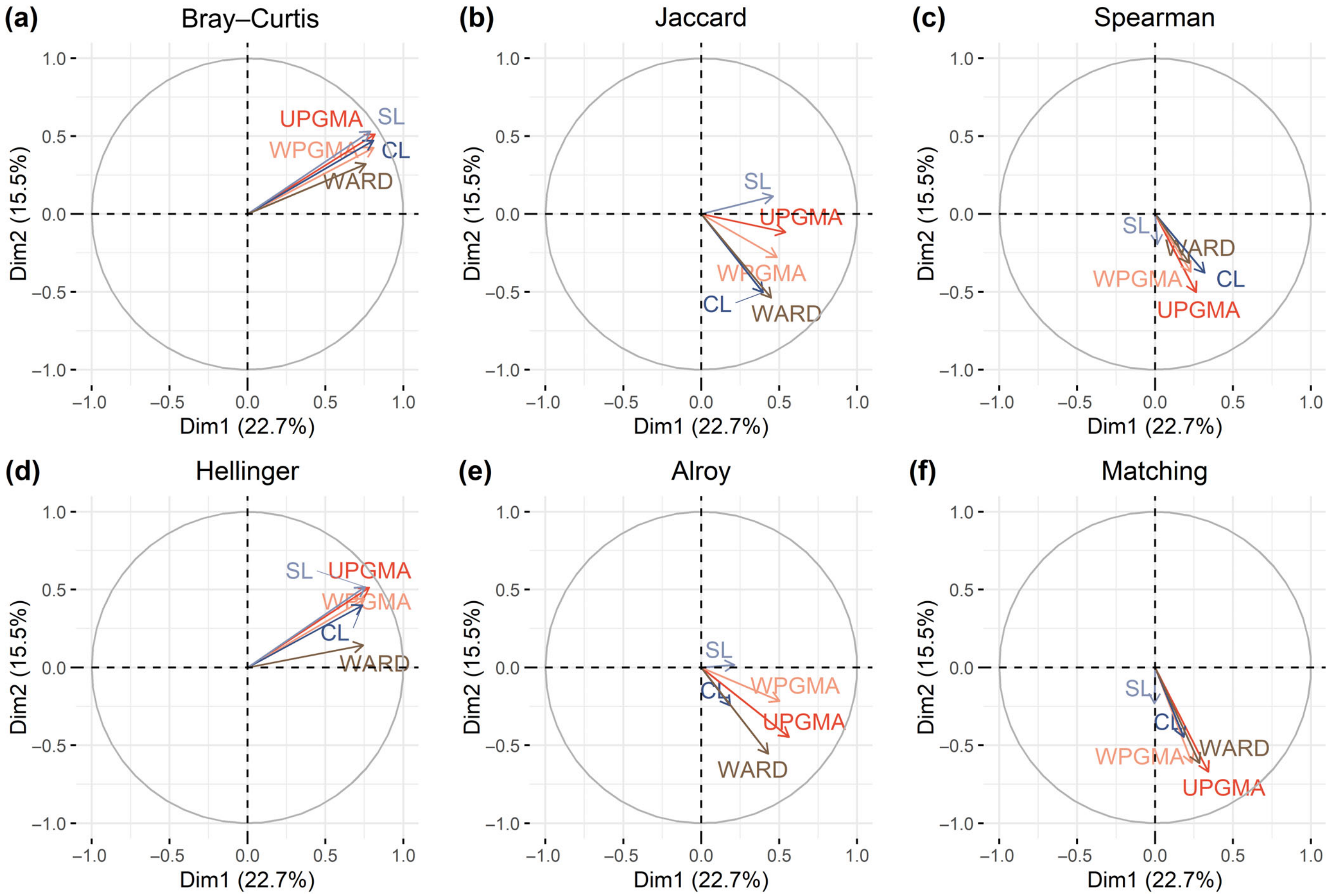

**FIGURE 3** Comparison of dendrograms through a principal components analysis (PCA, applied to all 84 candidate dendrograms). For clarity, results were separated and plotted for a subset of indices: (a) Bray–Curtis, (b) Jaccard, (c) Spearman, (d) Hellinger, (e) Alroy, and (f) matching. Each variable plot shows the first two PC axes and loadings of dendrograms resulting from five clustering algorithms (excluding unweighted pair-group method using centroid and weighted pair-group method using centroid because they led to reversals), which were based on a particular association index. Note, PC axes and loadings across panels were comparable as they were obtained from the same PCA. CL, complete linkage; SL, single linkage; UPGMA, unweighted pair-group method using arithmetic average; WARD, Ward's method; WPGMA, weighted pair-group method using arithmetic average.

Dendrogram outcomes indicated groups of association indices for which the resulting dissimilarity matrices exhibited low or high sensitivity to the choice of clustering algorithm. For example, variations among dendrograms based on Jaccard index and Alroy coefficient were moderately high, even among dendrograms resulting from the same clustering algorithm (Figure 3b,e). By comparison, dendrograms based on Bray–Curtis dissimilarity and Hellinger distance were generally less varied (Figure 3a,c). Dendrograms based on Spearman correlation and matching coefficient were also quite similar to each other (Figure 3c,f). Note that because we compared dendrograms using Baker's gamma, variance here pertained specifically to differences in dendrogram structure as defined by the relative positioning of their nodes (for comparisons using co-phenetic correlation, common nodes, or FM-index, see Appendix S1: Figures S6–S8 and S10–S12).

## Evaluations of dendrogram performance

Dendrogram performance varied substantially across association indices and clustering algorithms (Figure 4; for performances of all 84 candidate dendrograms, see Appendix S1: Figure S13). We first examined performances among clustering algorithms. Dendrograms most faithful to the original dissimilarity matrix (i.e., co-phenetic correlation) were generally those resulting from UPGMA, while the least faithful resulted from CL, SL, and WARD (Figure 4a). Cluster strength (i.e., agglomerative coefficient) and balance (i.e., tree balance) were highest for WARD and second highest for CL, but lowest for SL. Hence, Euclidean scores were generally higher among dendrograms resulting from WARD or CL because they performed better on two out of the three metrics (Figure 4b). Next, among dendrograms based on different association indices, dendrograms based on difference- and distance-based indices (bray and

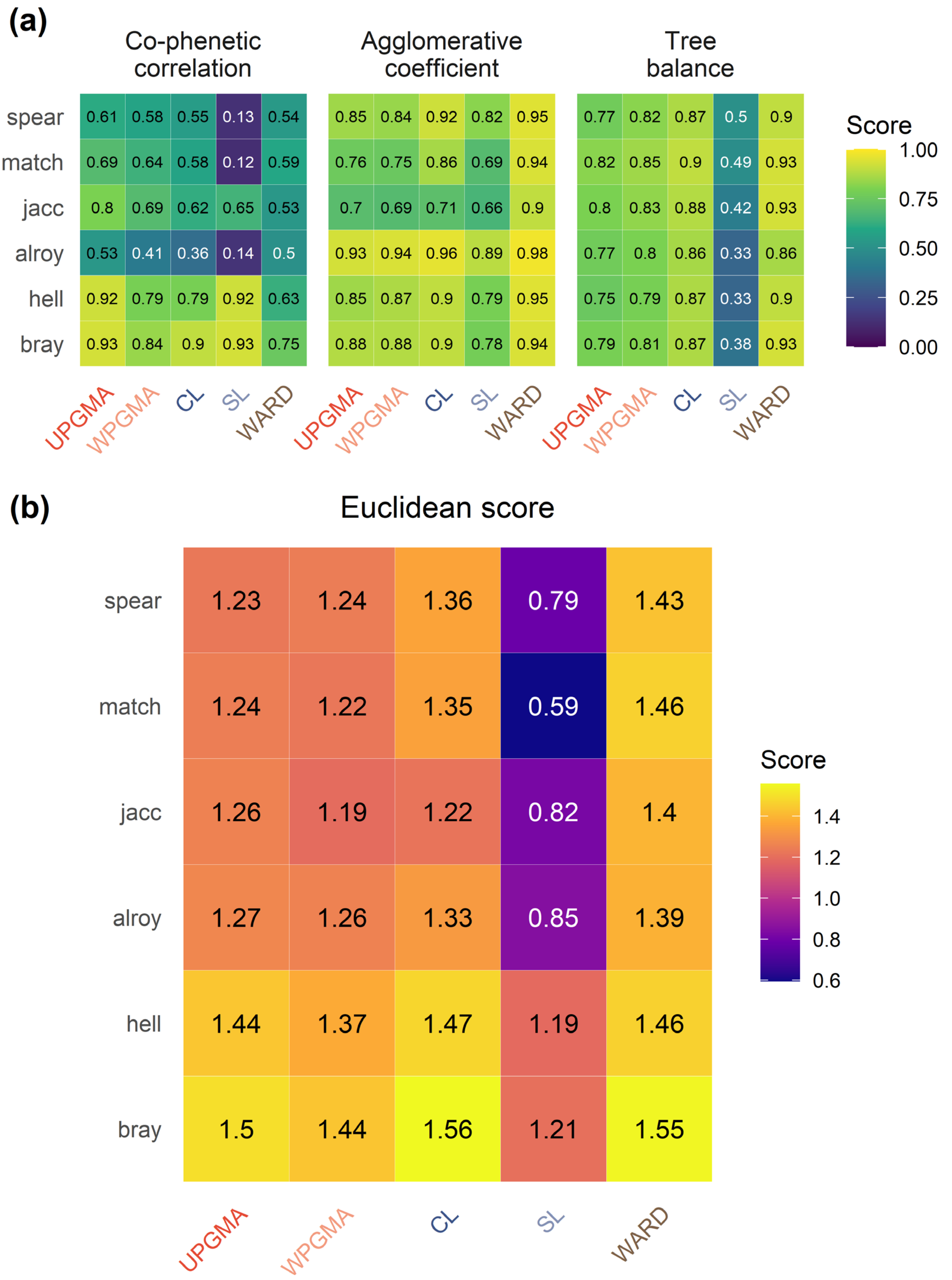

**FIGURE 4** The performance of dendrograms resulting from five clustering algorithms (excluding unweighted pair-group method using centroid and weighted pair-group method using centroid because they led to reversals) for a subset of association indices. (a) Dendrogram performance based on three metrics: co-phenetic correlation, agglomerative coefficient, and tree balance. (b) Dendrogram overall performance as the Euclidean distance (score) across the three metrics after applying a min–max scaling. CL, complete linkage; SL, single linkage; UPGMA, unweighted pair-group method using arithmetic average; WARD, Ward's method; WPGMA, weighted pair-group method using arithmetic average.

hell) had co-phenetic correlation scores that were higher (Figure 4a). Agglomerative coefficient scores were typically lower for dendrograms based on binary indices, except Alroy coefficient (Figure 4a) and scaled *C*-score (Appendix S1: Figure S13). However, we did not observe any clear differences in tree balance scores among indices. Hence, Euclidean scores were generally higher among dendrograms based on difference- and distance-based indices (bray and hell) because they performed much better in terms of co-phenetic correlation (Figure 4b). Overall, dendrograms with the highest Euclidean score were those based on Bray–Curtis dissimilarity and clustered using either the UPGMA, CL, or WARD algorithm (Figure 4b).

## Evaluations of NMDS plots and dendrograms

The NMDS stress levels were lowest for difference-based indices (bray = 0.103; Figure 5a–c), low for distance-based indices (hell = 0.157; Figure 5d,e), but extremely high for correlation-based indices (spear = 0.279; Figure 5f,g). Stress levels were also extremely high for binary indices, except those of proportional overlap (match = 0.23 and jacc = 0.227; Figure 5h–k). Higher stress levels suggest those indices had captured spatial relationships that were too complex to accurately represent in low-dimensional space. Thus, separations among clusters in highly stressed NMDS space may not be visually apparent, while visually overlapping clusters may be artifacts of imperfectly reduced axes.

The NMDS plots revealed distinct differences in the data structure that depended on the choice of association index. Difference- and distance-based indices resulted in points that formed a central aggregate with a scattering of outliers; the former's outliers were more unidirectional (Figure 5a–c), and the latter's more bidirectional (Figure 5d,e). Comparatively, other indices resulted in points that were spread more evenly (Figure 5f–i). To a lesser extent, binary indices that exclude co-absences also resulted in a central aggregate of points (jacc, dice, and alroy) (Figure 5j,k).

Cluster memberships showed how the spread of points affected clustering outcomes under different clustering algorithms. Clustering algorithms sensitive to outliers, such as UPGMA, and CL to some extent, tended to classify the central aggregation of points as one large cluster and outliers as multiple smaller clusters. We observed this for difference- and distance-based indices (Figure 5a,b,d) and binary indices that exclude co-absences (Figure 5j). Although accompanying dendrograms indicated that the central cluster could be partitioned at higher values of *k* (number of clusters), it would also result in the excessive partitioning of outlier clusters and higher instances of one-species clusters. Comparatively, WARD was less sensitive to outliers in general, wherein the central aggregation of points was partitioned while outliers formed moderately sized clusters (Figure 5c,e,k). As a result, WARD produced cluster memberships that were more balanced and thus more meaningful for subsequent spatial analyses. Problems with unbalanced cluster memberships were less pertinent when the association index used resulted in evenly spread points (Figure 5f–i).

## Final clusters of spatially associated species

As the final clustering outcome, we selected clusters based on Bray–Curtis dissimilarity resulting from the WARD clustering algorithm, which had the second-best Euclidean score (Figure 4b). We selected the second best rather than best clustering outcome because it produced a more balanced set of clusters and was thus more meaningful for subsequent analyses (Figure 5a,c). Moreover, its Euclidean score was only 0.01 lower than the best score (Figure 4b).

The optimal number of clusters *k* varied across stopping rules (Appendix S1: Table S2). Among diagnostic graphs, the L-method identified $k = 4$ for merging height and $k = 6$ for within-cluster variance, regardless of the cluster center used to quantify within-cluster variance. For between-cluster variance, the L-method identified $k = 5$ when aggregated distributions (centroids) were used as cluster centers and $k = 4$ when indicator distributions (medoids) were used instead. However, all diagnostic graphs showed a relatively smooth curvature (Appendix S1: Figure S2), suggesting the optimal number of clusters *k* to be above 6 as the L-method tends to underestimate *k* in such cases (Salvador & Chan, 2004). The bifurcation paired *t*-test identified $k = 34$ and 11, for cluster centers using aggregated and indicator distributions, respectively. The bifurcation paired *t*-test tended to identify a large *k* when aggregated distributions were used, particularly for clusters resulting from WARD, but identified *k* closer to the other three stopping rules when indicator distributions were used instead (Appendix S1: Table S2). This was likely because medoids typically represent image-type datasets (i.e., raster maps) better than centroids and are less sensitive to outliers that might inflate changes in within-cluster variance (Kaufman & Rousseeuw, 2005; Van der Laan et al., 2003). Hence, we set *k* as 11, as identified by the bifurcation paired *t*-test when using indicator distributions (medoids). Although only one value of *k* was selected, we acknowledged that a range of possible *k* values exists and explored other probable values of *k* in Appendix S1: Figures S14–S16 (Gauch & Whittaker, 1981).

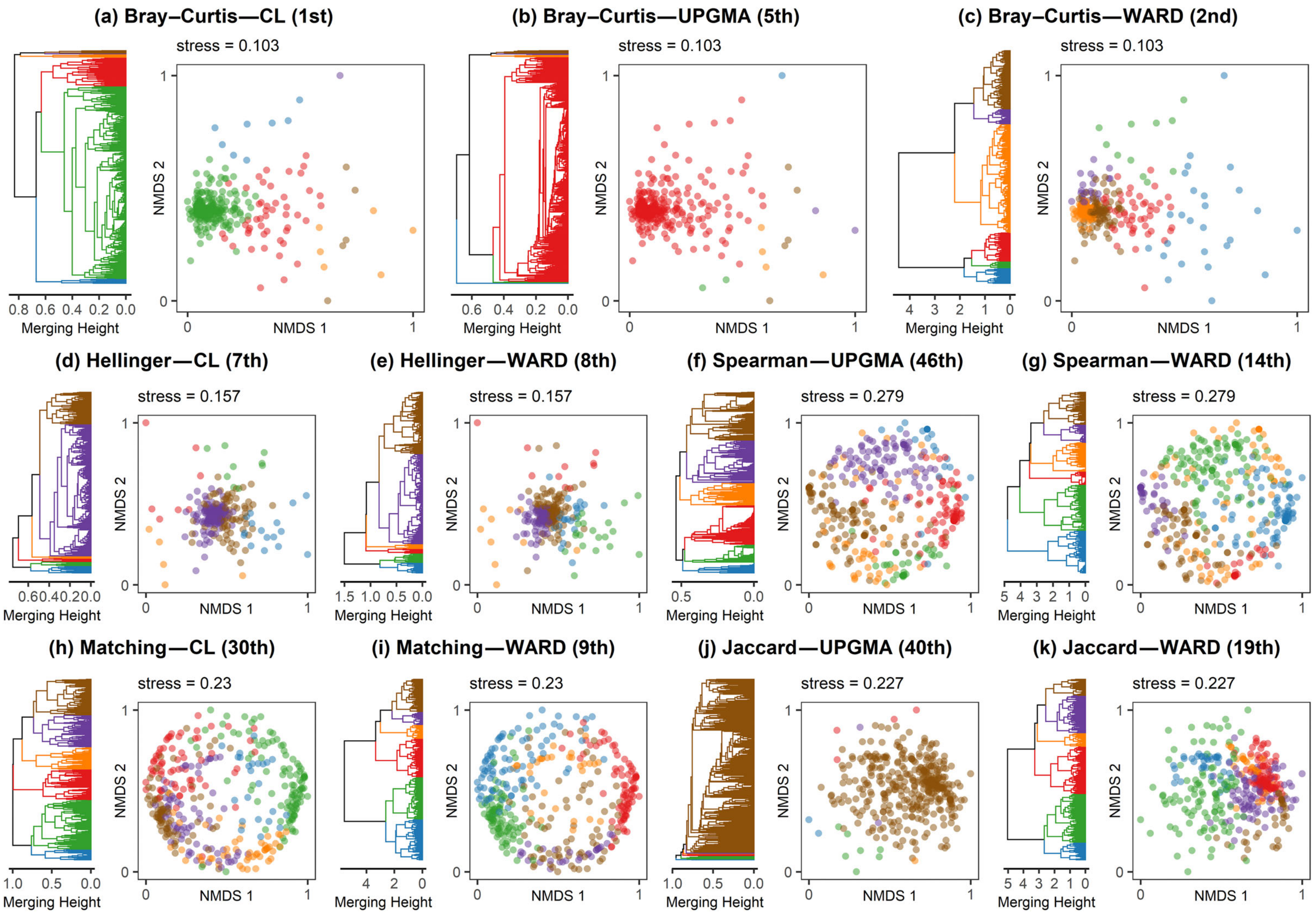

**FIGURE 5** The dendrograms of 11 clustering outcomes and their underlying dissimilarity matrix visualized through a nonmetric multidimensional scaling (NMDS) plot. Stress values are indicated above the NMDS plot. Each leaf of the dendrogram and each dot in the scatter plot represent a species' distribution. For comparisons across clustering outcomes, the number of clusters was fixed at $k = 6$ and differentiated using colors. Colors were neither comparable nor relevant between panels. Panel titles indicate the association index and clustering algorithm used and its overall rank, that is, association index—clustering algorithm (overall rank). Dendrograms and dissimilarity matrices were based on association indices Bray–Curtis dissimilarity (a, b, c), Hellinger distance (d, e), Spearman correlation (f, g), matching coefficient (h, i), and Jaccard index (j, k), resulting from clustering algorithms complete linkage (CL; a, d, f, h), unweighted pair-group method using arithmetic average (UPGMA; b, j), and Ward's method (WARD; c, e, g, i, k).

## Clusters of tree species distributions in Borneo

Resulting clusters and their representative distributions (i.e., the proportion of member species present) yielded spatially meaningful patterns of tree species distributions in Borneo (Figure 6a) (for aggregate and indicator species distributions, i.e., cluster centroids and medoids; see Appendix S1: Figures S17 and S18). Representative distributions delineated, to some extent, the geographical unit to which member species were endemic, and gradients indicate site suitability for supporting member species. The environmental conditions underlying each representative distribution were also extracted to characterize their habitats (Appendix S1: Figure S19). Interestingly, many of the representative distributions here were also observed in other well-performing dendrograms (Appendix S1: Figure S20), even when their dendrogram structure or cluster memberships differed greatly. This suggests that despite relatively varied clustering outcomes, well-performing dendrograms tended to identify clusters reflecting similar spatial patterns.

The first split among species distributions was between clusters 1–4 and clusters 5–11 and occurred early (i.e., high merging height in Figure 6b), indicating high dissimilarity between partitions. This first split separated clusters with highly range-restricted distributions (clusters 1–4) from the rest (clusters 5–11) (Figure 6a,c). Cluster 3 was distributed predominantly across Borneo's western coastal/peatland regions, and clusters 1, 2, and 4 were restricted to separate parts of Borneo's central montane region. Many peatland species are known, and were found here, to occur in montane habitats (e.g., *Litsea accedens* and *Timonius flavescens*; Slik, 2009), which may explain the grouping of cluster 3 with clusters 1, 2, and 4 (but see *Discussion*: *Challenges of clustering spatially associated species*).

The second split was between clusters 5–7 and clusters 8–11 (Figure 6b,c). Among clusters 8–11, cluster 8 was the most distinct since that cluster split off relatively early in the dendrogram and was distributed mainly across the western lowlands, like cluster 3, but more inland than coastal (Figure 6a). In comparison,

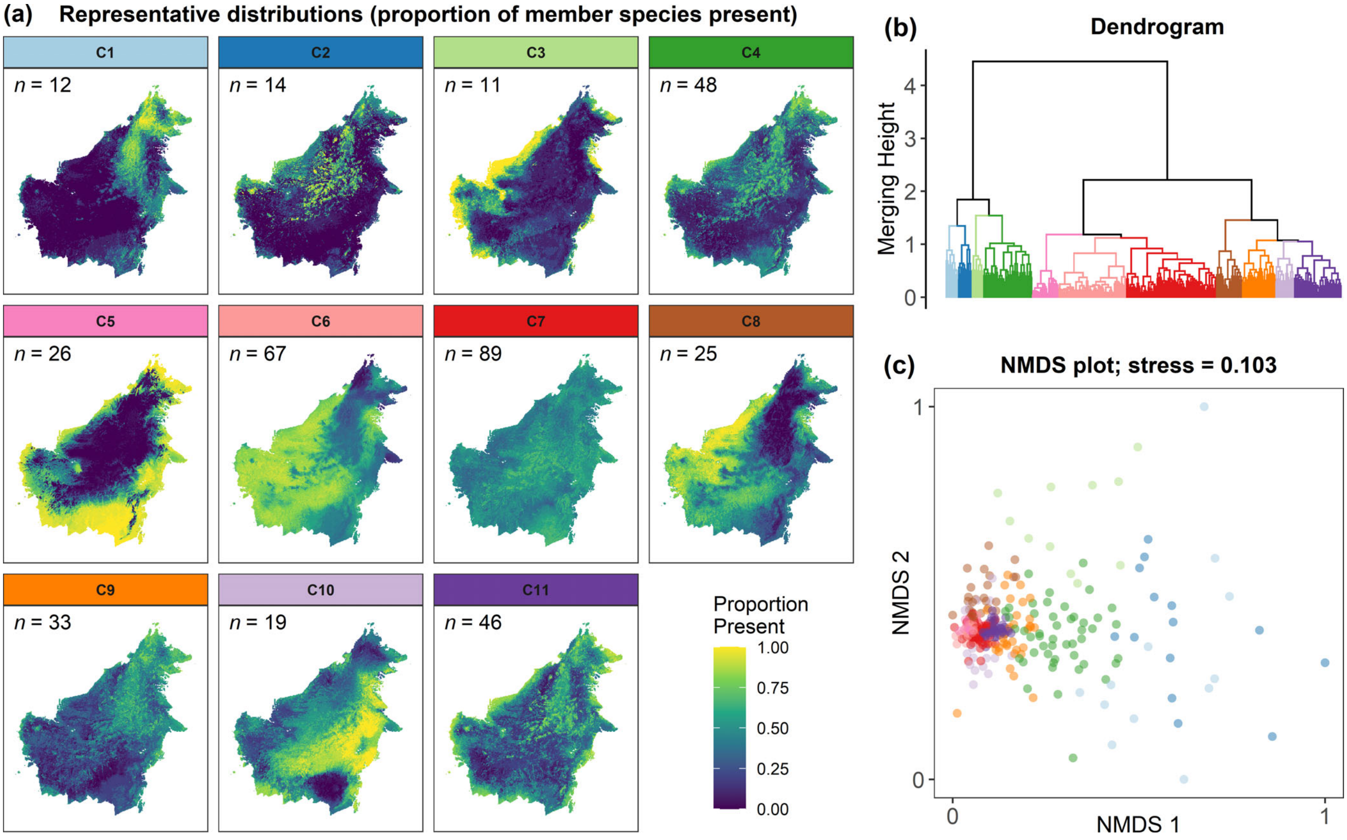

**FIGURE 6** The visualization of clusters obtained from the second-best performing clustering outcome (Bray–Curtis—WARD) for number of clusters $k = 11$. (a) The representative distribution of each cluster, where $n$ equals the number of member species. Representative distributions were generated by summing the binary data of member species at each pixel and dividing values by $n$ (i.e., the proportion of member species present). (b) The dendrogram of the clustering outcome and (c) its underlying dissimilarity matrix visualized as a nonmetric multidimensional scaling (NMDS) plot. Cluster memberships here ($k = 11$) differed from those in Figure 5c ($k = 6$). Clusters were differentiated by color, which were consistent across panels and for Figure 7.

clusters 9–11 split up much later and at near-identical levels (merging height = 1.04 and 1.07; Figure 6b), indicating low and even measures of between-cluster dissimilarities. While cluster 10 was distributed predominantly along the eastern lowland regions, clusters 9 and 11 were distributed across the mid-montane regions but over areas with vastly different underlying soil conditions (mainly available water and cation exchange capacity; Appendix S1: Figure S19).

Lastly, the remaining clusters, 5–7, split up late and at near-identical levels (merging height = 1.11 and 1.18; Figure 6b). Cluster 5 characterized the coastal/peatland forests of Indonesian Borneo (south and east Kalimantan), occupying areas south and east of Borneo's central mountain range (Figure 6a). Cluster 6 was broadly distributed across the lowland regions west and south of Borneo's central mountain range. Cluster 7 contained the most species ($n = 89$) and had a generally widespread distribution, though with slightly higher proportion values along the mid-montane and southeastern lowland regions of Borneo.

## Habitat loss due to land-cover changes

We found a substantial loss of habitat due to land-cover changes for all clusters (Figure 7). By 1992, habitat loss among clusters averaged 30% (Figure 7a)—highest for clusters 3 and 5 (38% and 44%, respectively) and lowest for cluster 2 (22%). Subsequent land-cover changes resulted in a cumulative mean habitat loss of 43% by 2020—again, habitat loss was highest for clusters 3 and 5 (56% and 61%, respectively) and lowest for cluster 2 (33%) (Figure 7a).

Annual trends made apparent the differences in habitat loss among clusters. Most striking was the severe and continued loss of coastal/peatland habitats supporting clusters 3 and 5, and the western lowland habitats supporting cluster 8 (Figure 7b). Rates of habitat loss from 1992 to 2020 increased for most clusters and were also highest for clusters 3, 5, and 8; only clusters 1, 9, and 10 experienced a decrease in rates of habitat loss (Figure 7c). Annual trends also revealed cycles in habitat loss, oscillating from a slump to a peak rate of habitat loss (Figure 7b,c). The first cycle started from a slump in 1994 to a peak in 1998, the second from 2003 to 2007, and the third from 2013 to 2016.

Differing rates of habitat loss between cycles indicate potential shifts in habitats targeted for land-cover change. Among montane-distributed clusters 1, 2, and 4 (Figure 6a), cluster 1 experienced higher rates of habitat loss than those of clusters 2 and 4 in the first cycle (1994–1998) (Figure 7c). The subsequent cycle saw a switch, with rates of habitat loss being greater for clusters 2 and 4 instead (2002–2007). Although comparatively lower in the third cycle, rates of habitat loss were still higher for clusters 2 and 4. This was likely a recent shift, as before 1992, clusters 2 and 4 were the clusters least impacted by land-cover changes (Figure 7a). Between coastal-/peatland-distributed clusters 3 and 5, rates of habitat loss surged for cluster 3 in the second cycle (2002–2007), but there was also an increase for cluster 5, albeit an increase of lower magnitude than for cluster 3. Moreover, the increased rate of habitat loss for cluster 3 in the second cycle was so great that its proportion of habitat remaining in 1992 that was lost by 2012 nearly matched that of cluster 5 (Figure 7b). Although we did not observe a similar surge in rates of habitat loss for cluster 3 in the third cycle (2012–2016), its overall increase from 1992 to 2020 was the highest among clusters (Figure 7c).

# DISCUSSION

To capitalize on the emerging wealth of distribution data, researchers must also contend with a corresponding increase in data complexity. We demonstrate how the CSAS offers a way to simplify that complexity to unravel meaningful patterns in species distributions, where resultant clusters provide a spatially explicit framework for investigating distribution-related questions in ecology, biogeography, and conservation (Clements, 1936; HilleRisLambers et al., 2012; Legendre & Legendre, 2012; Marquet et al., 2004). Importantly, clusters are quantitatively derived and easily reproducible, while the methodological framework promotes transparency and peer scrutiny of the methods and results. The primary advantage of the framework, however, is that it makes key steps of the clustering process more explicit, forcing practitioners to carefully consider their methodological choices in relation to their research objectives (Legendre & Legendre, 2012). Combined with steps that support practitioners in investigating different methods and interpreting results, our methodological framework encourages more informed decision-making and rigorous selection of final clustering outcomes.

## Clusters of spatially associated species with variable trends of habitat loss

Our methods identified 11 distinct clusters of tree species in Borneo, based on their spatial distributions, which provided valuable ecological and conservation insights. Montane species clusters 1, 2, and 4 had the most distinct distributions because their ranges were highly restricted. Their distributions also matched previous predictions by

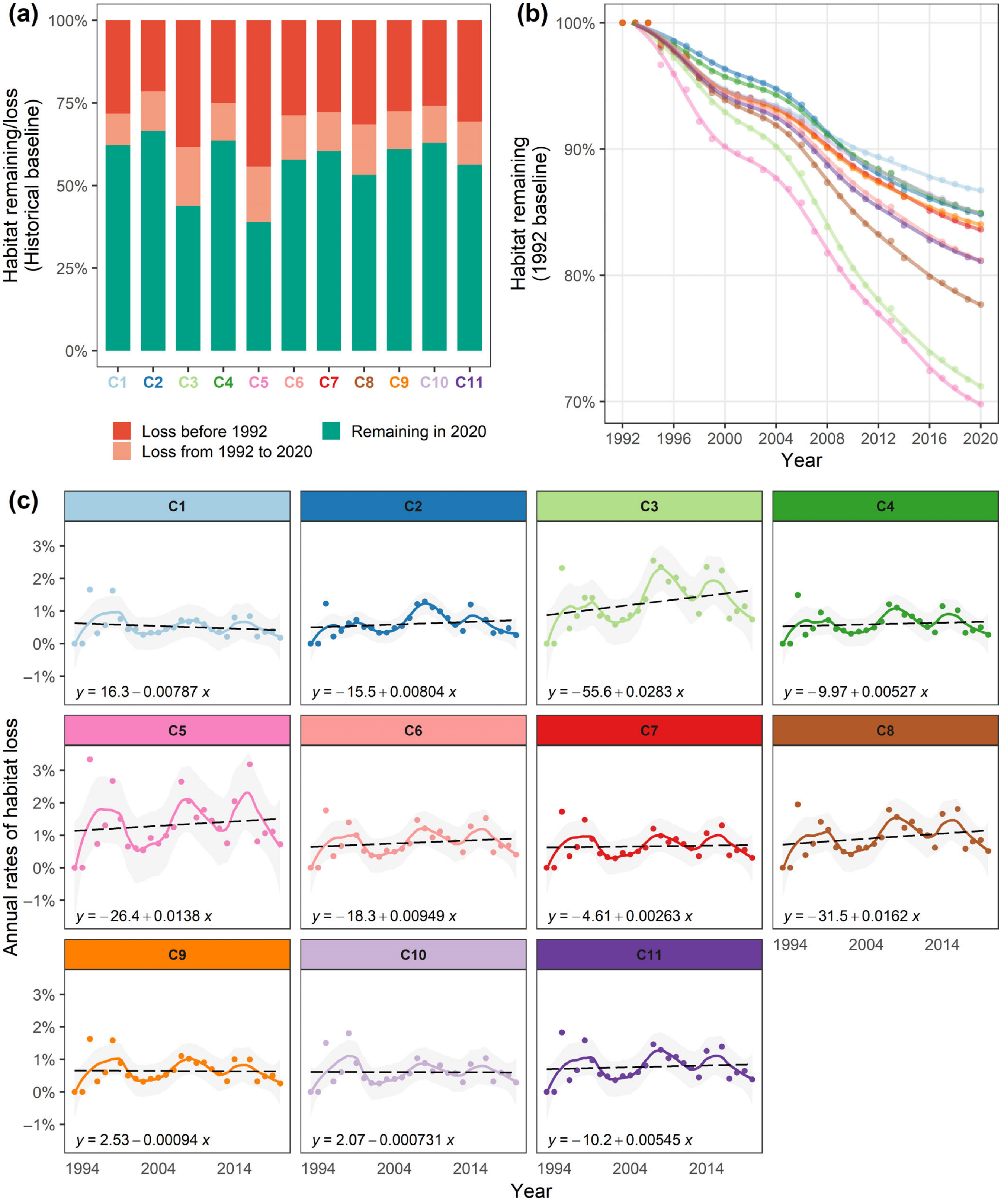

**FIGURE 7** The impact of deforestation on each cluster's representative distribution as changes in habitat availability. (a) Bar plots show the percentage of historically available habitats loss before 1992, loss from 1992 to 2020, and remaining in the year 2020. (b) Annual percentages of 1992 habitats remaining from 1992 to 2020 fitted using a generalized additive model. (c) Annual rates of habitat loss from 1993 to 2020 fitted using a local polynomial regression with α = 0.75 (solid, colored) and linear regression (dashed, black). Clusters are differentiated by color and are consistent across panels and for Figure 6.

Raes et al. (2009): in particular, the high richness of range-restricted species in the northern montane region of Borneo coincided with an area of overlap among our three montane—and some sub-montane—clusters. With narrow ranges and a lack of areas to migrate upwards in elevation under climate change (Bellard et al., 2014; Pang et al., 2021; Yanahan & Moore, 2019), these montane-distributed clusters are of great conservation value (Guisan et al., 2013; Struebig et al., 2015; Villalobos et al., 2013). We also observed clusters restricted to western Borneo: the coastal distributed cluster 3, slightly more inland cluster 8, and more widespread cluster 6 that extended further south as well. Because our SDMs relied on environmental variables, this restriction stems from a unique set of environmental conditions (i.e., high precipitation and low clay content; Appendix S1: Figure S19). Indeed, western Borneo is home to many endemics and is noted for its unique floral composition, with present-day environmental conditions previously also identified as probable drivers (Neo et al., 2021; Slik et al., 2003, 2011). Conversely, the more widely distributed clusters 5, 7, 9, and 10 seem to represent eastern Borneo, which is characterized by a combination of low precipitation, low available water capacity, and high soil clay content (Appendix S1: Figure S19).

Our general results of severe habitat loss corroborated previous assessments of deforestation in Borneo (Gaveau et al., 2014, 2016, 2019; Miettinen et al., 2011; Sloan et al., 2019; Wong et al., 2020), but our classification of species distributions allowed us to separate variable trends of habitat loss to discern especially threatened species groups. Clusters with coastal/peatland distributions (clusters 3 and 5) suffered the most severe loss of habitat, likely due to extensive oil palm expansions that primarily target coastal/peatland habitats (Gaveau et al., 2014, 2016; Miettinen et al., 2011). However, the western-restricted, inland cluster 8 also suffered severe habitat losses, losses not observed for its eastern-restricted, inland counterpart, cluster 10. Water stress greatly limits oil palm yields, where greater rainfall in western than southern/eastern Borneo may be facilitating oil palm plantation expansions into the more inland habitats (Carr, 2011; Sa'adi et al., 2021; Appendix S1: Figure S19). Although deforestation—in absolute terms—is less extensive in western Borneo (Gaveau et al., 2014, 2019; Miettinen et al., 2011), cluster 8 has a narrower distribution than most, resulting in it suffering the third most severe percentage loss of habitat.

Similarly, because our method separated montane habitats into three distinct clusters, we were able to identify a switch in habitat loss severity. The northern montane-distributed cluster 1 suffered habitat losses much greater in the first cycle (peak in 1998) than those after that (peaks in 2007 and 2016), potentially related to protected area implementation and enforcement in the Kinabalu montane alpine meadows ecoregion that coincides most with cluster 1's distribution (Olson et al., 2001; Phua et al., 2008). However, this protection seemed to not extend to the more central, western montane-distributed clusters 2 and 4, which suffered greater habitat losses in later cycles. Our results highlight the emerging threat of land-cover changes for tropical montane habitat that recent studies have also confirmed (Feng et al., 2021; Karger et al., 2021), but there exists variability in those trends of loss. Our separation of montane species clusters and their variable loss of habitat can facilitate more targeted conservation planning to better protect those more threatened in recent decades (Ashcroft, 2010; Guisan et al., 2013; Struebig et al., 2015).

The observed temporal oscillations in habitat losses were also of interest. Oscillations were temporally congruent across clusters, with peaks in 1995, 1998, 2007, and 2016 that coincided with notable El Niño periods of 1992–1995, 1997–1998, 2005–2007, and 2014–2016 (NOAA, 2022). Moreover, studies have found El Niño effects to compound with deforestation to increase forest fire severity and frequency (Chapman et al., 2020; Huijnen et al., 2016; Sloan et al., 2019; Wooster et al., 2012; but see Gaveau et al., 2015; Langner & Siegert, 2009), which may explain differing magnitudes of oscillation among clusters. Because a substantial proportion of habitat losses was tied to these oscillations, its exact driver warrants further investigation, for which our clusters provide the framework to accomplish.

## Further applications

The application of the CSAS goes beyond separating trends of habitat loss. We may use clustering outcomes to investigate structural changes in species associations, for instance, changes due to the spread of an invasive species or climate-induced range shifts, which are particularly underappreciated facets of global change impact on biodiversity (Early & Sax, 2014; Keil et al., 2021; Krosby et al., 2015). Alternatively, we might compare the resultant dendrogram against a phylogenetic tree to investigate speciation events linked to present-day spatial patterns (Villalobos et al., 2017), or against dendrograms of functional similarity to uncover coexistence or competitive mechanisms that underlie co-occurrence patterns (Rüger et al., 2020).

The representative spatial distribution of species from each group is also useful for biodiversity monitoring and management (Cousins, 1991; Webb, 1989). Commonly used criteria for evaluating the conservation value of sites, like absolute species richness or beta diversity, are prone to taxonomic and sampling biases (i.e., the so-called Linnaean and Wallacean shortfalls) (Lomolino, 2004; Possingham et al., 2007; Whittaker et al., 2005), which typically overrepresent widespread and easy-to-detect species (Boakes et al., 2010; Jetz & Rahbek, 2002; Lennon et al., 2003; Prendergast, 1993). However, in using species clusters and their representative distributions instead, site evaluations are unbiased by the relative number of species from each cluster (Possingham et al., 2007; Roberge & Angelstam, 2004). Widespread species will also form their own groups and not affect evaluations tied to range-restricted species. Moreover, representative distributions indicate the geographical unit to which member species reside and are, to some extent, restricted to that geographical unit. Representative distributions are thus highly informative when the goal is to detect specific ecosystems, protect species ranges, and assess extinction risks (Guisan et al., 2013; Hannah et al., 2020).

The difference in spatial biodiversity patterns identified through our use of species clustering (representative distributions) in comparison to the "sister analysis" of site clustering or bioregionalization (bioregions) remains a grossly underexplored aspect of biogeography and spatial ecology (Jongman et al., 1995; Keil et al., 2021; Legendre & Legendre, 2012). However, a key difference is likely that representative distributions are not spatially discrete (overlaps can and do occur) and better represent the diversity and distinction of distributional patterns among species than bioregions, which are by definition spatially discrete (Clements, 1936; Collins et al., 1993; Dufrêne & Legendre, 1997). Representative distributions and bioregions will only be comparable when representative distributions are spatially nonoverlapping and bioregions have nonoverlapping species, that is, a strict interpretation of Clement's community model that is unlikely because of the general lack of supporting evidence among real species communities (Clements, 1936; Roberts, 1987; Shipley & Keddy, 1987; Westman, 1985; Whittaker, 1951, 1953; but see Allen & Hoekstra, 1990; Collins et al., 1993; Hoekstra et al., 1991). Thus, there may be substantial differences in application between the two approaches. More research is still needed to evaluate the implications of using representative distributions versus bioregions for various ecological applications, such as developing essential biodiversity variables for informing conservation planning (Guisan et al., 2013; Jetz et al., 2019; Struebig et al., 2015). However, we expect species clusters and their representative distributions to be more appropriate for investigating and summarizing spatial relationships or phenomena among species, such as when investigating climate-induced distribution changes among lowland- and montane-distributed species or prioritizing the protection of sites that support particular species clusters (e.g., clusters with coastal/peatland distributions).

## On selecting association indices and clustering algorithms

Our findings highlight the value of our methodological framework, not just its individual steps but also the testing of multiple methods. Different association indices and clustering algorithms led to clustering outcomes that were highly varied in dendrogram structure, performance, and cluster memberships. Data-driven assessments and comparisons and ecological theory need to guide the selection of an appropriate method.

Exploring multiple association indices is especially crucial given the sequential nature of steps in our framework, in that the index influences how clustering algorithms work and perform (Jain et al., 1999; Legendre & Legendre, 2012). In our study, we explored a wide range of binary and continuous indices with different inherent mathematical properties. Like previous studies, indices varied greatly in their measurement of associations but generally separated into three groups as determined by their mathematical properties: (1) difference- and distance-based indices; (2) binary indices that exclude co-absences; and (3) correlation-based indices and binary indices that include co-absences (Hubálek, 1982; Keil et al., 2021). Given the computational intensity of quantifying associations across entire spatial extents (maps in lieu of plots), we advocate exploring at least one index from each group. Beyond the groups described, practitioners may consider other qualities when selecting (or excluding) indices to explore: the ability to recover simulated magnitudes of spatial attraction and repulsion, into which Keil et al. (2021) offer insight; adherence to the triangle inequality rule (e.g., Hellinger distance), an important axiom of distance matrices for geometric clustering; nonparametric approaches to allow measurements between noncomparable data without any prior transformation (e.g., Spearman correlation for occurrence probabilities between SDM algorithms, or count data between species with disparate raw abundances) (Warren et al., 2008, 2019); or popularity and simplicity to facilitate the interpretation of results (e.g., Jaccard index as proportional overlap).

Our framework emphasizes testing multiple clustering algorithms because algorithms vary in their capacity to simultaneously minimize within-cluster and maximize between-cluster variances. This was observed in the variation across dendrogram performances, which offer great insight into selecting clustering algorithms with certain characteristics (Fernández & Gómez, 2020; Legendre &

Legendre, 2012; Rousseeuw, 1986; Sokal & Rohlf, 1962). The relative importance of those characteristics, represented by each dendrogram performance metric, depends on the research objective. When examining changes in a community's spatial structure over time (e.g., due to invasive species introduction or climate-induced range shifts; Early & Sax, 2014; Krosby et al., 2015), it is important to minimize data distortions introduced by clustering. Thus, faithfulness to the original dissimilarity matrix may be the only priority, in which case UPGMA might be preferred. Alternatively, if the main objective is to decompose a large dataset into smaller but evenly sized subsets, cluster strength and balance may be more important, in which case WARD and CL might be selected instead. The flexibility of this step goes beyond our three metrics, where practitioners may incorporate others of importance, such as space distortion ratio, connectedness, or isolation (see Estabrook, 1966; Fernández & Gómez, 2020; Legendre & Legendre, 2012; Wirth et al., 1966). Practitioners must also consider the relevance of each candidate algorithm; the selected algorithms should have linkage functions that align with the study's theoretical expectation of how clusters form (see Erman et al., 2015; Roux, 2018; Seif, 2018). Here, we rejected dendrograms resulting from UPGMC and WPGMC to avoid reversals (Abe et al., 2017; Miyamoto, 2012; Wedley et al., 1993).

Although, for the case study of Borneo, we selected the dendrogram based on Bray–Curtis dissimilarity resulting from WARD because it performed (second) best, this does not suggest that practitioners should always base their selection on dendrogram performance scores. For example, Spearman correlation captured associations that were inherently more complex (i.e., highly stressed two-dimensional NMDS), and thus difficult to maintain during the clustering analysis (i.e., low co-phenetic correlation scores). However, this complexity likely stems from Spearman correlation capturing information on the directionality of associations—positive or negative correlation—which may be a point of ecological interest rather than a basis for rejection. Therefore, we reemphasize the need for ecological theory to guide comparisons of multiple indices. The resultant data structure of a chosen association index also determines whether the clustering algorithm achieves meaningful clusters (Jain et al., 1999; Legendre & Legendre, 2012; Rajalingam & Ranjini, 2011; Roux, 2018). In cases where it does not, practitioners might select an alternative—but comparably well-performing—algorithm instead, as we did for our case study. Indeed, dendrogram scores provide crucial information on each clustering outcome, but data visualizing tools like NMDS are equally vital aids for selecting methods most appropriate to one's study objective.

## Challenges of clustering spatially associated species

Despite substantial variations among candidate clustering outcomes, broad patterns in tree species distributions remained relatively consistent as many of those observed in our chosen clustering outcome were observed among other well-performing clustering outcomes as well (i.e., high dendrogram performance and meaningful cluster sizes). This consistency lends credibility to the spatial patterns of our chosen clustering outcome and the robustness of employing species clusters. However, distributional patterns characterizing widespread species were irregular, suggesting the classification of those species to be partly contingent on the association index and the clustering algorithm used (Dufrêne & Legendre, 1997; Jongman et al., 1995; Legendre & Legendre, 2012). Although most clustering outcomes were still able to discern some general patterns in widespread species distributions, our findings highlight that special attention is needed when clustering widely distributed species, more so because such clusters may represent hyper-abundant species that comprise the bulk of stem density and aboveground carbon (Fauset et al., 2015), where an understanding of their spatial patterns may improve the protection of high-carbon forests (Siman et al., 2021; Sullivan et al., 2017, 2020).

When interpreting clusters, we must also recognize the limits of the underlying methods and how that might affect ecological interpretations. In selecting Bray–Curtis dissimilarity, we define associations as absolute differences in occurrence probability. Hence, dissimilarities were low between species with probabilities close in value, even when their binary distributions were nonoverlapping or when their probabilities were uncorrelated (see clusters 10 and 11 representative distributions vs. their relatedness) (Figure 6). Because many lowland species exhibited this characteristic, we found many species with relatively low dissimilarities (large aggregate of points in the NMDS plot), which did not apply to other groups of associations indices (e.g., correlation-based indices). The inverse was also true for species with highly skewed probabilities (i.e., high dissimilarities due to sharp probability gradients), which was probably why highly range-restricted species were predominantly further dispersed in the NMDS plot. Thus, we acknowledge that dissimilarities among range-restricted species and their clusters might have been exaggerated to a degree, while more nuanced patterns among lowland widespread distributions might have been overlooked.

In selecting the WARD clustering algorithm, resultant cluster boundaries are typically oddly shaped and evenly sized (i.e., number of objects per cluster) (Erman et al., 2015; Legendre & Legendre, 2012; Seif, 2018; Ward, 1963). As a result, WARD was able to partition the large aggregate of points and form meaningful clusters. The trade-off, however, is that WARD often distorts dissimilarity space, especially for ecological data where objects often exist along a continuum (Fernández & Gómez, 2020; Holt et al., 2013; Kreft & Jetz, 2010; Legendre & Legendre, 2012). Hence, resulting hierarchical relationships were likely inaccurate in representing original dissimilarities and must be carefully interpreted. In summary, we emphasize that the most appropriate clustering outcome is context-dependent and study-specific. This need for a flexible combination of ecological theory and data-driven assessments—to guide practitioners in considering the benefits and drawbacks of each method—is embedded in our framework.

## CONCLUSION

Unraveling the complexity of species distribution data is necessary to understand the factors driving the diversity of distribution patterns among species (Collins et al., 1993; Keddy, 1992; Marquet et al., 2004). To that end, we present the CSAS as a way to simplify complex distribution data and identify distinct distributional patters as we demonstrated for Bornean tree species. A critical application of the CSAS is in uncovering the divergent impacts of spatially heterogeneous threats among species with dissimilar distributions, which we revealed through our analysis of cluster-specific trends of habitat loss due to land-cover change. The CSAS provides a timely tool addressing the urgent need to understand global change impacts on species with dissimilar distributions (Guisan et al., 2013; Marquet et al., 2004; Struebig et al., 2015).

We facilitate adoption of the CSAS through our methodological framework, which provides a clear and detailed structure to guide practitioners in developing species clusters and applying them for geographical, ecological, and conservation research. We emphasize the importance of (1) exploring multiple association indices and clustering algorithms, (2) selecting methods based on data-driven assessments of cluster performance/optimality (e.g., NMDS plots for visualizing the data structure and measures of dendrogram performance) and the relevance or appropriateness of the methods with respect to the study's objective (e.g., study's theoretical expectation of how clusters form and subsequent applications of derived species clusters), and (3) discussing the limitations of the chosen methods and their implications for ecologically interpreting species clusters and their representative distributions. Our methodological framework and publicly available codes will support practitioners in leveraging the ever-growing abundance of distribution data to better understand complex spatial patterns among species distributions and the disparate effects of global changes on biodiversity.

### AUTHOR CONTRIBUTIONS

Sean E. H. Pang and Edward L. Webb conceived the idea and designed the methodology. Sean E. H. Pang conducted all analyses and data visualizations and developed the methodological framework with key inputs from Edward L. Webb, J. W. Ferry Slik, and Damaris Zurell. All authors contributed to the interpretation of the results. Sean E. H. Pang and Edward L. Webb wrote the first draft of the manuscript. All authors provided feedback on the writing.

### ACKNOWLEDGMENTS

We thank Ryan A. Chisholm for his comments on the analysis and writing and A. Townsend Peterson for his feedback on the study, which greatly improved the final version of this manuscript.

### CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

### DATA AVAILABILITY STATEMENT

Occurrence data are available from the Global Biodiversity Information Facility (GBIF, 2019): https://doi.org/10.15468/dl.3sqcf4. The results data and R scripts used to analyze and visualize those data are available from FigShare: https://doi.org/10.6084/m9.figshare.21407379; these include the novel R functions developed to create and assess clusters of spatially associated species.

### ORCID

*Sean E. H. Pang* https://orcid.org/0000-0003-4574-6028
*J. W. Ferry Slik* https://orcid.org/0000-0003-3988-7019
*Damaris Zurell* https://orcid.org/0000-0002-4628-3558
*Edward L. Webb* https://orcid.org/0000-0001-5554-9955

### REFERENCES

Abe, R., S. Miyamoto, Y. Endo, and Y. Hamasuna. 2017. "Hierarchical Clustering Algorithms with Automatic Estimation of the Number of Clusters." Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS), 1–5. Otsu, Japan. June 27–30.

Aiello-Lammens, M. E., R. A. Boria, A. Radosavljevic, B. Vilela, and R. P. Anderson. 2015. "spThin: An R Package for Spatial Thinning of Species Occurrence Records for Use in Ecological Niche Models." *Ecography* 38: 541–5.

Allen, T. F. H., and T. W. Hoekstra. 1990. "The Confusion between Scale-Defined Levels and Conventional Levels of Organization in Ecology." *Journal of Vegetation Science* 1: 5–12.

Ashcroft, M. B. 2010. "Identifying Refugia from Climate Change." *Journal of Biogeography* 37: 1407–13.

Baatar, U.-O. 2019. “Evaluating Climatic Threats to Habitat Types Based on Co-occurrence Patterns of Characteristic Species.” *Basic and Applied Ecology* 38: 13–35.

Baker, F. B. 1974. “Stability of Two Hierarchical Grouping Techniques Case I: Sensitivity to Data Errors.” *Journal of the American Statistical Association* 69: 440–5.

Beech, E., M. Rivers, S. Oldfield, and P. P. Smith. 2017. “GlobalTreeSearch: The First Complete Global Database of Tree Species and Country Distributions.” *Journal of Sustainable Forestry* 36: 454–89.

Bellard, C., C. Leclerc, B. Leroy, M. Bakkenes, S. Veloz, W. Thuiller, and F. Courchamp. 2014. “Vulnerability of Biodiversity Hotspots to Global Change.” *Global Ecology and Biogeography* 23: 1376–86.

Blanchet, F. G., K. Cazelles, and D. Gravel. 2020. “Co-occurrence Is Not Evidence of Ecological Interactions.” *Ecology Letters* 23: 1050–63.

Boakes, E. H., P. J. K. McGowan, R. A. Fuller, D. Chang-qing, N. E. Clark, K. O’Connor, and G. M. Mace. 2010. “Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data.” *PLoS Biology* 8: e1000385.

Boria, R. A., L. E. Olson, S. M. Goodman, and R. P. Anderson. 2017. “A Single-Algorithm Ensemble Approach to Estimating Suitability and Uncertainty: Cross-Time Projections for Four Malagasy Tenrecs.” *Diversity and Distributions* 23: 196–208.

Boyce, P. C., and S. Y. Wong. 2019. “Borneo and Its Disproportionately Large Rheophytic Aroid Flora.” *Gardens’ Bulletin Singapore* 71: 497–524.

Calatayud, J., E. Andivia, A. Escudero, C. J. Melián, R. Bernardo-Madrid, M. Stoffel, C. Aponte, et al. 2020. “Positive Associations among Rare Species and Their Persistence in Ecological Assemblages.” *Nature Ecology & Evolution* 4: 40–5.

Caliński, T., and J. Harabasz. 1974. “A Dendrite Method for Cluster Analysis.” *Communications in Statistics—Theory and Methods* 3: 1–27.

Carr, M. K. V. 2011. “The Water Relations and Irrigation Requirements of Oil Palm (*Elaeis guineensis*): A Review.” *Experimental Agriculture* 47: 629–52.

Cayuela, L., Í. Granzow-de la Cerda, F. S. Albuquerque, and D. J. Golicher. 2012. “Taxonstand: An R Package for Species Names Standardisation in Vegetation Databases.” *Methods in Ecology and Evolution* 3: 1078–83.

Chapman, S., J. Syktus, R. Trancoso, A. Salazar, M. Thatcher, J. E. M. Watson, E. Meijaard, D. Sheil, P. Dargusch, and C. A. McAlpine. 2020. “Compounding Impact of Deforestation on Borneo’s Climate during El Niño Events.” *Environmental Research Letters* 15: 084006.

Chouikhi, H., M. Charrad, and N. Ghazzali. 2015. “A Comparison Study of Clustering Validity Indices.” In *2015 Global Summit on Computer & Information Technology (GSCIT)* 1–4. Sousse: IEEE.

Clements, F. E. 1936. “Nature and Structure of the Climax.” *The Journal of Ecology* 24: 252.

Collins, S. L., S. M. Glenn, and D. W. Roberts. 1993. “The Hierarchical Continuum Concept.” *Journal of Vegetation Science* 4: 149–56.

Corlett, R. T., and K. W. Tomlinson. 2020. “Climate Change and Edaphic Specialists: Irresistible Force Meets Immovable Object?” *Trends in Ecology & Evolution* 35: 367–76.

Cousins, S. H. 1991. “Species Diversity Measurement: Choosing the Right Index.” *Trends in Ecology & Evolution* 6: 190–2.

Cramér, H. 1924. “Remarks on Correlation.” *Scandinavian Actuarial Journal* 1924: 220–40.

Currie, D. J. 2019. “Where Newton Might Have Taken Ecology.” *Global Ecology and Biogeography* 28: 18–27.

Curtis, J. T. 1959. *The Vegetation of Wisconsin: An Ordination of Plant Communities*. Madison, WI: University of Wisconsin Press.

Di Febbraro, M., L. Sallustio, M. Vizzarri, D. De Rosa, L. De Lisio, A. Loy, B. A. Eichelberger, and M. Marchetti. 2018. “Expert-Based and Correlative Models to Map Habitat Quality: Which Gives Better Support to Conservation Planning?” *Global Ecology and Conservation* 16: e00513.

Duda, R. O., and P. E. Hart. 1973. *Pattern Classification and Scene Analysis*. New York: Wiley.

Dufrêne, M., and P. Legendre. 1997. “Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach.” *Ecological Monographs* 67: 345–66.

Early, R., and D. F. Sax. 2014. “Climatic Niche Shifts between species’ Native and Naturalized Ranges Raise Concern for Ecological Forecasts during Invasions and Climate Change.” *Global Ecology and Biogeography* 23: 1356–65.

Elith, J., and J. R. Leathwick. 2009. “Species Distribution Models: Ecological Explanation and Prediction across Space and Time.” *Annual Review of Ecology, Evolution, and Systematics* 40: 677–97.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. “A Statistical Explanation of MaxEnt for Ecologists.” *Diversity and Distributions* 17: 43–57.

Erman, N., A. Korosec, and J. Suklan. 2015. “Performance of Selected Agglomerative Hierarchical Clustering Methods.” *Innovative Issues and Approaches in Social Sciences* 8: 180–204.

ESA. 2017. “Land Cover CCI Product User Guide Version 2.” Technical Report. https://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf.

Estabrook, G. F. 1966. “A Mathematical Model in Graph Theory for Biological Classification.” *Journal of Theoretical Biology* 12: 297–310.

Faurby, S., and M. B. Araújo. 2018. “Anthropogenic Range Contractions Bias Species Climate Change Forecasts.” *Nature Climate Change* 8: 252–6.

Fauset, S., M. O. Johnson, M. Gloor, T. R. Baker, A. Monteagudo M., R. J. W. Brienen, T. R. Feldpausch, et al. 2015. “Hyperdominance in Amazonian Forest Carbon Cycling.” *Nature Communications* 6: 6857.

Feng, X., D. S. Park, C. Walker, A. T. Peterson, C. Merow, and M. Papeş. 2019. “A Checklist for Maximizing Reproducibility of Ecological Niche Models.” *Nature Ecology & Evolution* 3: 1382–95.

Feng, Y., A. D. Ziegler, P. R. Elsen, Y. Liu, X. He, D. V. Spracklen, J. Holden, X. Jiang, C. Zheng, and Z. Zeng. 2021. “Upward Expansion and Acceleration of Forest Clearance in the Mountains of Southeast Asia.” *Nature Sustainability* 4: 892–9.

Fernández, A., and S. Gómez. 2020. “Versatile Linkage: A Family of Space-Conserving Strategies for Agglomerative Hierarchical Clustering.” *Journal of Classification* 37: 584–97.

Fithian, W., and T. Hastie. 2013. “Finite-Sample Equivalence in Statistical Models for Presence-Only Data.” *The Annals of Applied Statistics* 7: 1917.

Gaston, K. J. 1996. “Species-Range-Size Distributions: Patterns, Mechanisms and Implications.” *Trends in Ecology & Evolution* 11: 197–201.

Gaston, K. J. 2003. *The Structure and Dynamics of Geographic Ranges*. Oxford: Oxford University Press.

Gauch, H. G., and R. H. Whittaker. 1981. “Hierarchical Classification of Community Data.” *The Journal of Ecology* 69: 537.

Gaveau, D. L. A., B. Locatelli, M. A. Salim, H. Yaen, P. Pacheco, and D. Sheil. 2019. “Rise and Fall of Forest Loss and Industrial Plantations in Borneo (2000–2017).” *Conservation Letters* 12: e12622.

Gaveau, D. L. A., M. A. Salim, K. Hergoualc’h, B. Locatelli, S. Sloan, M. Wooster, M. E. Marlier, et al. 2015. “Major Atmospheric Emissions from Peat Fires in Southeast Asia during Non-drought Years: Evidence from the 2013 Sumatran Fires.” *Scientific Reports* 4: 6112.

Gaveau, D. L. A., D. Sheil, M. Husnayaen, M. A. Salim, S. Arjasakusuma, M. Ancrenaz, P. Pacheco, and E. Meijaard. 2016. “Rapid Conversions and Avoided Deforestation: Examining Four Decades of Industrial Plantation Expansion in Borneo.” *Scientific Reports* 6: 32017.

Gaveau, D. L. A., S. Sloan, E. Molidena, H. Yaen, D. Sheil, N. K. Abram, M. Ancrenaz, et al. 2014. “Four Decades of Forest Persistence, Clearance and Logging on Borneo.” *PLoS One* 9: e101654.

GBIF. 2019. “GBIF Occurrence Download.” https://doi.org/10.15468/dl.3sqcf4.

GBIF. 2022. “Global Data Trends.” https://www.gbif.org/analytics/global.

Graham, C. H., and R. J. Hijmans. 2006. “A Comparison of Methods for Mapping Species Ranges and Species Richness.” *Global Ecology and Biogeography* 15: 578–87.

Guerra, L., V. Robles, C. Bielza, and P. Larrañaga. 2012. “A Comparison of Clustering Quality Indices Using Outliers and Noise.” *Intelligent Data Analysis* 16: 703–15.

Gueta, T., and Y. Carmel. 2016. “Quantifying the Value of User-Level Data Cleaning for Big Data: A Case Study Using Mammal Distribution Models.” *Ecological Informatics* 34: 139–45.

Guisan, A., and W. Thuiller. 2005. “Predicting Species Distribution: Offering More than Simple Habitat Models.” *Ecology Letters* 8: 993–1009.

Guisan, A., R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. T. Tulloch, T. J. Regan, et al. 2013. “Predicting Species Distributions for Conservation Decisions.” *Ecology Letters* 16: 1424–35.

Hannah, L., P. R. Roehrdanz, P. A. Marquet, B. J. Enquist, G. Midgley, W. Foden, J. C. Lovett, et al. 2020. “30% Land Conservation and Climate Action Reduces Tropical Extinction Risk by More than 50%.” *Ecography* 43: 943–53.

Hazzi, N. A., J. S. Moreno, C. Ortiz-Movliav, and R. D. Palacio. 2018. “Biogeographic Regions and Events of Isolation and Diversification of the Endemic Biota of the Tropical Andes.” *Proceedings of the National Academy of Sciences* 115: 7985–90.

Hengl, T., J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, et al. 2017. “SoilGrids250m: Global Gridded Soil Information Based on Machine Learning.” *PLoS One* 12: e0169748.

Hijmans, R. J., and J. V. Etten. 2012. “Geographic Analysis and Modeling with Raster Data.” R Package Version 2.1-25. https://cran.r-project.org/web/packages/raster/index.html.

HilleRisLambers, J., P. B. Adler, W. S. Harpole, J. M. Levine, and M. M. Mayfield. 2012. “Rethinking Community Assembly through the Lens of Coexistence Theory.” *Annual Review of Ecology, Evolution, and Systematics* 43: 227–48.

Hoekstra, T. W., T. F. H. Allen, and C. H. Flather. 1991. “Implicit Scaling in Ecological Research.” *Bioscience* 41: 148–54.

Holt, B. G., J.-P. Lessard, M. K. Borregaard, S. A. Fritz, M. B. Araújo, D. Dimitrov, P.-H. Fabre, et al. 2013. “An Update of Wallace’s Zoogeographic Regions of the World.” *Science* 339: 6.

Hubálek, Z. 1982. “Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation.” *Biological Reviews* 57: 669–89.

Hubert, L. J., and J. R. Levin. 1976. “A General Statistical Framework for Assessing Categorical Clustering in Free Recall.” *Psychological Bulletin* 83: 1072–80.

Huijnen, V., M. J. Wooster, J. W. Kaiser, D. L. A. Gaveau, J. Flemming, M. Parrington, A. Inness, D. Murdiyarso, B. Main, and M. van Weele. 2016. “Fire Carbon Emissions over Maritime Southeast Asia in 2015 Largest since 1997.” *Scientific Reports* 6: 26886.

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* 104: 13384–9.

Jaccard, P. 1901. “Étude comparative de la distribution florale dans une portion des Alpes et des Jura.” *Bulletin de la Société Vaudoise des Sciences Naturelles* 37: 547–79.

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. “Data Clustering: A Review.” *ACM Computing Surveys* 31: 264–323.

Jetz, W., M. A. McGeoch, R. Guralnick, S. Ferrier, J. Beck, M. J. Costello, M. Fernandez, et al. 2019. “Essential Biodiversity Variables for Mapping and Monitoring Species Populations.” *Nature Ecology & Evolution* 3: 539–51.

Jetz, W., and C. Rahbek. 2002. “Geographic Range Size and Determinants of Avian Species Richness.” *Science* 297: 1548–51.

Jetz, W., C. H. Sekercioglu, and J. E. M. Watson. 2008. “Ecological Correlates and Conservation Implications of Overestimating Species Geographic Ranges: Overestimation of Species Ranges.” *Conservation Biology* 22: 110–9.

Jongman, R. H., C. J. F. ter Braak, and O. F. R. van Tongeren. 1995. *Data Analysis in Community and Landscape Ecology*. Cambridge: Cambridge University Press.

Kahneman, D., and A. Tversky. 1972. “Subjective Probability: A Judgment of Representativeness.” *Cognitive Psychology* 3: 430–54.

Karger, D. N., O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, and M. Kessler. 2017. “Climatologies at High Resolution for the Earth’s Land Surface Areas.” *Scientific Data* 4: 170122.

Karger, D. N., M. Kessler, M. Lehnert, and W. Jetz. 2021. “Limited Protection and Ongoing Loss of Tropical Cloud Forest Biodiversity and Ecosystems Worldwide.” *Nature Ecology & Evolution* 5: 854–62.

Kass, J. M., R. Muscarella, P. J. Galante, C. L. Bohl, G. E. Pinilla-Buitrago, R. A. Boria, M. Soley-Guardia, and R. P.

Anderson. 2021. “ENMeval 2.0: Redesigned for Customizable and Reproducible Modeling of Species’ Niches and Distributions.” *Methods in Ecology and Evolution* 12: 1602–8.

Kaufman, L., and P. J. Rousseeuw. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: Wiley.

Keddy, P. A. 1992. “Assembly and Response Rules: Two Goals for Predictive Community Ecology.” *Journal of Vegetation Science* 3: 157–64.

Keil, P., T. Wiegand, A. B. Tóth, D. J. McGlinn, and J. M. Chase. 2021. “Measurement and Analysis of Interspecific Spatial Associations as a Facet of Biodiversity.” *Ecological Monographs* 91: e01452.

Kramer-Schadt, S., J. Niedballa, J. D. Pilgrim, B. Schröder, J. Lindenborn, V. Reinfelder, M. Stillfried, et al. 2013. “The Importance of Correcting for Sampling Bias in MaxEnt Species Distribution Models.” *Diversity and Distributions* 19: 1366–79.

Kreft, H., and W. Jetz. 2010. “A Framework for Delineating Biogeographical Regions Based on Species Distributions: Global Quantitative Biogeographical Regionalizations.” *Journal of Biogeography* 37: 2029–53.

Krosby, M., C. B. Wilsey, J. L. McGuire, J. M. Duggan, T. M. Nogeire, J. A. Heinrichs, J. J. Tewksbury, and J. J. Lawler. 2015. “Climate-Induced Range Overlap among Closely Related Species.” *Nature Climate Change* 5: 883–6.

Langner, A., and F. Siegert. 2009. “Spatiotemporal Fire Occurrence in Borneo over a Period of 10 Years.” *Global Change Biology* 15: 48–62.

Ledo, A. 2015. “Nature and Age of Neighbours Matter: Interspecific Associations among Tree Species Exist and Vary across Life Stages in Tropical Forests.” *PLoS One* 10: e0141387.

Legendre, P., and L. Legendre. 2012. *Numerical Ecology*. Oxford: Elsevier.

Lennon, J. J., P. Koleff, J. J. D. Greenwood, and K. J. Gaston. 2003. “Contribution of Rarity and Commonness to Patterns of Species Richness: Richness Patterns and Rarity/Commonness.” *Ecology Letters* 7: 81–7.

Leroy, B., M. S. Dias, E. Giraud, B. Hugueny, C. Jézéquel, F. Leprieur, T. Oberdorff, and P. A. Tedesco. 2019. “Global Biogeographical Regions of Freshwater Fish Species.” *Journal of Biogeography* 46: 2407–19.

Liu, C., G. Newell, and M. White. 2016. “On the Selection of Thresholds for Predicting Species Occurrence with Presence-Only Data.” *Ecology and Evolution* 6: 337–48.

Lomolino, M. V. 2004. “Conservation Biogeography.” In *Frontiers of Biogeography: New Directions in the Geography of Nature*, edited by M. V. Lomolino and L. R. Heaney, 293. Sunderland, MA: Sinauer Associates.

Lomolino, M. V., B. R. C. Riddle, and J. H. C. Brown. 2006. *Biogeography*. Sunderland, MA: Sinauer Associates, Inc.

Ludwig, J. A., J. F. Reynolds, L. Quartet, and J. F. Reynolds. 1988. *Statistical Ecology: A Primer in Methods and Computing*. New York: John Wiley & Sons.

Mainali, K., T. Hefley, L. Ries, and W. F. Fagan. 2020. “Matching Expert Range Maps with Species Distribution Model Predictions.” *Conservation Biology* 34: 1292–304.

Manchego, C. E., P. Hildebrandt, J. Cueva, C. I. Espinosa, B. Stimm, and S. Günter. 2017. “Climate Change versus Deforestation: Implications for Tree Species Distribution in the Dry Forests of Southern Ecuador.” *PLoS One* 13: e0190092.

Marquet, P. A., M. Fernandez, S. A. Navarrete, and C. Valdovinos. 2004. “Diversity Emerging: Toward a Deconstruction of Biodiversity Patterns.” In *Frontiers of Biogeography: New Directions in the Geography of Nature*, edited by M. V. Lomolino and L. R. Heaney, 191–209. Sunderland, MA: Sinauer Associates.

Marshall, L., J. C. Biesmeijer, P. Rasmont, N. J. Vereecken, L. Dvorak, U. Fitzpatrick, F. Francis, et al. 2018. “The Interplay of Climate and Land Use Change Affects the Distribution of EU Bumblebees.” *Global Change Biology* 24: 101–16.

Merow, C., M. J. Smith, and J. A. Silander. 2013. “A Practical Guide to MaxEnt for Modeling Species’ Distributions: What It Does, and Why Inputs and Settings Matter.” *Ecography* 36: 1058–69.

Miettinen, J., C. Shi, and S. C. Liew. 2011. “Deforestation Rates in Insular Southeast Asia between 2000 and 2010: Deforestation in Insular Southeast Asia 2000-2010.” *Global Change Biology* 17: 2261–70.

Milanesi, P., F. Della Rocca, and R. A. Robinson. 2020. “Integrating Dynamic Environmental Predictors and Species Occurrences: Toward True Dynamic Species Distribution Models.” *Ecology and Evolution* 10: 1087–92.

Milligan, G. W., and M. C. Cooper. 1985. “An Examination of Procedures for Determining the Number of Clusters in a Data Set.” *Psychometrika* 50: 159–79.

Minchin, P. R. 1987. “An Evaluation of the Relative Robustness of Techniques for Ecological Ordination.” *Vegetatio* 69: 89–107.

Miyamoto, S. 2012. “An Overview of Hierarchical and Non-hierarchical Algorithms of Clustering for Semi-Supervised Classification.” In *Modeling Decisions for Artificial Intelligence*, edited by V. Torra, Y. Narukawa, B. López, and M. Villaret, 1–10. Berlin: Springer.

Morales, N. S., I. C. Fernández, and V. Baca-González. 2017. “MaxEnt’s Parameter Configuration and Small Samples: Are We Paying Attention to Recommendations? A Systematic Review.” *PeerJ* 5: e3093.

Neo, L., H. T. W. Tan, and K. M. Wong. 2021. “Centres of Endemism in Borneo and their Environmental Correlates Revealed by Endemic Plant Genera.” *Flora* 285: 151966.

Newbold, T. 2018. “Future Effects of Climate and Land-Use Change on Terrestrial Vertebrate Community Diversity under Different Scenarios.” *Proceedings of the Royal Society B: Biological Sciences* 285: 20180792.

NOAA. 2022. “Southern Oscillation Index (SOI) | El Niño/Southern Oscillation (ENSO).” Monitoring. https://www.ncei.noaa.gov/access/monitoring/enso/soi.

Norberg, A., N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araújo, et al. 2019. “A Comprehensive Evaluation of Predictive Performance of 33 Species Distribution Models at Species and Community Levels.” *Ecological Monographs* 89: 1–24.

Oksanen, J., R. Kindt, P. Legendre, B. O’Hara, M. H. H. Stevens, M. J. Oksanen, and M. Suggests. 2007. “The Vegan Package.” *Community Ecology Package* 10: 719.

Olson, D. M., E. Dinerstein, E. D. Wikramanayake, N. D. Burgess, G. V. Powell, E. C. Underwood, J. A. D’amico, I. Itoua, H. E. Strand, and J. C. Morrison. 2001. “Terrestrial Ecoregions of the World: A New Map of Life on Earth.” *Bioscience* 51: 933–8.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. “How to

Make More out of Community Data? A Conceptual Framework and Its Implementation as Models and Software." *Ecology Letters* 20: 561–76.

Owen-Smith, N., J. Martin, and K. Yoganand. 2015. "Spatially Nested Niche Partitioning between Syntopic Grazers at Foraging Arena Scale within Overlapping Home Ranges." *Ecosphere* 6: 1–17.

Pang, S. E. H., J. D. T. De Alban, and E. L. Webb. 2021. "Effects of Climate Change and Land Cover on the Distributions of a Critical Tree Family in the Philippines." *Scientific Reports* 11: 276.

Pang, S. E. H., Y. Zeng, J. D. T. De Alban, and E. L. Webb. 2022. "Occurrence–Habitat Mismatching and Niche Truncation when Modelling Distributions Affected by Anthropogenic Range Contractions." *Diversity and Distributions* 28: 1327–43.

Peterson, A. T., ed. 2011. *Ecological Niches and Geographic Distributions*. Princeton, NJ: Princeton University Press.

Peterson, A. T., and J. Soberón. 2012. "Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right." *Natureza & Conservação* 10: 102–7.

Peterson, A. T., J. Soberón, J. Ramsey, and L. Osorio-Olvera. 2020. "Co-occurrence Networks Do Not Support Identification of Biotic Interactions." *Biodiversity Informatics* 15: 1–10.

Phillips, S. J., R. P. Anderson, M. Dudík, R. E. Schapire, and M. E. Blair. 2017. "Opening the Black Box: An Open-Source Release of Maxent." *Ecography* 40: 887–93.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. "Maximum Entropy Modeling of Species Geographic Distributions." *Ecological Modelling* 190: 231–59.

Phua, M.-H., S. Tsuyuki, N. Furuya, and J. S. Lee. 2008. "Detecting Deforestation with a Spectral Change Detection Approach Using Multitemporal Landsat Data: A Case Study of Kinabalu Park, Sabah, Malaysia." *Journal of Environmental Management* 88: 784–95.

Pompe, S., J. Hanspach, F.-W. Badeck, S. Klotz, H. Bruelheide, and I. Kühn. 2010. "Investigating Habitat-Specific Plant Species Pools under Climate Change." *Basic and Applied Ecology* 11: 603–11.

Possingham, H. P., H. Grantham, and C. Rondinini. 2007. "How Can You Conserve Species that Haven't Been Found?: Commentary." *Journal of Biogeography* 34: 758–9.

Prendergast, R. 1993. "Rare Species, the Coincidence of Diversity Hotspots and Conservation Strategies." *Nature* 365: 335–7.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing* 275–86. Vienna: R Foundation for Statistical Computing.

Raes, N., M. C. Roos, J. W. F. Slik, E. E. Van Loon, and H. ter Steege. 2009. "Botanical Richness and Endemicity Patterns of Borneo Derived from Species Distribution Models." *Ecography* 32: 180–92.

Rajalingam, D. N., and K. Ranjini. 2011. "Hierarchical Clustering Algorithm – A Comparative Study." *International Journal of Computer Applications* 19: 42–6.

Roberge, J.-M., and P. Angelstam. 2004. "Usefulness of the Umbrella Species Concept as a Conservation Tool." *Conservation Biology* 18: 76–85.

Roberts, D. W. 1987. "A Dynamical Systems Perspective on Vegetation Theory." *Vegetatio* 69: 27–33.

Rousseeuw, P. J. 1986. "A Visual Display for Hierarchical Classification." In *Data Analysis and Informatics 4*, edited by E. Diday, Y. Escoufier, L. Lebart, J. Pagès, Y. Schektman, and R. Tomassone, 743–8. North-Holland: Elsevier.

Rousseeuw, P. J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20: 53–65.

Roux, M. 2018. "A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms." *Journal of Classification* 35: 345–66.

Roxburgh, S. H., and P. Chesson. 1998. "A New Method for Detecting Species Associations with Spatially Autocorrelated Data." *Ecology* 79: 2180–92.

Royle, J. A., R. B. Chandler, C. Yackulic, and J. D. Nichols. 2012. "Likelihood Analysis of Species Occurrence Probability from Presence-Only Data for Modelling Species Distributions." *Methods in Ecology and Evolution* 3: 545–54.

Rüger, N., R. Condit, D. H. Dent, S. J. DeWalt, S. P. Hubbell, J. W. Lichstein, O. R. Lopez, C. Wirth, and C. E. Farrior. 2020. "Demographic Trade-Offs Predict Tropical Forest Dynamics." *Science* 368: 165–8.

Sa'adi, Z., S. Shahid, and M. S. Shiru. 2021. "Defining Climate Zone of Borneo Based on Cluster Analysis." *Theoretical and Applied Climatology* 145: 1467–84.

Salvador, S., and P. Chan. 2004. "Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms." In *16th IEEE International Conference on Tools with Artificial Intelligence* 576–84. Boca Raton, FL: IEEE Computer Society.

Santini, L., L. H. Antão, M. Jung, A. Benítez-López, G. Rapacciuolo, M. Di Marco, F. A. M. Jones, J. M. Haghkerdar, and M. González-Suárez. 2021. "The Interface between Macroecology and Conservation: Existing Links and Untapped Opportunities." *Frontiers of Biogeography* 13: e53025.

Seif, G. 2018. "The 5 Clustering Algorithms Data Scientists Need to Know." https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.

Shipley, B., and P. A. Keddy. 1987. "The Individualistic and Community-Unit Concepts as Falsifiable Hypotheses." In *Theory and Models in Vegetation Science* 47–55. Dordrecht: Springer.

Siman, K., D. A. Friess, M. Huxham, S. McGowan, J. Drewer, L. P. Koh, Y. Zeng, et al. 2021. "Nature-Based Solutions for Climate Change Mitigation: Challenges and Opportunities for the ASEAN Region." British High Commission and the COP26 Universities Network, 1–36.

Singh, N., and D. Singh. 2012. "Performance Evaluation of K-Means and Hierarchical Clustering in Terms of Accuracy and Running." *Time* 3: 3.

Slik, J. W. F. 2009. "Plants of Southeast Asia." https://asianplant.net/.

Slik, J. W. F., S.-I. Aiba, M. Bastian, F. Q. Brearley, C. H. Cannon, K. A. O. Eichhorn, G. Fredriksson, et al. 2011. "Soils on Exposed Sunda Shelf Shaped Biogeographic Patterns in the Equatorial Forests of Southeast Asia." *Proceedings of the National Academy of Sciences* 108: 12343–7.

Slik, J. W. F., A. D. Poulsen, P. S. Ashton, C. H. Cannon, K. A. O. Eichhorn, K. Kartawinata, I. Lanniari, et al. 2003. "A Floristic Analysis of the Lowland Dipterocarp Forests of Borneo." *Journal of Biogeography* 30: 1517–31.

Slik, J. W. F., N. Raes, S.-I. Aiba, F. Q. Brearley, C. H. Cannon, E. Meijaard, H. Nagamasu, et al. 2009. "Environmental

Correlates for Tropical Tree Diversity and Distribution Patterns in Borneo." *Diversity and Distributions* 15: 523–32.

Sloan, S., P. Meyfroidt, T. K. Rudel, F. Bongers, and R. Chazdon. 2019. "The Forest Transformation: Planted Tree Cover and Regional Dynamics of Tree Gains and Losses." *Global Environmental Change* 59: 101988.

Soberon, J., and A. T. Peterson. 2005. "Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas." *Biodiversity Informatics* 2: 1–10.

Sokal, R. R., and C. D. Michener. 1958. "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Scientific Bulletin* 38: 1409–38.

Sokal, R. R., and F. J. Rohlf. 1962. "The Comparison of Dendrograms by Objective Methods." *TAXON* 11: 33–40.

Stolar, J., and S. E. Nielsen. 2015. "Accounting for Spatially Biased Sampling Effort in Presence-Only Species Distribution Modelling." *Diversity and Distributions* 21: 595–608.

Struebig, M. J., A. Wilting, D. L. A. Gaveau, E. Meijaard, R. J. Smith, M. Fischer, K. Metcalfe, et al. 2015. "Targeted Conservation to Safeguard a Biodiversity Hotspot from Climate and Land-Cover Change." *Current Biology* 25: 372–8.

Sullivan, M. J. P., S. L. Lewis, K. Affum-Baffoe, C. Castilho, F. Costa, A. C. Sanchez, C. E. N. Ewango, et al. 2020. "Long-Term Thermal Sensitivity of Earth's Tropical Forests." *Science* 368: 869–74.

Sullivan, M. J. P., J. Talbot, S. L. Lewis, O. L. Phillips, L. Qie, S. K. Begne, J. Chave, et al. 2017. "Diversity and Carbon Storage across the Tropical Forest Biome." *Scientific Reports* 7: 39102.

Tikhonov, G., N. Abrego, D. Dunson, and O. Ovaskainen. 2017. "Using Joint Species Distribution Models for Evaluating How Species-to-Species Associations Depend on the Environmental Context." *Methods in Ecology and Evolution* 8: 443–52.

Titeux, N., K. Henle, J.-B. Mihoub, A. Regos, I. R. Geijzendorffer, W. Cramer, P. H. Verburg, and L. Brotons. 2017. "Global Scenarios for Biodiversity Need to Better Integrate Climate and Land Use Change." *Diversity and Distributions* 23: 1231–4.

Torres, R., N. I. Gasparri, P. G. Blendinger, and H. R. Grau. 2014. "Land-Use and Land-Cover Effects on Regional Biodiversity Distribution in a Subtropical Dry Forest: A Hierarchical Integrative Multi-Taxa Study." *Regional Environmental Change* 14: 1549–61.

Trabucco, A., and R. J. Zomer. 2010. "Global Soil Water Balance Geospatial Database." CGIAR Consortium for Spatial Information.

Trabucco, A., and R. J. Zomer. 2018. "Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2." CGIAR Consortium for Spatial Information 10.

Trisos, C. H., C. Merow, and A. L. Pigot. 2020. "The Projected Timing of Abrupt Ecological Disruption from Climate Change." *Nature* 580: 496–501.

Van der Laan, M., K. Pollard, and J. Bryan. 2003. "A New Partitioning around Medoids Algorithm." *Journal of Statistical Computation and Simulation* 73: 575–84.

VanDerWal, J., L. P. Shoo, C. Graham, and S. E. Williams. 2009. "Selecting Pseudo-Absence Data for Presence-Only Distribution Modeling: How Far Should You Stray from What you Know?" *Ecological Modelling* 220: 589–94.

Velazco, S. J. E., F. Villalobos, F. Galvão, and P. De Marco Júnior. 2019. "A Dark Scenario for Cerrado Plant Species: Effects of Future Climate, Land Use and Protected Areas Ineffectiveness." *Diversity and Distributions* 25: 660–73.

Villalobos, F., A. Lira-Noriega, J. Soberón, and H. T. Arita. 2013. "Range–Diversity Plots for Conservation Assessments: Using Richness and Rarity in Priority Setting." *Biological Conservation* 158: 313–20.

Villalobos, F., M. Á. Olalla-Tárraga, M. V. Cianciaruso, T. F. Rangel, and J. A. F. Diniz-Filho. 2017. "Global Patterns of Mammalian co-Occurrence: Phylogenetic and Body Size Structure within Species Ranges." *Journal of Biogeography* 44: 136–46.

Vollering, J., R. Halvorsen, I. Auestad, and K. Rydgren. 2019. "Bunching up the Background Betters Bias in Species Distribution Models." *Ecography* 42: 1717–27.

Ward, J. H. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58: 236–44.

Warren, D. L., L. J. Beaumont, R. Dinnage, and J. B. Baumgartner. 2019. "New Methods for Measuring ENM Breadth and Overlap in Environmental Space." *Ecography* 42: 444–6.

Warren, D. L., R. E. Glor, and M. Turelli. 2008. "Environmental Niche Equivalency versus Conservatism: Quantitative Approaches to Niche Evolution." *Evolution* 62: 2868–83.

Webb, N. R. 1989. "Studies on the Invertebrate Fauna of Fragmented Heathland in Dorset, UK, and the Implications for Conservation." *Biological Conservation* 47: 153–65.

Wedley, W. C., B. Schoner, and E. U. Choo. 1993. "Clustering, Dependence and Ratio Scales in AHP: Rank Reversals and Incorrect Priorities with a Single Criterion." *Journal of Multi-Criteria Decision Analysis* 2: 145–58.

Westman, W. E. 1985. "Xeric Mediterranean-Type Shrubland Associations of Alta and Baja California and the Community/Continuum Debate." In *Plant Community Ecology: Papers in Honor of Robert H. Whittaker*, edited by R. K. Peet, 79–95. Dordrecht: Springer Netherlands.

Whittaker, R. H. 1951. "A Criticism of the Plant Association and Climatic Climax Concepts." *Northwest Scientist* 25: 17–31.

Whittaker, R. H. 1953. "A Consideration of Climax Theory: The Climax as a Population and Pattern." *Ecological Monographs* 23: 41–78.

Whittaker, R. J., M. B. Araújo, P. Jepson, R. J. Ladle, J. E. M. Watson, and K. J. Willis. 2005. "Conservation Biogeography: Assessment and Prospect." *Diversity and Distributions* 11: 3–23.

Wirth, M., G. F. Estabrook, and D. J. Rogers. 1966. "A Graph Theory Model for Systematic Biology, with an Example for the Oncidiinae (Orchidaceae)." *Systematic Zoology* 15: 59–69.

Wong, C. J., D. James, N. A. Besar, K. U. Kamlun, J. Tangah, S. Tsuyuki, and M.-H. Phua. 2020. "Estimating Mangrove Above-Ground Biomass Loss Due to Deforestation in Malaysian Northern Borneo between 2000 and 2015 Using SRTM and Landsat Images." *Forests* 11: 1018.

Wooster, M. J., G. L. W. Perry, and A. Zoumas. 2012. "Fire, Drought and El Niño Relationships on Borneo (Southeast Asia) in the Pre-MODIS Era (1980–2000)." *Biogeosciences* 9: 317–40.

Wüest, R. O., N. E. Zimmermann, D. Zurell, J. M. Alexander, S. A. Fritz, C. Hof, H. Kreft, et al. 2020. "Macroecology in the Age of Big Data – Where to Go from Here?" *Journal of Biogeography* 47: 1–12.

Yanahan, A. D., and W. Moore. 2019. "Impacts of 21st-Century Climate Change on Montane Habitat in the Madrean Sky Island Archipelago." *Diversity and Distributions* 25: 1625–38.

Zurell, D., J. Franklin, C. König, P. J. Bouchet, C. F. Dormann, J. Elith, G. Fandos, et al. 2020. "A Standard Protocol for Reporting Species Distribution Models." *Ecography* 43: 1261–77.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Pang, Sean E. H., J. W. Ferry Slik, Damaris Zurell, and Edward L. Webb. 2023. "The Clustering of Spatially Associated Species Unravels Patterns in Tropical Tree Species Distributions." *Ecosphere* 14(6): e4589. https://doi.org/10.1002/ecs2.4589