Spatial occurrence records and distributions of tropical Asian butterflies

*Eugene Yu Hin Yau[1], *Emily E. Jones[1], Toby Pak Nok Tsang[1,2], Shuang Xing[1,3], Richard T. Corlett[4], Patrick Roehrdanz[5], David J. Lohman [6], Adam Kai Chi Lee[1], Catherine Wai Ching Hai[1], Shawan Chowdhury[7,8,9,10], Jane K. Hill[11], Jade A. T. Badon[12], Cheong Weei Gan[13], Yves Basset[14], I-Ching Chen[15], Suzan Benedick[16], Anuj Jain[13,17], Tiffany L.T. Ki[11,18], Krushnamegh Kunte[19], Akihiro Nakamura[20], Lien Van Vu[21], Sarah A. Scriven[11], Alice C. Hughes[1], Timothy C. Bonebrake[1#]

*authors contributed equally, #corresponding author

[1]School of Biological Sciences, The University of Hong Kong, Pokfulam, Hong Kong SAR, CN
[2]The University of Toronto Scarborough, 1265 Military Trail, Scarborough, ON M1C 1A4, CA
[3]School of Ecology, Shenzhen Campus of Sun Yat-sen University; Shenzhen 518107, China.
[4]Center for Integrative Conservation and Yunnan Key Laboratory for the Conservation of Tropical Rainforests and Asian Elephants, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences; Honorary Research Associate, Royal Botanic Gardens, Kew.
[5]Betty and Gordon Moore Center for Science, Conservation International, Arlington, VA, USA
[6]City College of New York, City University of New York, 160 Convent Avenue New York, NY 10031, USA; PhD Program in Biology, City University of New York, 365 Fifth Avenue, New York, NY 10016; Zoology Division, National Museum of Natural History, Rizal Park, Manila 1000, Philippines
[7]Institute of Biodiversity, Friedrich Schiller University Jena, Dornburger Straße 159, 07743 Jena, Germany
[8]Department of Biodiversity and People, Helmholtz Centre for Environmental Research – UFZ, Permoserstraße 15, 04318 Leipzig, Germany
[9]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstraße 4, 04103 Leipzig, Germany
[10]Czech University of Life Sciences Prague, Faculty of Environmental Sciences, Prague, Czech Republic
[11]Leverhulme Centre for Anthropocene Biodiversity, Department of Biology, University of York, York, YO10 5DD, UK
[12]Animal Biology Division, Institute of Biological Sciences, University of the Philippines Los Baños, Laguna 4031, Philippines
[13]Nature Society Singapore, 510 Geylang Road, 389466, Singapore
[14]Smithsonian Tropical Research Institute, Apartado 0843-03092, Balboa, Ancon, Panama
[15]Department of Life Sciences, National Cheng Kung University, Taiwan
[16]Faculty of Sustainable Agriculture, Universiti Malaysia Sabah, Locked Bag No. 3, 90509, Sandakan, Sabah, Malaysia
[17]bioSEA Pte Ltd., 68 Chestnut Avenue, 679521, Singapore
[18]Science Department, Natural History Museum, London SW7 5BD, United Kingdom
[19]National Centre for Biological Sciences (NCBS), Tata Institute of Fundamental Research (TIFR), GKVK Campus, Bellary Road, Bengaluru 560065, India
[20]CAS Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Menglun, China

[21]Vietnam National Museum of Nature, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Ha Noi, Vietnam

corresponding author(s): Timothy C. Bonebrake (tbone@hku.hk)

Abstract

Insect biogeography is poorly documented globally, particularly in the tropics. Recent intensive research in tropical Asia, combined with increasingly available records from citizen science, provides an opportunity to map the distributions of tropical Asian butterflies. We compiled a dataset of 724,247 occurrences of 3,591 tropical Asian butterfly species by aggregating records from GBIF (651,285 records), published literature (21,271), published databases (37,695), and unpublished data (13,993). Here, we present this dataset and single-species distribution maps of 1,520 species. Using these maps, along with records of the 2,071 remaining species, we identified areas of limited sampling (e.g., the Philippines, Myanmar, and New Guinea) and predicted areas of high diversity (Peninsular Malaysia and Borneo). This dataset can be leveraged for a range of studies on Asian and tropical butterflies, including 1) species biogeography, 2) sampling prioritization to fill gaps, 3) biodiversity hotspot mapping, and 4) conservation evaluation and planning. We encourage the continued development of this dataset and the associated code as a tool for the conservation of tropical Asian insects.

Background & Summary

Tropical Asia, home to multiple major global biodiversity hotspots, harbors a rich assemblage of highly range-restricted endemic species[1]. Unfortunately, reliable distribution data for many species in this region are scarce[2]. One prominent challenge for invertebrate conservation, known as the Wallacean shortfall, stems from our inadequate knowledge of species distributions[3]. Insufficient information on species distributions impedes the identification of vulnerable species and the efficient allocation of conservation resources across regions and species[3,4].

While recent global studies of butterfly biogeography have incorporated data from tropical Asia[5,6], they have primarily relied on coarse, country-level data to examine biogeographic patterns[5-7]. The distribution information summarized based on those data is largely influenced by political boundaries rather than relevant ecological areas and is less ideal for identifying important conservation/vulnerable areas, which requires fine-scale, biogeographic data with low bias[8]. There have also been attempts to map spatial phylogenetic diversity using range maps[9], but the quality of such spatial analyses is highly dependent on the range maps used, which often fail to capture distribution patterns at local scales, thereby limiting the resolution of the spatial pattern of interest. Although fine-scale geographic distributions of several Asian butterfly groups have been mapped (e.g., *Elymnias* in Wei et al.[10]; *Papilio* in Condamine et al.[11]; *Polyura* in Toussaint et al.[12]; range-restricted butterflies in Scriven et al.[13]), to date, no unified, fine-scale distribution dataset has been produced for the entire region – despite the importance of such a tool for examining patterns of diversity within this highly biodiverse region[1,6]. Existing locality data might not be readily accessible and frequently require aggregation and standardization. Fine-grained information on species distributions is an essential first step for understanding insect biodiversity patterns and conservation needs.

The creation of regional datasets of species distributions is aided by the recent development of large, open-source biodiversity data platforms such as the Global Biodiversity Information Facility (hereafter, GBIF), an online database that organizes crowd-sourced data from citizen science platforms, scientific literature, and specimen collections[14]. These data, however, often include large spatial biases due to uneven sampling and data mobilization efforts among regions[14,15]. Even if available, much of the fine-scale biogeographic data that could be employed to reduce these biases remains buried in literature and regional databases, requiring concerted efforts to make it analysis-ready[7]. Without unified and standardized datasets, it is difficult to test macroecological and macroevolutionary questions[16], produce high-quality species distribution models[17], and identify effective conservation targets[5,6,8,18].

The process of mapping species distributions can be accomplished either through data-driven modeling or by relying on expert knowledge. Range maps (expert range maps) solely based on expert knowledge tend to overestimate active areas of species at the local scale[15,17,19]. In addition, the quality of their source data, hence the uncertainty of the analysis, is often unknown[16]. The dependency of range maps on expert knowledge means this method is available to a small subset of well-studied species[7]. In contrast, data-driven distribution maps offer greater transparency and reproducibility[18,20]. Modern modeling techniques allow the interpolation of potential distributions into areas for which primary data collection may not be possible, enabling the production of more detailed and reliable distribution maps[3,21]. However, major data gaps exist for occurrence records of most taxa[16,22], particularly invertebrates, and the non-random distribution of these gaps necessitates careful treatment within models[23].

Species distribution maps facilitate the identification of species ranges and hotspots of diversity. This provides valuable insights for local conservation planning/prioritization[24,25] and policy-making, paving the way for future investigations into butterfly biogeography[5] and phylogeographic patterns[24]. Specifically, species distribution maps can guide the allocation of conservation resources, inform the strategic design of protected areas in high-suitability/biologically diverse areas, and identify low-suitability areas in need of management[25,26], enabling effective conservation interventions. In conjunction with SDMs, occurrence datasets can help inform species reintroduction programs by identifying potentially suitable areas[25,27] and optimal source populations[28], and expedite IUCN Red List assessment, which has poor species coverage in Asia. Additionally, applications of SDMs include the modeling of species and community-level responses to climate change[24,27,29] and the assessment of extinction risks[30].

The need for species conservation is particularly acute in tropical Asia, defined broadly here to include South Asia and Southeast Asia (see Fig. 1). The area is home to over 20,000 islands, many of which were repeatedly connected and separated from adjacent landmasses during drastic sea-level fluctuations over the past 4 Mya[31]. This dynamic past led to the evolution of numerous species endemic to single islands or island groups, and as such this region hosts some of the world's greatest biodiversity – an estimated 15-25% of all well-studied terrestrial taxa and a large proportion of undescribed taxa[32,33]. This highly biodiverse region is also one of the globe's most biologically threatened: it is estimated that 42% of Southeast Asia's biodiversity may be lost by 2100 as three quarters of its primary forests are lost to agriculture, urbanization, and mineral extraction[32,34,35].

We present a comprehensive dataset of tropical Asian butterflies, with more than half of the records possessing high spatial accuracy (uncertainty < 10 km). This fills a major sampling gap, given that Asia is poorly represented in global biodiversity data repositories[15,22,36], improved datasets are urgently needed to enable effective monitoring and management of biodiversity across the region. Leveraging the data along with tailored species distribution models (SDMs), we generate data-driven distribution maps at a resolution of 10 km x 10 km. These maps enhance a fundamental understanding of butterfly macroecological patterns in tropical Asia. Each butterfly species' distribution was individually modeled and, together with buffered occurrence points of unmodelled species, employed to assess regional patterns of species diversity. Combined with species distribution models, our aggregated data advances knowledge of butterfly macroecology and facilitates evidence-based decision-making for butterfly conservation in tropical Asia.

Methods

Occurrence data

We manually extracted GBIF records for tropical Asian Papilionoidea (Lepidoptera: Nymphalidae, Papilionidae, Lycaenidae, Pieridae, Hesperiidae, Riodinidae; -11.426 – 35.64 N, 67.588 – 174.990 E) for the years 1970-present on 15 April 2024 (Derived dataset GBIF.org[37]). The geographical extent of the study area was selected to encompass northern temperate Asia to secure sufficient data to capture the full niche breadth of all species in the subsequent SDMs. We included presence records derived from human observation, preserved specimens, material samples, or literature, provided they had associated coordinates. We omitted all records with >100,000 m coordinate uncertainty, so-called "fuzzy" taxon matches, and records for which the scientific name was missing or incomplete unless nomenclature could be extracted using a BOLD identifier (boldsystems.org/). This resulted in a final number of GBIF records equalling 651,285.

Roughly 73% (472,714) of these records are 'Research-grade' observations from iNaturalist. Information on how this designation is made is available at GBIF.org. The accuracy of opportunistically collected data from crowd-sourced platforms like GBIF is often diminished due to misidentifications, taxonomic, spatial, and temporal biases, as well as uneven taxonomic validation due to lack of standardized reference data[14,38-41]. Given these potential issues, and to fill geographic gaps, we supplemented these GBIF data with expert data (coauthor datasets, published literature) and harmonized binomials to a single expert dataset (Lamas, 2015. Catalogue of the butterflies (Papilionoidea), available from the author.; see below).

We extracted data from the B2D2 Database of Butterflies for Borneo provided by JKH/the Darwin Initiative (n = 19,417), a dataset for Bangladesh provided by SC (Chowdhury et al.[42]; n = 18,278), and unpublished datasets from coauthors AN, DJL, LVV, TK, and YB (n = 13,993). For geographic regions with relatively few records (e.g., China, Myanmar, Thailand) and for species with < 10 records, we conducted targeted searches of post-1970 published literature on Google Scholar in English and Chinese (simplified and traditional) (genus OR genus + species + country name), producing an additional 21,271 records. Although some publications lacked collection dates for records (e.g., checklists), we assume that the inclusion of species in recent publications is indicative of species' current localities.

For all records in published sources, we extracted coordinates, locality name, locality type (e.g., exact coordinates, city, national park, island, or province), country, and year of record (where available). If exact coordinates were not provided by the source, we used Google Earth Pro (v7.3.6.9345) to estimate the locality centroid for any record provided at the province level or below (e.g., national park or city). For records from islands ≤ 100 km at the widest dimension (e.g., localities within the Philippines and Indonesia), we estimated the island or archipelago centroid. If a range of coordinates was provided (e.g., records from The Butterflies of Vietnam), we selected a point within the range. Data sources for all records are provided in the reference column in Occurrence Records of Tropical Asian Butterflies: 1970-2024 (https://doi.org/10.6084/m9.figshare.25037645).

Final binomial harmonization, validation, and authority assignment were conducted by DJL using a taxonomic reference prepared by Gerardo Lamas (Lamas, 2015). Family names were aligned by hand to GBIF.

The resulting database consists of 724,247 occurrence records for 3,591 species from 546 genera. These records represent approximately 20% of all described butterfly species globally[43,44] (17,280-17,500 spp.; but see Pinkert et al., 2022[5]). Records of Nymphalidae (313,899; 1,324 spp.) comprise 43% of the dataset, followed by Lycaenidae (968 spp; 147,277 records), Papilionidae (264 spp.; 101,500 records), Pieridae (405 spp.; 97,120 records), Hesperiidae (611 spp.; 60,460 records), and Riodinidae (22 spp.; 3,991 records).

Of the 3,591 species in the database, 1,580 (< 31%) are represented by ≥ 10 records within the extent of 36˚ N to 10˚ S and 69˚ E to 161.6˚ E that are >10 km apart (see details on distribution modeling below). Most occurrence records are concentrated in a limited number of regions, for example, India (28.34% of all data), Taiwan (13.75% of all data), Singapore (8.92% of all data), Hong Kong (8.30% of all data), and Malaysia (7.16% of all data) (Fig. 1). Equatorial regions together with southern China are relatively underrepresented in our dataset. As much of the data is derived from GBIF, which contains a large proportion of citizen science data, we observed a clustering of our data in areas of high population and a general lack of data in more inaccessible regions.

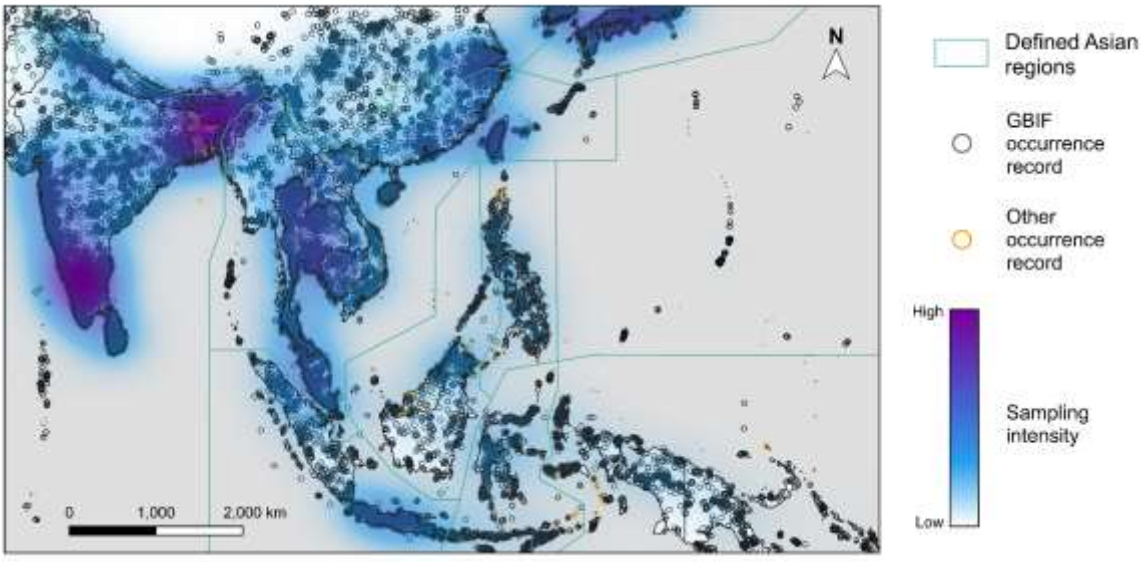


Figure 1. Distribution of GBIF and other occurrence records in our study area. Sampling intensity was estimated by running kernel density on the coordinates of all available occurrence data of every species. Regions of Asian landmasses based on the ecoregions and biogeographic realms as revised by Dinerstein et al.[45], as well as Wallace's Line, Huxley's Line, and Weber's Line.

SDM methods and results

Five algorithms, Generalized Linear Model (GLM), Maximum Entropy (MaxEnt), Multivariate Adaptive Regression Splines (MARS), Classification Tree Analysis (CTA), and eXtreme Gradient Boosting (XGBOOST) were selected to create an ensemble model for each butterfly species, using the ensemble platform "biomod2"[46] in R. We ensured that the underlying mechanism of our selection of algorithms was diverse and relatively balanced between the main categories of algorithms. We used 13 predictor variables for selection by individual models. All modeling was conducted at 10 km x 10 km resolution.

The Generalized Linear Model (GLM) is a regression-based algorithm widely used in SDMs[47]. They are not as flexible when fitting complex response curve shapes, but this also means that GLMs are less vulnerable to overfitting[47]. Maximum Entropy (MaxEnt) in our study was based on the "maxnet" R package[48], which uses penalized maximum likelihood for model fitting. MaxEnt is one of the computationally less expensive algorithms that perform well, making it a popular SDM algorithm[49]. MaxEnt is more capable of fitting complicated, non-linear response curves, enabling users to model more complex relationships by using progressively complex statistics based on the number of samples available[50]. The classification tree analysis (CTA) used by our SDM is based on the "rpart" R package[51]. The CTA algorithm recursively splits one group of data into two subgroups using one of the predictor variables given; therefore, the final model can be visualized as binary decision trees[51]. Finally, eXtreme Gradient Boosting (XGBoost) is one of the more computationally efficient gradient boosting algorithms implemented in R by the "xgboost" package[52]. Boosting algorithms feature an ensemble of weak models, each trained to minimize the errors of the previous models[47,53].

For the species distribution models, we used 13 predictor variables, which comprised 8 Bioclim variables extracted from WorldClim[54], three soil variables extracted from SoilGrids[55] through ISRIC (International Soil Reference and Information Centre)[56], and 2 vegetation variables derived from satellite data. The Bioclim variables employed included annual mean temperature (Bio 1), temperature seasonality (Bio 4), maximum temperature of warmest month (Bio 5), minimum temperature of coldest month (Bio 6), annual precipitation (Bio 12), precipitation of wettest month (Bio 13), precipitation of driest month (Bio 14), precipitation seasonality (Bio 15). The soil variables at a depth of 5-15 cm were used, including soil pH (phh2o), soil organic carbon content in the fine earth fraction (SOC), and total nitrogen (nitrogen). Nitrogen is generally recognized as one of the main limiting elements for plant growth[57], while soil organic carbon indicates soil quality[58]. In addition, soil pH exerts considerable influence on soil biogeochemical processes, ultimately impacting plant growth[59]. The selection of variables for our models was guided by expert knowledge to reflect/cover the key limitations and resources relevant to both butterflies and their host plants. Knowledge of the study region and biology/ecophysiology of the species being modeled allows the identification of the most ecologically relevant variables; therefore, it is the preferred approach for variable selection[47,49,60,61].

The vegetation variables used were the Normalized Difference Vegetation Index (NDVI) and Canopy Height. NDVI was calculated from the USGS Landsat 5 (Level 2, Collection 2, Tier 1, 1985 – 1999) and USGS Landsat 7 (Level 2, Collection 2, Tier 1, 2000 – 2020) datasets, with a customized script to filter satellite images by cloud cover (retaining images with 15% or less cloud cover over land) and to obtain the mean NDVI value. Canopy Height data was retrieved from the ETH Global Sentinel-2 10 m Canopy Height dataset[62]. These vegetation cover variables were directly used to model the land cover/habitat available to butterflies. Mean NDVI provided information on the general greenness of an area, while Canopy Height data offered structural details on vegetation to better identify different types of habitats. Together, these variables indicate resource availability and, to some extent, habitat structure. To address potential issues associated with negative values in NDVI data, an alternative variable, Corrected NDVI, which contains no negative values, was also examined. The Corrected NDVI is derived from the equation Corrected NDVI = NDVI +1. However, the SDMs using Corrected NDVI produced identical results to those using standard NDVI data, indicating that our models were unaffected by negative NDVI values.

The resolution of all environmental variables was set to 10 km x 10 km by averaging the values from contributing grid cells. This resolution was chosen as a result of balancing the spatial accuracy of available data and computational capabilities. Our data comprises 440,731 records with coordinate uncertainty data, while an additional 283,523 records that do not have coordinate uncertainty data. Among the records with known coordinate uncertainty, 73,372 (18.31% of records with uncertainty data) had uncertainties ranging from 1-10 km, and 39,302 (9.81% of records with uncertainty data) had uncertainties exceeding 10 km, thus 10 km seemed a reasonable compromise to reflect this. For the construction of SDMs, the map of the study area and predicting variables were formatted to share the same extent, resolution, and projection. We excluded entries with invalid species names (e.g., “NA” and “not present”) or outside of our study area and those recorded before 1970. Data entries published after 1970 but without date records were kept, assuming that their publication infers their validity at the time of publication. Next, the map of tropical Asia and all explanatory variable rasters were all projected to equal area projection EPSG:6933 and cropped to the extent of 36˚ N to 10˚ S and 69˚ E to 161.6˚ E to fully cover the study region. Our final cleaned database included 721,060 global records.

We used a function to further prepare the input files required by biomod2 and to generate SDMs individually for each species. Occurrence data of a species was first extracted from our butterfly occurrence dataset and used to produce a raster of resolution of 10 km x 10 km. A total of *n* cells in the raster were assigned a value of “1” to represent at least one occurrence record present in that cell, while cells with no record were assigned “n/a” instead of “0” since no true absence data is available.

Only species with $n \geq 10$ were modeled. It has been shown that SDMs based on ten occurrence points can reach 90% of the maximum possible accuracy[63], while recent studies suggest a minimum requirement of 3 to 13 occurrence points in virtual simulations and 14 to 25 occurrence points in real-world conditions to infer accurate SDMs[64]. Therefore, *n=10* was chosen as the lower limit of sample size for constructing SDMs to maximize the number of species modeled while maintaining a reasonably high predictive accuracy[63]. A total of 1,580 species met this qualification, whereas 1951 species had fewer records. For each species, occurrence records were split into three sets: 10% of the data was first reserved for model evaluation, and another 10% was then partitioned for model validation, leaving the remaining

80% of data for model calibration. The partitioning of model validation data was repeated 5 times to generate five different combinations of calibration and validation occurrence data.

Before SDM construction, pseudo-absence records were generated. Despite our efforts to fill the spatial data gaps, the sampling effort of our dataset is still spatially biased toward highly populated areas and roads due to the overwhelming number of records from GBIF and iNaturalist in our dataset (more than 80%). As part of our effort to account for biases in our data, we integrated the spatial bias of our dataset into the generation of pseudo-absence records, assuming that all species were sampled in areas with at least one occurrence record of any species. To capture such spatial bias, we created a raster layer of the spatial sampling effort for all species across our study area (shown as sampling intensity in Fig. 1), which is equivalent to the bias layer commonly used in the MaxEnt program. This was done by pooling occurrence data of all species used in our models and summarising them in a raster, then performing two-dimensional kernel density estimation (kde2d) using the R package "MASS"[65] with the default settings. We excluded cells with occurrence records and sampled the remaining study area for pseudo-absence records based on the bias layer, giving more weight to well-sampled areas, as suggested by Phillips et al.[66] and Ferrier et al.[67]. Following the recommendation of Barbet-Massin et al.[68], for calibration, validation, and evaluation data, we produced five sets of pseudo-absence data for each species, maintaining a 1:1 ratio between the number of pseudo-absence points and occurrence points in each set.

Subsequently, we constructed SDMs for each species using five different partitions of calibration and validation occurrence data, five selected algorithms, and five sets of pseudo-absence data. This resulted in a total of 125 SDM models (5 x 5 x 5). Both presence and pseudo-absence records were given equal weight during model construction to ensure a consistent prevalence of 0.5 among all species. We applied a generalized setting for all butterfly species for consistency across species, with adjustments made only to the learning rate and the number of decision trees for the XGBoost algorithm to address overfitting. Other model tuning options were retained at their default.

We generated binary outputs by maximizing True Skill Statistics (TSS), a widely used threshold-dependent index of model fit. Ensemble modeling was selected over single best models for its superior performance in rare species[69], and its robustness to uncertainties in individual models by capturing the central tendency among models[47,70,71]. We constructed an ensemble model using all single models with TSS values greater than 0.7, ensuring that only "substantial" models were included[72]. 1,520 species out of the 1,580 modeled species obtained one or more single models meeting such criteria, allowing the further construction of ensemble models. The ensemble model was generated using the mean algorithm[71], where all candidate models' probabilistic predictions were averaged without weighting. Finally, we projected the ensemble model to the current environment using the same variables when constructing the SDMs.

Ensemble models were evaluated using two metrics: TSS and Boyce index. TSS and Kappa are two of the most popular SDM threshold-dependent evaluation metrics. TSS was chosen over Kappa due to the inherent dependency of Kappa on species prevalence[73]. Since we are modeling thousands of species with differing degrees of rarity and prevalence, TSS is more appropriate for model comparison between species. TSS varies from +1 to -1, in which +1 indicates perfect agreement with evaluation data, while a TSS value close to or less than 0 indicates model performance comparable to a random model[73].

Following the suggestions of Hernandez et al.[74] and Breiner et al.[69] to use multiple evaluation measures when using presence-only data, we also calculated the Boyce index for all models built to supplement TSS. The Boyce index is capable of providing an accurate and reliable measure of model performance for models based on presence-only data[75], which is the key reason for its use in our study. Another reason for the use of the Boyce index is its lower sensitivity (correlation) to species prevalence relative to other metrics, including CVI, MaxKappa, and adjusted D2[75], while AUC and TSS also have a negative correlation with prevalence[73]. AUC was also found to produce inflated estimates of model quality when the modeled species is rare[76]. Boyce index ranges from +1 to -1, in which +1 indicates the model is of the highest quality and perfectly predicts evaluation data, while -1 indicates counter-prediction of evaluation data[75]. Boyce index with a value close to 0 indicates the model performs no better than a random model[75].

To factor biogeography into predictions and correct for biogeographic overprediction generated by our SDMs (and account for differences between fundamental and realised niches), we restrained the sampling of pseudo-absence records and distribution maps produced by our models to regions that hosted more than 1% of species points (as such regions fall within species biogeographic ranges). By incorporating biogeography into model predictions, we aimed to reflect the impact of oceans as dispersal barriers in the SDM outputs to give a more realistic estimate of species' distribution and reduce false positive predictions. We first divided the landmasses of tropical Asia into 11 regions (Fig. 1) based on the ecoregions and biogeographic realms as revised by Dinerstein et al.[45], as well as Wallace's Line, Huxley's Line, and Weber's Line. For each species, we identified regions that included at least 1% of the species occurrence records, considering them to be active regions. We then cropped the SDM-predicted distribution maps to include only the active regions specific to each species. These cropped distribution maps were stacked together to generate an alpha diversity map, which illustrates the number of species present in each 10 km x 10 km cell across tropical Asia. The stacked SDM predictions highlighted a number of locations with relatively high diversity, exceeding 600 species in some locations (Fig. 2).

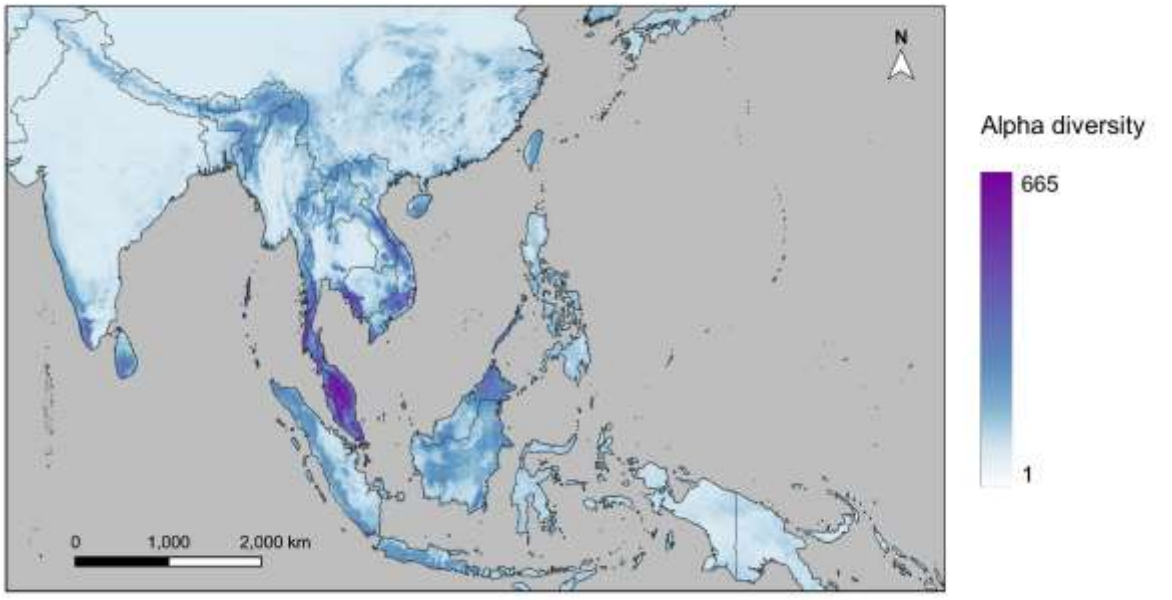


Figure 2. Projected distribution of butterfly diversity based on our species distribution models, using the mean algorithm for ensemble modeling.

Point buffer methods

For the 2,011 species (56% of all recorded species in our dataset) excluded from our species distribution modeling outputs either due to insufficient data or low quality of species distribution models, we plotted and buffered their occurrence records to infer alpha diversity. We first mapped their occurrence records and created 30 km-wide polygons (buffers) around these points to enhance clarity. Subsequently, the buffered occurrence points were converted into binary raster maps for each species and stacked to generate an additional alpha diversity map, representing species with limited occurrence records.

The diversity map derived from buffered occurrence points was then stacked with the species distribution model (SDM) projections to produce Fig. 3. This figure provides an overview of the alpha diversity of all species documented in our dataset. We identified two major butterfly diversity hotspots: peninsular Malaysia and the Sabah region of Borneo. We also found high levels of diversity predicted in Borneo, Sumatra, coastal Cambodia, southern Thailand, the Western Ghats in peninsular India, the Assam region of India, the Cardamom mountains in Cambodia, and Vietnam.

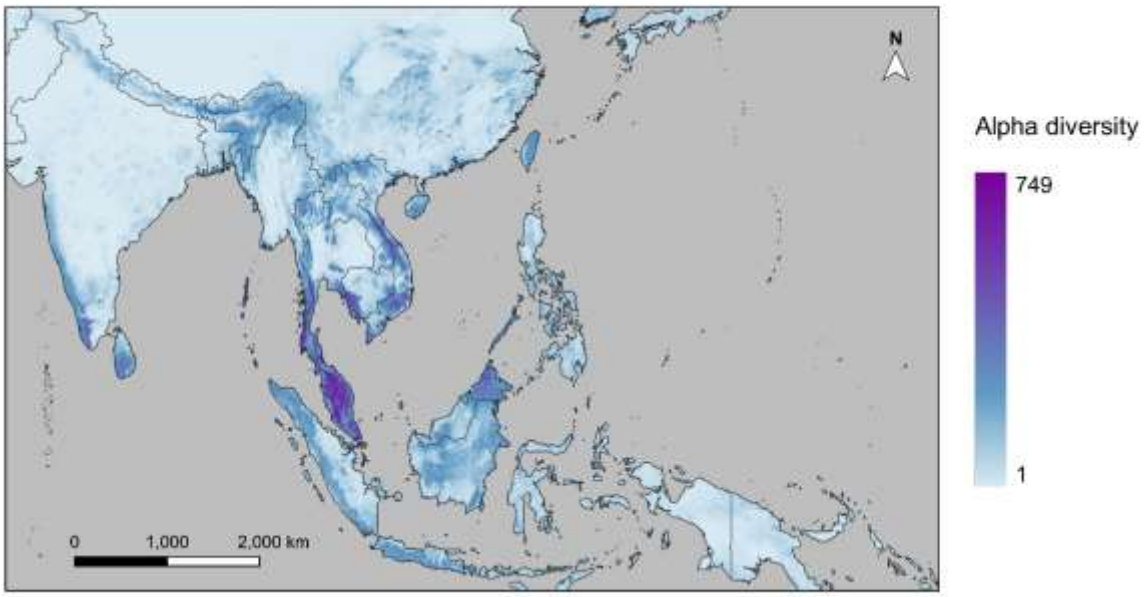


Figure 3. Estimated distribution of butterfly diversity based on our species distribution model projections and buffered occurrence points (for species not included in our SDM outputs).

Software

We calculated the SDMs in R[77], version 4.1.2. To construct and merge the SDMs into ensemble models, we utilized the "biomod2" package, version 4.2-4[46]. The high-performance computing cluster HPC2021 at The University of Hong Kong, operating on CentOS 8, was employed to run the SDMs.

Data Records

All project files are publicly available in a Figshare repository (https://doi.org/10.6084/m9.figshare.25037645). Users may access the referenced occurrence dataset and metadata as .csv files, the SDM rasters (as file type, e.g., Fig. 4), and links to R scripts for SDM construction and distribution map generation were published on GitHub: https://github.com/eugeneyau/Tropical-Asian-Butterfly-Distribution/tree/main
The GBIF-derived dataset is available at GBIF[37] (https://doi.org/10.15468/dd.nvw5wr).
These outputs are licensed under a CC BY 4.0 license.

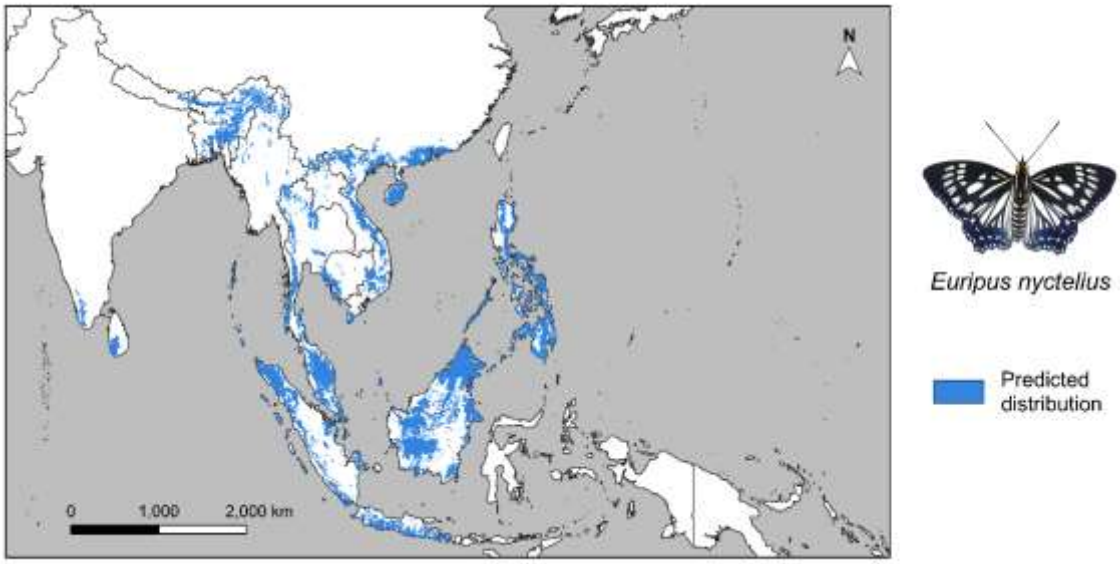


Figure 4. SDM-predicted distribution of *Euripus nyctelius* (Doubleday, 1845) (Nymphalidae: Apaturinae) based on our occurrence dataset.

Technical Validation

SDM model evaluation/verification
The mean TSS score of all ensemble models is 0.922, with a standard deviation of 0.155, while the Boyce index is 0.766, with a standard deviation of 0.305.

Both evaluation metrics indicate that the models constructed are of good quality. The mean TSS score of our ensemble models is higher than 0.8, falling into the category of "almost perfect" models according to the widely used division suggested by Landis & Koch[72] (e.g., Capinha et al.[78]; Jones et al.[79]). Since we only included models with TSS values of more than 0.7 in our ensemble models, a high mean TSS score among the ensemble models is expected. The mean Boyce index of our models is higher than 0.7, which has been considered an indicator of good models in other studies (e.g., Rupprecht et al.[80]). Boyce index value of 0.5 is usually considered a cutoff for acceptable performance[81].

Collaborator Evaluation
Our model outputs were also inspected by experts to evaluate their plausibility. Plausibility checks form an important part of model validation by making sure the modeling results

confine to the known range and possible range of the species modeled[49,82], serving as a supplement to evaluation metrics, which only measure the goodness of fit of models.

Experts (coauthors/collaborators) agreed that our model outputs are generally reasonable and informative. However, it is important to note that some of the sampling biases persisted in the final model outputs despite our efforts to address data gaps by incorporating additional datasets. We, therefore, encourage future data contributions to improve the coverage of our dataset, especially in the areas with identified data gaps.

Although the majority of data gaps can be attributed to insufficient sampling effort, the absence of data in the Philippines (and potentially other parts of tropical Asia) is primarily a result of the dominance of Facebook over other platforms like iNaturalist for citizen science data contribution. However, such data on Facebook contains limited information since EXIF data (containing GPS coordinates) of photos are removed when uploaded. Filling the Philippine data gap should be a priority, and mining Facebook data (e.g., Chowdhury et al.[18]) and other sources would be a good place to start.

Our modeling results identified the Cardamom Mountains on the Cambodian-Thai border as a butterfly diversity hotspot. During the Pleistocene when sea levels were up to 120 m lower than present, and this area was on the eastern edge of a paleoriver watershed that included the similarly diverse Malay peninsula and extended south to present-day Borneo[83,84]. The high diversity in this area is likely relictual[85]. Endemism in this area likely contributes to high butterfly diversity, which supports our models' prediction there.

Multiple experts pointed out the unexpected diversity differences between different parts of Borneo. While our models identified Sabah as a hotspot for butterfly diversity, lower diversity was predicted for other parts of Borneo, such as Sarawak and Kalimantan. This contradicted our expectations, as all these areas possess mountainous regions and endemic species, suggesting similar levels of butterfly diversity. The heart of Borneo, characterized by lower disturbance compared to other parts of the island, was also predicted to host a relatively lower diversity of butterflies by our models. Such a model prediction also contradicts our expectation of higher butterfly diversity in less disturbed areas. This inconsistency between expected and modeled butterfly diversity in Borneo is likely attributed to sampling bias, evident through the alignment of modeled butterfly diversity with political boundaries and sampling intensity (Fig. 1), and the lower modeled diversity in less accessible areas such as the heart of Borneo (Fig. 2 and 3). The lack of data in less accessible areas has been discussed by Hughes et al.[15] and Boakes et al.[86], while this trend is even more obvious in citizen science data[15], which constitutes a considerable proportion of our dataset.

While some of the spatial variations in the sampling effort of our dataset are reflected in the spatial bias of our modeling results, there are several notable discrepancies between the distribution of data and modeled diversity. Fig. 1 illustrates that Japan, Taiwan, and northern Thailand have a relatively high intensity of sampling effort compared to their predicted butterfly diversity in Fig. 2. Conversely, a reversed pattern is evident in Southern Borneo and Southern Sumatra, where our data shows low sampling effort but our models predict high butterfly diversity. These patterns demonstrate the robustness of the models to some of the spatial sampling biases present in our data.

To determine the variable importance in our SDMs, we calculated, for each variable, the mean variable importance throughout the ensemble models of all species. Temperature

seasonality (Bio 4) emerged as the most important variable (scoring 0.280 out of 1), followed by the minimum temperature of the coldest month (Bio 6, scoring 0.163 out of 1) and annual mean temperature (Bio 1, scoring 0.140 out of 1). Soil pH (phh2o, scoring 0.107 out of 1), precipitation of driest month (Bio 14, scoring 0.0973 out of 1), and Canopy Height (scoring 0.0907 out of 1) also exhibited high importance in the models. The ranking of variable importance in the SDMs conforms to the hierarchical framework of Pearson & Dawson[87], in which climatic variables exert greater control over species distribution at continental scales, while land cover and soil variables gain influence at more localized scales. In addition, the high importance of temperature variables, particularly temperature seasonality (Bio 4), is consistent with the results of Carvalho et al.[88], which highlighted the strong impact of temperature, especially temperature seasonality, on butterfly distribution and diversity.

Usage Notes

The predictor variables considered in our SDMs, which include the 8 Bioclim variables and the 3 SoilGrids variables, are products of interpolation between available point data[54,55]. As with most data collected without stratified sampling, these point data are likely to be spatially biased. Users should note that our SDMs inherit these biases, as well as uncertainties in the interpolation result.

By generating more pseudo-absences for SDMs in well-sampled areas with the use of the bias mask, we are essentially augmenting the weighting of extensively surveyed regions in our models, while unsampled habitats may be presumed as suitable. Consequently, the transferability of our models to unsampled areas is limited, especially when extrapolating in novel environments not covered by training data[66] or in areas where biogeographic barriers prevent dispersal. This is also one of the reasons for restraining our model predictions to the regions where a species is known to occur so that the results are not overly optimistic. Such an approach to pseudo-absence generation also assumes that the data collection method is consistent throughout the entire dataset[66], while our dataset is compiled from various sources. However, since a majority of our data is derived from a single source (GBIF), we can consider the data collection method consistent in terms of the observation method. To use our data and models for the prediction of future butterfly distribution under climate change, we suggest using the “random” method from the biomod2 package to generate pseudo-absence records.

Regarding uncertainty in model results, we have limited confidence in the model predictions in the Philippines and New Guinea. The scarcity of occurrence data in these two regions (Fig. 1) prevents strong inferences. Additionally, the presence of biogeographic barriers such as Wallace's Line and Huxley’s Line restricts the use of occurrence data from other regions to infer butterfly distribution in these specific areas.

Code Availability

All code used to conduct synonym harmonization, preprocess environmental variables for SDMs, execute SDMs, process SDM outputs, and conduct point buffer analysis can be accessed in our GitHub project repository: https://github.com/eugeneyau/Tropical-Asian-Butterfly-Distribution

under the Code directory. All data, including our butterfly occurrence dataset, SDM-predicted distribution maps, tropical Asian biogeographic regions shapefile, and buffered occurrence points for species excluded from our species distribution modeling are available from our FigShare repository (https://doi.org/10.6084/m9.figshare.25037645). Some data for Sulawesi, provided by TK, have not been included in the data release but are available upon request to TK.

Acknowledgments

Funding for this research was provided by a National Science Foundation China Excellent Young Scientist award to TCB. Additionally, preliminary data and analyses were supported by a Global Environmental Facility grant. The computations were performed using research computing facilities offered by Information Technology Services at The University of Hong Kong. Landsat-5 and Landsat-7 image courtesy of the U.S. Geological Survey. Special thanks to Kirsten Boehm, Ryan Leung, Xueying Wang, and Tracy Zhang for assistance with data extraction. We are very grateful to the Natural History Museum UK (NHMUK) and Zoologische Staatssammlung München (ZSM) for facilitating TK's access to the collection and would like to thank the curators for their kind support. We wish to thank B. Huertas, C. Beale, and R. Vane-Wright for their constant support of TK's work. LVV was supported by the Vietnam Ministry of Science and Technology (ĐTĐL.CN-113/21).

Author Contributions

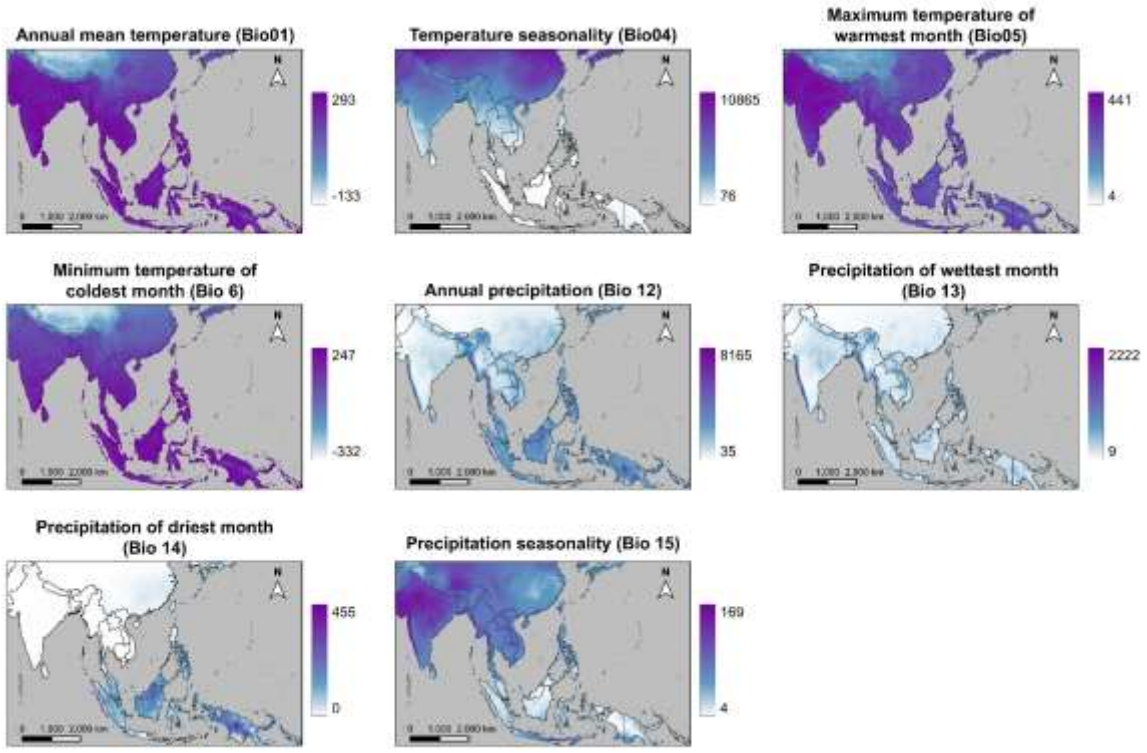
TCB, TPNT, SX, RTC, and PR conceptualized the study, with funding acquired by TCB, TPNT, and SX. The project methodology was developed by EEJ, EYHY, TPNT, SX, AKCL, RTC, PR, and ACH. TCB and ACH supervised the study, while TCB managed the project administration. The occurrence dataset was compiled by EEJ with assistance from CWCH, and data curation was carried out by EEJ, EYHY, and DJL. Data were contributed by DJL, SC, JKH, YB, TK, LVV, and AN. EYHY conducted the species distribution modeling and subsequent analyses. DJL validated scientific names in the dataset and, along with CWCH, SC, JKH, ALM, JAT, ICC, GC, GCW, SB, AJ, TK, KK, and SAS, provided insights on the plausibility of the species distribution model outputs. The initial draft of the manuscript was written by EEJ, EYHY, TCB, and ACH, with EYHY responsible for data visualization. All authors contributed input and suggestions to the draft and approved the final manuscript.

Additional information

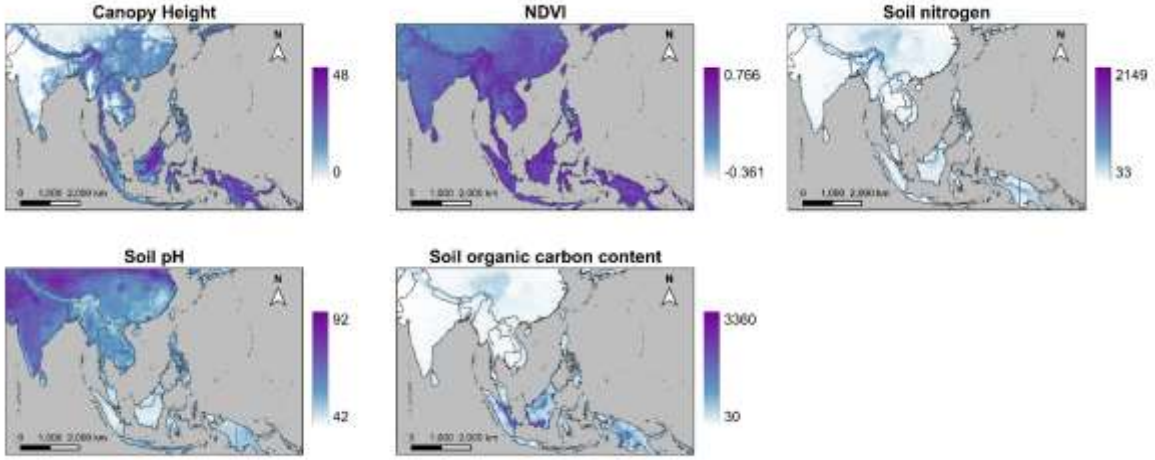
Competing interests

The authors declare no competing interests.

Supplementary Information



Supplementary Figure 1. Climatic predictor variables included in our SDMs.



Supplementary Figure 2. Non-climatic predictor variables included in our SDMs.

References

1. de Bruyn, M. et al. Borneo and Indochina are major evolutionary hotspots for Southeast Asian biodiversity. *Syst. Biol.* **63**, 879-901 (2014).
2. Verde Arregoitia, L. D. Biases, gaps, and opportunities in mammalian extinction risk research. *Mammal Rev.* **46**, 17-29 (2016).
3. Cardoso, P., Erwin, T. L., Borges, P. A. & New, T. R. The seven impediments in invertebrate conservation and how to overcome them. *Biol. Conserv.* **144**, 2647-2655 (2011).

4. Xing, Shuang. et al. Conservation of data deficient species under multiple threats: Lessons from an iconic tropical butterfly (*Teinopalpus aureus*). *Biol. Conserv.* **234**, 154-164 (2019).
5. Pinkert, S., Barve, V., Guralnick, R. & Jetz, W. Global geographical and latitudinal variation in butterfly species richness captured through a comprehensive country-level occurrence database. *Glob. Ecol. Biogeogr.* **31**, 830-839 (2022).
6. Kawahara, A. Y. et al. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nat. Ecol. Evol.* **7**, 903–913 (2023).
7. Pinkert, S., Sica, Y. V., Winner, K. & Jetz, W. The potential of ecoregional range maps for boosting taxonomic coverage in ecology and conservation. *Ecography* **2023**, (2023).
8. Whittaker, R. J. et al. Conservation biogeography: assessment and prospect. *Divers. Distrib.* **11**, 3-23 (2005).
9. Earl, C. et al. Spatial phylogenetics of butterflies in relation to environmental drivers and angiosperm diversity across North America. *IScience* **24**, (2021).
10. Wei, C.-H., Lohman, D. J., Peggie, D. & Yen, S.-H. An illustrated checklist of the genus Elymnias Hübner, 1818 (Nymphalidae, Satyrinae). *Zookeys* **676**, 47-152. (2017).
11. Condamine, F. L. et al. Fine-scale biogeographical and temporal diversification processes of peacock swallowtails (*Papilio* subgenus *Achillides*) in the Indo-Australian Archipelago. *Cladistics* **29**, 88–111 (2012).
12. Toussaint, E. F. et al. Comparative molecular species delimitation in the charismatic Nawab butterflies (Nymphalidae, Charaxinae, *Polyura*). *Mol. Phylogenet. Evol.* **91**, 194-209 (2015).
13. Scriven, S. A. et al. Assessing the effectiveness of protected areas for conserving range-restricted rain forest butterflies in Sabah, Borneo. *Biotropica* **52**, 380-391 (2020).
14. Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecol. Inform.* **19**, 10-15 (2014).
15. Hughes, A. C. et al. Sampling biases shape our view of the natural world. *Ecography* **44**, 1259-1269 (2021).
16. Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* **6**, 1-8 (2015).
17. Peterson, A. T., Navarro-Sigüenza, A. G. & Gordillo, A. Assumption-versus data-based approaches to summarizing species' ranges. *Conserv. Biol.* **32**, 568-575 (2016).
18. Chowdhury, S. et al. Using social media records to inform conservation planning. *Conserv. Biol.* **38**, (2024).
19. Jetz, W., Sekercioglu, C. H. & Watson, J. E. Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* **22**, 110-119 (2008).
20. Sofaer, H. R. et al. Development and delivery of species distribution models to inform decision-making. *Biosci.* **69**, 544-557 (2019).
21. Anderson, R. P. et al. *Ecological niches and geographic distributions* (Princeton University Press, 2011).
22. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, (2017).
23. Kramer-Schadt, S. et al. The importance of correcting for sampling bias in MaxEnt species distribution models. *Divers. Distrib.* **19**, 1366-1379 (2013).
24. Guillera-Arroita, G. et al. Is my species distribution model fit for purpose? Matching data and models to applications. *Glob. Ecol. Biogeogr.* **24**, 276-292 (2015).

25. Smeraldo, S. et al. Species distribution models as a tool to predict range expansion after reintroduction: A case study on Eurasian beavers (Castor fiber). *J. Nat. Conserv.* **37**, 12-20 (2017).
26. Guisan, A. et al. Predicting species distributions for conservation decisions. *Ecol. Lett.* **16**, 1424-1435 (2013).
27. Araújo, M. B. et al. Standards for distribution models in biodiversity assessments. *Sci. Adv.* **5**, (2019).
28. Maes, D. et al. The potential of species distribution modelling for reintroduction projects: the case study of the Chequered Skipper in England. *J. Insect Conserv.* **23**, 419-431 (2019).
29. Pacifici, M. et al. Assessing species vulnerability to climate change. *Nat. Clim. Change* **5**, 215-224 (2015).
30. Attorre, F. et al. How to include the impact of climate change in the extinction risk assessment of policy plant species? *J. Nat. Conserv.* **44**, 43-49 (2018).
31. Lohman, D. J. et al. Biogeography of the Indo-Australian archipelago. *Annu. Rev. Ecol. Evol. Syst.* **42**, 205-226 (2011).
32. Hughes, A. C. Understanding the drivers of Southeast Asian biodiversity loss. *Ecosphere* **8**, (2017).
33. Corlett, R. T. *The Ecology of Tropical East Asia* 3rd edn (Oxford University Press, 2019).
34. Sodhi, N. S., Koh, L. P., Brook, B. W. & Ng, P. K. L. Southeast Asian biodiversity: an impending disaster. *Trends Ecol. Evol.* **19**, 654-660 (2004).
35. Wilcove, D. S., Giam, X., Edwards, D. P., Fisher, B., & Koh, L. P. Navjot's nightmare revisited: logging, agriculture, and biodiversity in Southeast Asia. *Trends Ecol. Evol.* **28**, 531-540 (2013).
36. Orr, M. C. et al. Global patterns and drivers of bee distribution. *Curr Biol.* **31**, 451-458 (2021).
37. *Global Biodiversity Information Facility* https://doi.org/10.15468/dd.nvw5wr (2024).
38. Ball-Damerow, J. E. et al. Research applications of primary biodiversity databases in the digital age. *PloS one* **14**, (2019).
39. Gaiji, S. et al. Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. *Biodiversity Informatics* **8**, (2013).
40. Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z. Q. & Bourne, P. E. Biodiversity data should be published, cited, and peer reviewed. *Trends Ecol. Evol.* **28**, 454-461 (2013).
41. Goodwin, Z. A., Harris, D. J., Filer, D., Wood, J. R. & Scotland, R. W. Widespread mistaken identity in tropical plant collections. *Curr. Biol.* **25**, R1066-R1067 (2015).
42. Chowdhury, S. et al. Butterflies are weakly protected in a mega-populated country, Bangladesh. *Glob. Ecol. Conserv.* **26**, (2021).
43. Shields, O. World numbers of butterflies. *J. Lepid. Soc.* **43**, 178-183 (1989).
44. Robbins, R. K. & Opler, P. A. in *Biodiversity II: Understanding and Protecting Our Biological Resources* (eds. Reaka-Kudla, M. L., Wilson, D. E. & Wilson, E. O.) Ch. 6 (Joseph Henry Press, 1997).
45. Dinerstein, E. et al. An ecoregion-based approach to protecting half the terrestrial realm. *Biosci.* **67**, 534–545 (2017).
46. Thuiller, W. et al. biomod2: Ensemble Platform for Species Distribution Modeling. R package version 4.2-4. https://CRAN.R-project.org/package=biomod2 (2023).
47. Guisan, A., Thuiller, W. & Zimmermann, N. E. *Habitat Suitability and Distribution Models: With Applications in R* (Cambridge University Press, 2017).

48. Phillips, S. J. maxnet: Fitting 'Maxent' Species Distribution Models with 'glmnet'. R package version 0.1.4. https://CRAN.R-project.org/package=maxnet (2021).
49. Porfirio, L. L. et al. Improving the use of species distribution models in conservation planning and management under climate change. *PloS One* **9**, (2014).
50. Elith, J. & Graham, C. H. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography* **32**, 66-77 (2009).
51. Therneau, T., Atkinson, B. & Ripley, B. Rpart: Recursive Partitioning. R Package version 4.1-3. http://CRAN.R-project.org/package=rpart (2013).
52. Chen, T. et al. xgboost: Extreme Gradient Boosting. R package version 1.7.5.1. https://CRAN.R-project.org/package=xgboost (2023).
53. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **7**, (2013).
54. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965-1978 (2005).
55. Poggio, L. et al. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL* **7**, 217-240 (2021).
56. *International Soil Reference and Information Centre* https://www.isric.org/ (2024).
57. Ågren, G. I., Wetterstedt, J. M. & Billberger, M. F. Nutrient limitation on terrestrial plant growth–modelling the interaction between nitrogen and phosphorus. *New Phytol.* **194**, 953-960 (2012).
58. Lal, R. Soil health and carbon management. *Food Energy Secur.* **5**, 212-222 (2016).
59. Neina, D. The role of soil pH in plant nutrition and soil remediation. *Appl. Environ. Soil Sci.* **2019**, 1-9 (2019).
60. Barbet-Massin, M. & Jetz, W. A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling. *Divers. Distrib.* **20**, 1285-1295 (2014).
61. Zeng, Y., Low, B. W. & Yeo, D. C. Novel methods to select environmental variables in MaxEnt: A case study using invasive crayfish. *Ecol. Model.* **341**, 5-13 (2016).
62. Lang, N., Jetz, W., Schindler, K. & Wegner, J. D. A high-resolution canopy height model of the Earth. *Nat. Ecol. Evol.* **7**, 1778-1789 (2023).
63. Stockwell, D. R. & Peterson, A. T. Effects of sample size on accuracy of species distribution models. *Ecol. Model.* **148**, 1-13 (2002).
64. van Proosdij, A. S., Sosef, M. S., Wieringa, J. J. & Raes, N. Minimum required number of specimen records to develop accurate species distribution models. *Ecography* **39**, 542-552 (2016).
65. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S-PLUS* 4th edn (Springer, 2002).
66. Phillips, S. J. et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* **19**, 181-197 (2009).
67. Ferrier, S., Watson, G., Pearce, J. & Drielsma, M. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity Conserv.* **11**, 2275-2307 (2002).
68. Barbet-Massin, M., Jiguet, F., Albert, C. H. & Thuiller, W. Selecting pseudo-absences for species distribution models: How, where and how many?. *Methods Ecol. Evol.* **3**, 327-338 (2012).
69. Breiner, F. T., Guisan, A., Bergamini, A. & Nobis, M. P. Overcoming limitations of modelling rare species by using ensembles of small models. *Methods Ecol. Evol.* **6**, 1210-1218 (2015).

70. Araújo, M. B. & New, M. Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **22**, 42-47 (2007).
71. Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K. & Thuiller, W. Evaluation of consensus methods in predictive species distribution modelling. *Divers. Distrib.* **15**, 59-69 (2009).
72. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174 (1977).
73. Allouche, O., Tsoar, A. & Kadmon, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* **43**, 1223-1232 (2006).
74. Hernandez, P. A., Graham, C. H., Master, L. L. & Albert, D. L. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* **29**, 773-785 (2006).
75. Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. Evaluating the ability of habitat suitability models to predict species presences. *Ecol. Model.* **199**, 142-152 (2006).
76. Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145-151 (2008).
77. R Core Team. R: A language and environment for statistical computing, version 4.1.2. R Foundation for Statistical Computing https://www.R-project.org/ (2021).
78. Capinha, C., Rocha, J. & Sousa, C. A. Macroclimate determines the global range limit of Aedes aegypti. *Ecohealth* **11**, 420-428 (2014).
79. Jones, C. C., Acker, S. A. & Halpern, C. B. Combining local-and large-scale models to predict the distributions of invasive plant species. *Ecol. Appl.* **20**, 311-326 (2010).
80. Rupprecht, F., Oldeland, J. & Finckh, M. Modelling potential distribution of the threatened tree species Juniperus oxycedrus: how to evaluate the predictions of different modelling approaches? *J. Veg. Sci.* **22**, 647-659 (2011).
81. Pomoim, N., Hughes, A. C., Trisurat, Y. & Corlett, R. T. Vulnerability to climate change of species in protected areas in Thailand. *Sci. Rep.* **12**, (2022).
82. Zurell, D. A standard protocol for reporting species distribution models. *Ecography* **43**, 1261-1277 (2020).
83. Sholihah, A. et al. Impact of Pleistocene eustatic fluctuations on evolutionary dynamics in Southeast Asian biodiversity hotspots. *Syst. Biol.* **70**, 940-960 (2021).
84. Voris, H. K. Maps of Pleistocene sea levels in Southeast Asia: Shorelines, river systems and time durations. *J. Biogeogr.* **27**, 1153-1167 (2000).
85. Monastyrskii, A. L. & Vane-Wright, R. I. Identity of *Euploea orontobates* Fruhstorfer, 1910 (Lepidoptera: Nymphalidae), a milkweed butterfly from Thailand and Vietnam. *Zootaxa* **1991**, 43-50 (2009).
86. Boakes, E. H. et al. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PloS Biol.* **8**, (2010).
87. Pearson, R. G. & Dawson, T. P. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob. Ecol. Biogeogr.* **12**, 361-371 (2003).
88. Carvalho, A. P. S. et al. Comprehensive phylogeny of Pieridae butterflies reveals strong correlation between diversification and temperature. *iScience* **27**, (2024).