Appendix 1: Extended methods

This is an extended version of part 2.1 of the main text, including details and graphs that had to be omitted from the main text for space constraints. As we strive to provide here a comprehensive, detailed protocol of methods, some redundancies with the main text do occur.

Content

S1.1 Compiling and processing the distribution record database

Globally ca. 1470 species of Sphingidae are known (Kitching and Cadiou 2000, and recent updates), whereas outside the Americas (i.e., our study area) 981 autochtonous taxa are recognized. Due to a need to reduce workload for the project, we restricted our efforts to these species. There is almost no overlap with the Americas (only *Hyles galli* is autochtonous to both regions), which made this geographic split feasible. Sphingidae are an attractive group to both amateur collectors and taxonomists, and as a consequence more is known about their distribution, biology and taxonomy than for most other invertebrate groups.

We extracted distribution records for sphingid moths from published literature (>1300 references, ES2), supplemented by own field sampling, and we used distribution records from online data bases such as GBIF (www.gbif.org, Nov. 2009) and BOLD (www.barcodinglife.org, Aug. 2010). Furthermore, we retrieved specimen label data from >400 natural history collections (ES2), from amateur collectors, and from online searches. Distributional data from our earlier online projects on sphingids and their distribution were fully integrated in this data compilation.

We ignored all records that we considered erroneous (locality or taxon). Furthermore, we ignored records that were known to be rare vagrant specimens or where we could reasonably infer that single specimens were transported by human traffic far out of their native range. This included several New World species that only recently, aided by human transport, established populations in our study region (e.g., *Agrius cingulata* in Africa; Ballesteros-Mejia et al. 2011; *Darapsa myron* in Thailand; several taxa on Hawaii). Although we produced range maps for the autochtonous species of the Eastern Pacific, we restricted presentation and analyses of data to species ranges west of 180°E (i.e., excluding data for islands of the Eastern Pacific), due to lack of consistent environmental data for analysis. This implicitly excluded available data for 9 species endemic to that region from analyses (hence, data presented refer to 972 species).

For the majority of species, we followed the nomenclature of Kitching and Cadiou (2000) and more recent taxonomic publications. However, taxonomic findings can have a considerable time-lag until reaching official status according to the International Code of Zoological Nomenclature (ICZN 1999), so we allowed deviations for the purposes of this data compilation. In particular, we did not consider some recent descriptions where we were quite sure that they were erroneous (although they are not yet refuted in publication), whereas we accepted some recent splits and revisions based on compelling evidence even if not (yet) published. We adjusted all nomenclature to this system, but in some cases (e.g. *Hippotion boerhaviae*-complex) we did not consider distribution records that could not clearly be associated with a currently valid taxon (i.e., specimens inaccessible, no pictures or other backup data). ES3 lists species' names as utilized in this dataset.

For higher-taxon associations we followed the molecular phylogeny of Kawahara et al. (2009). Generally, in order to avoid errors due to misidentifications, identification of difficult taxa was checked by the taxonomist in our team (IJK; either on the specimen or by photograph) unless data stemmed from a renowned expert on sphingid identification.

A number of relatively common and highly dispersive (possibly migrating) taxa are known to establish summer populations in temperate latitudes, but do not survive the cold season. However, details on boundaries between permanent (i.e., overwintering) and non-permanent populations are only known for well-sampled Europe. We used these data to assign coldest-month isotherms as northern boundaries for records of these species and applied these thresholds across their entire range (i.e., to Central and East Asia) in order to model only permanent ranges. Local migrations may also occur in other, non-European taxa as well as in some arid regions, but insufficient knowledge prevented us from addressing these.

S1.2 Georeferencing

Based on locality data given in publications or associated to specimen labels, we assigned geographic coordinates to distribution records. If not given from GPS measurement, we found coordinates of localities in online gazetteers, Google Earth, Atlases, and local maps. For difficult localities (e.g., old records with changing names, small places, transliterations from non-Latin spellings) we also made use of historical atlases and travel itinerates of the collectors in questions. Hints towards localities were also found in broad internet searches, such as traveller blogs, sites on Christian missions or on military history. Uncertainties arose particularly from creative transliterations of Chinese localities, from incomplete data and from very common place names. In the latter case, we made educated guesses on the most likely locality based on other records for the species and the 'home range' or travel route of the collector.

We generally aimed at georeferencing with a resolution of 0.01° (ca. 1 km) or higher. However, sometimes this was not feasible either because we could not find sites precisely enough, or because no detailed locality was given in original sources. We estimated the spatial precision of records (i.e., ≤0.01°, ≤0.1°, ≤1° or 'unspecific' (e.g., "Southern India"; Wieczorek et al. 2004) to facilitate filtering our data depending on the resolution required. If data could not be localized precisely but altitude information was given, we set coordinates to a region of similar altitude (using Google Earth, which incorporates a 90 m resolution digital elevation model) to minimize environmental deviation from the true site.

We went to great lengths to identify georeferencing errors by mapping data and by checking consistency between records processed by different people (see Acknowledgements). We also subjected data already containing coordinates to this procedure (e.g., collectors' GPS-data or downloads from databases), and we found many errors (often apparently due to mistyping). We prioritized databasing

and georeferencing towards regions and taxa with relatively few records, hereby adding more information per work effort (Beck et al. 2013).

S1.3. Distribution modelling and its validation

We based all SDMs on climatic data provided by Worldclim (averaged over 50 years, 1950-2000) and on remote sensing vegetation cover data from MODIS (vegetation continuous fields, 2000-2001; Fig. S1). Environmental data stems from a different period in time than some of our species records, which might theoretically be a source of erroneous model specification if the environment has changed between recording a species and the environment of the site of this recording. On a global scale, this is probably not highly relevant for the pattern of climatic variation (over the last 200 years), but it may well be for vegetation cover (e.g., deforestation). However, as Fig. S1D indicates, vegetation data turned out to be of minor predictive importance in most SDMs after accounting for climatic variation.

Modelling was carried out at a spatial resolution of 2.5 arcmin (ca. 5 km). SDM was based on implementing the maximum entropy algorithm (Maxent; Phillips et al. 2006). In a preliminary comparison (unpublished, available upon request) of eight commonly applied SDM algorithms (using 64 species chosen to be representative for our dataset) we found Maxent to perform best, judged by cross-validation and expert-evaluated plausibility.

For species with a large number of available records we used only data that were georeferenced with high precision (i.e., if more than 50 spatially independent records where available at modelling resolution, records with a precision >0.1° were excluded). However, we included records up to a precision of 1° for data-deficient species. Unspecific (regional) records were not used for modelling, but we considered them when validating and editing range estimates for dispersal barriers (see below). For species known from fewer than five spatially independent localities we reviewed SDM data (if models could be run at all) rigorously as if they were tentative, expert-drawn range estimates.

We used Maxent software (v. 3.3.3e) with default settings, except that we used 10,000 background points chosen from a bias file to account for undersampled regions (Phillips et al. 2009). The bias file was produced from kernel densities of raw records (100 km radius). As a result, background sampling of environmental conditions is down-weighted in regions where no distinction at all is possible between a species' absence and lack of sampling. Modelling regions were restricted a priori according to the known biogeographic association of species (e.g., excluding Europe and Asia when modelling a species restricted to sub-Saharan Africa). However, if in doubt on the potential spread of a species, we erred on the side of modelling more inclusive extents.

We evaluated raw SDMs according to two criteria: (1) We followed standard SDM procedures by randomly splitting available data and using 75% for model fitting or "training", and the remaining 25% for testing. This cross-validation procedure retrieved a metric akin to the area under the receiver-operating characteristic (AUC), based on averaging five replicate model runs. Because Maxent software replaces commission error with relative predicted area, we denote these as AUC_{Mx} to avoid confusion with true AUC based on commission assessments. We do not rest our quality assessment on AUC_{Mx} alone because the lack of commission data, as well as other issues, complicates its interpretation (Merckx et al. 2011, Jiménez-Valverde 2012). (2) Range predictions were checked for plausibility based on record data (including those not used for modelling), inventories at very well-sampled sites (see paragraph below; providing information on true absences of a species) and our knowledge of the species' biology (e.g., host plant associations). We tried to find and fix sources of error for obviously bad models (e.g., input data biases; Beck et al. 2014). If this did not lead to improvement, we excluded species from SDM and provided expert-drawn range estimates instead.

As an operational definition of "well-sampled" 5 x 5 km cells, we extracted information on the maximum species richness (at 5 km grain) within each of the "Ecoregions of the World" (Olsen et al. 2001). We demanded a well-sampled cell to contain at least twice as many records as the maximum cell-wide species richness of its ecoregion. This procedure allowed distinguishing between regions of low and high diversity, and the associated demands in sampling intensity to reach a near-complete inventory. Most of these well-sampled cells had many more than the minimum required number of records (>200 records in all but one chosen cell).

S1.4. Range editing, thresholds, and expert estimates

Current SDM methods are unable to account for dispersal limitation. Rather, the output is a measure of environmental suitability, irrespective of whether the species has reached a region or not. Beyond the first step of limiting modelling regions if the broad extent of occurrence of a species was certain, we applied expert-opinion edits of modelled distributions based on known biogeographic barriers for sphingids (Beck et al. 2006a) or for other taxa (Wallace 1869), as well as large gaps in suitable habitat inferred from model output. We assumed that species did not cross such potential barriers unless demonstrated by records. For these assessments we considered, in particular, the existence of reliable records that were not used for modelling (e.g., due to too high spatial uncertainty/imprecise georeferencing), the absence of a species from locations known to be well-sampled (as indicated by large amounts of available data, or personal knowledge of intense sampling by colleagues), and relevant biological traits of species (if known; e.g., host plant specificity and range of the host plant). For each species, we (in particular, authors IJK and JB) discussed existing hints and evidence, and concluded a likely maximum extent of the species. The extent of model predictions was limited by these conclusions.

For transforming continuous suitability into binary presence-absence predictions we set a threshold so that at least 90% of recorded presences were predicted correctly (Engler et al. 2004; cf. Bean et al. 2012, Radosavljevic and Anderson 2014). Other, recently recommended thresholding rules (Bean et al. 2012) require confirmed absences, which are not available in any presence-only dataset. However, by taking absences from 'well-sampled' cells (see below) we could investigate model quality and threshold selection for a small test set of widespread species. For these, we calculated alternative thresholds, resulting range estimates, and 'true' AUC for comparison.

For some species we were unable to model distributions, e.g. due to lack of sufficient occurrence data. In these cases we created 'expert-opinion' range estimates by plotting records on maps and drawing estimated extents of occurrence. We intersected these with our assessment of habitat restrictions (elevation, temperature, precipitation, tree cover; cf. Hurlbert and Jetz 2007) and converted them to a presence-absence grid of the same resolution as SDM-based maps.

S1.5. Grain size, stacking, and validation of richness patterns

For further analysis we projected range data into Mollweide World equal area projection. To investigate effects of grain size we up-scaled thresholded range estimates (5 km grain) to grid cell sizes of 15, 25, 50, 100, 200 and 400 km. For each grain size, we estimated species richness by summing predicted species for each cell.

Conceptual problems and a trend towards overprediction from thresholded range data have been noted (Calabrese et al. 2014), leading to the recommendation that raw SDM data be used. To investigate this, we estimated species richness as the sum of (logistic) output from Maxent models (after

dispersal editing, 5 km grain), adding presence-predictions of the non-modelled species to produce richness data consistent with thresholded data.

While independent evaluation data on species' presence and absence can be used for selected, common species in methodological studies, they do not exist for a data set that contains many rare taxa. Instead, we provide a quasi-independent validation of emergent species richness patterns by defining 'well-sampled' locations, which were then used to test model-derived richness predictions.

(1) We correlated observed species numbers from local field surveys across the Malay Archipelago (Beck et al. 2006b) with richness estimates of the containing 5 km-cells.

(2) We defined 52 'well-sampled' 5 km cells (Fig. 3) and correlated observed species richness (disregarding records for non-permanent populations) against predictions from stacked range maps.

(3) We used observed species richness of larger cells containing those 'well-sampled' 5 km cells to test predictions at larger grains. Record density showed strong spatial clumping, so at larger grain size these regions were also thoroughly sampled

Because there is potential error in model-prediction as well as observed data (due to possible undersampling), we fitted regression slopes with reduced major axis (RMA) regression. We tested the significance of relationships with spatial correlations (adjusted degree of freedoms; Clifford et al. 1989). Due to up-scaling, the number of test cells decreased with increasing cell size (adjusted R_{adj}^2's are used). Our procedure kept locations of test cells constant across scales, which eased interpretation of the data.

S1.6 Environmental models and predictors

As environmental predictors of species richness at analysis resultion we used Actual Evapotranspiration (AET; Ahn & Tateishi, 1994) as a proxy of net primary productivity. Mean Annual Temperature was taken from Hijmans et al. (2005) and topographic range was calculated from the GTopo30 Digital Elevation Model (https://lta.cr.usgs.gov/GTOPO30). These continuous predictors did not exhibit collinearity (linear regressions using log-transformed data: r^2 <0.05). Data on biogeographic realms were taken from G200 (Olsen et al., 2001) and coded as binary dummy variables. These data were aggregated in the 1° equal area polygons used for analyses. The same data had been used in earlier, large-scale gridded analyses (e.g., Kreft & Jetz, 2007).

We log-transformed and z-transformed all data, and then built univariate and multivariate linear models using (z-transformed) richness as response. Multivariate models followed the structure

```
S~AET+Temperature+TopoRange+Afrotropics[0/1]+Indomalaya[0/1]+Australasia[0/1]
```

This implied that the remaining realm in our research region, Palaearctic, had a model coefficient of zero by default. We applied ordinarly least square (OLS) models and simultaneous spatially autoregressive (SAR) models (choosing a 1500 km neighbourhood after comparing fits of various neighbourhood distances; R package *spdep*).

Table A1 Validation of modelled species richness patterns with observed data at 'well-sampled' sites (summarized in Fig. 2 of main text).

Obs. spec. (y) [grain size]	**Predictor (x)** [grain size]	**N**	$R^2_{adj.,\ (low.,\ up.\ 95\%CI)}$	$Intercept_{\pm SE}$	$Slope_{\pm SE}$	F_{corr}	df_{corr}	p_{corr}
Local site	T, 5 km	15	$0.368_{(0.029,\ 0.707)}$	$-1.660_{\pm 8.700}$	$0.500_{\pm 0.110}$	8.260	11.740	0.014
Local site	ΣM, 5 km	15	$0.594_{(0.317,\ 0.871)}$	$-23.140_{\pm 10.460}$	$1.220_{\pm 0.210}$	16.970	10.250	0.002
5 km	T, 5 km	52	$0.675_{(0.535,\ 0.814)}$	$-0.140_{\pm 3.642}$	$0.793_{\pm 0.063}$	9.527	4.469	0.032
5 km	ΣM, 5 km	52	$0.700_{(0.569,\ 0.831)}$	$1.673_{\pm 3.385}$	$1.307_{\pm 0.100}$	10.202	4.250	0.031
15 km	T, 15 km	48	$0.752_{(0.635,\ 0.868)}$	$0.244_{\pm 3.547}$	$0.712_{\pm 0.052}$	13.189	4.226	0.021
25 km	T, 25 km	44	$0.800_{(0.700,\ 0.901)}$	$-1.848_{\pm 3.555}$	$0.729_{\pm 0.050}$	18.108	4.374	0.011
50 km	T, 50 km	41	$0.835_{(0.747,\ 0.922)}$	$-0.321_{\pm 3.500}$	$0.683_{\pm 0.044}$	19.557	3.766	0.014
100 km	T, 100 km	39	$0.786_{(0.674,\ 0.899)}$	$0.670_{\pm 4.362}$	$0.661_{\pm 0.050}$	16.229	4.272	0.014
200 km	T, 200 km	37	$0.809_{(0.704,\ 0.913)}$	$1.084_{\pm 4.578}$	$0.650_{\pm 0.047}$	19.282	4.395	0.010
400 km	T, 400 km	30	$0.877_{(0.800,\ 0.954)}$	$0.239_{\pm 4.836}$	$0.721_{\pm 0.047}$	32.225	4.346	0.004

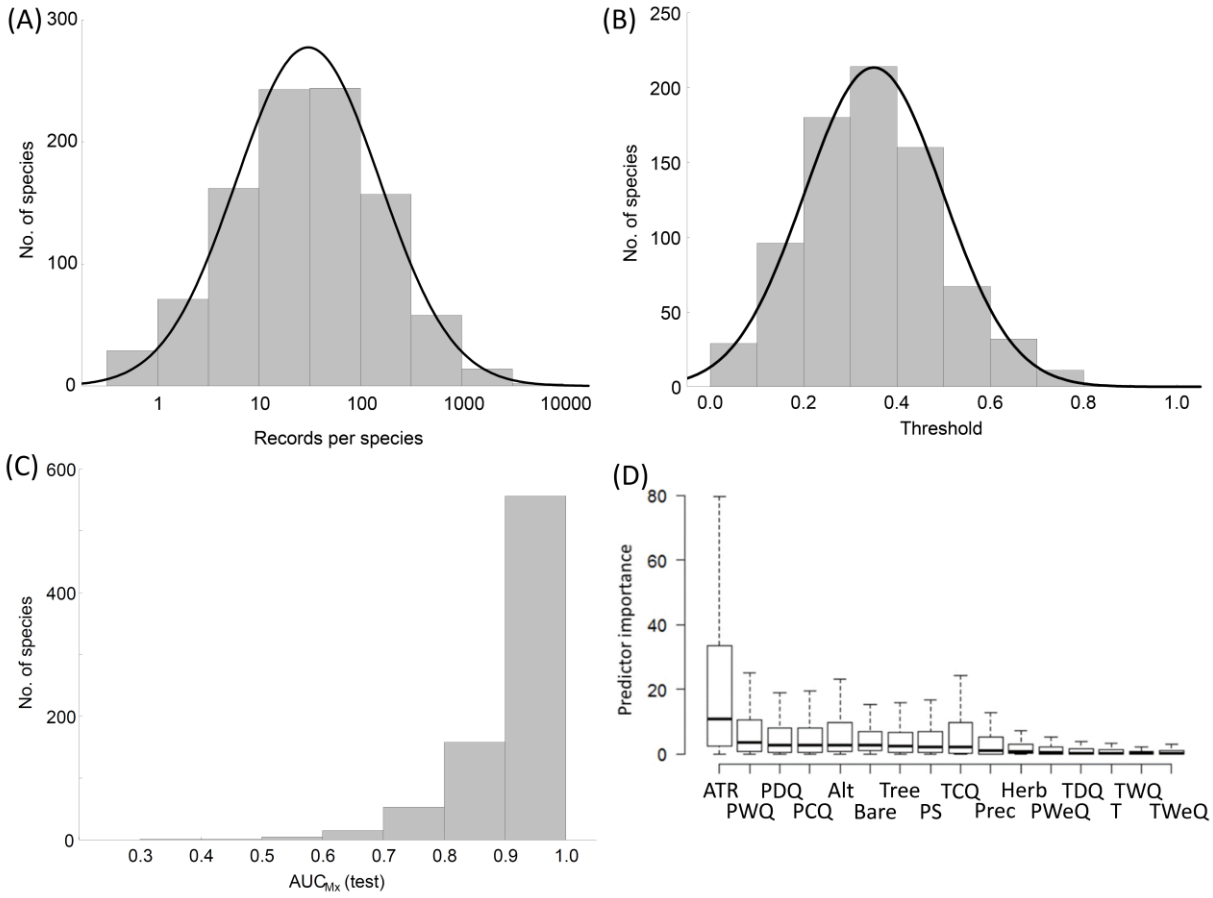

Figure A1 Properties of records data and SDMs. **[A]** Frequency distribution of records per species (log_{10}-scale; 981 species, 109'880 records). A Gaussian distribution is fitted. Note the substantial number of species only known from few or even only one record (i.e., only holotypes known). **[B]** Frequency distribution of thresholds for converting logistic Maxent output into binary presence-absence predictions (789 species), using the 'minimum predicted area' rule (i.e., set to predict at least 90% of records). A Gaussian distribution was fitted. Thresholds are negatively related to total numbers of records (Spearman rank correlation: R = -0.422, p <0.001). **[C]** Frequency distribution of Maxent-derived AUC_{Mx} of test data (789 modelled species). Median = 0.94, 25-75 percentiles = 0.88-0.973; minimum= 0.3598; maximum = 0.997. **[D]** Percent importance of variables in Maxent models (median, quartiles & range across 789 species). Acronyms: ATR = annual temperature range; PWQ = precipitation during the warmest quarter; PDQ = precipitation during the dryest quarter; PCQ = precipitation during the coldest quarter; Alt = altitude; Bare = percentage of bare ground; Tree = percentage of tree cover; PS = precipitation seasonality; Herb = percentage of herb cover; PWeQ = precipitation during the wettest quarter; TDQ = temperature during the driest quarter; T = mean annual temperature; TWQ = temperature during the warmest quarter; TWeQ = temperature during the wettest quarter; sources and documentation of climate data: www.worldclim.org; of land cover data: http://modis-land.gsfc.nasa.gov/vcc.html).

"T" is based on stacked presence-absence predictions after applying a threshold to SDM data, "ΣM" is the sum of logistic Maxent output for 789 species, plus the presence predictions for 183 non-modelled species (very few of the latter were predicted at any of the 'well-sampled' sites (max. 3)). Strong over-prediction (slopes <<1) is expected for tests of 5 km cells against local sites which necessarily do not include most parts of a 5 km cell. All data are strongly autocorrelated. Because autocorrelation may change with grain size (i.e., become stronger), the trend towards higher R^2 with coarser grain size may be an artefact, but results from spatial correlation (Clifford *et al.* 1989, three last columns) indicate increasing effect sizes with grain size after correcting for such spatial effects. Slopes were fitted by reduced major axis regression (RMA) to account for error on both axes. Note that data for comparisons with local sampling data are omitted from the main text (incl. Fig. 2).

S1.7 References for ES1

Ahn, C. H. &Tateishi, R. (1994) Estimation of Potential Evapotranspiration from global data sets. *International Archives of Photogrammetry and Remote Sensing*, **30**, 586-593.

Ballesteros-Mejia, L., I. J. Kitching & J. Beck (2011) Projecting the potential invasion of the Pink Spotted Hawkmoth (*Agrius cingulata*) across Africa. *International Journal of Pest Management,* **57**, 153–159.

Bean, W. T., R. Stafford & J . S. Brashares (2012) The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, **35**, 250–258.

Beck, J., I. J. Kitching & K. E. Linsenmair (2006a) Wallace's line revisited: has vicariance or dispersal shaped the distribution of Malesian hawkmoths (Lepidoptera: Sphingidae)? *Biological Journal of the Linnean Society*, **89**, 455–468.

Beck, J., I. J. Kitching & K. E. Linsenmair (2006b) Extending the study of range – abundance relations to tropical insects: sphingid moths in Southeast Asia. *Evolutionary Ecology Research*, **8**, 677–690.

Beck, J., L. Ballesteros-Mejia, P. Nagel & I. J. Kitching (2013) Online solutions and the "Wallacean shortfall": What does GBIF contribute to our knowledge of species' ranges? Diversity and Distribution, **19**, 1043–1050.

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

Calabrese, J. M., G. Certain, C. Kraan & C. F. Dormann (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, **23**, 99–112.

Clifford, P., S. Richardson & D. Hemon (1989) Assessing the significance of the correlation between 2 spatial processes. *Biometrics***, 45**, 123–134.

Engler, R., A. Guisan & L. Rechsteiner (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. and Jarvis, A. (2005), Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.,* **25**, 1965–1978.

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

ICZN (1999) *International Code of Zoological Nomenclatur*e. International Trust for Zoological Nomenclature, 4th edition. London, UK.

Jiménez-Valverde, A. (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography, **21**, 498–507.

Kawahara, A. Y., A. A. Mignault, J. C. Regier, I. J. Kitching & C. Mitter (2009) Phylogeny and biogeography of hawkmoths (Lepidoptera: Sphingidae): evidence from five nuclear genes. *PloS One,* **4**, e5719.

Kitching, I. & J. M. Cadiou (2000) *Hawkmoths of the world*. The Natural History Museum (Ed). Cornell University Press, London (UK).

Kreft, H. & Jetz, W. (2007) Global patterns and determinants of vascular plant diversity. *Proc. Natl. Acad. Sci. (B)*, **104**, 5925–30.

Olsen, D. M. et al. (2001) Terrestrial ecoregions of the world: a new map of life on Earth. *BioScience*, **51**, 933-938.

Merckx, B., M. Steyaert, A. Vanreusel, M. Vincx & J. Vanaverbeke (2011) Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, **222**, 588–597.

Phillips, S., R. Anderson & R. Schapire (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick & S. Ferrier (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

Radosavljevic, A. & R. P. Anderson (2014) Making better Maxent models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, **41**, 629–643.

Wallace, A. (1869) *The Geographical Distribution of Animals*. Cambridge University Press, Cambridge (UK).

Wieczorek, J., Q. Guo & R. Hijmans (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, **18**, 745–761.