

Exploring snake occurrence records: Can accessible social media help fill the gaps and improve species distribution models?

Benjamin M Marshall ^{Corresp., 1}, **Colin T Strine** ^{Corresp. 1}

¹ School of Biology, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima, Nakhon Ratchasima, Thailand

Corresponding Authors: Benjamin M Marshall, Colin T Strine

Email address: benjaminmichaelmarshall@gmail.com, strine.conservations@gmail.com

A species' distribution provides fundamental information on: climatic niche, biogeography, and conservation status. Species distribution models often use occurrence records from biodiversity databases, subject to spatial and taxonomic biases. Deficiencies in occurrence data can lead to incomplete species distribution estimates. We can incorporate other data sources to supplement occurrence datasets. The general public is creating –via GPS-enabled cameras to photograph wildlife– incidental occurrence records that may present an opportunity to improve species distribution models. We investigated 1) occurrence data of a cryptic group of animals: non-marine snakes, in a biodiversity database: Global Biodiversity Information Facility (GBIF). And determined 2) whether incidental occurrence records from social media (Flickr) could improve distribution models for 18 tropical snake species. We show the biodiversity database's 302,386 records disproportionately originate from North America, Europe and Oceania (250,063, 82.7%), with substantial gaps in tropical areas that host the highest snake diversity. North America, Europe and Oceania averaged several hundred records per species; whereas Asia, Africa and South America averaged less than 35 per species. Occurrence density showed similar patterns; Asia, Africa and South America have roughly ten-fold fewer records per 100 km² than other regions. Social media provided 44,687 additional records. However, including them in distribution models only marginally impacted niche estimations; niche overlap indices were consistently over 0.9. Similarly, we show negligible differences in Maxent model performance between models trained using GBIF-only and Flickr-supplemented datasets. Variation in model performance appeared dependent on species, rather than number of occurrences or training dataset. We suggest that researchers studying species with low detectability or limited data consider integrating social media records into occurrence datasets (if available), so long as photographic identification is viable.

Exploring snake occurrence records: Can accessible social media help fill the gaps and improve species distribution models?

Benjamin M. Marshall^{1*}, Colin T. Strine^{1**}

¹ School of Biology, Suranaree University of Technology, Nakhon Ratchasima, Nakhon Ratchasima, Thailand

Corresponding author:

Benjamin M. Marshall

Colin T. Strine

Email address:

* benjaminmichaelmarshall@gmail.com

** strine.conservation@gmail.com

Abstract

A species' distribution provides fundamental information on: climatic niche, biogeography, and conservation status. Species distribution models often use occurrence records from biodiversity databases, subject to spatial and taxonomic biases. Deficiencies in occurrence data can lead to incomplete species distribution estimates. We can incorporate other data sources to supplement occurrence datasets. The general public is creating –via GPS-enabled cameras to photograph wildlife– incidental occurrence records that may present an opportunity to improve species distribution models. We investigated 1) occurrence data of a cryptic group of animals: non-marine snakes, in a biodiversity database: Global Biodiversity Information Facility (GBIF). And determined 2) whether incidental occurrence records from social media (Flickr) could improve distribution models for 18 tropical snake species. We show the biodiversity database's 302,386 records disproportionately originate from North America, Europe and Oceania (250,063, 82.7%), with substantial gaps in tropical areas that host the highest snake diversity. North America, Europe and Oceania averaged several hundred records per species; whereas Asia, Africa and South America averaged less than 35 per species. Occurrence density showed similar patterns; Asia, Africa and South America have roughly ten-fold fewer records per 100 km² than other regions. Social media provided 44,687 additional records. However, including them in distribution models only marginally impacted niche estimations; niche overlap indices were consistently over 0.9. Similarly, we show negligible differences in Maxent model performance between models trained using GBIF-only and Flickr-supplemented datasets. Variation in model performance appeared dependent on species, rather than number of occurrences or training dataset. We suggest that researchers studying species with low detectability or limited data consider integrating social media records into occurrence datasets (if available), so long as photographic identification is viable.

Introduction

A species' conservation status is in large part determined by their known distribution (IUCN Standards and Petitions Subcommittee, 2017). To accurately determine the potential distribution for a species, researchers require occurrence records covering the entire breadth of a species climatic or ecological niche –the absolute niche. Relying on limited or partial information risks under estimating or misrepresenting a species' niche and therefore the potential distribution (Monsarrat et al., 2019). An under estimated range can mask the impacts human activity is having on species distributions and contribute to shifting baseline syndrome –where progressively eroded species distributions or populations are accepted as normal or healthy (Cromsigt, Kerley & Kowalczyk, 2012).

Issues stemming from insufficient information and shifting baseline syndrome are likely more pronounced for difficult to detect species. Snakes are both cryptic and have limited detectability, presenting issues when collecting broad scale ecological data (Steen, 2010; Durso & Seigel, 2015). Thus, we would expect snakes to have more limited global occurrence records compared to more conspicuous taxa (birds, butterflies and macrofaunal mammals).

Insufficient occurrence records is a concern: snakes can play important regulatory and keystone roles in ecosystem functioning (Willson & Winne, 2016; Miranda, 2017); there are major gaps in reptile conservation assessments (Bland & Böhm, 2016; Tingley, Meiri & Chapple, 2016; Hughes, 2017); and snakes are frequently involved in human-wildlife conflicts worldwide (Whitaker & Shine, 2000; Akani et al., 2002; Meek, 2012; Miranda, Ribeiro- & Strüssmann, 2016; Marshall et al., 2018). Protecting snakes and mitigating human-snake conflict requires solid baseline data, occurrence records present the most basic form of these data

Technological advances, global survey effort and digitisation of museum records have developed large biodiversity databases, pulling together disparate data sources to make global occurrence records more accessible. However, considerable gaps in biodiversity databases result from detection difficulties, inconsistent surveying, and inadequate (or sometimes inaccurate) locality data for museum specimens (Yesson et al., 2007; Beck et al., 2013). How are occurrence reporting deficiencies impacting snake occurrence records? And are there ways we can mitigate the deficiencies?

Supplementary data sources could help fill biodiversity database gaps (Toivonen et al., 2019). With the proliferation of GPS enabled devices, the public largely has the technology to record occurrence data and make the data available via social media websites (Barve, 2014; ElQadi et al., 2017; Jiménez-Valverde et al., 2019). Geo-tagged images provide a vehicle for communicating species identity and location; When the images enter searchable social media platforms new occurrence records are available. Such incidental biodiversity records have already improved data availability for invertebrates and birds (Barve, 2014; ElQadi et al., 2017; Jiménez-Valverde et al., 2019) –their utility for snakes is unexplored.

We describe the current state of snake occurrence records in the Global Biodiversity Information Facility (GBIF) database, highlighting gaps in surveying. We then explore the potential utility of a supplementary data source, a photography sharing platform Flickr, in the modelling of tropical snake distributions.

Materials & Methods

We completed all data analysis in R v.3.5.3 (R Core Team, 2019) and R Studio v.1.2.1335 (R Studio Team, 2019). For data visualisation we used ggplot2 (Wickham, 2016), ggrepel (Slowikowski, 2018) and scico (Pedersen & Crameri, 2018) packages. We have included a directory of scripts, packages (using packrat (Ushey et al., 2018)), and data used in analysis at: doi:10.5281/zenodo.3243983.

Data Retrieval

GBIF

Before acquiring data from the GBIF database (GBIF.org, 2019) we generated a comprehensive list of snake species. We used the taxize package (Chamberlain et al., 2018) to access GBIF and National Center for Biotechnology Information (NCBI; Benson, 2008; Sayer, 2009) records for

all squamates families, filtering for those mentioning Serpentes in downstream classification. For each Serpentes family we then queried the GBIF and downloaded occurrence records, on a per genus basis, using the dismo (Hijmans et al., 2017) and rgbif packages (Chamberlain et al., 2019). Once we had downloaded all records, we queried the GBIF database a second time to ensure that downloaded files were complete and included all available occurrences. All GBIF downloads and metadata (including a list of data sources) are available at doi:10.5281/zenodo.3243983.

After downloading, we compiled the resulting genus occurrence files, filtering out marine snake families (data manipulation performed with the dplyr (Wickham et al., 2017), data.table (Dowle & Srinivasan, 2019) and reshape2 (Wickham, 2007) packages). Due to the size of the dataset, we were forced to automate cleaning. We opted to use the CoordinateCleaner package to clean each species individually (Zizka et al., 2019). Following the process outlined in Zizka et al. 2019, we removed records with locations as NAs, zeros, identical values, near GBIF headquarters, near biodiversity institutions, within oceans, and that were extreme outliers for that species (using interquartile range outlier detection). Species with over 15,000 records (e.g. *Thamnophis spp.*, *Natrix spp.* and *Vipera berus*) failed or produced erroneous results so we examined these species manually, removing outlying occurrences.

Flickr

To acquire data from Flickr we generated a list of species names, both common and binomial as search terms. We used the taxize package (Chamberlain et al., 2018) to query the GBIF and NCBI databases for all species downstream of the Serpentes families. We then compiled results from both databases into a single list, removing duplicates. Common name queries of GBIF and NCBI databases were inadequate or failed for many species. Therefore, we created a query system that accessed The Reptile Database (Uetz et al., 2019)(using XML (Lang & CRAN Team, 2019a), xml2 (Wickham, Hester & Ooms, 2018) and rvest (Wickham, 2019)). For each species we pulled all common names the Reptile Database listed. We parsed each set of common names to separate them to generate a list of search terms for each species, attempting to anticipate as many notation styles as possible used by Reptile Database. We relied on the stringr package (Wickham, 2018) to tackle recurring character patterns.

We then accessed Flickr's API (Flickr Development Team, 2019), via R packages XML (Lang & CRAN Team, 2019a), RCurl (Lang & CRAN Team, 2019b) and httr (Wickham, 2017), using each species' search terms to retrieve search results for images. Photos had to be tagged as *snake* and geo-tagged so that a location was evident. During this process we saved a URL and year (extracted with the lubridate package (Grolemund & Wickham, 2011)) for each photo to later manually verify species identification. We then manually reviewed each image for selected tropical species, removing records of non-target species or images taken within captive settings (captive settings were identified by the presence of artificial substrate, white-balances associated with artificial lighting and geographic proximity to zoos).

Raster Layers

We used the raster package (Hijmans, 2017) to retrieve climatic raster data from WorldClim Fick & Hijmans, 2017). To guard against over-parameterisation and over-fitting during species distribution modelling (Fourcade, Besnard & Secondi, 2018), we discarded 14 of WorldClim's bioclimatic layers. We discarded layers until between-layer correlations with an R value >0.6 were removed (Merow, Smith & Silander, 2013; Castellanos et al., 2019). We explored different combinations that reduced the correlation, and opted for a set of five variables covering a variety of climatic aspects likely important to snake range delimitation (Kearney, Shine & Porter, 2009; Fourcade, Besnard & Secondi, 2018). The remaining layers were: BIO1 (annual mean temperature), BIO2 (mean diurnal temperature range), BIO7 (temperature annual range), BIO12 (annual precipitation), and BIO15 (precipitation seasonality).

We limited the remaining WorldClim data to three regions of interest: Tropical Asia (longitude: 50, 150; latitude: -25, 50), Africa (longitude: -40, 40; latitude: -25, 75) and South America (longitude: -120, -25; latitude: -60, 25). We also downloaded global elevation data (Danielson & Gesch. 2011; U.S. Geological Survey, 2016.) and human footprint index (Venter et al. 2016a; Venter et al. 2016b). Then we downscaled elevation and footprint layers to conform with our regional WorldClim layers.

Point Process Analysis

We examined the distribution of GBIF occurrences via several point process analyses. We set the data within a landmass polygon (to account for water bodies during calculations) downloaded from natural earth using the rnatualearth package (South, 2017). Then using the spatstat package (Baddeley & Turner, 2005) we tested for spatial conformity. We performed four types of spatial tests. We ran quadrat tests (with quadrats roughly equivalent to 10 degrees squared) to examine the spatial randomness of points. We calculated nearest neighbour distance functions (G) with Kaplan-Meier, border, and hazard corrections to examine the distribution of distances from points to their nearest neighbour. We estimated the empty space function (F) with Kaplan-Meier, border, and Chiu-Stoyan correction to examine how empty space is distributed between points. Finally, we estimated Ripley's reduced second moment function (K), with no correction applied due to the prohibitively large dataset, for further examination of spatial non-randomness.

We estimated continental area and occurrence density using the rnatualearth landmass. To estimate continental area, we projected the landmass for each continent using the closest Albers equal area conic projection (specifications obtained from <https://epsg.io>) with the rgeos (Bivand & Rundel, 2018) and sp packages (Pebesma et al., 2017). For standard error calculations we used the prisma (Borchers, 2018).

Modelling

Species selection

We manually selected nine taxa based on relative taxonomic stability, their charismatic appearance and the ease of photographic identification: *Bitis arietans* MERREM, 1820; *Bothriechis schlegelii* (BERTHOLD, 1846); *Bungarus fasciatus* (SCHNEIDER, 1801);

Calloselasma rhodostoma (KUHL, 1824); *Coelognathus radiatus* (BOIE, 1827); *Dendroaspis polylepis* GÜNTHER, 1864; *Eunectes murinus* (LINNAEUS, 1758); *Malayopython reticulatus* (SCHNEIDER, 1801); and *Ophiophagus hannah* (CANTOR, 1836). Our manual selection represented all three tropical regions (Tropical Asia: 5, Africa: 2, South America: 2).

We then randomly selected a further nine species using the `sample_n` function in `dplyr` (Wickham et al., 2017). The random species had to have occurrences entirely within one of the three tropical regions and be considered taxonomically stable. We defined the second criteria using the names listed on Reptile Database. Any species with a single binomial name listed since 2000, we considered stable. Once we had filtered the list of species, we randomly selected nine species from 25 species with the most Flickr results. We had to repeat the random selection removing species with too few points to model or an insufficiently sized distribution to be estimated with the resolution of raster layers. *Porthidium spp.* also had to be excluded because of the difficulties verifying identity in images. The final nine randomly selected species were: *Aplopeltura boa* (BOIE, 1828); *Atheris nitschei* TORNIER, 1902; *Boiga cynodon* (BOIE, 1827); *Boiga kraepelini* STEJNEGER, 1902; *Chironius carinatus* (LINNAEUS, 1758); *Echis coloratus* GÜNTHER, 1878; *Enhydryis enhydryis* (SCHNEIDER, 1799); *Hydrodynastes gigas* (DUMÉRIL, BIBRON & DUMÉRIL, 1854); and *Sinonatrix percarinata* (BOULENGER, 1899).

Model Settings

We created four training datasets per species. First, we used `SPthin` (Aiello-Lammens et al., 2014) with a grid size equal to the raster cell to thin the data, ensuring only a single occurrence per cell. We split the GBIF occurrences into five randomly assigned groups in geographic space, limiting nonindependence in environmental space (Roberts et al., 2017; Castellanos et al., 2019). We used the `BlockCV` package (Valavi et al., 2018) with the recommended block size based on the climatic and elevation raster layers (using 100,000 samples, group assignment was optimised across 500 iterations). Where the recommended block size failed to assign at least one occurrence to every group, we decreased the block size by 5% and re-ran the assignment until all groups were represented. Once groups had been successfully assigned, we set aside the median sized group of points from testing. We used the remaining points to train the geo-independent model. We generated a second GBIF data-only training set with a random subset of the original data removed. We removed this subset with no space weighting (to replicate random k-folds frequently used in the modelling literature), and the size was equal to the subset removed for the geo-independent model training dataset. We refer to the second model as the GBIF random model. The final models used the two GBIF training datasets described above supplemented by the Flickr data collected for that species.

We generated an array of 10,000 background points for each species, the array was consistent between model runs and training datasets. We bounded background point generation with a minimum convex polygon around all species occurrence records (Castellanos et al., 2019), plus a buffer equal to half the mean distance between occurrences. Whereas studies usually chose a fixed buffer to create the bounding area, the disparity in our 18 species distributions required

species-specific buffers based on relative occurrence record spread. Relying on a compromised fixed buffer for all species could under estimate AUC scores for species with large distributions, while inflating AUC scores for species with small distributions (Anderson & Raza, 2010). Because survey effort is undocumented and unequal (Tulloch et al., 2013), we weighted background point distribution using a bias layer to areas that are likely to have had increased survey effort (Phillips et al., 2009; Merow, Smith & Silander, 2013). We chose human footprint as proxy for survey likelihood, under the assumption that increased access and human presence will lead to greater occurrence records.

We used the ENMeval package (Muscarella et al., 2014) to run Maxent models across varying model settings. We chose Maxent because of its flexibility and performance relative to other methods (Elith et al., 2006). We used combinations of linear and quadratic feature classes and ran models using a sequence of regularization values from 1 to 8 to reduce the chances of overfitting (Shcheglovitova & Anderson, 2013; Merow, Smith & Silander, 2013; Radosavljevic & Anderson, 2014) and set internal cross validation to user groups defined with BlockCV (Valavi et al., 2018).

Model Evaluation

The metrics used to assess species distribution model performance are debated. Due to their reliance on pseudo-absences some of the ways of evaluating models are unhelpful. We chose to follow Castellanos et al.'s, (2019) advice and use multiple metrics. We selected receiver operating characteristic AUC (ROC AUC) because of its wide use and ability to compare models based on different datasets. Use of ROC AUC has drawbacks (Lobo, Jiménez-valverde & Real, 2008): it is sensitive to background area (Anderson & Raza, 2010), and is liable to overestimate model performance (Fernandes, Scherrer & Guisan, 2018). To supplement ROC AUC evaluation, we use precision-recall values (PRRC) –recently recommended as metrics insensitive to background area and species rarity because they ignore correctly predicted absences (Sofaer, Hoeting & Jarnevich, 2019). For every model created by the four training datasets we calculated ROC AUC and PRRC values for all three test datasets with the PRROC package (Grau, Grosse & Keilwagen, 2015).

As an additional measure of the Flickr data's contribution to models, we examined the niche overlap between models trained on only GBIF records and those trained on datasets supplemented with Flickr occurrences. We estimated niche overlap using Schoener's D measure with the ENMeval package (Muscarella et al., 2014).

We explored Maxent model performance using GLM and GLMMs with the lme4 package (Bates et al., 2015). We created models using combinations of number of occurrences, species and training dataset as predictors of PRRC and ROC AUC values. Full list of models tested can be found in Table S1. We used spearman's rank test to explore the relationships between area and occurrence count, after testing for normality with qqplots (from the car package (Fox & Weisberg, 2011)) and Shapiro-Wilk tests.

Results

Data Summary

Our assessments of GBIF snake occurrences reveal strong spatial bias in the 302,386 unique locations of non-marine snakes. Flickr data searches produced only 44,689 images tagged with snakes and location information; Flickr data was also spatially nonuniform.

All point process analysis showed that the distribution of GBIF and Flickr points are not randomly distributed: multiple metrics suggest spatial clustering (GBIF data Quadrat test: $X^2 = 2425600$, $df = 288$, $p\text{-value} < 2.2e-16$; Flickr: Quadrat test: $X^2 = 426820$, $df = 288$, $p\text{-value} < 2.2e-16$; G-function: Fig.S1; F-function: Fig.S2; K-function: Fig.S3). The clustering is apparent in Fig.1 and Fig.2, illustrating points concentrated in North America, Europe and Australia –both GBIF and Flickr appear to follow similar distributions.

Examining the GBIF results per continent reveals the scale of spatial bias (Table 1). The number of occurrence records are considerably lower in Africa, Asia and South America, despite their large area and diversity of snake species. This pattern is particularly apparent in the density of occurrence record estimates that are approximately ten-fold lower.

The data available for our 18 selected snake species varied dramatically (Fig.3), and appeared to only be weakly associated with the size of the minimum convex polygon (MCP) of occurrence points (Fig.4).

Modelling Results

Overall, we found that models trained on GBIF supplemented with Flickr results were marginally better at predicting both randomly selected and geographically selected GBIF records when assessed using ROC AUC (Fig.5). Precision-Recall values only saw the Flickr supplemented models perform better when predicting geographically independent sample of GBIF records (Fig.6).

Models trained on randomly and geographically independent GBIF data performed similarly when tested against the Flickr data. The randomly subset GBIF models showed more variable results both for ROC AUC and PRRC. The respectable ability to predict Flickr results from only GBIF records suggest that Flickr results have little in the way of new climatic information.

The limited new information provided by Flickr datasets is further supported by the high levels of niche overlap between models trained on GBIF-only and Flickr-supplemented datasets, albeit with variation between species (Fig.7).

When we investigated which variable predicts model performance the mixed-models using the training dataset and species as random predictors were superior based on AIC. The resulting model agrees with Fig.1 and Fig.2 indicating variability between species and a weak trend driven by the training dataset. While the model investigations seem to support species as the driver behind Maxent model performance, the residuals from the models remain highly structured and

non-normal. The best performing model using training set and species as random effects produced non-normal residuals (Sharpiro-Wilk test: PRRC as response, $W = 0.7526$, $p\text{-value} < 2.2e-16$; ROC AUC as response, $W = 0.94194$, $p\text{-value} < 2.2e-16$). Our models exploring change in model evaluation metrics, suggested that the difference in sample size played a very small role (negative relationship with PRRC values: -0.0095 ± 0.0039 , $p = 0.015$; positive relationship with ROC AUC values: 0.0260 ± 0.0042 , $p < 0.001$) and the changes were largely dependent on the species (model specification and AIC values can be found in Table S1).

Discussion

Spatial bias

Our results show a strong spatial bias in GBIF's occurrence records for non-marine snakes. The lack of records in the critical snake hotspots mirrors investigations into other taxonomic groups (Yesson et al., 2007; Roll et al., 2017). The results support calls to make use of more diverse data sources: by filling gaps in GBIF coverage and boosting sample sizes, supplementary data sources can reduce the chances of underestimating species distributions and ecological niches (Beck et al., 2013; Monsarrat et al., 2019). The gaps in GBIF records are likely not the results of lack of knowledge in these locations (Tantipisanuh & Gale, 2018), but lack of digitisation and submission to biodiversity databases.

While other studies had highlighted the potential of social media photographs to supplement existing occurrence records (Barve, 2014; ElQadi et al., 2017), they stopped short of exploring how the records would impact distribution modelling and model predictive power. Studies that explored the impact on models predictive power, targeted more readily photographed species in a region with greater interaction with biodiversity recording keeping (Jiménez-Valverde et al., 2019). Tropical snakes provide a harsher assessment of the utility of community generated geo-tagged images. Our findings suggest that while there is a growing potential for social media to supplement biodiversity databases, the benefits are currently relatively minimal for species with low-detectability and can vary dramatically between species. However, the model evaluation metrics can fail to reflect whether a model is realistic (Fourcade, Besnard & Secondi, 2018; Sofaer, Jarnevich & Flather, 2018), and simulations suggest improvements to sample size will lead to more accurate predictions (Fernandes, Scherrer & Guisan, 2018). Therefore, the benefits to even marginal increases in occurrence records may justify the effort to retrieve them from social media platforms.

Supplementary data sources limitations and potential

We highlight three limitations to implementing social media occurrence into species distribution efforts.

First is the number of geo-tagged images for low detectability species. The species with the most photographs relative to GBIF records tended to be more striking, either in size or colouration (e.g. *Eunectes murinus*, *Malayopython reticulatus* and *Bothriechis schlegelii*); a pattern reflected in GBIF records overall (Troudet et al., 2017). Limitations associated with the quantity of photos

will lessen over time as GPS enabled cameras become more common and the growth in geo-tagged images continue to increase (Fig.S4). Accessing other social media platforms containing geo-tagged images could additionally bolster occurrence datasets. However, current terms and conditions on several potential platforms prohibit data mining or have significant barriers to data access (Toivonen et al., 2019). Reliance on manual curation of occurrence records may be feasible when focusing on a single species but will become prohibitively time-consuming when assessing a wider clade.

The second limitation is the need to verify the identity of species depicted. While community science projects can have good identification rates for non-professional participants (Austen et al., 2016; Kosmala et al., 2016), species distribution modelling can be sensitive to false-positives (Fernandes, Scherrer & Guisan, 2018). Eliminating false-positives currently requires manual verification by the researchers, but there is significant progress being made in automated species identification (Botella et al., 2018; Wäldchen & Mäder, 2018; Toivonen et al., 2019). For snakes, a reliable system may be difficult to perfect given their crypsis and current taxonomic fluidity. Even if automated photographic verification can become reasonably reliable, it would be prudent to explicitly integrate the confidence of species identification into the distribution models, a practice that has already been demonstrated to improve predictions (Louvrier et al., 2018; Johnston et al., 2018).

Finally, researchers must consider the drivers behind different data sources distributions (Li, Goodchild & Xu, 2013). The use of bias layers in presence only modelling is the primary way to mitigate the impacts of an unknowable survey effort (Phillips et al., 2009; Merow, Smith & Silander, 2013). However, bias layers derived from the spatial patterns of one dataset may be inappropriate for another. This is why we opted for a bias layer, human footprint, that likely is connected to the overall distribution wildlife observations. With larger datasets from more sources there may be a need to account for sampling bias on a per-dataset basis. Alternatively, social media derived datasets could be used only in model validation, proving a “semi-independent” dataset to supplement cross-validation (Gegr et al., 2019).

Conservation implications

Numerous reptiles lack proper conservation assessment due to deficient data (Bland & Böhm, 2016). Discovering ways to fill data gaps without having to fund additional surveying efforts would be valuable at a time when natural history investigations are under appreciated but macro-ecological questions are popular (Ríos-Saldaña, Delibes-Mateos & Ferreira, 2018; McCallen et al., 2019). Overcoming data deficiencies should be prioritised; delays could result in occurrence data derived from distributions defined by human activity (realised niche), rather than the climatic or absolute niche of a species (Monsarrat et al., 2019). However, while improvements in occurrence data may help identify current distributions, unstructured occurrence data cannot help quantify population trends sorely needed for many reptile species (Bland & Böhm, 2016; Bayraktarov et al., 2019).

The quantity and accessibility of social media species occurrence records is open for abuse. In herpetology there have been several cases of species being negatively affected by the scientific publication of location data (Stuart et al., 2006; Lindenmayer & Scheele, 2017) even though journals allow masked or partial publication (Lowe et al., 2017). While there is understandable fear in publishing the locations of new and desirable species in scientific literature, how long does it take for that information to enter the public sphere via geo-tagged photography? With the rapid growth geo-tagged images, being able to keep a desirable species protected by secrecy or gate-keeping may become increasingly difficult.

Conclusion

We have highlighted that there is significant spatial bias in the GBIF records for non-marine snakes, with gaps in tropical regions that house exceptionally high diversity of snakes (Roll et al., 2017). While we encourage the investigation of supplementary data sources to help fill gaps in biodiversity databases, currently accessible social media occurrence records only improve species distribution models marginally. The data availability for snakes is highly variable between species and emphasises the difficulties researchers face when studying low detectability species. Both GBIF and social media data sources are growing exponentially; tapping the full potential of these resources may be best realised with integration of image recognition and identification confidence.

Acknowledgements

We thank the Suranaree University of Technology for providing the resources required to undertaken this research. We thank Inês Silva and Matt Crane for enduring long discussions on model evaluation metrics. We thank the Flickr team for creating an API that is accessible and searchable. Finally, we thank countless photographers across the globe for their enthusiasm for wildlife.

References

- Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP. 2014. *spThin: Functions for Spatial Thinning of Species Occurrence Records for Use in Ecological Models*.
- Akani GC, Eyo E, Odegbune E, Eniang EA, Luiselli L. 2002. Ecological patterns of anthropogenic mortality of suburban snakes in an African tropical region. *Israel Journal of Zoology* 48:1–11. DOI: 10.1092/NL55-UK13-XXQ9-NCYE.

- Anderson RP, Raza A. 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: Preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography* 37:1378–1393. DOI: 10.1111/j.1365-2699.2010.02290.x.
- Austen GE, Bindemann M, Griffiths RA, Roberts DL. 2016. Species identification by experts and non-experts: comparing images from field guides. *Scientific Reports* 6:33634. DOI: 10.1038/srep33634.
- Baddeley A, Turner R. 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software* 12:1–42.
- Barve V. 2014. Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecological Informatics* 24:194–199. DOI: 10.1016/j.ecoinf.2014.08.008.
- Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67:1–48. DOI: 10.18637/jss.v067.i01.
- Bayraktarov E, Ehmke G, O’Connor J, Burns EL, Nguyen HA, McRae L, Possingham HP, Lindenmayer DB. 2019. Do Big Unstructured Biodiversity Data Mean More Knowledge? *Frontiers in Ecology and Evolution* 6:239. DOI: 10.3389/fevo.2018.00239.
- Beck J, Ballesteros-Mejia L, Nagel P, Kitching IJ. 2013. Online solutions and the ‘Wallacean shortfall’: what does GBIF contribute to our knowledge of species’ ranges? *Diversity and Distributions* 19:1043–1050. DOI: 10.1111/ddi.12083.
- Benson, DA, Karsch-Mizrachi, I, Lipman, DJ, Ostell, J, & Wheeler, DL. (2008). GenBank *Nucleic Acids Res.* Jan, 1, 33.
<https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>

413 Bivand R, Rundel C. 2018. *rgeos: Interface to Geometry Engine - Open Source ('GEOS')*.
 414 Bland LM, Böhm M. 2016. Overcoming data deficiency in reptiles. *Biological Conservation*
 415 204:16–22. DOI: 10.1016/j.biocon.2016.05.018.
 416 Borchers HW. 2018. *pracma: Practical Numerical Math Functions*.
 417 Botella C, Joly A, Bonnet P, Monestiez P, Munoz F. 2018. Species distribution modeling based
 418 on the automated identification of citizen observations. *Applications in Plant Sciences*
 419 6:e1029. DOI: 10.1002/aps3.1029.
 420 Castellanos AA, Huntley JW, Voelker G, Lawing AM. 2019. Environmental filtering improves
 421 ecological niche models across multiple scales. *Methods in Ecology and Evolution*
 422 10:481–492. DOI: 10.1111/2041-210X.13142.
 423 Chamberlain S, Barve V, Mcglinn D, Oldoni D, Desmet P, Geffert L, Ram K. 2019. *rgbif:*
 424 *Interface to the Global Biodiversity Information Facility API*.
 425 Chamberlain S, Szoecs E, Foster Z, Arendsee Z, Boettiger C, Ram K, Bartomeus I, Baumgartner
 426 J, O'Donnell J, Oksanen J, Tzovaras BG, Marchand P, Tran V. 2018. *taxize: Taxonomic*
 427 *information from around the web*.
 428 Cromsigt JPGM, Kerley GIH, Kowalczyk R. 2012. The difficulty of using species distribution
 429 modelling for the conservation of refugee species - the example of European bison.
 430 *Diversity and Distributions* 18:1253–1257. DOI: 10.1111/j.1472-4642.2012.00927.x.
 431 Danielson, JJ, and Gesch, DB. 2011. Global multi-resolution terrain elevation data 2010
 432 (GMTED2010): U.S. Geological Survey Open-File Report 2011-1073.
 433 <http://pubs.usgs.gov/of/2011/1073/pdf/of2011-1073.pdf>
 434 Dowle M, Srinivasan A. 2019. *data.table: Extension of 'data.frame'*.

435 Durso AM, Seigel RA. 2015. A Snake in the Hand is Worth 10,000 in the Bush. *Journal of*
 436 *Herpetology* 49:503–506. DOI: 10.1670/15-49-04.1.

437 Elith J, H. Graham C, P. Anderson R, Dudík M, Ferrier S, Guisan A, J. Hijmans R, Huettmann F,
 438 R. Leathwick J, Lehmann A, Li J, G. Lohmann L, A. Loiselle B, Manion G, Moritz C,
 439 Nakamura M, Nakazawa Y, McC. M. Overton J, Townsend Peterson A, J. Phillips S,
 440 Richardson K, Scachetti-Pereira R, E. Schapire R, Soberón J, Williams S, S. Wisz M, E.
 441 Zimmermann N. 2006. Novel methods improve prediction of species’ distributions from
 442 occurrence data. *Ecography* 29:129–151. DOI: 10.1111/j.2006.0906-7590.04596.x.

443 ElQadi MM, Dorin A, Dyer A, Burd M, Bukovac Z, Shrestha M. 2017. Mapping species
 444 distributions with social media geo-tagged images: Case studies of bees and flowering
 445 plants in Australia. *Ecological Informatics* 39:23–31. DOI: 10.1016/j.ecoinf.2017.02.006.

446 Fernandes RF, Scherrer D, Guisan A. 2018. Effects of simulated observation errors on the
 447 performance of species distribution models. *Diversity and Distributions*:1–14. DOI:
 448 10.1111/ddi.12868.

449 Flickr Development Team. 2019. Flickr API. <https://www.flickr.com/services/api/>

450 Fick, SE, Hijmans, RJ. 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for
 451 global land areas. *International Journal of Climatology*.

452 Fourcade Y, Besnard AG, Secondi J. 2018. Paintings predict the distribution of species, or the
 453 challenge of selecting environmental predictors and evaluation statistics. *Global Ecology*
 454 *and Biogeography* 27:245–256. DOI: 10.1111/geb.12684.

455 Fox J, Weisberg S. 2011. *An R Companion to Applied Regression*. Thousand Oaks CA: Sage.

456 GBIF.org. 2019. Global Biodiversity Information Facility Website. <https://www.gbif.org> [05-08
 457 April 2019].

458 Grau J, Grosse I, Keilwagen J. 2015. PRROC: computing and visualizing precision-recall and
459 receiver operating characteristic curves in R. *Bioinformatics* 31:2595–2597.

460 Gregg EJ, Palacios DM, Thompson A, Chan KMA. 2019. Why less complexity produces better
461 forecasts: an independent data evaluation of kelp habitat models. *Ecography* 42:428–443.
462 DOI: 10.1111/ecog.03470.

463 Grolemund G, Wickham H. 2011. Dates and Times Made Easy with lubridate. *Journal of*
464 *Statistical Software* 40:1–25.

465 Hijmans RJ. 2017. *raster: Geographic Data Analysis and Modeling*.

466 Hijmans RJ, Phillips S, Leathwick J, Elith J. 2017. *dismo: Species Distribution Modeling*.

467 Hughes AC. 2017. Mapping priorities for conservation in Southeast Asia. *Biological*
468 *Conservation* 209:395–405. DOI: 10.1016/j.biocon.2017.03.007.

469 IUCN Standards and Petitions Subcommittee. 2017. Guidelines for Using the IUCN Red List
470 Categories and Criteria. Version 13.

471 Jiménez-Valverde A, Peña-Aguilera P, Barve V, Burguillo-Madrid L. 2019. Photo-sharing
472 platforms key for characterising niche and distribution in poorly studied taxa. *Insect*
473 *Conservation and Diversity*:icad.12351. DOI: 10.1111/icad.12351.

474 Johnston A, Fink D, Hochachka WM, Kelling S. 2018. Estimates of observer expertise improve
475 species distributions from citizen science data. *Methods in Ecology and Evolution* 9:88–
476 97. DOI: 10.1111/2041-210X.12838.

477 Kearney M, Shine R, Porter WP. 2009. The potential for behavioral thermoregulation to buffer
478 “cold-blooded” animals against climate warming. *Proceedings of the National Academy*
479 *of Sciences* 106:3835–3840. DOI: 10.1073/pnas.0808913106.

- Kosmala M, Wiggins A, Swanson A, Simmons B. 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment* 14:551–560. DOI: 10.1002/fee.1436.
- Lang DT, CRAN Team. 2019a. *XML: Tools for Parsing and Generating XML Within R and S-Plus*.
- Lang DT, CRAN Team. 2019b. *RCurl: General Network (HTTP/FTP/...) Client Interface for R*.
- Li L, Goodchild MF, Xu B. 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science* 40:61–77. DOI: 10.1080/15230406.2013.777139.
- Lindenmayer D, Scheele B. 2017. Do not publish. *Science* 356:800–801. DOI: 10.1126/science.aan1362.
- Lobo JM, Jiménez-valverde A, Real R. 2008. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17:145–151. DOI: 10.1111/j.1466-8238.2007.00358.x.
- Louvrier J, Molinari-Jobin A, Kéry M, Chambert T, Miller D, Zimmermann F, Marboutin E, Molinari P, Müeller O, Černe R, Gimenez O. 2018. Use of ambiguous detections to improve estimates from species distribution models. *Conservation Biology* 33:185–195. DOI: 10.1111/cobi.13191.
- Lowe AJ, Smyth AK, Atkins K, Avery R, Belbin L, Brown N, Budden AE, Gioia P, Guru S, Hardie M, Hirsch T, Hobern D, La Salle J, Loarie SR, Miles M, Milne D, Nicholls M, Rossetto M, Smits J, Sparrow B, Terrill G, Turner D, Wardle GM. 2017. Publish openly but responsibly. *Science* 357:141–141. DOI: 10.1126/science.aao0054.
- Marshall BM, Strine CT, Jones MD, Theodorou A, Amber E, Waengsothorn S, Suwanwaree P, Goode M. 2018. Hits Close to Home: Repeated Persecution of King Cobras

(*Ophiophagus hannah*) in Northeastern Thailand. *Tropical Conservation Science* 11:194008291881840. DOI: 10.1177/1940082918818401.

McCallen E, Knott J, Nunez-Mir G, Taylor B, Jo I, Fei S. 2019. Trends in ecology: shifts in ecological research themes over the past four decades. *Frontiers in Ecology and the Environment*. DOI: 10.1002/fee.1993.

Meek R. 2012. Anthropogenic sources of mortality in the western whip snake, *Hierophis viridiflavus*, in a fragmented landscape in Western France. *Herpetological Bulletin* 120:4–8.

Merow C, Smith MJ, Silander JA. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* 36:1058–1069. DOI: 10.1111/j.1600-0587.2013.07872.x.

Miranda EBP de. 2017. The Plight of Reptiles as Ecological Actors in the Tropics. *Frontiers in Ecology and Evolution* 5:159. DOI: 10.3389/fevo.2017.00159.

Miranda EBP, Ribeiro- RP, Strüssmann C. 2016. The ecology of human-anaconda conflict: a study using internet videos. *Tropical Conservation Science* 9:43–77. DOI: 10.1177/194008291600900105.

Monsarrat S, Novellie P, Rushworth I, Kerley GIH. 2019. Shifted distribution baselines: neglecting long-term biodiversity records risks overlooking potentially suitable habitat for conservation management. *bioRxiv*. DOI: 10.1101/565929.

Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, Anderson RP. 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution* 5:1198–1205. DOI: 10.1111/2041-210X.12261.

526 Pebesma E, Bivand R, Rowlingson B, Gomez-Rubio V, Hijmans R, Sumner M, MacQueen D,
527 Lemon J, O'Brien J. 2017. Package 'sp.'

528 Pedersen TL, Crameri F. 2018. scico: Colour Palettes Based on the Scientific Colour-Maps.

529 Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S. 2009. Sample
530 selection bias and presence-only distribution models: implications for background and
531 pseudo-absence data. *Ecological Applications* 19:181–197. DOI: 10.1890/07-2153.1.

532 R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria:
533 R Foundation for Statistical Computing.

534 R Studio Team. 2019. *RStudio: Integrated Development Environment for R*. Boston, MA:
535 RStudio, Inc.

536 Radosavljevic A, Anderson RP. 2014. Making better Maxent models of species distributions:
537 Complexity, overfitting and evaluation. *Journal of Biogeography* 41:629–643. DOI:
538 10.1111/jbi.12227.

539 Ríos-Saldaña CA, Delibes-Mateos M, Ferreira C. 2018. Are fieldwork studies being relegated to
540 second place in conservation science? *Global Ecology and Conservation*:e00389. DOI:
541 10.1016/j.gecco.2018.e00389.

542 Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-
543 Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF.
544 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or
545 phylogenetic structure. *Ecography* 40:913–929. DOI: 10.1111/ecog.02881.

546 Roll U, Feldman A, Novosolov M, Allison A, Bauer AM, Bernard R, Böhm M, Castro-Herrera
547 F, Chirio L, Collen B, Colli GR, Dabool L, Das I, Doan TM, Grismer LL, Hoogmoed M,
548 Itescu Y, Kraus F, LeBreton M, Lewin A, Martins M, Maza E, Meirte D, Nagy ZT, de C.

Nogueira C, Pauwels OSG, Pincheira-Donoso D, Powney GD, Sindaco R, Tallowin OJS, Torres-Carvajal O, Trape J-F, Vidan E, Uetz P, Wagner P, Wang Y, Orme CDL, Grenyer R, Meiri S. 2017. The global distribution of tetrapods reveals a need for targeted reptile conservation. *Nature Ecology & Evolution* 1:1677–1682. DOI: 10.1038/s41559-017-0332-2.

Sayers, E. W. 2009. Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. Database resources of the National Center for Biotechnology Information. *Nucl Acids Res.* 37, D5-D15. <https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>

Shcheglovitova M, Anderson RP. 2013. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling* 269:9–17. DOI: 10.1016/j.ecolmodel.2013.08.011.

Slowikowski K. 2018. *ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2.”*

Sofaer HR, Hoeting JA, Jarnevich CS. 2019. The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution* 10:565–577. DOI: 10.1111/2041-210X.13140.

Sofaer HR, Jarnevich CS, Flather CH. 2018. Misleading prioritizations from modelling range shifts under climate change. *Global Ecology and Biogeography* 27:658–666. DOI: 10.1111/geb.12726.

572 South A. 2017. *rnaturalearth: World Map Data from Natural Earth*.

573 Steen DA. 2010. Snakes in the grass: Secretive natural histories defy both conventional and
574 progressive statistics. *Herpetological Conservation and Biology* 5:183–188.

575 Stuart BL, Rhodin AGJ, Grismer LL, Hansel T. 2006. Scientific description can imperil species.
576 *Science* 312:1137. DOI: 10.1126/science.312.5777.1137b.

577 Tantipisanuh N, Gale GA. 2018. Identification of biodiversity hotspot in national level –
578 Importance of unpublished data. *Global Ecology and Conservation* 13:e00377. DOI:
579 10.1016/j.gecco.2018.e00377.

580 Tingley R, Meiri S, Chapple DG. 2016. Addressing knowledge gaps in reptile conservation.
581 *Biological Conservation* 204:1–5. DOI: 10.1016/j.biocon.2016.07.021.

582 Toivonen T, Heikinheimo V, Fink C, Hausmann A, Hiippala T, Järvi O, Tenkanen H, Di Minin
583 E. 2019. Social media data for conservation science: A methodological overview.
584 *Biological Conservation* 233:298–315. DOI: 10.1016/j.biocon.2019.01.023.

585 Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017. Taxonomic bias in
586 biodiversity data and societal preferences. *Scientific Reports* 7:9132. DOI:
587 10.1038/s41598-017-09084-6.

588 Tulloch AIT, Mustin K, Possingham HP, Szabo JK, Wilson KA. 2013. To boldly go where no
589 volunteer has gone before: predicting volunteer activity to prioritize surveys at the
590 landscape scale. *Diversity and Distributions* 19:465–480. DOI: 10.1111/j.1472-
591 4642.2012.00947.x.

592 U.S. Geological Survey, 2016, TopoTools.
593 https://topotools.cr.usgs.gov/GMTED_viewer/gmted2010_global_grids.php [01
594 December, 2018]

595 Uetz, P, Freed, P, Hošek, J (eds.). 2019. The Reptile Database, <http://www.reptile-database.org>
 596 [17 April 2019]

597 Ushey K, McPherson J, Cheng J, Atkins A, Allaire JJ. 2018. *packrat: A Dependency*
 598 *Management System for Projects and their R Package Dependencies*.

599 Valavi R, Elith J, Lahoz-Monfort J, Guillera-Arroita G. 2018. *blockCV: Spatial and*
 600 *environmental blocking for k-fold cross-validation*.

601 Venter O, Sanderson EW, Magrath A, Allan JR, Beher J, Jones KR, Possingham HP, Laurance
 602 WF, Wood P, Fekete BM, Levy MA, Watson JE. 2016a. Global terrestrial Human
 603 Footprint maps for 1993 and 2009. *Scientific Data* 3: 160067.
 604 <https://doi.org/10.1038/sdata.2016.67>

605 Venter O, Sanderson EW, Magrath A, Allan JR, Beher J, Jones KR, Possingham HP, Laurance
 606 WF, Wood P, Fekete BM, Levy MA, Watson JEM. 2016b. Data from: Global terrestrial
 607 Human Footprint maps for 1993 and 2009. *Dryad Digital Repository*.
 608 <https://doi.org/10.5061/dryad.052q5.2>

609 Wäldchen J, Mäder P. 2018. Machine learning for image based species identification. *Methods in*
 610 *Ecology and Evolution* 9:2216–2225. DOI: 10.1111/2041-210X.13075.

611 Whitaker PB, Shine R. 2000. Sources of Mortality of Large Elapid Snakes in an Agricultural
 612 Landscape. *Journal of Herpetology* 34:121–128. DOI: 10.2307/1565247.

613 Wickham H. 2007. Reshaping Data with the reshape Package. *Journal of Statistical Software*
 614 21:1–20.

615 Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

616 Wickham H. 2017. *httr: Tools for Working with URLs and HTTP*.

617 Wickham H. 2018. *stringr: Simple, Consistent Wrappers for Common String Operations*.

- Wickham H. 2019. *rvest: Easily Harvest (Scrape) Web Pages*.
- Wickham H, Francois R, Henry L, Müller K. 2017. dplyr: A Grammar of Data Manipulation.
- Wickham H, Hester J, Ooms J. 2018. *xml2: Parse XML*.
- Willson JD, Winne CT. 2016. Evaluating the functional importance of secretive species: A case study of aquatic snake predators in isolated wetlands. *Journal of Zoology* 298:266–273. DOI: 10.1111/jzo.12311.
- Yesson C, Brewer PW, Sutton T, Caithness N, Pahwa JS, Burgess M, Gray WA, White RJ, Jones AC, Bisby FA, Culham A. 2007. How Global Is the Global Biodiversity Information Facility? *PLoS ONE* 2:e1124. DOI: 10.1371/journal.pone.0001124.
- Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R, Svanteson S, Wengström N, Zizka V, Antonelli A. 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*. DOI: 10.1111/2041-210X.13152.

Table 1(on next page)

Summary information of GBIF snake records.

Species = Number of species appearing in GBIF data, not the actual number of species known to exist across the continent. # Occurrences = Number of occurrence records in GBIF downloads. Mean occurrences per species = Total number of occurrences records in a continent divided by number of species in GBIF data. SE = Standard error associated with the mean occurrences per species. Area = Area in millions of km² estimated using Alber's equal area conic projection. Occurrences per 100 km² = Total number of occurrence records divided by the estimated continental area.

Continent	# Species	# Occurrences	Mean occurrences per species	SE	Area (million km ²)	Occurrences per 100 km ²
Africa	513	17,108	33.35	3.38	29.89	0.006
Asia	576	19,187	33.31	5.69	44.67	0.004
Europe	99	42,892	433.25	169.78	8.97	0.048
North America	680	157,923	232.24	33.62	24.64	0.064
Oceania	236	49,247	208.67	32.56	8.92	0.055
South America	633	16,029	25.32	2.21	17.91	0.009

1

Figure 1(on next page)

Global distribution of all GBIF non-marine snake records displayed against continental divisions.

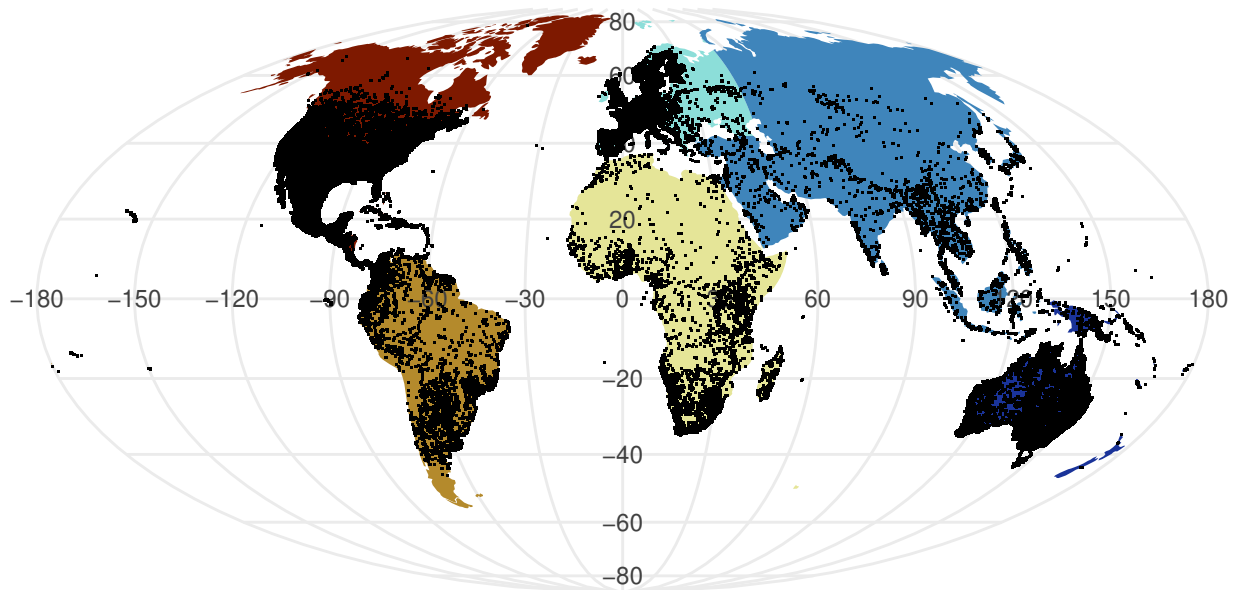


Figure 2(on next page)

Global distribution of geo-tagged Flickr photos that appeared across all searches

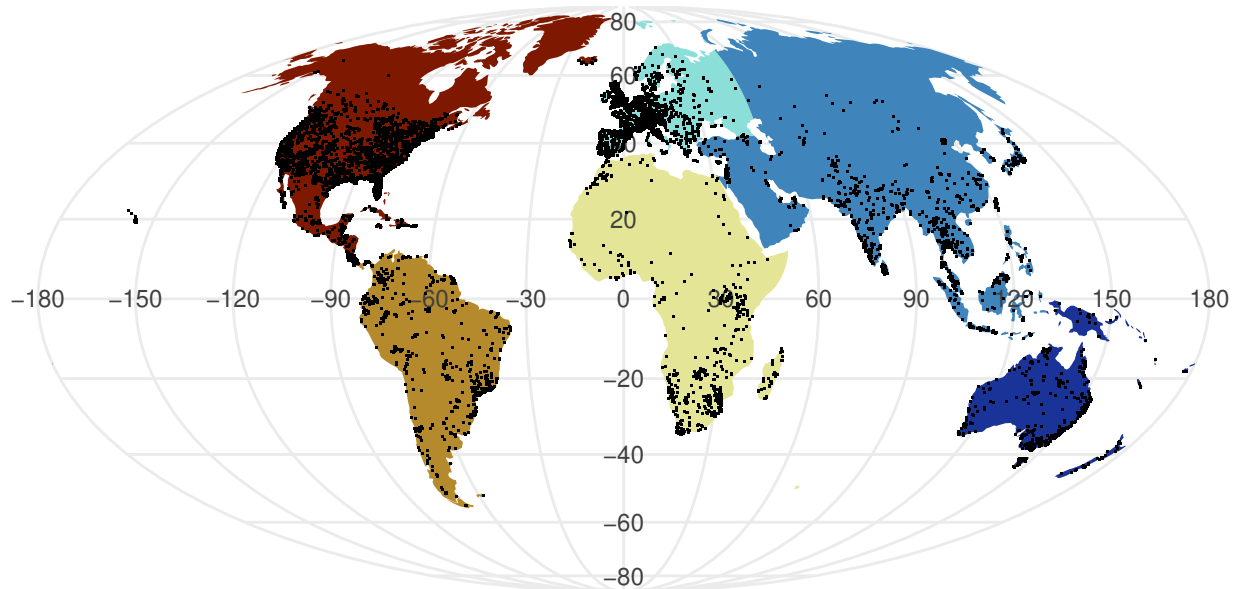


Figure 3(on next page)

Number and source of selected species occurrence records.

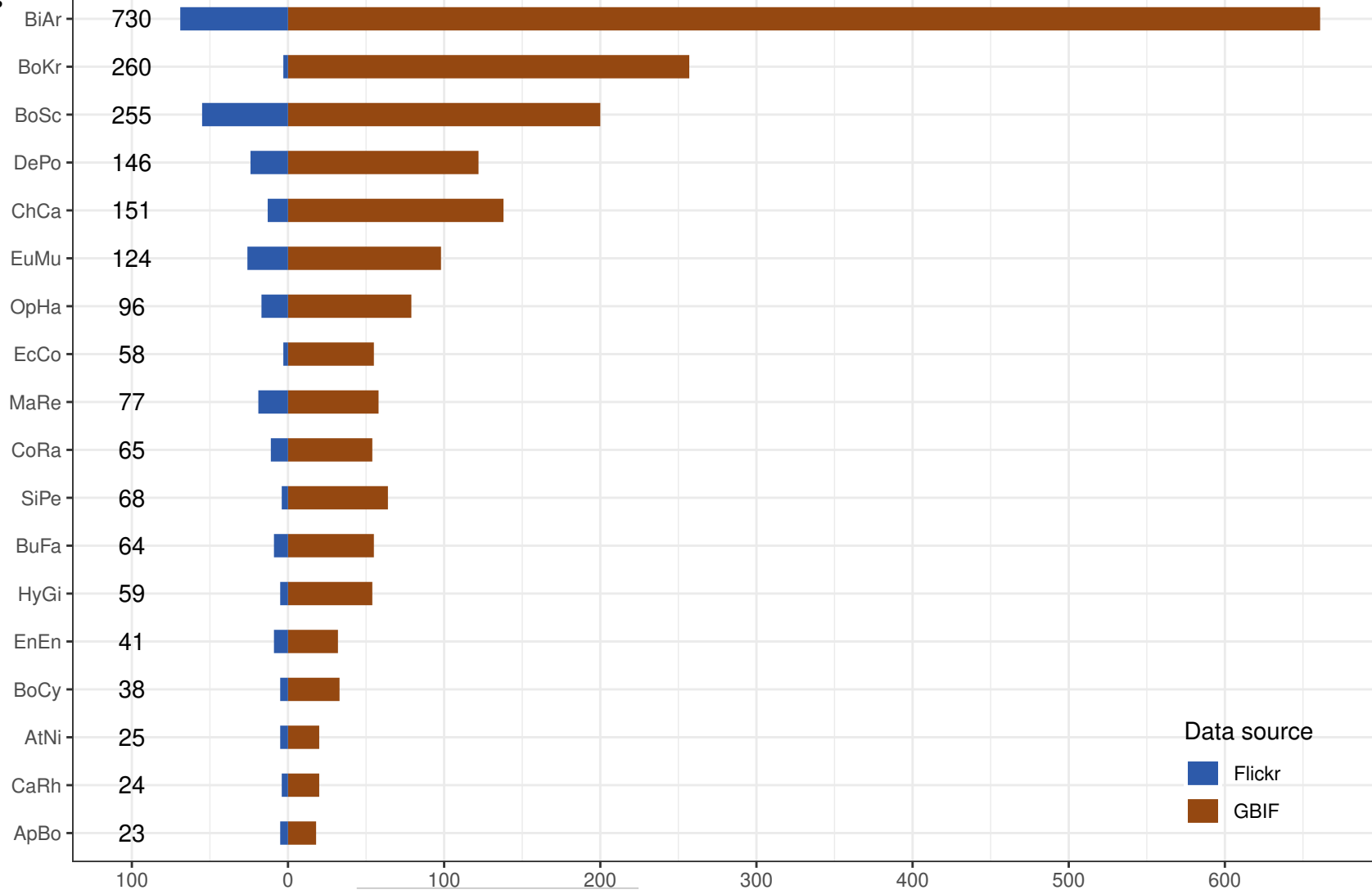
Species codes in order of appearance: BiAr = *Bitis arietans*, BoKr = *Boiga kraepelini*, BoSc = *Bothriechis schlegelii*, DePo = *Dendroaspis polylepis*, ChCa = *Chironius carinatus*, EuMu = *Eunectes murinus*, OpHa = *Ophiophagus hannah*, EcCo = *Echis coloratus*, MaRe = *Malayopython reticulatus*, CoRa = *Coelognathus radiatus*, SiPe = *Sinonatrix percarinata*, BuFa = *Bungarus fasciatus*, HyGi = *Hydrodynastes gigas*, EnEn = *Enhydris enhydris*, BoCy = *Boiga cynodon*, AtNi = *Atheris nitschei*, CaRh = *Calloselasma rhodostoma*, ApBo = *Aplopeltura boa*.

Total

PeerJ

Manuscript to be reviewed

Species



Data source

Flickr

GBIF

Figure 4(on next page)

Relationship between number of occurrences and the minimum convex polygon (MCP) area cover by occurrence points.

Minimum convex polygon are clipped to exclude oceans. Both scales are presented as logs to make individual species visible. Species codes from left to right: BoKr = *Boiga kraepelini*, CaRh = *Calloselasma rhodostoma*, AtNi = *Atheris nitschei*, BoSc = *Bothriechis schlegelii*, ApBo = *Aplopeltura boa*, SiPe = *Sinonatrix percarinata*, BoCy = *Boiga cynodon*, MaRe = *Malayopython reticulatus*, CoRa = *Coelognathus radiatus*, BuFa = *Bungarus fasciatus*, EnEn = *Enhydris enhydris*, DePo = *Dendroaspis polylepis*, EcCo = *Echis coloratus*, OpHa = *Ophiophagus hannah*, HyGi = *Hydrodynastes gigas*, ChCa = *Chironius carinatus*, EuMu = *Eunectes murinus*, BiAr = *Bitis arietans*.

S = 619.779, p-value = 0.14, rho = 0.36

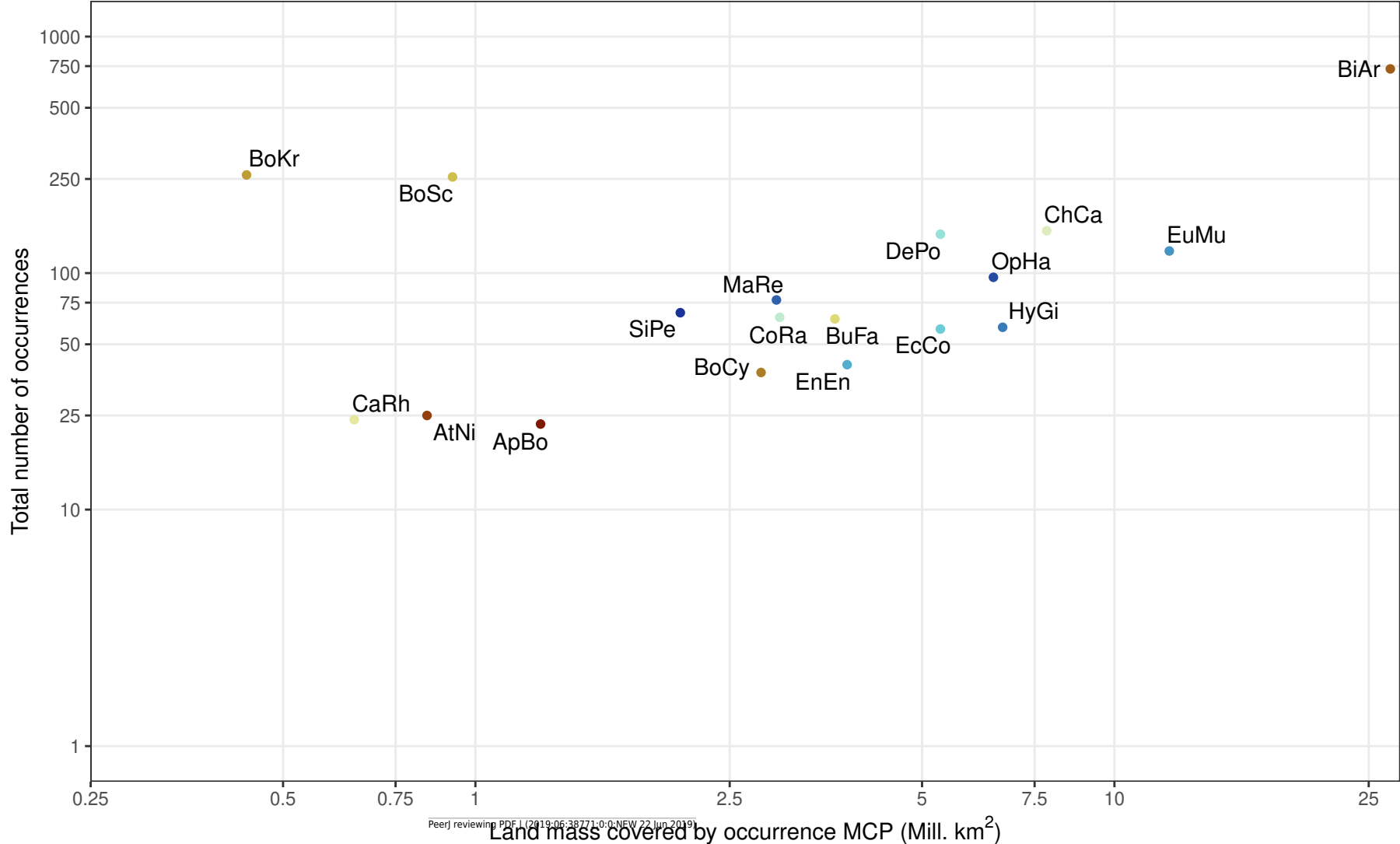


Figure 5 (on next page)

Receiver Operating Characteristic results for the three models when tested against the three independent test datasets.

Species codes in alphabetical order: ApBo = *Aplopeltura boa*, AtNi = *Atheris nitschei*, BiAr = *Bitis arietans*, BoCy = *Boiga cynodon*, BoKr = *Boiga kraepelini*, BoSc = *Bothriechis schlegelii*, BuFa = *Bungarus fasciatus*, CaRh = *Calloselasma rhodostoma*, ChCa = *Chironius carinatus*, CoRa = *Coelognathus radiatus*, DePo = *Dendroaspis polylepis*, EcCo = *Echis coloratus*, EnEn = *Enhydris enhydris*, EuMu = *Eunectes murinus*, HyGi = *Hydrodynastes gigas*, MaRe = *Malayopython reticulatus*, OpHa = *Ophiophagus hannah*, SiPe = *Sinonatrix percarinata*.

Test:
GBIF random sample

Test:
GBIF geo-independent sample

Test:
Flickr data

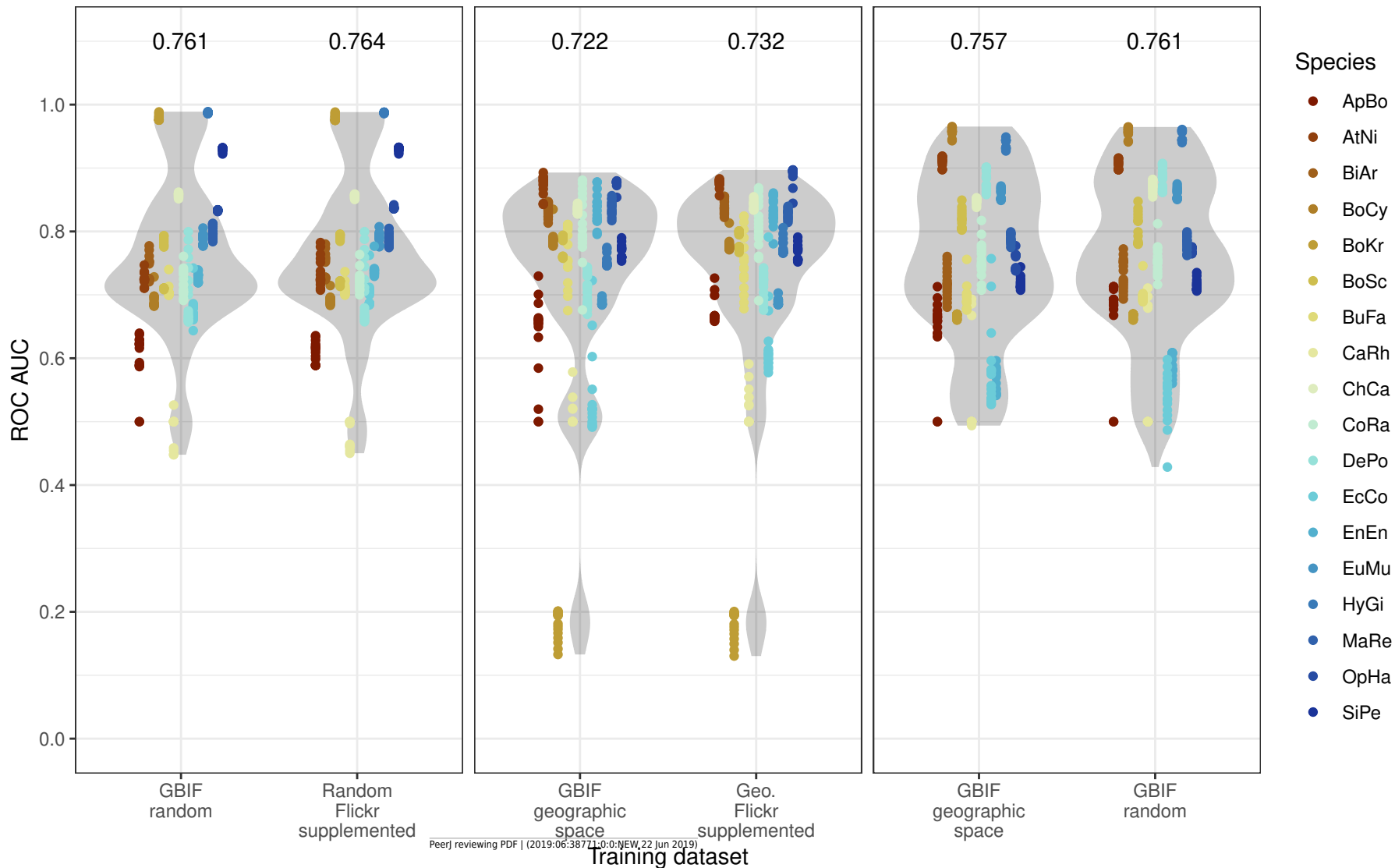


Figure 6(on next page)

Precision-Recall results for the three models when tested against the three independent test datasets.

Species codes in alphabetical order: ApBo = *Aplopeltura boa*, AtNi = *Atheris nitschei*, BiAr = *Bitis arietans*, BoCy = *Boiga cynodon*, BoKr = *Boiga kraepelini*, BoSc = *Bothriechis schlegelii*, BuFa = *Bungarus fasciatus*, CaRh = *Calloselasma rhodostoma*, ChCa = *Chironius carinatus*, CoRa = *Coelognathus radiatus*, DePo = *Dendroaspis polylepis*, EcCo = *Echis coloratus*, EnEn = *Enhydris enhydris*, EuMu = *Eunectes murinus*, HyGi = *Hydrodynastes gigas*, MaRe = *Malayopython reticulatus*, OpHa = *Ophiophagus hannah*, SiPe = *Sinonatrix percarinata*.

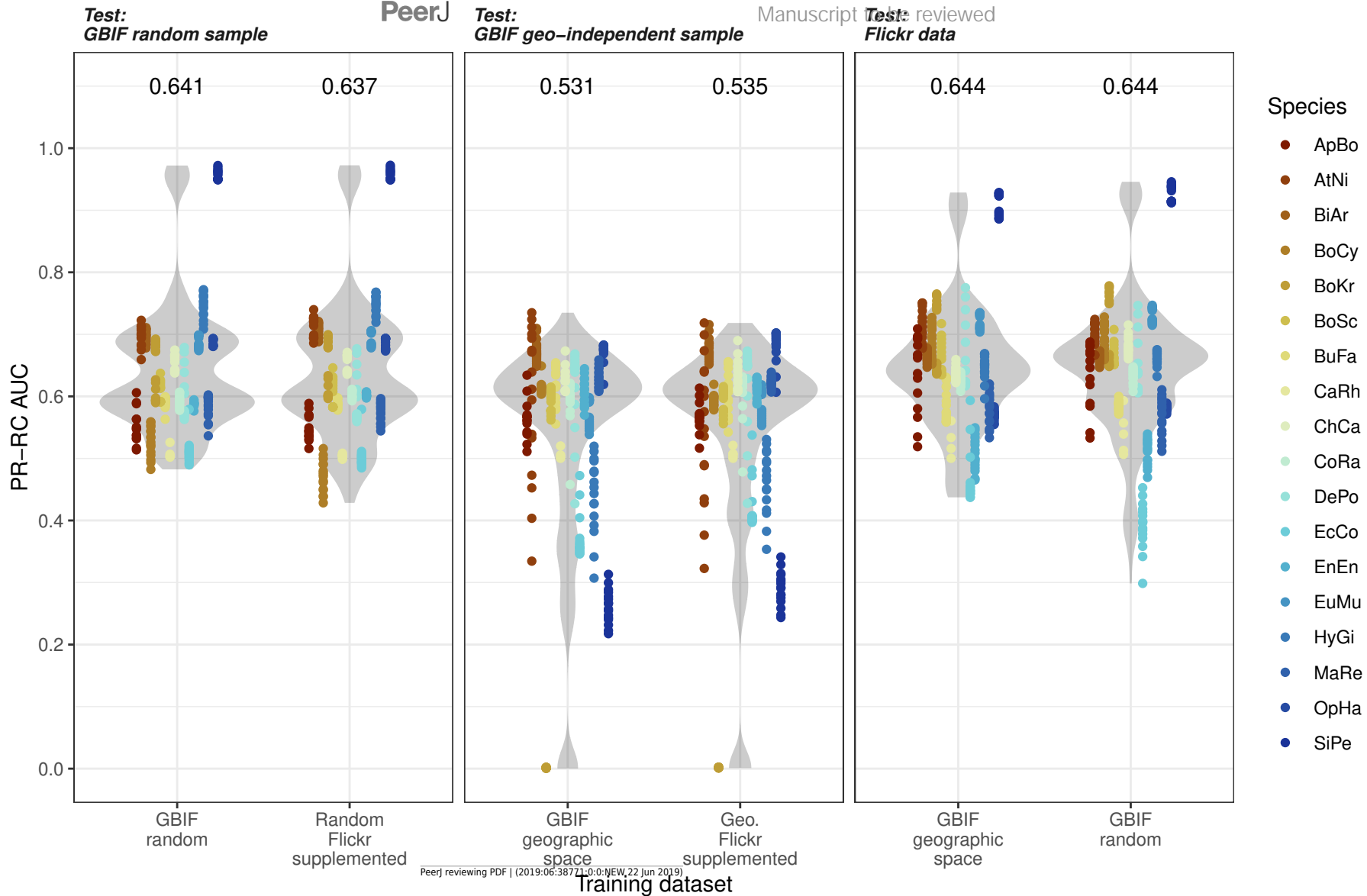


Figure 7 (on next page)

Schoener's D measure of niche overlap for models trained on GBIF-only and Flickr supplemented datasets

Right-hand side values show the overall niche overlap mean per species and the standard error. Species codes top to bottom: BoCy = *Boiga cynodon*, EnEn = *Enhydryis enhydryis*, AtNi = *Atheris nitschei*, OpHa = *Ophiophagus hannah*, CoRa = *Coelognathus radiatus*, SiPe = *Sinonatrix percarinata*, EcCo = *Echis coloratus*, MaRe = *Malayopython reticulatus*, EuMu = *Eunectes murinus*, HyGi = *Hydrodynastes gigas*, BiAr = *Bitis arietans*, ApBo = *Aplopeltura boa*, DePo = *Dendroaspis polylepis*, BoKr = *Boiga kraepelini*, BoSc = *Bothriechis schlegelii*, ChCa = *Chironius carinatus*, BuFa = *Bungarus fasciatus*, CaRh = *Calloselasma rhodostoma*.

