Journal of Tropical Ecology

www.cambridge.org/tro

When big data misleads: limitations of large-scale online biodiversity databases for species distribution modelling

Fabio Petrozzi[1] and Luca Luiselli[2,3,4]

[1]OrViCa Int, Italy; [2]Institute for Development, Ecology, Conservation and Cooperation, via G. Tomasi di Lampedusa 33, I-00144 Rome, Italy; [3]Department of Animal and Environmental Biology, Rivers State University, Port Harcourt, Nigeria and [4]Laboratoire d'Ecologie et d'Ecotoxicologie, Faculté des Sciences, Université de Lomé, Lomé, Togo

Research Article

Cite this article: Petrozzi F and Luiselli L (2026). When big data misleads: limitations of large-scale online biodiversity databases for species distribution modelling. *Journal of Tropical Ecology*. **42**(e24), 1–10. doi: https://doi.org/10.1017/S0266467426100686

Received: 14 March 2026 UTC
Revised: 7 July 2026 UTC
Accepted: 8 July 2026 UTC

Keywords:
Bitis nasicornis; data quality; GBIF; MaxEnt; rainforest ecology; species distribution models; tropical reptiles

Corresponding author:
Luca Luiselli; Email: l.luiselli@ideccngo.org

Abstract

Large-scale online biodiversity databases, particularly the Global Biodiversity Information Facility (GBIF), have become indispensable resources for macroecology, biogeography, and conservation biology. Their accessibility and extensive coverage have promoted widespread application in species distribution modelling, often without adequate assessment of data quality or ecological reliability. We evaluated the limitations of opportunistically collected occurrence records for ecological niche modelling using the West and Central African viper *Bitis nasicornis* as a case study. We compared MaxEnt models generated from GBIF records with models based on our own field observations collected between 1996 and 2025 across Liberia, Côte d'Ivoire, Ghana, Togo, and Nigeria, and with models integrating both datasets. Models built exclusively from GBIF data produced ecologically unrealistic predictions, identifying suitable habitats across rainforest, moist savanna, and dry savanna ecosystems despite the species' well-documented restriction to lowland tropical rainforests, typically near permanent water bodies. By contrast, models based on field-verified records closely matched the species' known ecological requirements and the current IUCN Red List distribution. Predicted suitable areas overlapped by only 8.6% between GBIF- and field-based models, revealing striking disagreement. Our analyses further suggest that this problem extends beyond B. nasicornis, with a substantial proportion of erroneous West African forest records originating from Benin, potentially affecting numerous rainforest taxa. We discuss how taxonomic uncertainty, spatial sampling bias, temporal mismatch, and the loss of ecological context compromise the reliability of large online datasets, and argue that robust field validation remains essential for ecological niche modelling and conservation decision-making.

Introduction

Over the last two decades, biodiversity science has undergone a profound transformation driven by the rapid expansion of online, open-access databases (e.g. Edwards 2004; Hampton *et al.* 2013; Soberón and Peterson 2004). Platforms such as the Global Biodiversity Information Facility (GBIF), VertNet, and iNaturalist now host hundreds of millions of georeferenced species occurrence records, providing unprecedented opportunities for large-scale ecological and biogeographical analyses. These resources have become especially influential in the field of species distribution modelling (SDM), where presence-only algorithms such as MaxEnt are routinely applied to online datasets to infer ecological niches, potential distributions, and responses to climate change (Anderson *et al.* 2011; Nelson and Ellis 2019; Sánchez-Mercado *et al.* 2010).

Despite their undeniable value, the increasing reliance on large-scale databases has raised concerns regarding data quality, ecological validity, and the uncritical use of "big data" in ecological inference (Beck *et al.* 2014; Hortal *et al.* 2015; Marcer *et al.* 2022). Occurrence records aggregated from heterogeneous sources (e.g., museum collections, citizen science platforms, environmental impact assessments, and unpublished surveys) are often treated as equivalent, even though they differ greatly in accuracy, precision, and ecological meaning. For species with narrow ecological requirements, cryptic behaviour, or limited detectability, these issues may lead to profoundly misleading conclusions (Boakes *et al.* 2010; Graham *et al.* 2004; Hughes *et al.* 2024; Soberón and Peterson 2004). Recent studies have increasingly emphasized the effects of sampling bias, spatial autocorrelation, uneven observer effort, and occurrence-data quality on SDM performance, particularly in tropical systems and for habitat specialists (Beck *et al.* 2014; Feng *et al.* 2019; Fourcade *et al.* 2014; Kramer-Schadt *et al.* 2013; Meyer *et al.* 2016; Yates *et al.* 2018). However, most of these studies have relied on multispecies datasets or broad taxonomic assemblages. By contrast, species-specific analyses focusing on ecologically well-known tropical specialists remain comparatively rare, despite their particular value for assessing the ecological

© The Author(s), 2026. Published by Cambridge University Press

realism of SDM outputs. Because predictions for such species can be evaluated directly against detailed natural-history knowledge accumulated through long-term field research, species-focused approaches may provide especially robust tests of ecological plausibility in SDMs. We implement this approach here using a detailed case study of a well-documented tropical specialist.

In this paper, we explore these limitations through a detailed case study of *Bitis nasicornis*, the nose-horned viper, a very large and highly distinctive viperid snake endemic to the lowland rainforests of West and Central Africa (IUCN 2025). This species is no doubt one of the most charismatic species in the world because of its massive size, bright colouration and bizarre horns on the head. Therefore, it is widely represented in online databases and frequently used in ecological and educational contexts. However, its strict association with humid forest environments (e.g., see Luiselli 2006a; Spawls and Branch 2018) makes it particularly sensitive to potential errors in occurrence data and model extrapolation. Indeed, because *B. nasicornis* is a strict rainforest specialist with a narrow realised niche (Luiselli 2006a), species distribution models for this species are particularly sensitive to spatially imprecise or ecologically inconsistent occurrence records, which can disproportionately inflate inferred niche breadth and lead to ecologically unrealistic predictions (Aubry *et al.* 2017; Moudrý *et al.* 2024). Moreover, this species has a widely overlapping range in West Africa with other rainforest specialists (for instance, the tortoises *K. homeana* and *Kinixys erosa*, see the respective maps in IUCN 2025 and Luiselli 2006b), and therefore can be used as a proxy for the rainforest taxonomic communities inhabiting this African region.

By comparing SDMs based on GBIF records with models dcerived from long-term, field-verified occurrence data collected over nearly three decades, we aim not only to evaluate the effects of occurrence-data provenance on model outputs but also to illustrate the value of species-specific approaches for assessing ecological realism in SDMs. Because the ecology and habitat requirements of *B. nasicornis* are well documented (see citations above), this species provides an ideal framework for identifying biologically implausible predictions that may remain undetected in broader multi-species analyses. Specifically, we aim to (1) illustrate how large online databases can produce ecologically nonsensical predictions, (2) identify the underlying causes of these discrepancies, and (3) discuss broader implications for the use of big biodiversity data in ecology and conservation.

Study species: *Bitis nasicornis*

B. nasicornis (Shaw 1802) is a large, terrestrial viper belonging to the family Viperidae. It is distributed across the Guineo-Congolian region of Africa, from southeastern Guinea through Liberia, Côte d'Ivoire, Ghana, and Nigeria, and eastwards into Central Africa (IUCN 2025; Spawls and Branch 2018). The species is easily recognizable by its vivid coloration, prominent nasal horns, and cryptic patterning adapted to the leaf-littered forest floor (Figure 1).

Ecologically, *B. nasicornis* is a habitat specialist. Numerous field studies and natural history observations describe it as strongly associated with lowland tropical rainforests (Barlow *et al.* 2019; Bombi *et al.* 2011; Spawls and Branch 2018; Vaughan 2023), particularly primary and mature secondary forests with high humidity, dense canopy cover, and deep leaf litter (Luiselli 2006a; Shine and Spawls 2020). The species is rarely, if ever, encountered in open habitats or dry forests, and its physiological and behavioural traits are consistent with a rainforest-adapted ambush predator. This well-defined ecological niche makes *B. nasicornis* an ideal model organism for testing the ecological plausibility of SDMs generated from different data sources.



Figure 1. Juvenile nose-horned viper (*Bitis nasicornis*) from Kpalimé area, south-western Togo. Photo: Luca Luiselli.

Materials and methods

Occurrence data

Two independent datasets were used to model the potential distribution of *B. nasicornis*. The first dataset consisted of occurrence records downloaded from GBIF (Online Supplementary Table S1). GBIF-derived data are available at https://doi.org/10.15468/dd.yv4g5z. Records were initially screened to remove obvious spatial and taxonomic inconsistencies following current biodiversity-data cleaning recommendations (Zizka *et al.* 2019). Specifically, we removed: (i) duplicated coordinates, (ii) records located in oceans or major water bodies, (iii) country-centroid records, (iv) records corresponding to museum or biodiversity-institution headquarters, (v) records with coordinate uncertainty exceeding 10 km, and (vi) records lacking temporal metadata. However, no direct ecological or specimen-level verification was possible for most records, as is typical in SDM studies relying on aggregated online databases.

The second dataset comprised original occurrence records collected by the authors between 1996 and 2025 during systematic and opportunistic field surveys in Liberia, Côte d'Ivoire, Ghana, Togo, and Nigeria (Table S2). All records were based on direct observations or captures, georeferenced with GPS, and accompanied by detailed habitat descriptions. Most records occurred in rainforests or in forest-plantation mosaics and were also used for other ecological studies that were published on the target species (Luiselli and Akani 2003; Luiselli 2006a, 2006b).

Environmental variables and species distribution modelling

We downloaded the shapefile of the African continent from the OpenAfrica repository (https://bulk.openafrica.net/en_AU/) and subsequently extracted the following countries to delimit the study area: Guinea, Sierra Leone, Côte d'Ivoire, Ghana, Togo, Benin, and Nigeria. The major vegetation zones of West Africa, which provide context for the observed distributions and model predictions, are

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press

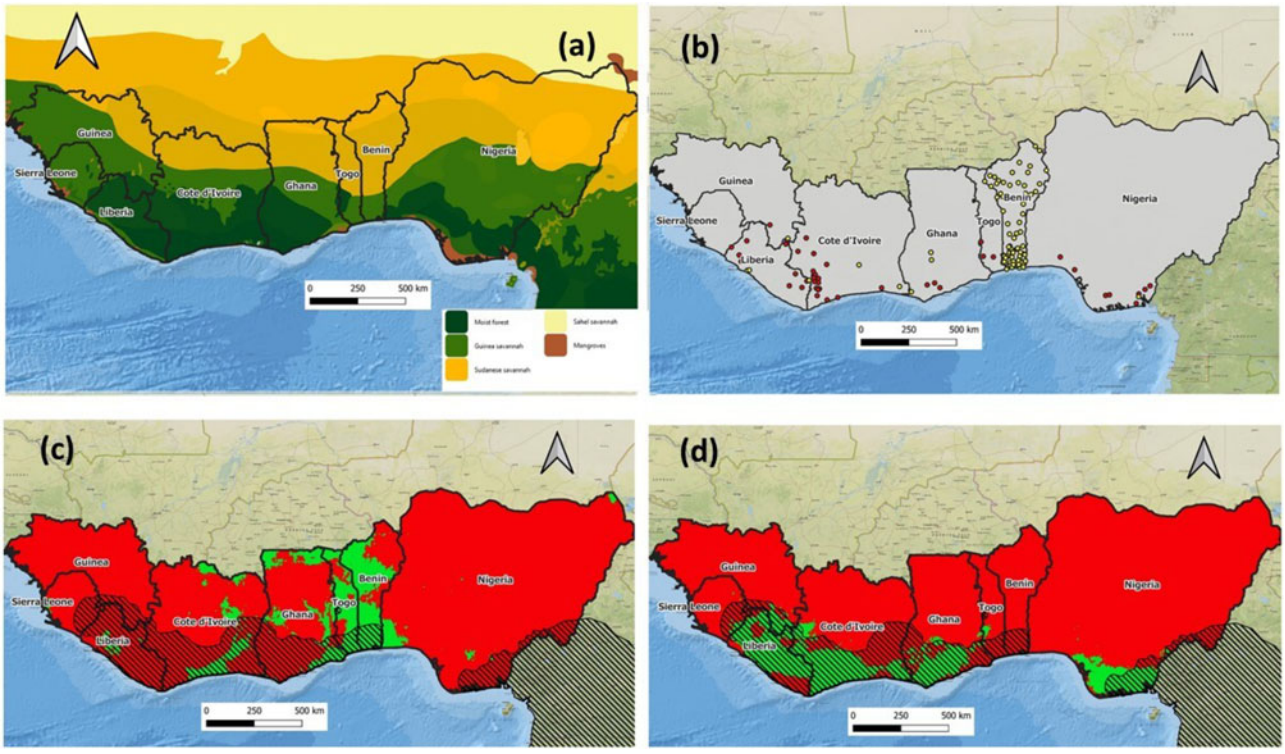


Figure 2. (a) Major vegetation types across the study area. (b) Geographical extent of the study area in West Africa, encompassing Liberia, Côte d'Ivoire, Ghana, Togo, and Nigeria, with occurrence records of *Bitis nasicornis*. GBIF records are shown as yellow dots, while field-verified records collected by the authors between 1996 and 2025 are shown as red dots. GBIF records exhibit a broad and scattered distribution, including areas dominated by savannah and forest–savannah mosaics, whereas field-verified records are tightly clustered within lowland tropical rainforest regions. (c) Predicted potential distribution of *Bitis nasicornis* based on MaxEnt models using GBIF occurrence records, compared with the IUCN Red List range map (shaded area). Green indicates areas of predicted presence, and red indicates areas of predicted absence. The model predicts extensive suitable habitat across moist and dry savannas, forest–savannah mosaics, and regions lacking closed-canopy rainforest, suggesting that heterogeneous and ecologically inconsistent input data may weaken the ecological signal. (d) Predicted potential distribution of *Bitis nasicornis* based on MaxEnt models using field-verified occurrence records, compared with the IUCN Red List range map (shaded area). Green indicates areas of predicted presence, and red indicates areas of predicted absence. High suitability is largely restricted to lowland tropical rainforest regions characterized by high annual precipitation and low climatic seasonality, with savannah and dry forest regions consistently predicted as unsuitable.

shown in Figure 2a. We selected 19 bioclimatic variables as covariates, obtained from WorldClim 2.1 (Fick and Hijmans 2017) at a spatial resolution of 2.5 arc-minutes. To minimize multicollinearity and enhance model interpretability, highly correlated predictors were excluded prior to model calibration. Specifically, we computed the pairwise Pearson correlation coefficient for all potential bioclimatic variables and adopted a conservative threshold of $|r| > 0.7$, frequently used in species distribution studies to indicate problematic collinearity (e.g., Bombi *et al.* 2011). When two variables exceeded this threshold, we removed the one with the highest mean absolute correlation with all remaining predictors. This procedure was repeated until all pairwise correlations among retained variables were ≤ 0.7, resulting in a final set of 10 bioclimatic variables that provide independent information, enhancing both model stability and interpretability.

Species distribution models were generated using MaxEnt 3.4.3, a presence-background algorithm widely applied in ecological studies. Occurrence records from GBIF were retrieved using QGIS 3.34.11. The study area was delineated using a shapefile of the African continent, from which the countries included in this study (Guinea, Sierra Leone, Côte d'Ivoire, Ghana, Togo, Benin, and Nigeria) were extracted to generate a study-area shapefile. GBIF records were downloaded via the "GBIF Occurrences" tool, with the temporal range restricted to 1990–2020 to match the period of our field surveys. Records were spatially filtered by intersecting specimen localities with the study-area shapefile, retaining only those within the study area. Attribute tables were subsequently cleaned to retain only species names and geographic coordinates and were exported in CSV format.

Occurrence records from both the field-verified dataset and the GBIF dataset were spatially thinned prior to modelling to reduce spatial autocorrelation and mitigate sampling bias. A minimum nearest-neighbour distance of 10 km was applied, ensuring that spatially clustered records from intensively surveyed localities did not disproportionately influence model calibration. To address sampling bias in environmental space, background points were generated within a biologically meaningful accessible area (M region sensu Barve *et al.* 2011), defined as the Upper Guinean forest block and adjacent forest–savanna transition zones potentially accessible to *B. nasicornis* over relevant ecological timescales. This restriction avoids unrealistic extrapolation beyond the species' dispersal and biogeographical context. Within this accessible area, background points were randomly sampled to capture environmental heterogeneity while maintaining spatial representativeness across climatic gradients.

Species distribution models were developed separately for the field-verified dataset and the GBIF dataset, employing identical environmental variables and model settings. Additionally, two further models were developed: one combining both datasets and another calibrated using all available field records, with

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press

background points randomly sampled from the six countries where occurrence records had been collected (that is, excluding Benin and Sierra Leone as we did not collect original field data from these two countries). The final habitat suitability map was projected across the entire eight-country region, including Benin and Sierra Leone, to illustrate the potential distribution of the species in areas without occurrence data. An additional GBIF-based model excluding Benin was considered but not produced due to the low number of records. We also did not implement a true presence–absence modelling approach because our datasets consisted exclusively of presence-only occurrence records derived from field observations and GBIF data, without systematically collected absence data. Under these conditions, presence-only approaches such as MaxEnt are generally considered more appropriate and statistically robust for estimating species distributions, particularly when sampling effort is heterogeneous across the study area (Campos *et al.* 2023). Although pseudo-absence or background approaches can be incorporated into alternative modelling frameworks, these methods may introduce additional uncertainty when true absences are unavailable or when survey effort is spatially uneven. In our study, we therefore retained the MaxEnt framework because it is specifically designed for presence-background data and is widely applied in ecological niche and species distribution modelling (SDM) studies under comparable data conditions

All MaxEnt modelling procedures were conducted in R (RStudio) using the *dismo* package (Hijmans *et al.* 2017) (see Table S3 for the script). Model performance was assessed through five-fold cross-validation. Multiple MaxEnt parameterizations were tested, including adjustments to feature classes, to account for differences in occurrence record numbers, and to improve model robustness while reducing overfitting. Models were calibrated using presence-only occurrence data and background points randomly sampled within the study area. Because the field-verified dataset and the GBIF dataset differed substantially in sample size and spatial heterogeneity, different background-point densities were used: for the field-verified dataset, the number of background points was scaled proportionally to occurrence records; for the GBIF dataset, which contained substantially more records across a broader environmental space, a maximum of 5,000 background points was used to adequately characterize environmental availability without increasing model complexity (Merow *et al.* 2013; Muscarella *et al.* 2014) (Table 1). Model outputs were visualized as continuous suitability maps and binary predictions, interpreted in the context of known habitat preferences and vegetation types.

The continuous habitat suitability raster (values ranging from 0 to 1) generated by the final MaxEnt model was converted into a binary suitable–unsuitable habitat map through thresholding. The threshold was defined as the median suitability value predicted at the occurrence points used for model calibration. This approach approximately retains 50% of the training occurrence records as predicted suitable habitat. Raster cells with suitability values equal to or greater than the median threshold were assigned a value of 1 (suitable), and those below the threshold were assigned a value of 0 (unsuitable). The resulting binary raster was exported in GeoTIFF format. In QGIS, binary rasters were summed using the Raster Calculator to produce a composite layer with values of 0, 1, or 2. Cells with a value of 2 represent areas where both models predicted suitable habitat (model agreement), whereas cells with a value of 1 represent areas predicted as suitable by only one model (model disagreement). The overlapping suitable habitat raster was extracted to quantify spatial concordance between models. Zonal statistics were then used to calculate the number of overlapping pixels relative to the total number of suitable pixels in the GBIF-based model, providing an estimate of percentage spatial overlap.

All models were therefore calibrated using spatially thinned occurrence data and background points restricted to the accessible area, reducing the effects of spatial autocorrelation and sampling bias.

Table 1. Summary of MaxEnt modelling parameters and settings for *Bitis nasicornis* using field-verified and GBIF occurrence datasets. Both models used the same set of 10 bioclimatic variables and identical feature classes (linear and quadratic). The number of occurrence records differed between datasets, and background points were assigned proportionally to the number of presences (field data) or capped at 5,000 (GBIF data). Model performance was evaluated using 5-fold cross-validation, and predictive accuracy was assessed with AUC, TSS, sensitivity, and specificity. Thresholds for presence–absence classification were determined using the optimal Youden's J statistic

Parameter/step	Model with field data	Model with GBIF data
Number of occurrence records	44	167
Environmental variables	10 bioclim layers	10 bioclim layers
Feature classes used	Linear, quadratic	Linear, quadratic
Background points	4,400	5,000
Cross-validation	5-fold	5-fold
Evaluation metrics	AUC, TSS, sensitivity, specificity	AUC, TSS, sensitivity, specificity
Threshold selection	Optimal Youden's J	Optimal Youden's J

Results

Occurrence records from both GBIF and field surveys are given in Figure 2b. MaxEnt models based on GBIF data predicted extensive suitable habitat, including regions outside the species' known rainforest preference, indicating a weak ecological signal likely driven by heterogeneous and ecologically inconsistent input data (Figure 2c). In contrast, models based on field-verified occurrences produced predictions closely aligned with the species' known ecological niche, with high suitability largely restricted to lowland tropical rainforest and consistently low suitability in savanna and dry forest regions (Figure 2d). The GBIF-based model predicted potential presences not only in lowland rainforest regions but also across large areas of moist savanna, forest–savanna mosaics, and relatively dry savanna zones. Consequently, this model showed little correspondence with major biogeographical boundaries, such as the Dahomey Gap, and predicted suitability in areas known to lack closed-canopy forest. Conversely, the field-verified model correctly identified the Dahomey Gap as largely unsuitable while capturing the species' ecological preference for regions characterized by high precipitation, low seasonality, and dense forest cover. Overall, the field-verified predictions exhibited remarkable congruence with the IUCN Red List range for West Africa.

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press

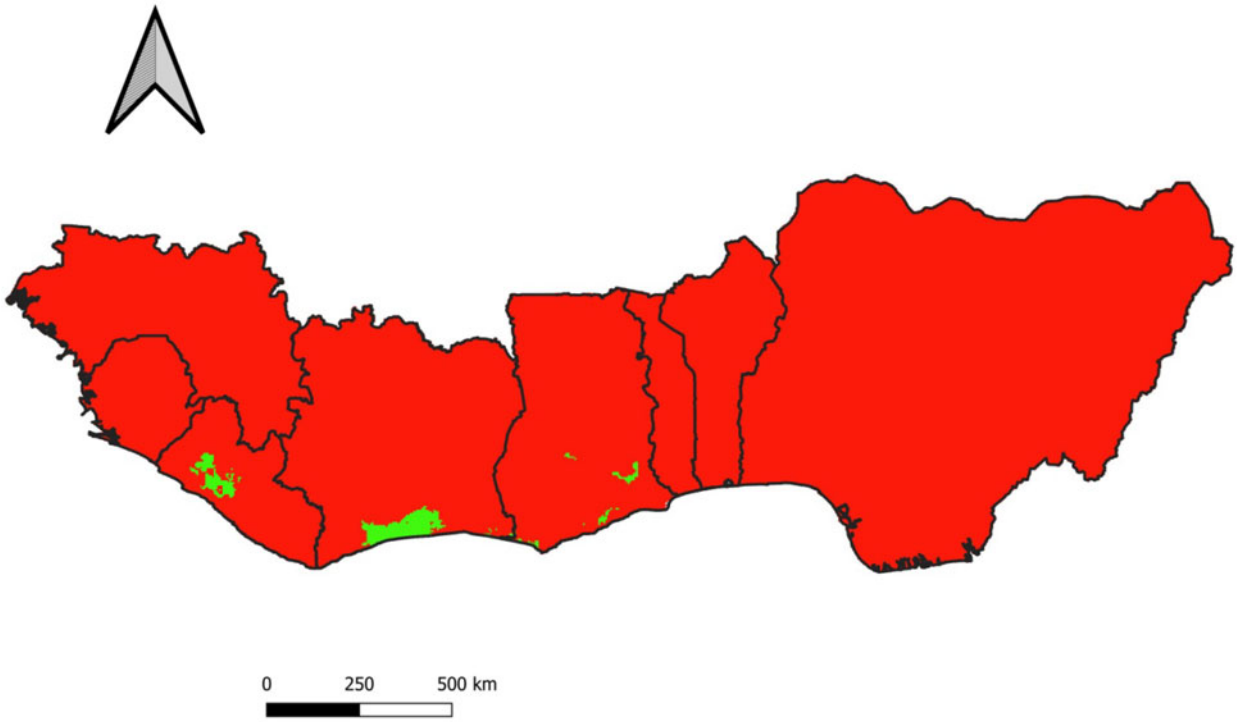


Figure 3. Geographic areas where predictions from the GBIF-based and field-verified MaxEnt models overlap, indicated in green.

Binary presence–absence maps generated from the GBIF-based and field-verified MaxEnt models were compared to assess spatial concordance. Only 8.6% of predicted presence areas overlapped between the two models, indicating a very limited agreement in predicted suitable habitats (Figure 3).

Model accuracy was assessed using the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and the True Skill Statistic (TSS). The field-verified model outperformed the GBIF-based model across most indicators, achieving a slightly higher AUC (0.95 vs. 0.916), considerably greater sensitivity (0.932 vs. 0.805), and much higher TSS (0.804 vs. 0.694) (Table 2). Although specificity was slightly lower in the field-verified model (0.872 vs. 0.889), cross-validation confirmed superior overall predictive performance for the field dataset, with consistently higher mean AUC and sensitivity.

Bioclimatic variable contributions differed markedly between datasets (Table 3). The GBIF-based model was primarily driven by temperature-related variables, particularly minimum temperature of the coldest month (BIO6), whereas the field-verified model was overwhelmingly dominated by precipitation of the driest month (BIO14), with only minor contributions from temperature extremes and climatic seasonality.

The model constructed by combining our field data with those derived from GBIF (Figure 4), despite exhibiting strong apparent statistical performance (Table 2), did not produce an ecologically meaningful pattern. In particular, the entire Dahomey Gap savannah region is predicted as suitable for the target species, whereas the adjacent Guinea savannahs in Ghana and Nigeria are not (Figure 4), resulting in an ecologically inconsistent prediction and failing to reflect a reliable signal of forest habitat use.

The fourth model, calibrated using all available field records and with background points randomly sampled from the six countries where occurrence records had been collected, is shown in Figure 5. This model not only exhibited strong statistical performance (Table 2) but also produced a macroecological pattern consistent with the IUCN distribution map and the known ecology of the species. In particular, predicted suitability closely follows the distribution of the Upper and Lower Guinean forest blocks (Figure 5).

Overall, 52% of the model generated using GBIF data overlaps with the model built using the combined datasets (Online Supplementary Fig. S1) and only 37% of the suitable area predicted by the model based on field-collected data overlaps with the model constructed using the combined datasets (Fig. S2). When comparing the model based on field data with background points sampled across the entire study area to the model based on field data with background points restricted to the countries where the species is present, the latter model shows nearly identical spatial extent to the former, with a 98% overlap relative to the previous model (Fig. S3).

Discussion

The GBIF-based model predicted potential occurrences of *B. nasicornis* not only within lowland rainforest regions—its documented habitat (Barlow *et al.* 2019; Bombi *et al.* 2011; Luiselli 2006a; Spawls and Branch 2018; Vaughan 2023)—but also across extensive moist savannas, forest–savanna mosaics, and comparatively drier savanna zones, where the species does not occur at all. Consequently, the model exhibited limited concordance with major biogeographical barriers, such as the Dahomey Gap, and identified areas of suitability in regions lacking closed-canopy forest. In contrast, the field-verified model accurately represented the Dahomey Gap as largely unsuitable and captured the species' ecological association with areas characterized by high precipitation, low seasonality, and dense forest cover. Its predictions closely

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press

Table 2. Performance metrics of MaxEnt models for *Bitis nasicornis* using GBIF and field-verified occurrence datasets. Model accuracy was assessed using the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and the True Skill Statistic (TSS). Optimal thresholds for converting continuous suitability scores into presence–absence predictions were determined using Youden's J statistic. Cross-validation (5-fold) results are presented as mean ± standard deviation, highlighting differences in predictive performance and model reliability between the GBIF-based and field-verified datasets. Standard errors are indicated in the cross-validation statistics

Metric	Model with GBIF data	Model with field-verified data	Model with both dataset	Model without Benin and Sierra Leone background points
AUC	0.916	0.95	0.89	0.933
Sensitivity	0.805	0.932	0.901	0.977
Specificity	0.889	0.872	0.723	0.779
TSS	0.694	0.804	0.623	0.773
Optimal threshold	0.382	0.39	0.349	0.271
Cross-validation (5-fold)				
AUC	0.870 ± 0.059	0.936 ± 0.023	0.866 ± 0.015	0.909 ± 0.027
Sensitivity	0.793 ± 0.090	1.000 ± 0.000	0.832 ± 0.069	1.000 ± 0.000
Specificity	0.912 ± 0.043	0.821 ± 0.052	0.790 ± 0.089	0.791 ± 0.058
TSS	0.705 ± 0.077	0.821 ± 0.052	0.623 ± 0.038	0.791 ± 0.058

Table 3. Performance metrics of MaxEnt models for *Bitis nasicornis* based on GBIF and field-verified occurrence records. Values include overall model evaluation metrics (AUC, TSS, sensitivity, specificity, and optimal threshold) and cross-validation results (mean ± SD) from five-fold cross-validation. The field-verified model outperformed the GBIF-based model across most indicators, indicating more reliable predictions of species' ecological preferences

Bioclimatic variable contributions GBIF data			Bioclimatic variable contributions field verified data		
Code	Variable description	Percent contribution (%)	Code	Variable description	Percent contribution (%)
BIO6	Minimum temperature of the coldest month	45.1	BIO14	Precipitation of the driest month	86.3
BIO19	Precipitation of the coldest quarter	12.4	BIO5	Maximum temperature of the warmest month	5.5
BIO3	Isothermality	12.2	BIO4	Temperature seasonality (standard deviation × 100)	3.4
BIO12	Annual precipitation	11.0	BIO9	Mean temperature of the driest quarter	2.1
BIO13	Precipitation of the wettest month	5.6	BIO8	Mean temperature of the wettest quarter	1.4
BIO14	Precipitation of the driest month	5.3	BIO3	Isothermality	0.4
BIO15	Precipitation seasonality (coefficient of variation)	4.4	BIO17	Precipitation of the driest quarter	0.4
BIO 17	Precipitation of the driest quarter	1.7	BIO13	Precipitation of the wettest month	0.2
BIO18	Precipitation of the warmest quarter	0.8	BIO18	Precipitation of the warmest quarter	0.1
BIO9	Mean temperature of the driest quarter	0.5	BIO1	Annual mean temperature	0.1

aligned with the West African range delineated by the IUCN (2025) Red List. A similar pattern was observed when the GBIF-based model was overlaid with the model derived from combined datasets, indicating that the ecological signal was substantially weakened in models relying, even partially, on GBIF data. These results indicate that models derived from verified field records provide more reliable and ecologically coherent predictions than those based on heterogeneous GBIF occurrences, which can generate misleading patterns and overestimate suitable habitat outside the species' true ecological range.

Overall, the two main models (original data versus GBIF data) showed only 8.6% agreement in presence–absence predictions, underscoring the limitations of relying exclusively on large aggregated occurrence datasets for species distribution modeling, especially for habitat specialists linked to tropical rainforests. Whereas the GBIF-based model generated an ecologically implausible broad distribution, the field-verified model restricted suitability to lowland tropical rainforests, consistent with the species' known ecological requirements.

The primary source of error appears to stem from records attributed to Benin, which include several potentially unreliable occurrence entries. In this country, forested patches are scarce and highly fragmented, typically restricted to small sacred groves. During our long-term field surveys conducted over multiple decades, we also

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press

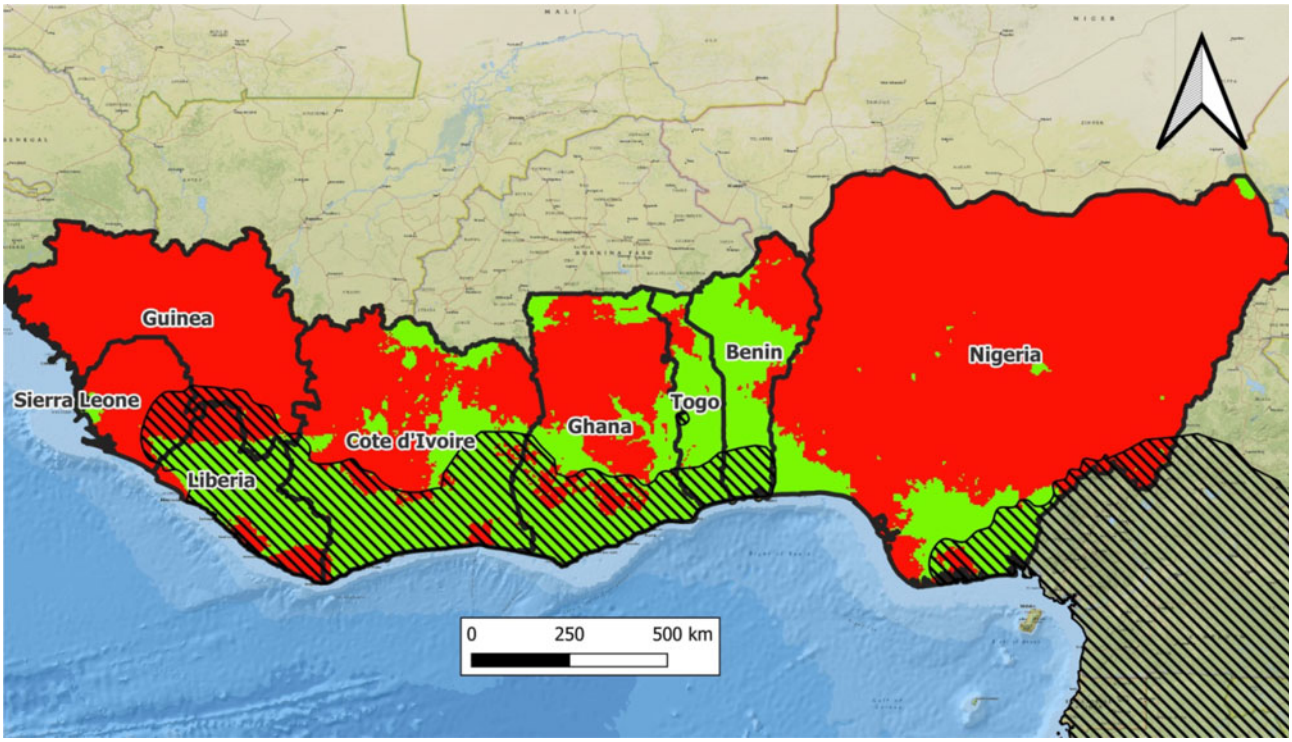


Figure 4. Predicted potential distribution of *Bitis nasicornis* based on MaxEnt models using both original field data and GBIF occurrence records, compared with the IUCN Red List range map (shaded area). Green indicates areas of predicted presence, and red indicates areas of predicted absence.

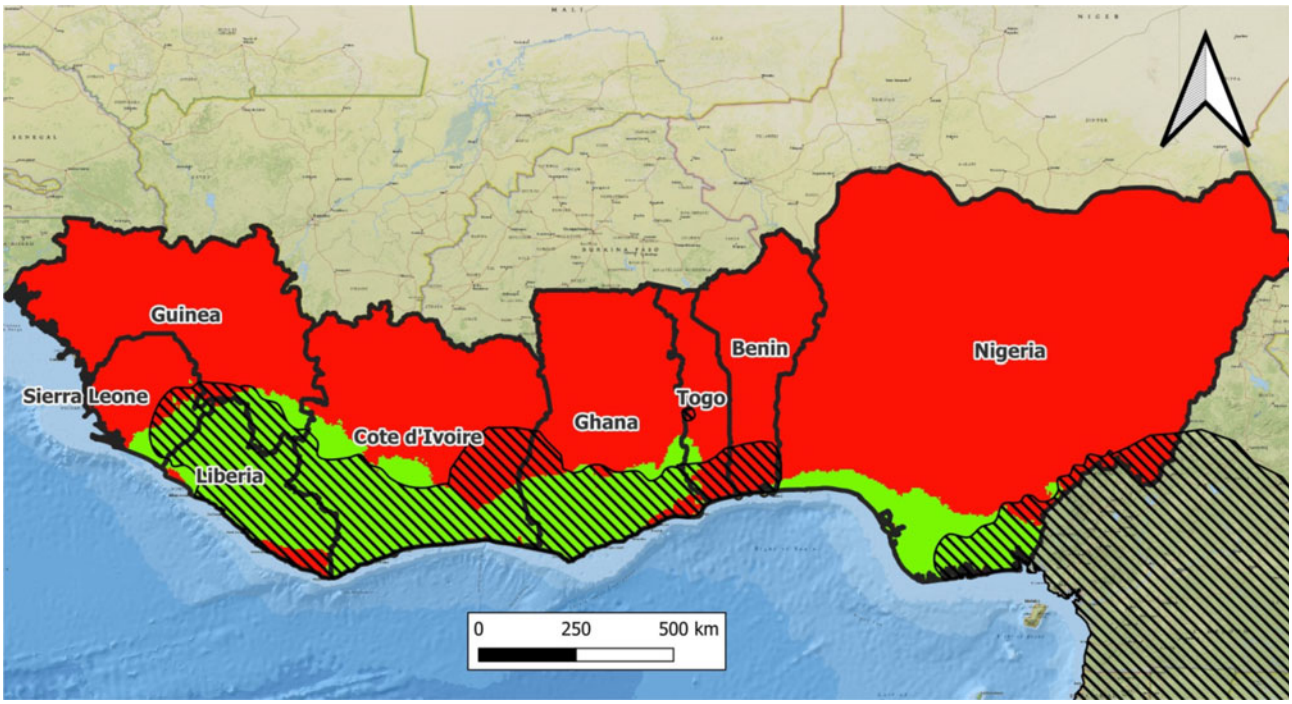


Figure 5. Predicted potential distribution of *Bitis nasicornis* based on MaxEnt models calibrated using all available field records and with background points randomly sampled from the six countries where occurrence records had been collected.

investigated areas of Benin where the presence of *B. nasicornis* is reported in the GBIF database; however, we did not detect any individuals despite targeted searches in suitable habitats. These observations suggest that several occurrence records within aggregated biodiversity databases may be affected by mis-identification, geo-referencing errors, or outdated information, potentially leading to misleading model outputs when not carefully validated. The same problem arose for several other species, implying

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press

that Benin records may compromise the reliability of GBIF database for a potentially high number of taxa, and not just for reptiles.

From an ecological perspective, it cannot be excluded that small remnant populations of *B. nasicornis* may persist in residual sacred grove forests. However, if this is the case, such occurrences would introduce a substantial bias into the MaxEnt modelling framework. Specifically, the modelling procedure may still be driven by the environmental signal of the extensive surrounding deforested matrix, thereby attributing habitat suitability to large areas across West Africa that are, in reality, unsuitable for the species and not used even during dispersal movements (our unpublished radio-tracking studies).

Sources of error in large-scale online databases

The discrepancies between the two models highlight several well known but often underestimated limitations of large online biodiversity databases. First, taxonomic uncertainty remains a major issue. Misidentifications, outdated taxonomy, and confusion with similar species can introduce records from inappropriate habitats. In the case of *B. nasicornis*, confusion with other *Bitis* species adapted to more open habitats cannot be excluded but seems unlikely. In fact, the only other similar species is *Bitis gabonica* in Nigeria and *Bitis rhinoceros* in Togo, Ghana, Cote d'Ivoire, and Liberia (see Spawls and Branch 2018), but also these two taxa are also forest-dwelling species that could not occur in savannah or dry lands.

Second, temporal mismatches between occurrence records and environmental predictors can potentially distort species distribution models (Araújo *et al.* 2005; Hijmans 2012). Climate regimes and land-cover patterns in West Africa have undergone substantial changes over the past century, and historical occurrence records might therefore reflect ecological conditions that no longer exist. In the present study, however, the temporal span of our field surveys (1994–2025) largely overlaps with that of the GBIF records used for modelling (1990–2020), whereas the bioclimatic layers represent long-term climatic averages (1970–2000). Despite this partial temporal mismatch, it is unlikely to be the primary source of error in the GBIF-based model, as the model predicts the presence of *B. nasicornis* in regions (such as dry Sudanian savana) where the species is known to have never occurred and where environmental conditions have remained broadly stable for centuries.

Finally, and perhaps most critically, large online biodiversity databases are characterized by a substantial loss of ecological context. Presence-only records typically lack information on habitat type, microhabitat use, and species detectability. Consequently, species distribution models based on such data tend to capture broad climatic envelopes while overlooking fine-scale ecological constraints that are essential for habitat-specialist taxa. We argue that this loss of ecological information represents the primary cause of the ecologically unrealistic patterns observed in the GBIF-based model for *B. nasicornis*. Similar modelling artefacts have been observed by us when modelling the niche of other West African forest specialists (e.g. the tortoises *Kinixys homeana*, *K. erosa*, the snakes *B. gabonica*, *B. rhinoceros,* the crocodiles *Mecistops cataphractus,* and *Osteolaemus tetraspis*) using GBIF data, showing that the problem is far wider than just the nose-horned viper. Although these additional cases are not presented here in order to maintain the focus and conciseness of the present study, in this study, we uncovered a likely broad problem with big data for West African tropical forest taxa in general. For instance, the use of GBIF data in this modelling technique did not work well even considering a multispecies assemblage of West African butterflies (Petrozzi and Luiselli 2026).

Importantly, our study differs from many previous assessments of SDM bias because it focuses on a single, **ecologically well-characterized rainforest specialist** rather than a multispecies assemblage. Most recent evaluations of SDM bias have focused on large multispecies datasets or broad taxonomic assemblages, where ecological inconsistencies may remain difficult to detect because detailed natural-history knowledge is unavailable for many taxa. By contrast, *B. nasicornis* is among the best-known rainforest snakes in West Africa (Efenakpo *et al.* 2026; Spawls and Branch 2018), with a remarkably consistent ecological association with humid lowland forests documented across decades of field studies (Luiselli 2006a). This ecological specificity allows unusually robust evaluation of model realism and makes biologically implausible predictions immediately recognizable. Our results therefore demonstrate how species-focused analyses can complement broader multispecies assessments by revealing errors and ecological artefacts that may otherwise remain hidden within large-scale comparative datasets. In other words, this species-specific approach using a well-known taxon allows direct evaluation of ecological realism of Maxent modelling approaches against decades of field experience and natural-history knowledge, thereby providing a particularly robust framework for identifying biologically implausible model outputs.

The marked differences in bioclimatic variable importance between the GBIF-based and field-verified models indicate substantial divergence in the estimated climatic niche inferred from the two datasets. The field-based model is primarily structured by precipitation-related variables, consistent with the species' strong ecological association with humid rainforest environments. In contrast, the GBIF-based model shows a stronger influence of temperature-related variables, likely reflecting the inclusion of environmentally heterogeneous and ecologically inconsistent occurrence records. These inconsistencies appear to broaden the estimated climatic space and generate an apparent inflation of climatic tolerance, resulting in predictions that extend into ecologically implausible areas outside the species' known habitat associations.

Implications for species distribution modelling and the value of field-based data

Our results demonstrate that even widely accepted modelling approaches such as MaxEnt cannot compensate for fundamentally flawed input data. When occurrence records violate basic ecological assumptions, model outputs may appear statistically robust (like with AUC statistics in our study case) while being biologically meaningless. This issue is particularly acute in tropical regions, where sampling is uneven, habitats are highly heterogeneous, and many species have narrow ecological niches. Blind reliance on large online datasets risks producing misleading large-scale maps that may influence conservation planning, environmental impact assessments, and policy decisions.

The strong concordance between the field-based model and the IUCN Red List distribution (IUCN 2025) undoubtedly shows the continued importance of long-term, field-based research. Although such datasets are often spatially limited and resource intensive to obtain, they retain a level of ecological validity that is frequently lost in aggregated online databases. Rather than viewing large-scale biodiversity data and field-based observations as

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press

competing approaches, we advocate for their complementary integration. Online databases should be used primarily as hypothesis-generating resources, with model outputs critically evaluated and refined through expert knowledge and targeted field verification. Furthermore, we suggest that authors conducting large-scale spatial analyses should more carefully assess the reliability and ecological consistency of presence data for their focal taxa, as our results indicate that predictions for West African rainforest specialists may be substantially biased or ecologically misleading, thereby compromising the robustness of large-scale biogeographical inferences in this region, and likely at the continental level too. For habitat specialists, particularly in tropical regions, uncritical use of big data may obscure rather than illuminate ecological patterns. Robust SDM requires not only sophisticated algorithms but also careful data curation, ecological expertise, and, whenever possible, field-verified observations. As biodiversity science continues to embrace big data, a renewed emphasis on ecological realism and data quality will be essential to ensure that our models reflect biological reality rather than statistical artefacts.

For tropical habitat specialists, we recommend that SDM studies place greater emphasis on species-specific ecological validation rather than relying exclusively on statistical performance metrics. In particular, model outputs should be systematically evaluated against independent natural-history knowledge, vegetation structure, and known biogeographical barriers. Species-focused validation approaches may be especially valuable in tropical regions, where incomplete sampling and heterogeneous occurrence databases can substantially inflate estimated niche breadth and obscure biologically meaningful distributional limits.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0266467426100686.

Data availability statement. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. All the dataset concerning the original distribution records is available in Online Supplementary Table S2. GBIF-derived data are available at https://doi.org/10.15468/dd.yv4g5z.

Acknowledgements. We are indebted with two anonymous referees for the helpful comments on the submitted draft.

Author contribution. Conceptualization: FP, LL; Data curation: FP; Data collection: LL; Formal analysis: FP; Methodology: FP; Supervision: LL; Validation: LL; Writing – original draft: LL; Writing – review & editing: FP, LL

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors. However, field data on *Bitis nasicornis* were collected during expeditions funded by IDECC (Rome, Italy), Turtle Survival Alliance (USA), Mohamed Bin Zayed Species Conservation Fund (United Arab Emirates), and National Geographic (USA). Grant numbers are not available.

Competing interests. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics standards. The target species is not protected by any laws in any of the surveyed countries, so no permit is required. Anyway, we did not damage any of the observed specimens.

References

Anderson RP, Martínez-Meyer E, Nakamura M, Araújo MB, Peterson AT, Soberón J and Pearson RG (2011) *Ecological Niches and Geographic Distributions (MPB-49)*. Princeton: Princeton University Press. https://doi.org/10.1515/9781400840670.

Araújo MB, Pearson RG, Thuiller W and Erhard M (2005) Validation of species–climate impact models under climate change. *Global Change Biology* **11**(9), 1504–1513. https://doi.org/10.1111/j.1365-2486.2005.01000.x.

Aubry KB, Raley CM and McKelvey KS (2017) The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. *PLoS One* **12**(6), e0179152. https://doi.org/10.1371/journal.pone.0179152.

Barlow A, Wüster W, Kelly CM, Branch WR, Phelps T and Tolley KA (2019) Ancient habitat shifts and organismal diversification are decoupled in the African viper genus *Bitis* (Serpentes: Viperidae). *Journal of Biogeography* **46**(6), 1234–1248.

Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT and Villalobos F (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling* **222**(11), 1810–1819.

Beck J, Böller M, Erhardt A and Schwanghart W (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* **19**, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Boakes EH, McGowan PJ, Fuller RA, Chang-qing D, Clark NE, OConnor K and Mace GM (2010) Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology* **8**(6), e1000385. https://doi.org/10.1371/journal.pbio.1000385.

Bombi P, Akani GC, Ebere N and Luiselli L (2011) Potential effects of climate change on high-and low-abundance populations of the Gaboon viper (*Bitis gabonica*) and the nose-horned viper (*B. nasicornis*) in southern Nigeria. *The Herpetological Journal* **21**(1), 59–64.

Campos JC, Garcia N, Alírio J, Arenas-Castro S, Teodoro AC and Sillero N (2023) Ecological niche models using MaxEnt in Google Earth Engine: Evaluation, guidelines and recommendations. *Ecological Informatics* **76**, 102147.

Edwards JL (2004) Research and societal benefits of the Global Biodiversity Information Facility. *BioScience* **54**(6), 485–486. https://doi.org/10.1641/0006-3568(2004)054[0486:rasbot]2.0.co;2.

Efenakpo OD, Nneji LM, Jacob DE, Ringim AS, Eniang E, Petrozzi F and Luiselli LM (2026) Ecological research and knowledge gaps in the snake fauna of Nigeria: A systematic review with conservation implications. *Biodiversity*. https://doi.org/10.1080/14888386.2026.2646915.

Feng X, Park DS, Walker C, Peterson AT, Merow C and Papeş M (2019) A checklist for maximizing reproducibility of ecological niche models. *Nature Ecology & Evolution* **3**(10), 1382–1395.

Fick SE and Hijmans RJ (2017) WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302–4315. https://doi.org/10.1002/joc.5086.

Fourcade Y, Engler JO, Rödder D and Secondi J (2014) Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS One* **9**(5), e97122.

Graham CH, Ferrier S, Huettman F, Moritz C and Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution* **19**(9), 497–503. https://doi.org/10.1016/j.tree.2004.07.006.

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS and Porter JH (2013) Big data and the future of ecology. *Frontiers in Ecology and the Environment* **11**(3), 156–162. https://doi.org/10.1890/120103.

Hijmans RJ (2012) Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology* **93**(3), 679–688. https://doi.org/10.1890/11-0826.1.

Hijmans RJ, Phillips S, Leathwick J and Elith J (2017) dismo: Species Distribution Modeling. R package version 1.1-4. Available at https://doi.org/10.32614/cran.package.dismo.

Hortal J, De Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM and Ladle RJ (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46**(1), 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400.

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press

Hughes AC, Dorey JB, Bossert S, Qiao H and Orr MC (2024) Big data, big problems? How to circumvent problems in biodiversity mapping and ensure meaningful results. *Ecography* **2024**, e07115. https://doi.org/10.1111/ecog.07115.

IUCN (2025) The IUCN Red List of Threatened Species. Available at www.iucnredlist.org (accessed 21 December 2025).

Kramer-Schadt S, Niedballa J, Pilgrim JD, Schröder B, Lindenborn J, Reinfelder V and Wilting A (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions* **19**(11), 1366–1379.

Luiselli L (2006a) Site occupancy and density of sympatric Gaboon viper (*Bitis gabonica*) and nose-horned viper (*Bitis nasicornis*). *Journal of Tropical Ecology* **22**(5), 555–564. https://doi.org/10.1017/s0266467406003397.

Luiselli L (2006b) Food niche overlap between sympatric potential competitors increases with habitat alteration at different trophic levels in rain-forest reptiles (omnivorous tortoises and carnivorous vipers). *Journal of Tropical Ecology* **22**(6), 695–704. https://doi.org/10.1017/s0266467406003634.

Luiselli L and Akani GC (2003) Diet of sympatric Gaboon vipers (*Bitis gabonica*) and nose-horned vipers (*Bitis nasicornis*) in southern Nigeria. *African Journal of Herpetology* **52**(2), 101–106. https://doi.org/10.1080/21564574.2003.9635485.

Marcer A, Chapman AD, Wieczorek JR, Xavier Picó F, Uribe F, Waller J and Ariño AH (2022) Uncertainty matters: Ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. *Ecography* **2022**, e06025. https://doi.org/10.1111/ecog.06025.

Merow C, Smith MJ and Silander JA Jr (2013) A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography* **36**(10), 1058–1069.

Meyer C, Weigelt P and Kreft H (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**(8), 992–1006.

Moudrý V, Bazzichetto M, Remelgado R, Devillers R, Lenoir J, Mateo RG, Lembrechts JJ, Sillero N, Lecours V, Cord AF, Barták V, Balej P, Rocchini D, Torresani M, Arenas-Castro S, Man M, Prajzlerová D, Gdulová K, Prošek J, Marchetto E, Zarzo-Arias A, Gábor L, Leroy F, Martini M, Malavasi M, Cazzolla Gatti R, Wild J and Šímová P (2024) Optimising occurrence data in species distribution models: Sample size, positional uncertainty, and sampling bias matter. *Ecography* **2024**(12), e07294. https://doi.org/10.1111/ecog.07294.

Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M and Anderson RP (2014) ENM eval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution* **5**(11), 1198–1205.

Nelson G and Ellis S (2019) The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B* **374**(1763), 20170391. https://doi.org/10.1098/rstb.2017.0391.

Petrozzi F and Luiselli L (2026) A probabilistic habitat–suitability overlap framework (HMI) reveals spatial bias in MaxEnt models of West African forest butterflies. *Folia Oecologica* **53**(2), 108–122. https://doi.org/10.2478/foecol-2026-0010.

Sánchez-Mercado AY, Ferrer-Paris JR and Franklin J (2010) Mapping species distributions: Spatial inference and prediction. *Oryx* **44**(4), 615. https://doi.org/10.1017/S0030605310001201.

Shine R and Spawls S (2020) An ecological analysis of snakes captured by CJP Ionides in eastern Africa in the mid-1900s. *Scientific Reports* **10**(1), 5096.

Soberón J and Peterson T (2004) Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**(1444), 689–698. https://doi.org/10.1098/rstb.2003.1439.

Spawls S and Branch B (2018) *The Dangerous Snakes of Africa: Natural History, Species Directory, Venom, and Snakebite*, 2nd Edn. London: Bloomsbury Publishing.

Vaughan ER (2023) Evolution of forest Bitis (Macrocerastes) in Africa: Phylogeography, population genetics, and potential as a model for Batesian mimicry (Doctoral dissertation, The University of Texas at El Paso).

Yates KL, Bouchet PJ, Caley MJ, Mengersen K, Randin CF, Parnell S and Sequeira AM (2018) Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution* **33**(10), 790–802.

Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D and Antonelli A (2019) CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* **10**(5), 744–751.

https://doi.org/10.1017/S0266467426100686 Published online by Cambridge University Press