Boosting biodiversity monitoring using smartphone-driven, rapidly accumulating community-sourced data

Keisuke Atsumi[1]*, Yuusuke Nishida[1], Masayuki Ushio[2,3,4], Hirotaka Nishi[5], Takanori Genroku[1], Shogoro Fujiki[1,4]*

[1]Biome Inc, Kyoto, Japan; [2]Department of Ocean Science, Hong Kong University of Science and Technology, Kowloon, Hong Kong; [3]Hakubi Center, Kyoto University, Kyoto, Japan; [4]Center for Ecological Research, Kyoto University, Shiga, Japan; [5]Toyohashi Museum of Natural History, Aichi, Japan

Abstract Comprehensive biodiversity data is crucial for ecosystem protection. The *Biome* mobile app, launched in Japan, efficiently gathers species observations from the public using species identification algorithms and gamification elements. The app has amassed >6 million observations since 2019. Nonetheless, community-sourced data may exhibit spatial and taxonomic biases. Species distribution models (SDMs) estimate species distribution while accommodating such bias. Here, we investigated the quality of *Biome* data and its impact on SDM performance. Species identification accuracy exceeds 95% for birds, reptiles, mammals, and amphibians, but seed plants, molluscs, and fishes scored below 90%. Our SDMs for 132 terrestrial plants and animals across Japan revealed that incorporating *Biome* data into traditional survey data improved accuracy. For endangered species, traditional survey data required >2000 records for accurate models (Boyce index ≥ 0.9), while blending the two data sources reduced this to around 300. The uniform coverage of urban-natural gradients by *Biome* data, compared to traditional data biased towards natural areas, may explain this improvement. Combining multiple data sources better estimates species distributions, aiding in protected area designation and ecosystem service assessment. Establishing a platform for accumulating community-sourced distribution data will contribute to conserving and monitoring natural ecosystems.

eLife assessment

This **important** study presents findings of great practical value, offering fresh insights into natural species distributions across Japan. By combining multiple data sources (including those from non-academic sectors, aka citizen scientists), the manuscript also presents a **compelling** new tool that can be used to aid conservation agendas, detect species distribution changes, and testing of ecological theories.

***For correspondence:**
k.atsumi115@gmail.com (KA);
fujiki@biome.co.jp (SF)

Competing interest: See page 19

Funding: See page 19

Preprint posted
16 October 2023
Sent for Review
24 October 2023
Reviewed preprint posted
26 March 2024
Reviewed preprint revised
17 May 2024
Version of Record published
20 June 2024

Reviewing Editor: David A Donoso, National Polytechnic School, Ecuador

Introduction

Nature underpins human society, and the conservation of ecosystems and associated ecosystem services contributes to the sustainable development of human society, yet these services have been rapidly declining in recent years (***IPBES, 2019***; ***Loh et al., 2005***; ***Newbold et al., 2016***; ***Scholes and Biggs, 2005***). The Kunming-Montreal Global Biodiversity Framework (KM-GBF) by the United Nations envisions reversing the nature loss by 2030. As direct means for nature conservation, KM-GBF targeted making 30% of Earth's land and ocean area as protected areas by 2030 (i.e. 30 by 30). As

eLife digest The internet has allowed people to share their experiences through images, videos or audio recordings. This has led to the creation of online communities around a variety of topics, including biodiversity. In 2019, a smartphone app, called Biome, was created to fuel biodiversity engagement by making wildlife surveying an easy and fun activity via gamification and assisted species identification through image recognition and ecological analyses.

These types of observations are essential for understanding biological communities and species habitats, and they can indicate where and when species occur. Across Japan, Biome has gathered over 6.5 million observations of different species. For biologists, this type of data is extremely useful because it is continuous and enables advanced statistical estimations of species distributions. The fact that the approach is enjoyable to the user also means more people are willing to participate, lowering the barriers to collecting data about biodiversity loss.

However, questions remain regarding whether community-sourced data is robust enough for scientific purposes. To address this, Atsumi et al. investigated the quality of occurrence data collected in Biome. The researchers found that community identification of birds, reptiles, mammals and amphibians all exceeded 95% in accuracy. However, the accuracy fell for harder-to-judge seed plants, molluscs and fish species, ranging below 90%.

Atsumi et al. also compared how estimated distributions of each species changed when only scientific data was used, versus when it was combined with community data. To perform this analysis, the scientists recognized variations in observation efforts across different locations and individuals and adjusted for these biases in their estimations. They found that adding community-sourced data significantly improved the accuracy of species distribution estimations, including endangered species.

Atsumi et al. demonstrate that Biome data is useful when deciding which areas to designate as protected in terms of biodiversity. Additionally, these data can provide guidance for stakeholder-informed ecosystem service assessments. The element of rapid and reliable data collection can contribute to growing positive attitudes towards nature and biodiversity, The platform's community-driven nature also indicates an increase in biodiversity awareness and may link to crafting informative socio-environmental policy commitments.

an indirect but influential way, KM-GBF requires companies to “monitor, assess, and transparently disclose their risks, dependencies and impacts on biodiversity through their operations, supply and value chains and portfolios,” which is guided by the Taskforce on Nature-related Financial Disclosures (TNFD) (***TNFD, 2023***). To achieve these goals, it is imperative to assess the state of biodiversity with a sufficient spatiotemporal resolution to support conservation planning, adaptive management, and companies’ annual nature-related financial disclosures. The basis for such assessments lies in our knowledge of species distributions (***Gonzalez et al., 2023***; ***Newbold et al., 2016***). Traditionally, distribution data was acquired through on-site surveys by experts (people have expertise about biodiversity), but collecting distribution data with sufficient spatiotemporal resolution is challenging if we rely only on such limited human resources (***Miya et al., 2022***; ***Mori et al., 2023***; ***Pocock et al., 2018***).

Since the emergence of digital devices and the internet, people have been able to share their observations through various media, such as images and video/audio recordings. Such community-sourced data have significantly contributed to the accumulation of ecosystem information. These datasets have been instrumental in assessing the impacts of climate change and urbanisation on phenology (***Fuccillo Battle et al., 2022***; ***Klinger et al., 2023***), detecting distribution changes including invasive alien species (***Larson et al., 2020***; ***Roy et al., 2023***; ***Wallace and Bargeron, 2014***), exploring large-scale geographic variations in traits (***Atsumi and Koizumi, 2017***; ***Leighton et al., 2016***), and estimating species distributions (***Chandler et al., 2017***; ***Feldman et al., 2021***; ***Johnston et al., 2018***; ***Steen et al., 2019***). Moreover, the utilisation of machine learning to describe population trends based on community-sourced data (***Fink et al., 2023***) offers opportunities for conducting time-series analyses. These analyses can help us understand community assembly processes, unravel species interaction networks, and assess ecosystem stability (***Cornwell and Ackerly, 2009***; ***Tilman et al., 2006***; ***Ushio et al., 2018***), capitalising on the spatiotemporally dense sampling effort facilitated by community-sourced data (***Chandler et al., 2017***; ***Kobori et al., 2016***; ***Pocock et al., 2017***). Such analytical approaches enable us to make informed predictions about changes in species distribution, population dynamics, and ecosystem stability in the face of climate change (**Bury et al., 2021**; **Pennekamp et al., 2019**; **Urban et al., 2016**). In essence, community-sourced data, owing to its extensive sampling across time and space, has the potential to test existing ecological theories, expand our comprehension of ecosystems and the underlying processes, eventually allowing us to forecast ecological dynamics in the context of climate change.

When people photograph organisms using digital devices with GPS capabilities, the images often contain timestamps and location details. Such images, when accompanied by species identifications, serve as evidence for tracking phenology and species occurrences. This crowdsourcing approach has been particularly successful on web- or mobile-based platforms such as eBird and iNaturalist (**Chandler et al., 2017**; **Wood et al., 2011**). Individuals submit records to these platforms for various reasons, including a desire to contribute to science and engage with cutting-edge technologies (**Herodotou et al., 2024**; **Kaplan Mintz et al., 2023**). By making the process more enjoyable (i.e. gamification), we can potentially gather even more biological data from the public (**Bowser et al., 2013**; **Ponti et al., 2015**). Yet, the collection process of Community-sourced data is usually not well-designed (e.g. spatially biased 'presence-only' data) (**Feldman et al., 2021**; **Steen et al., 2019**) and its interpretation is challenging without proper statistical modelling. Thus, although much effort has

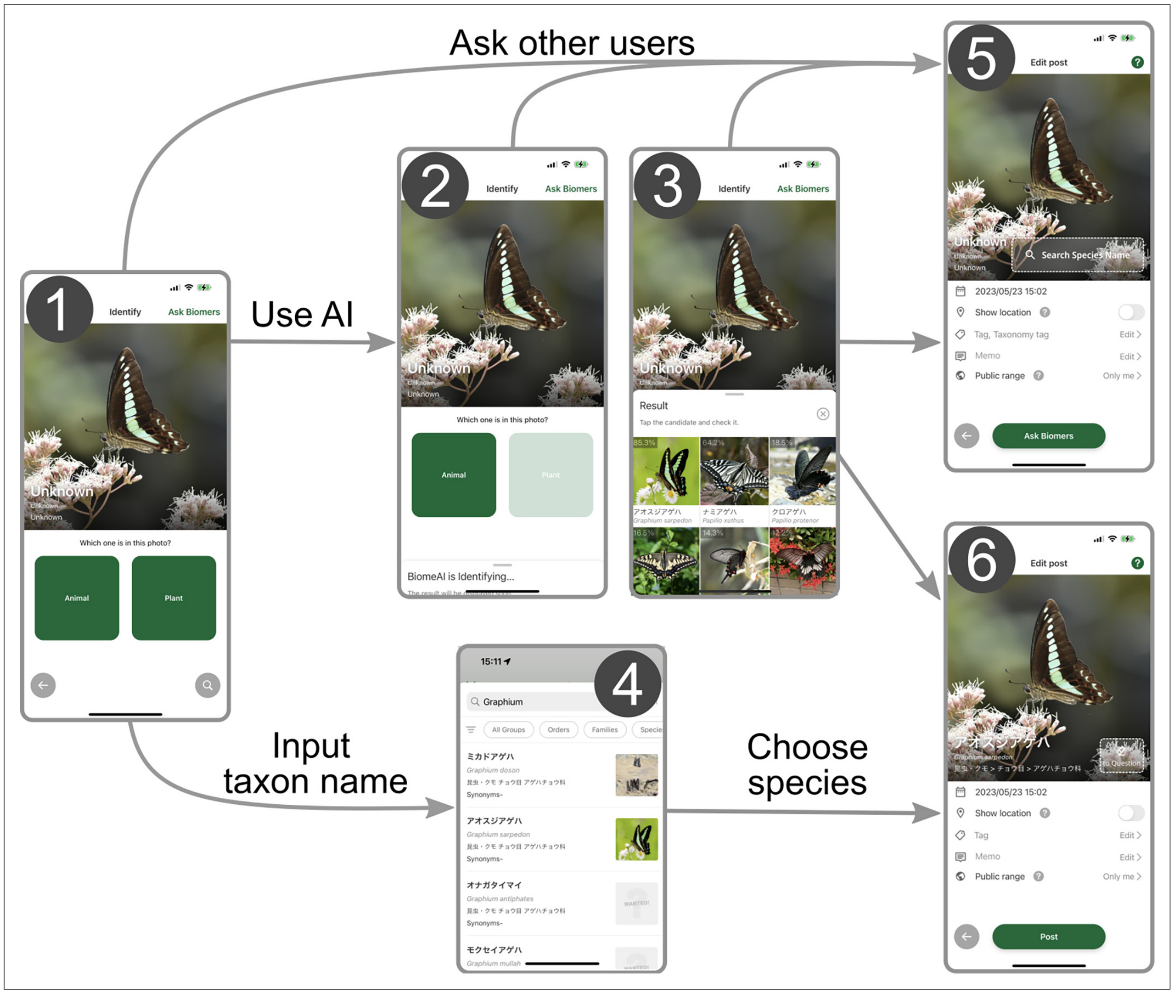


Figure 1. Workflow of submitting records to *Biome*. (1) Users can upload images that were taken by the smartphone camera or import existing images from the storage, including those imported from external devices. (2) Users select whether the image is about animals or plants to activate the species identification artificial intelligence (AI). (3) The AI analyses the image and its metadata to generate a candidate species list. (4) Alternatively, users can input the taxon name manually and obtain a list of candidate species. To submit the occurrence record, users can either (5) seek identification assistance from other users through the 'ask Biomers' feature, or (6) identify the species from the list. To the records, users can add memos and tags indicating phenology, life stage, sex, and whether the individual is wild or captive.

been invested in developing effective monitoring and modelling methods for biodiversity assessment, current approaches can be further improved by incorporating (i) more enjoyable community-based survey platforms using mobile applications and (ii) employing an advanced statistical modelling framework in estimating species distribution.

To fuel communities' engagement in biodiversity surveys and environmental education, we launched the mobile application *Biome* in 2019 in Japan (***Fujiki and Tatsuno, 2021***). For supporting species identification, *Biome* implements artificial intelligence (AI) algorithms that generate lists of potential species and enable users to seek help/suggestions from others for species identification (***Figure 1***) as in other applications such as iNaturalist and eBird. The unique feature of *Biome* is gamification which offers enjoyable experiences and facilitates communication among users (***Fujiki and Tatsuno, 2021***; ***Koide et al., 2023***). For example, users can earn 'points' by contributing in various ways such as submitting records and suggesting species identifications to others, and their levels are determined based on the total points earned. The inclusion of networking and gamification elements can attract a wider user base, including those who may not typically engage in community science (***Bowser et al., 2013***; ***Groom et al., 2021***). Consequently, *Biome* has accumulated data rapidly. Since its launch, 6 million records have been collected through the app (by 17 October 2023). This is more than four times greater than the number of records accumulated by the Global Biodiversity Information Facility (GBIF) from any data sources including iNaturalist and eBird during the same period in Japan (ca. 1.3 million). The data gathered through the app has been used for conservation planning and facilitating companies' financial disclosures by supplying and analysing species occurrence records.

Species distribution models (SDMs) are effective statistical tools for assessing biodiversity at specific sites while accounting for biases in survey efforts. SDMs use species occurrence records and environmental conditions to estimate the potential geographic ranges and suitable habitats for species (***Booth et al., 2014***; ***Box, 1981***; ***Elith et al., 2011***; ***Hutchinson, 1957***; ***Phillips et al., 2006***). These models play a crucial role in conservation and restoration planning by helping predict how changes in land use and climate impact species distributions (***Kindt, 2023***; ***Porfirio et al., 2014***; ***Urban et al., 2016***). While species presence/absence data—which needs extensive surveys by experts—is limited, presence-only data—which can be obtained from communities' observations—is much more available. MaxEnt (***Phillips et al., 2006***; ***Phillips and Dudík, 2008***) is one of the most popular SDM methods due to its computational efficiency and estimation accuracy (***Valavi et al., 2022***). It can estimate species distribution from presence-only data by maximising the entropy of the probability distribution while satisfying constraints based on the available information (***Elith et al., 2011***; ***Phillips and Dudík, 2008***). Since MaxEnt only requires occurrence records, it is well-suited for empowering community-based observations to predict species distributions. Also, while community-sourced data often suffer from spatially biased sampling efforts (i.e. sampling tends to concentrate in densely populated or touristic areas; ***Kendal et al., 2020***; ***Reddy and Dávalos, 2003***), SDMs such as MaxEnt can account for such spatial biases by considering the spatial distribution of sampling efforts when selecting pseudo-absence (background) locations (***Milanesi et al., 2020***; ***Phillips et al., 2009***). When sampling efforts are adequately controlled, adding community-sourced data improves the accuracy of SDMs (***Johnston et al., 2018***; ***Robinson et al., 2020***; ***Steen et al., 2019***). This implies that SDMs may be substantially improved by utilising rapidly accumulating *Biome*'s species occurrence records if we adequately control the sampling efforts.

Here, we show the quality of community-based data gathered through the smartphone app *Biome* and how the data improves the prediction accuracy of species distribution. First, we assess the quality of occurrence records by investigating the fractions of non-wild and misidentified records. Second, we built SDMs based on two types of data: (i) traditional survey data (e.g. forest inventory census, museum specimens, and records extracted from published researches) only and (ii) a mixture of traditional survey and *Biome* data. We then compare the performance of the two SDMs. We modelled the distributions of 132 terrestrial animals and seed plants in the Japanese archipelago which covers subtropical to boreal areas. We finally discuss how our SDMs relying on community-sourced data may contribute to meeting the goals of GBF.

Results

The amount and quality of *Biome* data

By 7 July 2023, *Biome* had accumulated 5,275,457 occurrence records of 40,957 species across the Japanese archipelago (***Figure 2A***). The amount of occurrence records submitted to *Biome* has increased across the years (***Figure 2B***). On average, in 2022, users submitted 5407 records per day. The distribution of data along environmental gradients somewhat differs between *Biome* and Traditional survey data. To elucidate this distinction, we employed principal component (PC) analysis to summarise all environmental variables. The two datasets demonstrated divergent distribution patterns along PC1 (***Figure 2C***). This component, accounting for 6.1% of the total variation, is primarily influenced by land use, topography, and climate (***Supplementary file 1***). Among the environmental variables, a notable contrast between the datasets was observed in relation to the natural-urban gradient. The *Biome* data exhibited a relatively uniform distribution encompassing the entire gradient, while Traditional survey data was substantially biased towards natural areas (***Figure 2C***). The majority of records are attributed to insects (31.2%) and seed plants (41.8%), which are relatively accessible and can be easily photographed using smartphones (***Figure 2D***).

Out of all the records submitted to *Biome*, a total of 2,373,303 records (45.0%) successfully passed through the automatic filtering process. This dataset, referred to as the *Biome* data, is utilised for subsequent investigations. The quality of *Biome* data varied across taxa and the rarities of species (***Table 1***). The fraction of the records of wild individuals exceeded 97% in insects and birds, while it was lower than 90% in molluscs, seed plants, mammals and fishes. Among the records of wild individuals, at the species level, identification accuracy was higher than 95% in birds, reptiles, mammals, and amphibians but less than 90% in insects, fishes, and seed plants. At the genus level, identification accuracy was higher than 90% in all taxa except for insects. In the case of fishes and seed plants, identifications became 5–6% more accurate at the genus level compared to the species level. The family was correctly identified in more than 94% of records in all taxa examined. Common species had higher identification accuracy than rare species (average value, 95% vs. 87%). This tendency was prominent in insects and seed plants, but less in the other taxa. These results suggest that identifying rare species in taxonomically diverse taxa (i.e. seed plants and insects) is a challenging task.

The performance of SDMs

SDMs using *Biome* + Traditional data, including *Biome* data at 50%, were more accurate than those modelled only using Traditional survey data when the two datasets have the same amount of occurrence records (***Figure 3***). Our analysis revealed that although the intercept of the Boyce index (BI, model accuracy metric that ranges between –1 and 1) did not differ between the two datasets (generalised linear mixed model, see 'Methods': $\beta = 0.02 \pm 0.03$, $t = 0.60$, p=0.55), *Biome* + Traditional data consistently led to a more rapid increase in SDM accuracy as the amount of data increased, compared to models solely relying on Traditional survey data ($\beta = 0.02 \pm 0.01$, $t = 3.72$, p<0.001).

When compared to SDMs using Traditional survey data, those using Biome + Traditional data achieved a high level of accuracy with a much smaller amount of data. For instance, BI, which ranges from –1 to 1, exceeds 0.9 with 294 ± 471 records (mean ± SD across all species) in the Biome + Traditional data, whereas the Traditional survey data requires 2129 ± 4157 records to achieve the same accuracy. This was also true in endangered species (included in Japanese national or prefectural red lists); although 2336 ± 3718 Traditional survey records were required to exceed 0.9 of BI, only 338 ± 571 were required for Biome + Traditional data.

Because we controlled the proportion of *Biome* data within the Biome + Traditional data as 50%, the amount of records of the Biome + Traditional data is often limited. In cases where a species had less *Biome* data compared to Traditional survey data, the total amount of records of Biome + Traditional data ends up being smaller than that of Traditional survey data alone. Therefore, the two datasets did not differ in the best model performances in each species (BIs of Biome + Traditional data: 0.81 ± 0.20; Traditional survey data: 0.83 ±0.20).

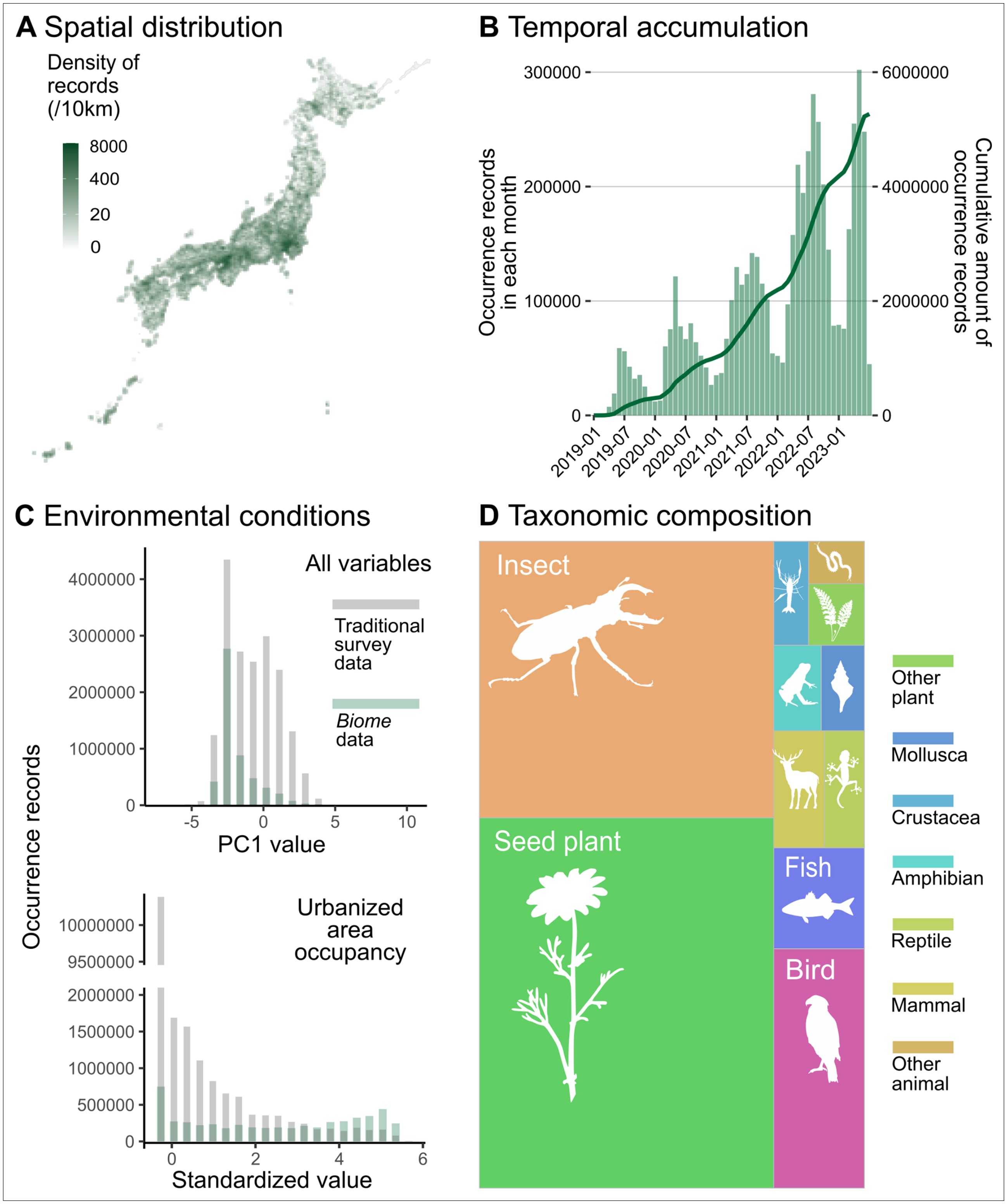


Figure 2. Description of data accumulated by *Biome*. Data distributions are shown based on all records submitted to *Biome* by 7 July 2023 (N = 5,275,457). (**A**) Spatial distribution of records across Japan. (**B**) Accumulation of records through time. The barplot represents the number of records each month and the line shows the cumulative amount of records. (**C**) Distributions of records along with PC1 of all environmental variables and standardised area occupancy of urban-type land uses. Grey and green represent distributions of Traditional and *Biome* data, respectively. (**D**) Taxonomic composition of records is shown as the area sizes. 'Other plant' consists of non-seed terrestrial plants; 'insects' include Arachnids and Insects; 'arthropods' cover any Arthropod not included in insects; 'other animals' covers all invertebrates not included in the taxa above.

Table 1. Data quality of *Biome*.

The fraction of records documenting wild individuals, and identification accuracy at species, genus, and family levels among the records documenting wild individuals are shown. Species were identified only for records documenting wild individuals.

Species group	Species rarity	*N*	Wild/total (%)	Species correct/wild (%)	Genus correct/ wild (%)	Family correct/ wild (%)
Total	Total	1420	81.6	91	93.6	96.9
Seed plant	Total	290	86.2	89.6	94.4	97.2
Mollusca	Total	140	87.9	90.2	91.1	96.7
Insect	Total	290	100	83.4	86.9	94.1
Fish	Total	140	73.6	87.4	93.2	96.1
Amphibian	Total	140	93.6	96.2	96.2	98.5
Reptile	Total	140	91.4	97.7	100	100
Bird	Total	140	98.6	98.6	99.3	99.3
Mammal	Total	140	80.7	95.6	95.6	96.5
Total	Rare	710	88.7	87	91	95.6
Total	Common	710	91	95	96.3	98.3
Seed plant	Rare	145	80.7	82.9	91.5	94.9
Seed plant	Common	145	91.7	95.5	97	99.2
Mollusca	Rare	70	82.9	86.2	87.9	96.6
Mollusca	Common	70	92.9	93.8	93.8	96.9
Insect	Rare	145	100	75.2	80	91.7
Insect	Common	145	100	91.7	93.8	96.6
Fish	Rare	70	74.3	88.5	94.2	94.2
Fish	Common	70	72.9	86.3	92.2	98
Amphibian	Rare	70	95.7	95.5	95.5	98.5
Amphibian	Common	70	91.4	96.9	96.9	98.4
Reptile	Rare	70	94.3	95.5	100	100
Reptile	Common	70	88.6	100	100	100
Bird	Rare	70	97.1	98.5	100	100
Bird	Common	70	100	98.6	98.6	98.6
Mammal	Rare	70	81.4	91.2	91.2	93
Mammal	Common	70	80	100	100	100

Discussion

Biome: The amount and quality of submitted data

Since its launch in 2019, the app *Biome* has accumulated species occurrence data rapidly (***Figure 2***). Despite our concerted efforts to engage non-expert users through gamification features, it is important to acknowledge that an excessive influx of non-expert users could potentially compromise the quality of the collected data. This could manifest in misidentifications or incomplete documentation, such as failing to appropriately label non-wild individuals. We thus have developed algorithms to exclude such suspicious records based on the features of records and users' behaviour on the app. The implementation of automatic data filtering techniques is expected to enhance the quality of the data, although further refinement is necessary. Notably, for insects and birds, which encompass numerous

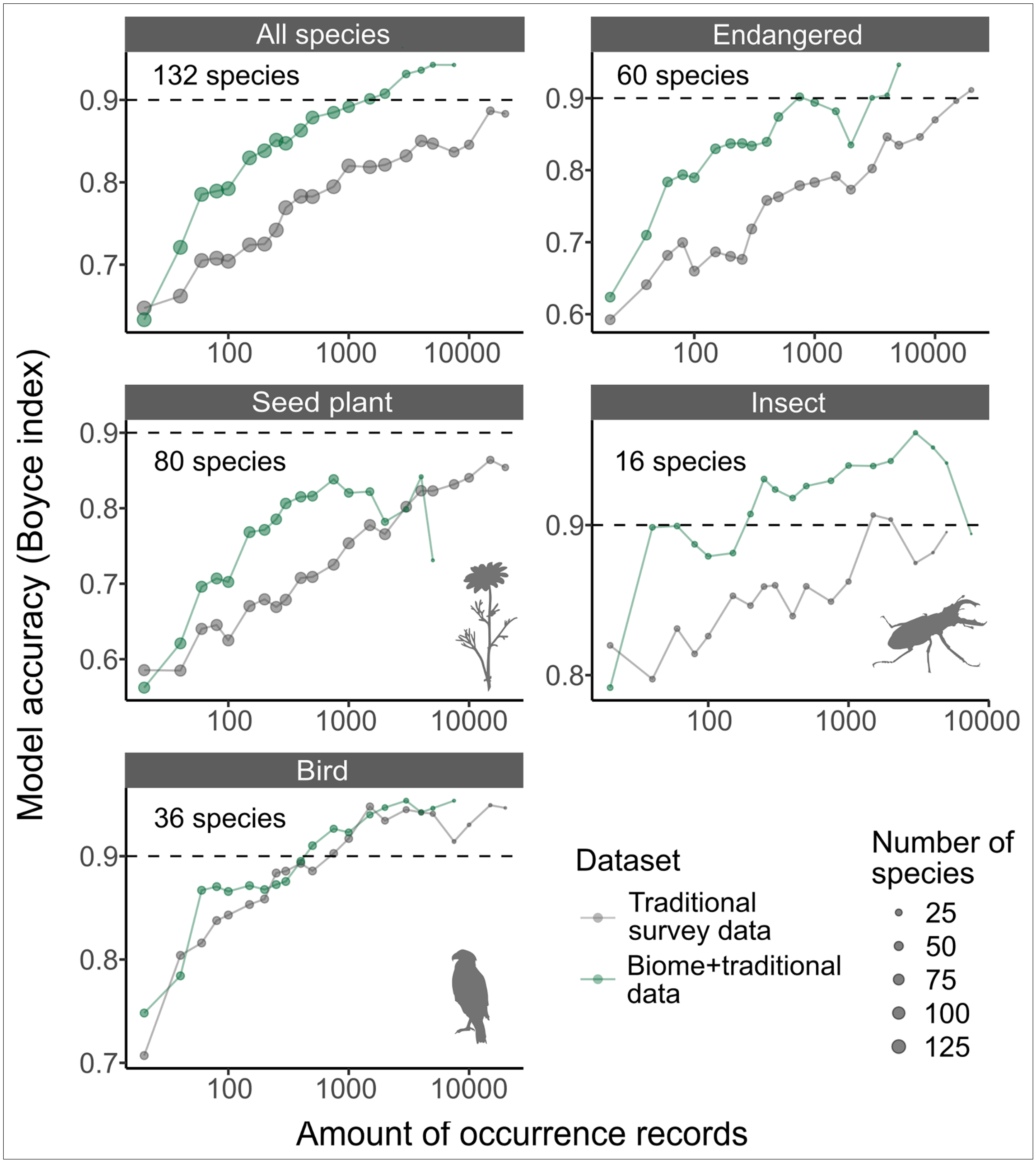


Figure 3. The accuracy of species distribution models. Accuracy of species distribution models (SDMs) using Traditional survey data (grey dots and lines) and Biome + Traditional data (i.e. 50% of *Biome* data: green). Each SDM was performed with a specific dataset, species, and the amount of records. For each species and amount of records, we computed the average model accuracy (Boyce index) from three replicated runs. Subsequently, we calculated the median model accuracy across species for each amount of records. These medians were then illustrated for each taxon in the strip of each respective panel. The 'Endangered' category includes species that are listed as endangered on Japan's national or prefectural red lists.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Accuracy of species distribution models (SDMs) using Traditional survey data (grey dots and lines) and Biome + Traditional data (i.e. 50% of *Biome* data: green), evaluated against test data only consisting of Traditional survey data.

species that can be kept in captivity, the majority of records that underwent filtering procedures were restricted to observations of wild individuals. Yet, the fraction of non-wild individuals is high in several taxa such as fishes and seed plants. In response, we have updated the posting flow in the app to prompt users to differentiate between non-wild and wild individuals. Further analysis is warranted to evaluate the impact of this update on data quality.

Once we could exclude non-wild individuals, species identification accuracy exceeded 95% in taxa with moderate species diversity (amphibians, reptiles, birds, and mammals). In seed plants, *Biome*'s species identification accuracy was 90%, which is higher than the accuracy of auto-suggest identification by commonly used apps for plants (69%, PlantNet, PlantSnap, LeafSnap, iNaturalist, and Google Lens; ***Hart et al., 2023***). During the invasive plants survey in the United States, the reports by non-professional volunteers were 72% correct (***Crall et al., 2011***). The higher accuracy of species identification in *Biome* data can be attributed to two key factors. Firstly, the vigilant oversight of the user community through the 'suggest identification' feature plays a crucial role. *Biome* encourages users to participate in suggesting identifications by offering 'points' as rewards for their contributions. Secondly, the species identification AI algorithm leverages past occurrence data from nearby areas, resulting in increasingly accurate automatic identifications as the data accumulates. Given these, as a community science app, the data quality of *Biome* is decent. Yet, rare species generally showed lower identification accuracy, which would require identification by experts and further improvement of species identification AI algorithm.

Species distribution modelling

The inclusion of *Biome* data resulted in improved accuracy of SDMs (***Figure 3***). The most accurate model predictions were obtained when the training data consisted of 50–70% *Biome* data (Appendix 1), highlighting the necessity of incorporating both traditional surveys and citizen observations for a comprehensive understanding of species distributions (***Miller et al., 2019***; ***Pacifici et al., 2017***; ***Robinson et al., 2020***).

The improvement can be attributed to introducing data with different biases compared to the Traditional survey data. Indeed, when controlling for the number of occurrence records, the model performance was higher in the Biome + Traditional data compared to the Traditional survey data. The variation in performance can be attributed to the distribution of data in relation to environmental conditions. Traditional survey data exhibits a strong bias towards natural areas, whereas *Biome* data is well balanced across the natural-urban habitat gradients (***Figure 2C***). Therefore, incorporating *Biome* data could significantly enhance modelling accuracy in urban and suburban landscapes, which are typically underrepresented in traditional survey data. As pseudo-absences are selected based on search effort, our models utilise numerous pseudo-absences from these areas. Consequently, this might lead to better estimation of species absence in such areas, not just presence, resulting in an overall increase in model accuracy across a wider range of species. A balanced distribution, along with the natural-urban gradient, is noteworthy because community science data is typically biased towards human population centres (***Kendal et al., 2020***; ***Reddy and Dávalos, 2003***). This could be influenced by the distribution of users' residencies, although we do not have specific information about the users' locations. The app has collaborated with numerous local governments across Japan, including 9 prefectures and 29 local municipalities such as cities and towns. Through these collaborations, the user base may be widely dispersed, enriching the geographical coverage of *Biome* data.

The *Biome* data also can improve SDM accuracy by simply increasing the overall amount of data. Essentially, SDM accuracy is enhanced with an increased amount of data (***Figure 3***; ***Erickson and Smith, 2023***; ***Stockwell and Peterson, 2002***). In our analysis, we maintained a fixed proportion of 50% for *Biome* data within the Biome + Traditional dataset, which in turn restricted the amount of available Biome + Traditional data. However, our preliminary analysis (Appendix 1) demonstrates that the enhancement of SDM accuracy occurs across a range of proportion variations for *Biome* data blending. This implies that the proportion of *Biome* data does not necessarily need to be controlled. Therefore, in practical application scenarios, the incorporation of *Biome* data predominantly serves to augment the overall volume of training data.

The impact of community-sourced data on SDMs has primarily been investigated using birds, with a limited focus on plants (***Feldman et al., 2021***). In our investigation, we observed that incorporating *Biome* data improved SDM accuracy for seed plants and insects, while the impact on birds remained unclear (***Figure 3***). This ambiguity is likely because community-sourced data from platforms such as eBird are already incorporated in Traditional data through GBIF. In comparison to other taxonomic groups, our results indicate that seed plants exhibited lower model accuracy when evaluated against both Biome + Traditional survey data (***Figure 3***) and Traditional survey data alone (***Figure 3—figure supplement 1***). The variation in model accuracy among taxonomic groups may be attributed to data quality issues in both *Biome* and Traditional survey data. For instance, in *Biome* data, while the fractions of wild individuals were high in birds and insects, it was lower for seed plants (***Table 1***). Compared with other taxa, distinguishing between wild and non-wild individuals can be particularly difficult in plants when they are planted outside. In addition, identifying plant species may be challenging in certain taxa, primarily due to the absence of key identification traits on leaves and stems. This becomes especially problematic when flowers are not present. These difficulties could potentially impact the quality of Traditional data as well. Although few studies have simultaneously assessed the quality of community-sourced data and its impact on SDMs across different taxa, it is important to recognise that data quality can vary among taxa.

Importantly, SDMs for endangered species, which often suffer from data deficit (***Erickson and Smith, 2023***; ***Wisz et al., 2008***), became accurate in a much fewer amount of records by blending *Biome* data (***Figure 3***). Specifically, a threshold of >0.9 BI could be reached with only around 300 records when using *Biome* data, whereas over six times of data is required when using Traditional survey data only. This finding highlights the importance of community-sourced data not only for monitoring the dynamics of endangered species (***Chandler et al., 2017***; ***Zapponi et al., 2017***) but also for modelling purposes. Considering the rapid accumulation of *Biome* data, *Biome* data would make a significant contribution to the more effective distribution modelling of endangered species.

Limitations of this study

In assessing data quality, reidentification was impossible for records that did not photograph key traits for species identification. To address this limitation, further app improvements can include allowing users to submit multiple images. Encouraging users to document various body parts of organisms through multiple images would make capturing key identification traits much easier. This will make reidentification easier and possibly improve automatic species identification accuracy.

Given the absence of a comprehensive, environmentally unbiased occurrence dataset spanning a wide range of taxa, we assessed SDM accuracy not relying on an independent test dataset. In this evaluation, the test data was meticulously crafted to include 25% *Biome* data, serving as an intermediary proportion between Biome + Traditional (50%) and Traditional survey data (0%). By leveraging the distinct distribution patterns of *Biome* and Traditional survey data along environmental variables (***Figure 2C***), the test data would better encapsulate the actual species distribution compared to datasets composed solely of either *Biome* or Traditional survey data. It is noteworthy that, even when the test data exclusively consisted of Traditional survey data (i.e. unfavourable conditions for Biome + Traditional data SDMs), the accuracy of SDMs derived from Biome + Traditional and Traditional survey data did not differ (***Figure 3—figure supplement 1***). This result further supports our conclusions that *Biome* provides valuable data for SDM in terms of the amount and quality, and that blending *Biome* data improves SDM accuracy.

We evaluated SDMs based on spatial transferability using the central Japan region, which encompasses a range of environmental conditions. However, the evaluation results may not necessarily indicate transferability across the entire Japanese archipelago. Instead, in the near future, we anticipate that we can evaluate SDM accuracy using temporal transferability. The rapid accumulation of *Biome* data will allow us to evaluate the temporal transferability using the occurrence dataset from different time periods, and thus enable assessing their performance in much wider regions. In addition, limited data availability for certain taxa hindered the assessment in those taxa (e.g. molluscs, amphibians, reptiles, and mammals), but *Biome* would be a platform to overcome the data limitation for many taxa.

Finally, our SDMs do not directly indicate the species' presence probability. The output from presence-only SDMs usually deviates from the probability of presence when species prevalence (i.e. the proportion of area where the species occupied, requiring presence/absence data throughout the area) is unavailable (***Elith et al., 2011***; ***Ward et al., 2009***). Due to the unavailability of absence data, SDM outputs in this work are indirect measures of species presence and thus are not directly comparable across different species. Nonetheless, they are comparable within a species, providing useful information for understanding species distributions.

Future directions

By blending data from traditional surveys and communities, we improved the accuracy of species distribution estimates. This enhanced estimation lays the groundwork for more precise subsequent analyses. For instance, estimated distributions will be useful in selecting new protected areas or areas with Other Effective area-based Conservation Measures (OECMs): allowing a wider range of land use as long as biodiversity and ecosystem services are sustained/improved. Using estimated distributions of each species, hotspots of species or evolutionary diverse taxa can be inferred. Such sites will be good candidates for protected areas (***Jones et al., 2016***) or OECMs (***Shiono et al., 2021***). Further, estimated distributions can be used as input for spatial conservation prioritisation tools (e.g. Marxanl ***Ball et al., 2009***).

In our experience, stakeholders—including corporate social responsibility managers and conservation practitioners—often seek the list of species potentially inhabiting their locations. Due to the uncertainty of SDMs and their thresholding into presence/absence, on-site surveys remain essential for assessing biodiversity status. Yet, SDMs can make such surveys cost-effective by screening important locations for on-site assessment (e.g. Locate phase in TNFD framework) and narrowing down the target species for surveying. Improved estimation through SDMs can mitigate the risks associated with their use in society and enable more informed decision-making for conservation efforts.

The rapid accumulation of data from diverse locations holds the potential to unveil valuable ecological patterns. The accumulated data enables early detection capabilities for range expansions of invasive species (Sakai et al., in preparation). For instance, *Biome* data has hinted at potential range expansions in several insect species, including butterflies, dragonflies, and stink bugs, as well as changes in wintering areas for birds (***Biome Inc, 2023***). Given the diverse taxonomic coverage of *Biome* data (***Figure 2D***), detecting phenological changes across various taxa may be possible. This, in turn, is useful in uncovering phenological mismatches exacerbated by climate change, which can significantly change the dynamics of interacting species (***Renner and Zohner, 2018***; ***Visser and Gienapp, 2019***). Moreover, *Biome* data is well-suited for assessing the effects of urbanisation on ecosystems since it comprehensively spans both urban and natural habitats (***Figure 2C***). The benefit of rapidly accumulating data, combined with recent advancements in machine learning methods, opens up opportunities for conducting time-series analyses. Community science data has rarely been used for time-series population analysis due to its notable spatiotemporal bias in sampling efforts (***Feldman et al., 2021***; ***Zhang et al., 2021***). However, the two-step machine learning approach, as demonstrated by Fink and colleagues in estimating bird population trends using eBird data (***Fink et al., 2023***), sets a precedent. In the future, *Biome* data may facilitate the inference of population dynamics for multiple taxa. This will enable various time-series analyses to unveil ecosystem stability and interaction strength, which holds potential for forecasting ecosystem dynamics (***Laubmeier et al., 2020***; ***Pennekamp et al., 2019***; ***Ushio et al., 2018***).

For financial disclosures, companies will assess how their activities rely on ecosystem services and their opportunities for protecting/recovering nature (***TNFD, 2023***). By incorporating taxon-specific ecosystem services, multifaceted ecosystem services can be preliminarily screened (***Kass et al., 2024***). For example, based on estimated distributions of bumblebees or insectivorous animals, the functioning of pollination services or pest regulation services might be inferred. Using counts of ‘likes’ or records from *Biome* data, the charismatic species can be determined. By identifying places with a high estimated richness of charismatic species, potential areas for ecotourism can be screened. Because SDMs allow us to simulate the impacts of changes in landuse and climate (***Porfirio et al., 2014***; ***Urban et al., 2016***), we will be able to forecast how those changes may influence local biodiversity and/or ecosystem functioning. Hence, estimated distributions provide the basis of nature-related financial disclosures.

Our platform facilitates collaboration among diverse stakeholders, including local communities, landowners, and employees from both private companies and government agencies. Engaging a broader spectrum of stakeholders is crucial for effective biodiversity assessment, nature management planning, and nature-related financial disclosures: this inclusivity allows for the incorporation of traditional knowledge into planning processes, mitigates conflicts among stakeholders, and ultimately supports more seamless and informed decision-making (**Chan et al., 2021**; **Keough and Blahna, 2006**; **Linsley et al., 2023**; **Roy et al., 2023**; **TNFD, 2023**). Supporting natural experiences for a wide range of people is also expected to contribute to changing people's minds towards nature. By experiencing nature, people become familiar with it and subsequently make pro-nature decisions (**Soga and Gaston, 2023**). We believe that community science can significantly contribute to KM-GBF and create a sustainable society by fostering nature-positive awareness in society and providing data tools that enable effective action.

Methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Software, algorithm	R 4.1.3; MaxEnt (using ENMeval 2.0 package on R)	R 4.1.3 (**R Core Team, 2021**); MaxEnt (**Phillips et al., 2006**; **Phillips and Dudík, 2008**); ENMeval 2.0 package (**Kass et al., 2021**)		
Other	Species occurrence data	Biome app, GBIF and others (see 'Methods')	For DOIs of GBIF data, see **Supplementary file 2**	For details, see section 'Occurrence data'

Occurrence record accumulation through the mobile app *Biome*

In April 2019, a free smartphone app called *Biome* was launched for the Japanese markets. The app has been downloaded 839,844 times by 13 September 2023. The app allows users to collect data on the distribution of plants and animals using their mobile devices. Users can post photographs of the plants and animals they find, and the app automatically records the location and timestamp from EXIF data. If the EXIF data is unavailable, users can manually input the locality and timestamp.

To support species identification, the app provides users with two options. First, the app provides a list of candidate species based on the image and metadata (e.g. location and timestamp). *Biome* employs a synergistic approach that integrates image recognition technology and geospatial data to facilitate species identification. The image recognition algorithm, constructed upon convolutional neural networks, classifies species at higher taxonomic levels. Subsequently, these candidates are refined based on their frequency of recent occurrences in the geographical area. Consequently, as the correctly identified records accumulate for a given area, species identification AI will improve the accuracy. Second, users can seek help from other users. If a user selects the 'ask Biomers' button, their occurrence record is added to a waiting list that appears on the home screen. Other users can suggest possible identifications for the records, as in other records of which species was already identified.

Users can view and comment on other users' records. However, for conservation purposes, *Biome* automatically conceals the geolocations of endangered species that are listed on the Japanese national or prefectural red lists. This feature sets it apart from iNaturalist, where users must manually choose to hide the location of endangered species (**Koide et al., 2023**). The social networking function provides opportunities for communication among users, including non-experts (**Fujiki and Tatsuno, 2021**). Users earn 'points' through their contributions, including record submissions and identification suggestions to other users, and progress to higher levels based on their total points. The points awarded depend on the rarity, conservation status, and societal impact of the species submitted, meaning that users earn more points when submitting records of rare, endangered, or invasive species. The app occasionally offers 'Quests' events that provide users with an opportunity to earn additional points by submitting records from specific locations or of particular species, crucial for monitoring phenology. Through the variety of gamification features, we stimulate people to participate in biological surveys as a fun activity.

We obtained occurrence records submitted to *Biome* by 7 July 2023. The raw data collected through *Biome* contains invalid presence records which we defined in the present study as unclear images, documenting non-wild individuals and misidentifications, and images including some privacy issues. To improve data quality, we excluded records deemed to be invalid mainly based on location metadata and users' reactions to the record is as detailed below. This filtered *Biome* data is used in the subsequent investigations.

Filtering suspicious occurrence record in *Biome* data

Occurrence records of non-wild individuals were eliminated as much as possible by using the information provided by users and location of records. *Biome* users sometimes report inappropriate records (e.g. unclear images and images from websites or books), and we excluded all of those reported records. All private records were excluded because they can harbour inappropriate and misidentified records not being screened by other users. We also excluded occurrence records that users had marked as non-wild individuals: users have an option to label their records as photographing bred or cultivated individuals, or specimens. Records from cultural centres (i.e. zoos, botanical gardens, museums, and aquariums) and large pet stores were removed as well. During the data correction process, we prioritise the suggestions provided by *certified users* (see below for the definition), regardless of the decisions made by the users who originally created the record. Furthermore, we excluded records that have not been posted by *certified users* or have not received identification suggestions from *certified users*.

Certified users are defined as users who achieved the higher accuracy of species identification (<15% of public occurrence records were suggested as misidentification by other users), submitted few inappropriate records (<0.5% of public records), and have created >20 public records. We also defined *specialist users*, a subset of *certified users* identified in each taxa (see ***Figure 2*** for the classification), who made a total of >30 records or identification suggestions with high identification accuracy (the fraction of suggested records is less than the average of *certified users* in the taxa). *Specialist users* are used in determining pseudo-absence for SDMs.

Assessing the accuracy of records

We investigated the proportion of occurrence records within the *Biome* data that were suitable for SDMs. Since SDMs are influenced by invalid presence records, we assessed the quality of *Biome* data based on a total of 1420 records from rare and common species of seed plants, molluscs, insects (including Arachnid and Insecta), fishes, mammals, birds, reptiles, and amphibians (***Figure 4***). We defined rare species as those with less than or equal to 10 occurrences in *Biome* data, and common species as those with the highest 15% of records in each taxonomic category. In each of the seed plant and insect species which account for the majority of *Biome* data (***Figure 2D***), we randomly selected 145 records of each rare and common species. For the other taxonomic categories, we chose each of the 70 records from rare and common species.

Records were first screened whether they targeted organisms (images with no organisms were discarded) and contained wild individuals. To assess the accuracy of species identification, species in the records that documented wild individuals were manually reidentified by experts with taxonomic knowledge (***Figure 4***). These experts have professional backgrounds, serving as a technician at a prefectural research institute (fish), highly experienced field survey conductors (plants and insects, respectively), a post-doctoral researcher (amphibians and reptiles, and mammals, respectively), and a museum curator (molluscs) specialising in the focal taxa. Then, by comparing species identifications by the experts and on *Biome* data, the results were classified into two categories: (1) correct based on the image and locality—based on the image, identification was probably correct, and the image locality matches with habitat/range of the species; (2) misidentification—records were reidentified by experts if possible. We also examined whether the identification was correct at genus and family levels.

Species distribution models

Occurrence data

To evaluate the impact of *Biome* data on SDM prediction accuracy, we compiled two datasets: 'Traditional survey data' and 'Biome + Traditional data'. The Traditional survey data comprised records collected through conventional survey techniques (e.g. riverine census, forest inventory census, and museum specimens) primarily sourced from the National Census on River and Dam Environments (NCRE) and GBIF. In contrast, the Biome + Traditional data encompassed records submitted to *Biome* that passed filtering methods, in addition to the Traditional survey data. To control the relative proportion of *Biome* data, we constrained the fraction of *Biome* data within the Biome + Traditional data to 50% for each species. Our preliminary results showed that blending 50–70% of *Biome* data in training data improved prediction accuracy (Appendix 1). For traditional survey data, we downloaded occurrence records of relevant taxa from GBIF between 20 April 2023. To prevent significant differences

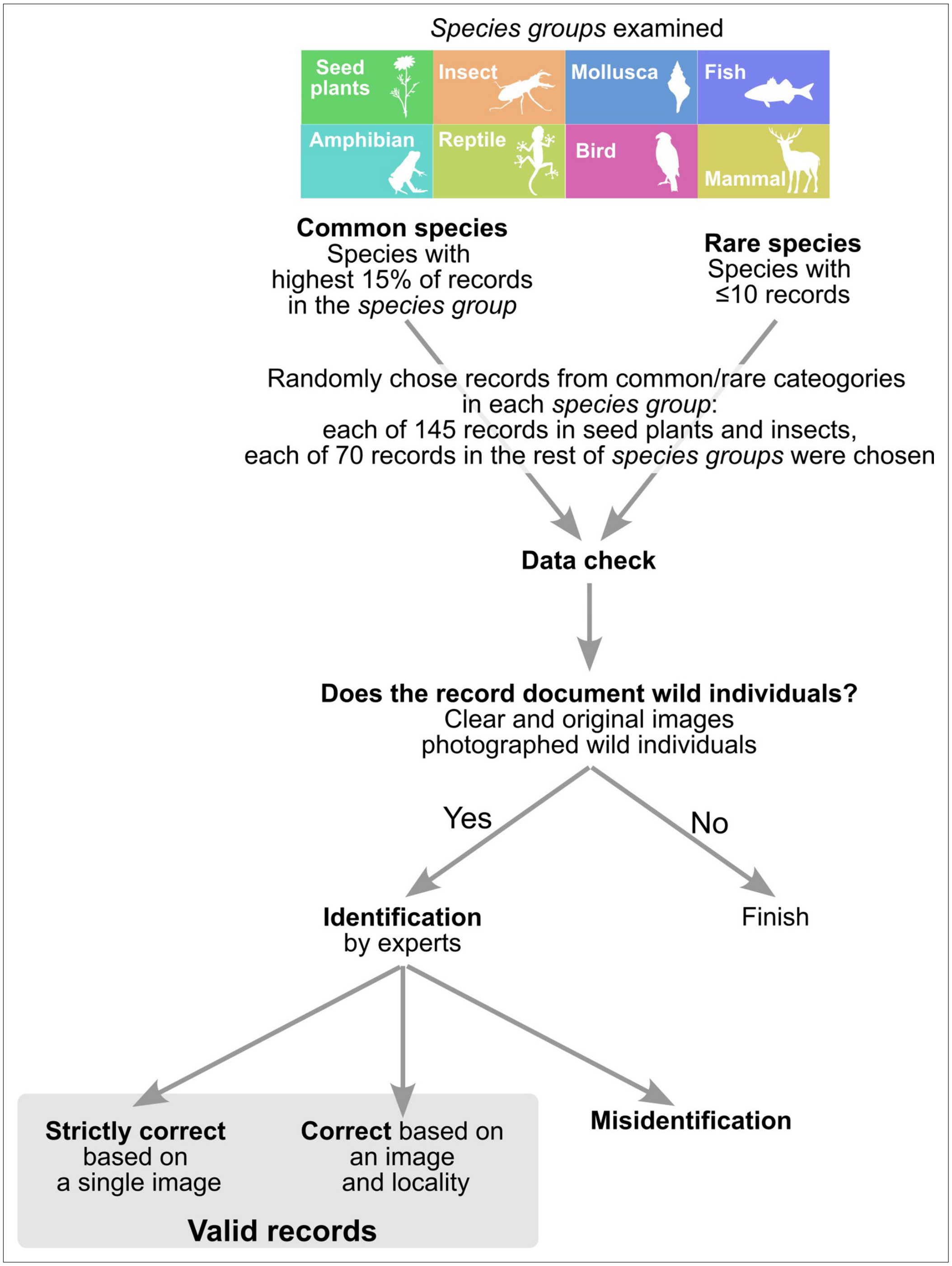

Species groups examined
Seed plants
Insect
Mollusca
Fish
Amphibian
Reptile
Bird
Mammal
Common species
Species with highest 15% of records in the *species group*
Rare species
Species with ≤10 records
Randomly chose records from common/rare cateogories in each *species group*:
each of 145 records in seed plants and insects,
each of 70 records in the rest of *species groups* were chosen
Data check
Does the record document wild individuals?
Clear and original images photographed wild individuals
Yes
No
Identification by experts
Finish
Strictly correct based on a single image
Correct based on an image and locality
Misidentification
Valid records

Figure 4. The workflow of checking accuracy of *Biome* data.

between the sampling periods of the GBIF records and environmental data, we used the GBIF sampled after 1970. The clean_coordinates function of the R package 'CoordinateCleaner' was used to remove records with erroneous coordinates such as records from country capitals and centroids, and biodiversity institutions. We obtained occurrence data from the large occurrence datasets such as the NCRE and Forest Ecosystem Diversity Basic Survey. For the areas or taxa where occurrences were scarce, we further compiled the literature with detailed locality information, such as local species inventories. The amount of occurrence records in the modelled species and species coverage of each dataset is summarised in ***Table 2***. For the species analysed (S9 Table), traditional survey data contains

Table 2. List of species occurrence datasets used for constructing species distribution models (SDMs).

To compare *Biome* dataset with the other datasets, iNaturalist and eBird data based on community science were classified as 'Traditional survey' data.

	Occurrence records of modelled species						
Original dataset	***N***	**Occupancy**	**Species coverage among modelled species**	**Survey method**	**Data group in SDM**	**Download date**	**Availability**
Biome (filtering applied)	201,114	8.6	132/132	Citizen science through smartphone app	*Biome*	7 July 2023	https://biome.co.jp/
National Census on River and Dam Environments (NCRE)	1,413,541	60.2	126/132	Traditional survey on freshwater and its adjacent ecosystems	Traditional survey	10 January 2023	http://www.nilim.go.jp/lab/fbg/ksnkankyo/
Institute records registered at GBIF	530,952	22.6	116/132	Traditional survey and museum specimens	Traditional survey	7 July 2023	GBIF*
iNaturalist and eBird	118,050	5	110/132	Citizen science through smartphone app and web service	Traditional survey*	7 July 2023	GBIF*
Forest Ecosystem Diversity Basic Survey	80,929	3.4	42/132	Traditional survey on forest trees	Traditional survey	30 March 2023	http://forestbio.jp/
Literature	3293	0.1	130/132	Traditional survey	Traditional survey	31 March 2023	Refs*

*For the list of GBIF download doi and literature, see ***Supplementary file 2***.

a negligible portion of community-sourced data (5.5%) because GBIF contains community-sourced data from iNaturalist and eBird.

Predictor variables

Predictors encompass a range of environmental variables recognised to impact species distribution (***Table 3***): land use (***Newbold et al., 2015***), climate (bioclim variables; ***Booth et al., 2014***), vegetation (***Abe, 2018***), lithology (***Ott, 2020***), and elevational range (***Udy et al., 2021***). Additionally, categorical

Table 3. Environmental data used for constructing species distribution models (SDMs).

Years indicate the data collection period. Usage in the SDM shows how the variables were converted before using in the species distribution modelling.

Data	Variables	Year	Usage in the SDM	Available at
Land use	The area sizes of forests, rice fields, farms, wastelands, inland waters, beaches, ocean, golf courses, urbanised areas, and others	2016	Extracted six principal components (PCA) explained ≧ 80% of total variation. PCs were converted into linear, quadratic and hinge terms.	The Ministry of Land, Infrastructure, Transport and Tourism of Japan (MLIT) (https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-L03-a.html)
Forest type	Forest type (planted and natural)	1998	Converted into linear, quadratic, and hinge terms.	The Biodiversity Centre of Japan (http://gis.biodic.go.jp/webgis/index.html)
Climate	Monthly average, minimum and maximum temperature and precipitation	11981–2010	Transformed into 19 bioclimatic variables (***Booth et al., 2014***), then extracted three PCs explained ≧ 80% of total variation. Converted into linear, quadratic, and hinge terms.	MLIT (https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-G02-v3_0.html)
Elevation-al range	Differences between maximum and minimum elevation, and maximum slope	1981	Converted into linear, quadratic, and hinge terms.	MLIT (https://nlftp.mlit.go.jp/ksj/jpgis/datalist/KsjTmplt-G04-a.html)
Vegetation	The area sizes	1998	Transformed into 37 PCs of which total variation explained was more than 80%. Converted into linear, quadratic and hinge terms.	MOE (http://gis.biodic.go.jp/webgis/index.html)
Geology	The area sizes of limestone and serpentinite	2022	Converted into linear, quadratic and hinge terms	The Research Institute of Geology and Geoinformation (https://gbank.gsj.jp/seamless/use.html)
Geohistory	Blakiston's Line (***Dobson, 1994***; ***Saitoh et al., 2015***), oceanic islands (***Wepfer et al., 2016***; ***Yamasaki, 2017***)		Categorical variables	

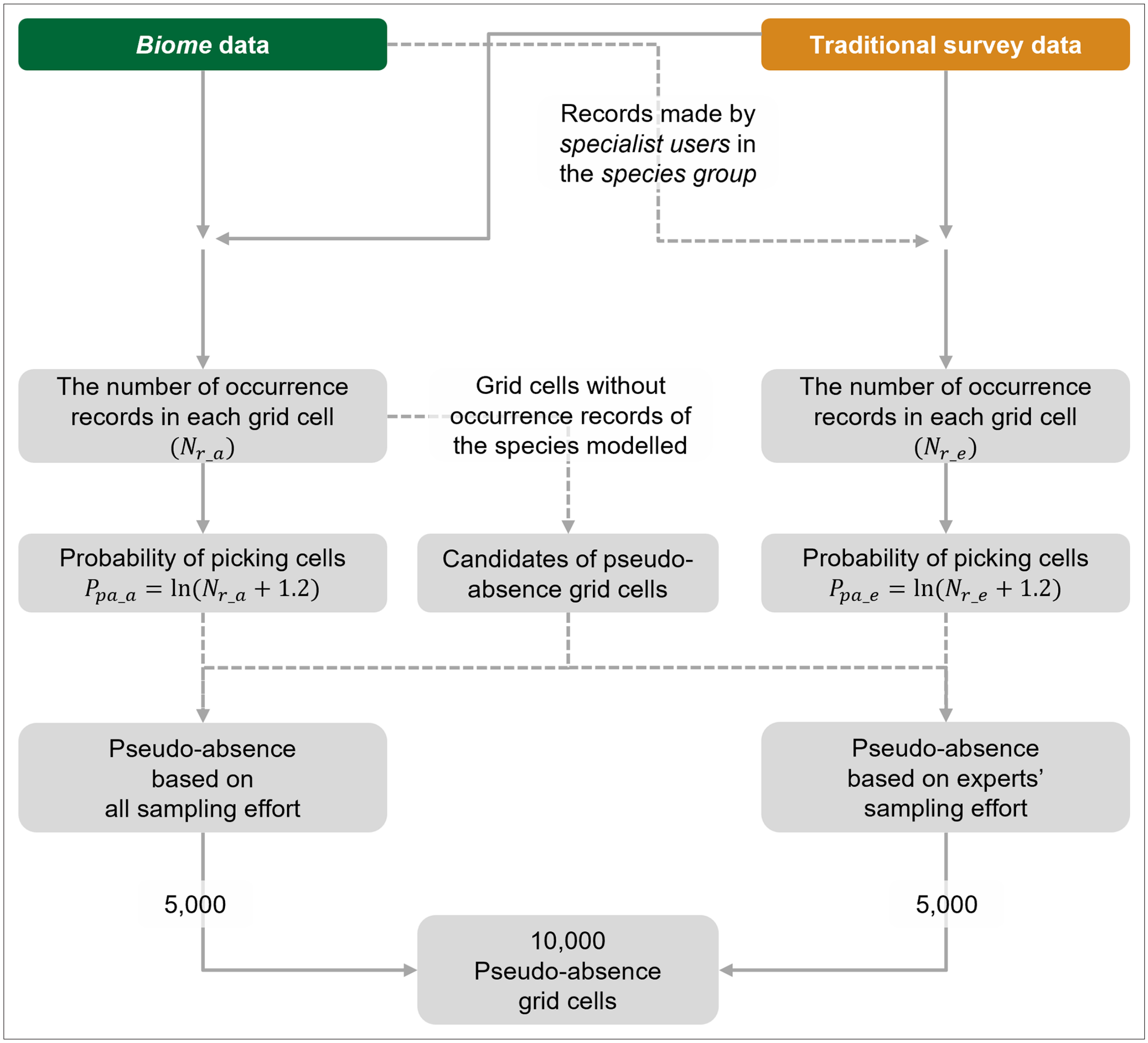

Biome data
Traditional survey data
Records made by *specialist users* in the *species group*
The number of occurrence records in each grid cell (N_{r_a})
Grid cells without occurrence records of the species modelled
The number of occurrence records in each grid cell (N_{r_e})
Probability of picking cells $P_{pa\_a} = \ln(N_{r\_a} + 1.2)$
Candidates of pseudo-absence grid cells
Probability of picking cells $P_{pa\_e} = \ln(N_{r\_e} + 1.2)$
Pseudo-absence based on all sampling effort
Pseudo-absence based on experts' sampling effort
5,000
5,000
10,000 Pseudo-absence grid cells

Figure 5. The workflow for selecting pseudo-absence (background) grid cells for species distribution models (SDMs) using the *Biome*-Traditional dataset. In this process, both *Biome* data and Traditional dataset are utilised to determine the suitable locations for pseudo-absence grid cells. However, when constructing SDMs using the Traditional dataset exclusively, *Biome* data is not involved in the selection of pseudo-absence points.

variables representing known biogeographic regions, reflecting geological history, were included. We applied Blakiston's Line—Tsugaru straits dividing the northern and main islands of Japan (i.e. Hokkaido and Honshu islands)— reflecting a significant historical migration barrier for mammals and birds (**Dobson, 1994**; **Saitoh et al., 2015**). Due to the distinct fauna (**Wepfer et al., 2016**; **Yamasaki, 2017**), we also specified oceanic islands (i.e. Ogasawara and Daito isles) which have never been connected with the Asiatic continents. Continuous environmental variables were transformed into linear, quadratic, and hinge feature classes to illustrate nonlinear associations between environments and species occurrence (**Phillips et al., 2017**). The regularisation multiplier was set at 2.5, falling within the established optimal range of 1.5–4 (**Elith et al., 2010**; **Moreno-Amat et al., 2015**).

Pseudo-absence reflecting search effort

We considered sampling efforts when selecting a total of 10,000 pseudo-absence locations. To accommodate biases in sampling efforts, we assigned picking probabilities as an increasing function of the amount of occurrence records of all and relevant taxa at the grid cell (an index of sampling efforts) (**Milanesi et al., 2020**; **Phillips et al., 2009**). That is, grid cells with rich occurrence records of relevant taxa are more likely to be chosen as pseudo-absences than cells with few records, as detailed below (see also ***Figure 5***).

To generate pseudo-absence (i.e. background) data, we employed two approaches considering different sampling efforts. The first approach incorporated all observers and taxa, while the second approach focused on experts and relevant taxa (***Figure 4***). In both cases, pseudo-absences were selected from grid cells that lacked any occurrence records of the species being modelled. However, due to variations in sampling efforts across locations, it was important to address potential bias. To mitigate this bias, we adjusted the picking probability based on the number of occurrences of other species in each grid cell (***Milanesi et al., 2020***; ***Phillips et al., 2009***).

In the first approach, we assumed that the users of *Biome* submit records of any taxon without specifically selecting species from particular taxa. The picking probability was simply determined by the total number of records from all taxa in the *Biome* data in every grid. In the second approach, we considered the expertise of observers (***Milanesi et al., 2020***) and the sampling effort for relevant taxa (***Phillips et al., 2009***). We also assumed that Traditional surveys targeted particular taxa. Under this approach, we selected records from *Biome* data contributed by *specialist users* and all records from the Traditional survey data. From this subset of data, we calculated the number of records for the taxa (e.g. seed plant, insect, and amphibian) to which the modelled species belonged. This information was then used to calculate the picking probability for each grid cell. To account for the variability in record counts among locations, we applied a logarithmic transformation to the number of records. We also added a value of 1.2 before taking logarithms to allow for the selection of pseudo-absences with low probabilities, particularly in locations with only one or no records of other species. Pseudo-absences were not chosen from the spatial block used as test data, but otherwise, there were no geographical restrictions on their selection.

Using the described approaches, we obtained a total of 10,000 pseudo-absences for our analyses. The amount of pseudo-absences follows the default setting of MaxEnt (***Elith et al., 2011***). For the models using Biome + Traditional dataset (also in *Biome*-blended dataset in Appendix 1), pseudo-absences were generated by merging each of the 5000 points identified through the two approaches. Meanwhile, for SDMs using the Traditional survey data only, we obtained 10,000 pseudo-absences by exclusively using the second approach without incorporating *Biome* data.

Modelling

We modelled distributions of terrestrial seed plants and animals at a scale of 1 × 1 km grid cell, based on Traditional survey data and *Biome* + Traditional data. To model species distributions from presence-only data, several algorithms have been utilised, including generalised additive models, random forest, and neural networks (***Norberg et al., 2019***; ***Valavi et al., 2022***). In our study, we opted for MaxEnt (***Phillips and Dudík, 2008***) due to its high estimation accuracy and relatively low computational burden (***Valavi et al., 2022***). We performed MaxEnt via ENMeval 2.0 package (***Kass et al., 2021***) on R 4.1.3 (***R Core Team, 2021***).

Model evaluation

We evaluated the model by examining spatial transferability because we could not find occurrence data that are environmentally unbiased and independent from training data. To minimise spatial autocorrelation between training and test data, we set a spatial block for splitting data (***Araújo et al., 2019***; ***Santini et al., 2021***). As the spatial block, we chose the central Japan region (latitude, 33.7°–37.7° N; longitude, 136.2°–137.6° E: ***Figure 6***) which covers various environments—alpine to coastal lowlands, metropolis to highly intact areas.

To ensure a fair and balanced assessment of the accuracy of SDMs built from Traditional survey data (0% *Biome* data) and Biome + Traditional data (50% *Biome* data), we compiled a test dataset that embodies characteristics intermediate between these two datasets. This composite test dataset encompasses 25% *Biome* data and 75% Traditional data, effectively bridging the differences between the two original datasets and providing a comprehensive basis for evaluating SDM accuracy.

Due to the presence of invalid records, *Biome* records were used as test data only when multiple users recorded the same species within an identical 1 km grid cell. Although *Biome* data may include invalid records (i.e. non-wild individuals or misidentification), if multiple users recorded the same species at the same place, any one of the records from the place is likely to be valid. As we know

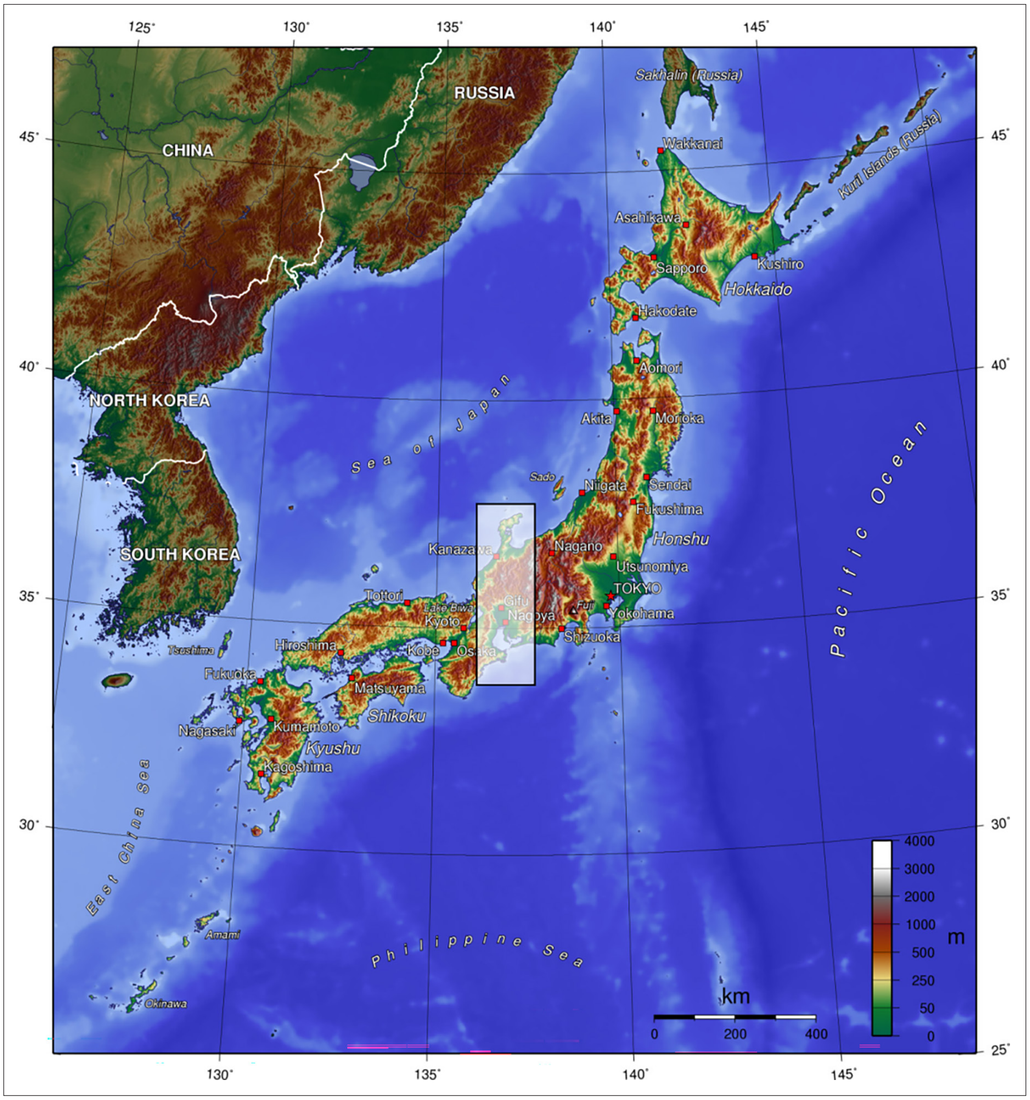


Figure 6. Japanese archipelago, coloured by altitude. Shaded area shows spatial block of test data. Retrieved from Wikipedia (2023, May 30), licensed under Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0).

the fraction of valid records within the *Biome* dataset in each taxon (see Results), we can calculate the probability of the true presence in a given location as follows, by assuming that records made by different users were independent:

$$p_{tp} = 1 - (1 - p_{valid})^{n_{users}}$$

The probability of valid records at a given taxon is shown as p_{valid} , and the number of users reported given species at the place is indicated as n_{users} . If p_{tp} exceeds 99%, we deemed that the species occurred in the location.

To reduce spatial sampling bias, we downsampled a dataset within Traditional survey data, NCRE with massive records from freshwaters, to match the number of records from the remaining Traditional survey data. This procedure is applied to all test datasets in both the main analysis and preliminary analyses documented in ***Figure 3—figure supplement 1*** and Appendix 1.

BI was used to measure model performance because it was designed to evaluate presence-only SDMs (**Hirzel et al., 2006**). In short, BI measures the correlation between estimated habitat preference and the frequency of actual presence, and ranges from –1 to 1. A high BI indicates high SDM accuracy that presence data points tend to be located in grids with higher habitat suitability values. To reliably calculate BI, at least 50 occurrences should be needed in test data (**Hirzel et al., 2006**). Thus, we used 132 species that have more than 50 occurrences in test data for calculating BI (**Supplementary file 3**).

Examining influences of blending *Biome* data on SDM accuracy

Given that the accuracy of SDMs is affected by the amount and quality of data (**Araújo et al., 2019**; **Erickson and Smith, 2023**; **Stockwell and Peterson, 2002**), blending *Biome* data in SDMs may affect the model performances in two possible ways: by increasing the overall amount of data and/or by introducing data with different information than the original data. We analysed to distinguish between these effects. We prepared two different datasets: 'Traditional survey data' and 'Biome + Traditional data'. Then, we separately trained SDMs using these two datasets. We further varied the data size by performing random downsampling, ranging from a minimum of 20 to a maximum of 20,000 records, in order to evaluate its impact on the model. As for the 'Biome + Traditional data' category, the proportion of *Biome* data was kept at 50%. For each condition, we conducted three iterations of training and testing to reduce the impact of random sampling stochasticity. Because the modelling was performed for each species, we obtained BI for each species, amount of records, and dataset (i.e. two datasets consisted of 132 species, each with a maximum of 123 conditions for the amount of records, and the models were replicated three times, resulting in a total of 12,351 individual model runs).

After obtaining BIs for each run, we evaluated the effects of data type (i.e. Biome + Traditional data or Traditional survey data) and species on BI while accounting for the amount of records. For each species and under each amount of records, the mean BI was calculated across the three iterations. Given that BI is a correlation coefficient, we applied the Fisher z-transformation to these BIs to approximate their distribution as a normal distribution. To the transformed BIs, we fitted a generalised linear mixed model that accounted for both the fixed and interaction effects of data type and amount of records. This model accommodated species identity as a random effect. The model was implemented and tested using R packages lme4 (**Bates et al., 2015**) and lmerTest (**Kuznetsova et al., 2017**), respectively.

Acknowledgements

We are grateful to Jamie M Kass for advice on the construction and evaluation of SDMs. We thank Midzuho Tatsuno, Kazumichi Morishita, Hironori Tanaka, Kotaro Takai and Shuhei Tochino for species identification; Akira Sawada and Yuko Maegawa for their advice on the analysis; Trevor H Booth and Dalelan Anderson for comments on the manuscript. We appreciate the invaluable contributions of Biome users. This research was partly supported by the collaborative research agreement between MU at Kyoto University and Biome Inc from 2020 to 2021.

Additional information

Competing interests

Keisuke Atsumi, Yuusuke Nishida: Employed by Biome Inc, but not financially benefit directly from the publication of this paper. Takanori Genroku: CTO of Biome Inc, and an inventor of the species-identification-AI-algorithm JPN patents 6590417 and US patents 11048969, but not financially benefit directly from the publication of this paper. Shogoro Fujiki: CEO of Biome Inc, and an inventor of the species-identification-AI-algorithm JPN patents 6590417 and US patents 11048969, but not financially benefit directly from the publication of this paper. The other authors declare that no competing interests exist.

Funding

No external funding was received for this work.

Author contributions

Keisuke Atsumi, Conceptualization, Data curation, Investigation, Visualization, Methodology, Writing – original draft, Writing – review and editing; Yuusuke Nishida, Data curation, Formal analysis, Investigation, Methodology; Masayuki Ushio, Supervision, Methodology; Hirotaka Nishi, Data curation; Takanori Genroku, Conceptualization, Software; Shogoro Fujiki, Conceptualization, Software, Supervision

Author ORCIDs

Keisuke Atsumi https://orcid.org/0000-0002-8206-4977
Masayuki Ushio http://orcid.org/0000-0003-4831-7181
Shogoro Fujiki https://orcid.org/0000-0002-9778-9532

Peer review material

Reviewer #1 (Public Review): https://doi.org/10.7554/eLife.93694.3.sa1
Author response https://doi.org/10.7554/eLife.93694.3.sa2

Additional files

Supplementary files

- Supplementary file 1. Distributions of occurrence records along with environmental variables.
- Supplementary file 2. List of GBIF data doi and literature compiled in occurrence data.
- Supplementary file 3. List of species for constructed species distribution models.
- MDAR checklist

Data availability

Our analytic code and data are posted on Figshare (https://doi.org/10.6084/m9.figshare.25572462). However, the occurrence data of red-listed species are available upon request for research or application purposes.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Atsumi K, Nishida Y, Ushio M, Nishi H, Genroku T, Fujiki S	2024	Scirpts and data of the article "Boosting biodiversity monitoring using smartphone-driven, rapidly accumulating community-sourced data"	https://figshare.com/articles/dataset/_b_Scirpts_and_data_of_the_article_Boosting_biodiversity_monitoring_b_b_using_smartphone-driven_b_b_rapidly_accumulating_b_b_community-sourced_data_b_/25572462	figshare, 10.6084/m9.figshare.25572462

References

Abe S. 2018. Habitat classification for 69 near threatened plants based on national vegetation survey data. *Veg Sci* **35**:67–88. DOI: https://doi.org/10.15031/vegsci.35.67

Araújo MB, Anderson RP, Márcia Barbosa A, Beale CM, Dormann CF, Early R, Garcia RA, Guisan A, Maiorano L, Naimi B, O'Hara RB, Zimmermann NE, Rahbek C. 2019. Standards for distribution models in biodiversity assessments. *Science Advances* **5**:eaat4858. DOI: https://doi.org/10.1126/sciadv.aat4858, PMID: 30746437

Atsumi K, Koizumi I. 2017. Web image search revealed large-scale variations in breeding season and nuptial coloration in a mutually ornamented fish, *Tribolodon hakonensis*. *Ecological Research* **32**:567–578. DOI: https://doi.org/10.1007/s11284-017-1466-z

Ball IR, Possingham HP, Watts ME. 2009. *Marxan and Relatives: Software for Spatial Conservation prioritizationSpatial Conservation Prioritisation: Quantitative Methods and Computational Tools*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780199547760.001.0001

Bates DM, Maechler M, Bolker B, Walker S. 2015. lme4: linear mixed-effects models using S4 classes. *Journal of Statistical Software* **67**:1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Biome Inc. 2023. The report of Climate Change Biosurvey in 2022. Biome Inc.

Booth TH, Nix HA, Busby JR, Hutchinson MF. 2014. bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions* **20**:1–9. DOI: https://doi.org/10.1111/ddi.12144

Bowser A, Hansen D, He Y, Boston C, Reid M, Gunnell L, Preece J. 2013. Using gamification to inspire new citizen science volunteers. Gamification '13: Proceedings of the First International Conference on Gameful Design, Research, and Applications. 18–25. DOI: https://doi.org/10.1145/2583008.2583011

Box EO. 1981. *Macroclimate and Plant Forms.* Springer. DOI: https://doi.org/10.1007/978-94-009-8680-0

Bury TM, Sujith RI, Pavithran I, Scheffer M, Lenton TM, Anand M, Bauch CT. 2021. Deep learning for early warning signals of tipping points. *PNAS* **118**:e2106140118. DOI: https://doi.org/10.1073/pnas.2106140118, PMID: 34544867

Chan L, Hillel O, Werner P, Holman N, Coetzee I, Galt R, Elmqvist T. 2021. Handbook on the Singapore index on cities Biodiversity (also known as the city Biodiversity index). Chan L (Ed). *The Secretariat of the Convention on Biological Diversity*. Singapore: Secretariat of the Convention on Biological Diversity. p. 1–80.

Chandler M, See L, Copas K, Bonde AMZ, López BC, Danielsen F, Legind JK, Masinde S, Miller-Rushing AJ, Newman G, Rosemartin A, Turak E. 2017. Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation* **213**:280–294. DOI: https://doi.org/10.1016/j.biocon.2016.09.004

Cornwell WK, Ackerly DD. 2009. Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California. *Ecological Monographs* **79**:109–126. DOI: https://doi.org/10.1890/07-1134.1

Crall AW, Newman GJ, Stohlgren TJ, Holfelder KA, Graham J, Waller DM. 2011. Assessing citizen science data quality: an invasive species case study. *Conservation Letters* **4**:433–442. DOI: https://doi.org/10.1111/j.1755-263X.2011.00196.x

Dobson M. 1994. Patterns of distribution in Japanese land mammals. *Mammal Review* **24**:91–111. DOI: https://doi.org/10.1111/j.1365-2907.1994.tb00137.x

Elith J, Kearney M, Phillips S. 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution* **1**:330–342. DOI: https://doi.org/10.1111/j.2041-210X.2010.00036.x

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* **17**:43–57. DOI: https://doi.org/10.1111/j.1472-4642.2010.00725.x

Erickson KD, Smith AB. 2023. Modeling the rarest of the rare: a comparison between multi-species distribution models, ensembles of small models, and single-species models at extremely low sample sizes. *Ecography* **2023**:06500. DOI: https://doi.org/10.1111/ecog.06500

Feldman MJ, Imbeau L, Marchand P, Mazerolle MJ, Darveau M, Fenton NJ. 2021. Trends and gaps in the use of citizen science derived data as input for species distribution models: a quantitative review. *PLOS ONE* **16**:e0234587. DOI: https://doi.org/10.1371/journal.pone.0234587, PMID: 33705414

Fink D, Johnston A, Strimas-Mackey M, Auer T, Hochachka WM, Ligocki S, Oldham Jaromczyk L, Robinson O, Wood C, Kelling S, Rodewald AD. 2023. A double machine learning trend model for citizen science data. *Methods in Ecology and Evolution* **14**:2435–2448. DOI: https://doi.org/10.1111/2041-210X.14186

Fuccillo Battle K, Duhon A, Vispo CR, Crimmins TM, Rosenstiel TN, Armstrong-Davies LL, de Rivera CE. 2022. Citizen science across two centuries reveals phenological change among plant species and functional groups in the northeastern US. *Journal of Ecology* **110**:1757–1774. DOI: https://doi.org/10.1111/1365-2745.13926

Fujiki S, Tatsuno M. 2021. Practice of citizen science for developing biodiversity monitoring methods using mobile devices. *Jpn J Ecol* **71**:85–90. DOI: https://doi.org/10.18960/seitai.71.2_85

Gonzalez A, Chase JM, O'Connor MI. 2023. A framework for the detection and attribution of biodiversity change. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **378**:20220182. DOI: https://doi.org/10.1098/rstb.2022.0182, PMID: 37246383

Groom Q, Pernat N, Adriaens T, de Groot M, Jelaska SD, Marčiulynienė D, Martinou AF, Skuhrovec J, Tricarico E, Wit EC, Roy HE. 2021. Species interactions: next-level citizen science. *Ecography* **44**:1781–1789. DOI: https://doi.org/10.1111/ecog.05790

Hart AG, Bosley H, Hooper C, Perry J, Sellors-Moore J, Moore O, Goodenough AE. 2023. Assessing the accuracy of free automated plant identification applications. *People and Nature* **5**:929–937. DOI: https://doi.org/10.1002/pan3.10460

Herodotou C, Ismail N, I. Benavides Lahnstein A, Aristeidou M, Young AN, Johnson RF, Higgins LM, Ghadiri Khanaposhtani M, Robinson LD, Ballard HL. 2024. Young people in iNaturalist: a blended learning framework for biodiversity monitoring. *International Journal of Science Education, Part B* **14**:129–156. DOI: https://doi.org/10.1080/21548455.2023.2217472

Hirzel AH, Le Lay G, Helfer V, Randin C, Guisan A. 2006. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling* **199**:142–152. DOI: https://doi.org/10.1016/j.ecolmodel.2006.05.017

Hutchinson GE. 1957. *Concluding Remarks Cold Spring Harbor Symposia on Quantitative Biology*. Cold Spring Harbor Laboratory Press. DOI: https://doi.org/10.1101/SQB.1957.022.01.039

IPBES. 2019. Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem ServicesIPBES.

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

Jones KR, Watson JEM, Possingham HP, Klein CJ. 2016. Incorporating climate change into spatial conservation prioritisation: A review. *Biological Conservation* **194**:121–130. DOI: https://doi.org/10.1016/j.biocon.2015.12.008

Kaplan Mintz K, Arazy O, Malkinson D. 2023. Multiple forms of engagement and motivation in ecological citizen science. *Environmental Education Research* **29**:27–44. DOI: https://doi.org/10.1080/13504622.2022.2120186

Kass JM, Muscarella R, Galante PJ, Bohl CL, Pinilla-Buitrago GE, Boria RA, Soley-Guardia M, Anderson RP. 2021. ENMeval 2.0: redesigned for customizable and reproducible modeling of species' niches and distributions. *Methods in Ecology and Evolution* **12**:1602–1608. DOI: https://doi.org/10.1111/2041-210X.13628

Kass JM, Fukaya K, Thuiller W, Mori AS. 2024. Biodiversity modeling advances will improve predictions of nature's contributions to people. *Trends in Ecology & Evolution* **39**:338–348. DOI: https://doi.org/10.1016/j.tree.2023.10.011, PMID: 37968219

Kendal D, Egerer M, Byrne JA, Jones PJ, Marsh P, Threlfall CG, Allegretto G, Kaplan H, Nguyen HKD, Pearson S, Wright A, Flies EJ. 2020. City-size bias in knowledge on the effects of urban nature on people and biodiversity. *Environmental Research Letters* **15**:124035. DOI: https://doi.org/10.1088/1748-9326/abc5e4

Keough HL, Blahna DJ. 2006. Achieving integrative, collaborative ecosystem management. *Conservation Biology* **20**:1373–1382. DOI: https://doi.org/10.1111/j.1523-1739.2006.00445.x, PMID: 17002755

Kindt R. 2023. TreeGOER: A database with globally observed environmental ranges for 48,129 tree species. *Global Change Biology* **29**:6303–6318. DOI: https://doi.org/10.1111/gcb.16914, PMID: 37602408

Klinger YP, Eckstein RL, Kleinebecker T. 2023. iPhenology: Using open-access citizen science photos to track phenology at continental scale. *Methods in Ecology and Evolution* **14**:1424–1431. DOI: https://doi.org/10.1111/2041-210X.14114

Kobori H, Dickinson JL, Washitani I, Sakurai R, Amano T, Komatsu N, Kitamura W, Takagawa S, Koyama K, Ogawara T, Miller-Rushing AJ. 2016. Citizen science: a new approach to advance ecology, education, and conservation. *Ecological Research* **31**:1–19. DOI: https://doi.org/10.1007/s11284-015-1314-y

Koide D, Tsujimoto S, Kumagai N, Ikegami M, Nishihiro J. 2023. Species' spatiotemporal distribution platform based on citizen science through desirable circulation between the real and digital worlds. *Jpn J Conserv Ecol* **01**:2217. DOI: https://doi.org/10.18960/hozen.2217

Kuznetsova A, Brockhoff PB, Christensen RHB. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* **82**:i13. DOI: https://doi.org/10.18637/jss.v082.i13

Larson ER, Graham BM, Achury R, Coon JJ, Daniels MK, Gambrell DK, Jonasen KL, King GD, LaRacuente N, Perrin-Stowe TI, Reed EM, Rice CJ, Ruzi SA, Thairu MW, Wilson JC, Suarez AV. 2020. From eDNA to citizen science: emerging tools for the early detection of invasive species. *Frontiers in Ecology and the Environment* **18**:194–202. DOI: https://doi.org/10.1002/fee.2162

Laubmeier AN, Cazelles B, Cuddington K, Erickson KD, Fortin MJ, Ogle K, Wikle CK, Zhu K, Zipkin EF. 2020. Ecological dynamics: integrating empirical, statistical, and analytical methods. *Trends in Ecology & Evolution* **35**:1090–1099. DOI: https://doi.org/10.1016/j.tree.2020.08.006, PMID: 32933777

Leighton GRM, Hugo PS, Roulin A, Amar A. 2016. Just Google it: assessing the use of Google Images to describe geographical variation in visible traits of organisms. *Methods in Ecology and Evolution* **7**:1060–1070. DOI: https://doi.org/10.1111/2041-210X.12562

Linsley P, Abdelbadie R, Abdelbadie R. 2023. The Taskforce on Nature-related Financial Disclosures must engage widely and justify its market-led approach. *Nature Ecology & Evolution* **7**:1343–1346. DOI: https://doi.org/10.1038/s41559-023-02113-w, PMID: 37386084

Loh J, Green RE, Ricketts T, Lamoreux J, Jenkins M, Kapos V, Randers J. 2005. The Living Planet Index: using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B* **360**:289–295. DOI: https://doi.org/10.1098/rstb.2004.1584

Milanesi P, Mori E, Menchetti M. 2020. Observer-oriented approach improves species distribution models from citizen science data. *Ecology and Evolution* **10**:12104–12114. DOI: https://doi.org/10.1002/ece3.6832, PMID: 33209273

Miller DAW, Pacifici K, Sanderlin JS, Reich BJ. 2019. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution* **10**:22–37. DOI: https://doi.org/10.1111/2041-210X.13110

Miya M, Sado T, Oka S, Fukuchi T. 2022. The use of citizen science in fish eDNA metabarcoding for evaluating regional biodiversity in A coastal marine region: A pilot study. *Metabarcoding and Metagenomics* **6**:e80444. DOI: https://doi.org/10.3897/mbmg.6.80444

Moreno-Amat E, Mateo RG, Nieto-Lugilde D, Morueta-Holme N, Svenning JC, García-Amorena I. 2015. Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. *Ecological Modelling* **312**:308–317. DOI: https://doi.org/10.1016/j.ecolmodel.2015.05.035

Mori AS, Suzuki KF, Hori M, Kadoya T, Okano K, Uraguchi A, Muraoka H, Sato T, Shibata H, Suzuki-Ohno Y, Koba K, Toda M, Nakano S, Kondoh M, Kitajima K, Nakamura M. 2023. Perspective: sustainability challenges, opportunities and solutions for long-term ecosystem observations. *Philosophical Transactions of the Royal Society B* **378**:20220192. DOI: https://doi.org/10.1098/rstb.2022.0192

Newbold T, Hudson LN, Hill SLL, Contu S, Lysenko I, Senior RA, Börger L, Bennett DJ, Choimes A, Collen B, Day J, De Palma A, Díaz S, Echeverria-Londoño S, Edgar MJ, Feldman A, Garon M, Harrison MLK, Alhusseini T, Ingram DJ, et al. 2015. Global effects of land use on local terrestrial biodiversity. *Nature* **520**:45–50. DOI: https://doi.org/10.1038/nature14324, PMID: 25832402

Newbold T, Hudson LN, Arnell AP, Contu S, De Palma A, Ferrier S, Hill SLL, Hoskins AJ, Lysenko I, Phillips HRP, Burton VJ, Chng CWT, Emerson S, Gao D, Pask-Hale G, Hutton J, Jung M, Sanchez-Ortiz K, Simmons BI, Whitmee S, et al. 2016. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science* **353**:288–291. DOI: https://doi.org/10.1126/science.aaf2201, PMID: 27418509

Norberg A, Abrego N, Blanchet FG, Adler FR, Anderson BJ, Anttila J, Araújo MB, Dallas T, Dunson D, Elith J, Foster SD, Fox R, Franklin J, Godsoe W, Guisan A, O'Hara B, Hill NA, Holt RD, Hui FKC, Husby M, et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs* **89**:e01370. DOI: https://doi.org/10.1002/ecm.1370

Ott RF. 2020. How lithology impacts global topography, vegetation, and animal biodiversity: a global-scale analysis of mountainous regions. *Geophysical Research Letters* **47**:e2020GL088649. DOI: https://doi.org/10.1029/2020GL088649

Pacifici K, Reich BJ, Miller DAW, Gardner B, Stauffer G, Singh S, McKerrow A, Collazo JA. 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology* **98**:840–850. DOI: https://doi.org/10.1002/ecy.1710, PMID: 28027588

Pennekamp F, Iles AC, Garland J, Brennan G, Brose U, Gaedke U, Jacob U, Kratina P, Matthews B, Munch S, Novak M, Palamara GM, Rall BC, Rosenbaum B, Tabi A, Ward C, Williams R, Ye H, Petchey OL. 2019. The intrinsic predictability of ecological time series and its potential to guide forecasting. *Ecological Monographs* **89**:e01359. DOI: https://doi.org/10.1002/ecm.1359

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**:231–259. DOI: https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips SJ, Dudík M. 2008. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography* **31**:161–175. DOI: https://doi.org/10.1111/j.0906-7590.2008.5203.x

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19**:181–197. DOI: https://doi.org/10.1890/07-2153.1, PMID: 19323182

Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME. 2017. Opening the black box: an open-source release of maxent. *Ecography* **40**:887–893. DOI: https://doi.org/10.1111/ecog.03049

Pocock MJO, Tweddle JC, Savage J, Robinson LD, Roy HE. 2017. The diversity and evolution of ecological and environmental citizen science. *PLOS ONE* **12**:e0172579. DOI: https://doi.org/10.1371/journal.pone.0172579, PMID: 28369087

Pocock MJO, Chandler M, Bonney R, Thornhill I, Albin A, August T, Bachman S, Brown PMJ, Cunha DGF, Grez A, Jackson C, Peters M, Rabarijaon NR, Roy HE, Zaviezo T, Danielsen F. 2018. Chapter six - A vision for global Biodiversity monitoring with citizen science in. Bohan DA, Dumbrell AJ, Woodward G, Jackson M (Eds). *Advances in Ecological Research, Next Generation Biomonitoring: Part 2*. Academic Press. p. 169–223. DOI: https://doi.org/10.1016/bs.aecr.2018.06.003

Ponti M, Hillman T, Stankovic I. 2015. Science and Gamification: The Odd Couple?. CHI PLAY '15: Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in PLAY. 679–684. DOI: https://doi.org/10.1145/2793107.2810293

Porfirio LL, Harris RMB, Lefroy EC, Hugh S, Gould SF, Lee G, Bindoff NL, Mackey B. 2014. Improving the use of species distribution models in conservation planning and management under climate change. *PLOS ONE* **9**:e113749. DOI: https://doi.org/10.1371/journal.pone.0113749, PMID: 25420020

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Found Stat Comput Vienna Austria.

Reddy S, Dávalos LM. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography* **30**:1719–1727. DOI: https://doi.org/10.1046/j.1365-2699.2003.00946.x

Renner SS, Zohner CM. 2018. Climate change and phenological mismatch in trophic interactions among plants, insects, and vertebrates. *Annual Review of Ecology, Evolution, and Systematics* **49**:165–182. DOI: https://doi.org/10.1146/annurev-ecolsys-110617-062535

Robinson OJ, Ruiz-Gutierrez V, Reynolds MD, Golet GH, Strimas-Mackey M, Fink D. 2020. Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions* **26**:976–986. DOI: https://doi.org/10.1111/ddi.13068, PMID: 36960319

Roy HE, Pauchard A, Stoett P, Renard Truong T, Bacher S, Galil BS, Hulme PE, Ikeda T, Sankaran KV, McGeoch MA, Meyerson LA, Nuñez MA, Ordonez A, Rahlao SJ, Schwindt E, Seebens H, Sheppard AW, Vandvik V. 2023. IPBES invasive alien species assessment: summary for policymakers. Version 2. Zenodo. DOI: https://doi.org//10.5281/zenodo.8314303

Saitoh T, Sugita N, Someya S, Iwami Y, Kobayashi S, Kamigaichi H, Higuchi A, Asai S, Yamamoto Y, Nishiumi I. 2015. DNA barcoding reveals 24 distinct lineages as cryptic bird species candidates in and around the Japanese Archipelago. *Molecular Ecology Resources* **15**:177–186. DOI: https://doi.org/10.1111/1755-0998.12282, PMID: 24835119

Santini L, Benítez-López A, Maiorano L, Čengić M, Huijbregts MAJ. 2021. Assessing the reliability of species distribution projections in climate change research. *Diversity and Distributions* **27**:1035–1050. DOI: https://doi.org/10.1111/ddi.13252

Scholes RJ, Biggs R. 2005. A biodiversity intactness index. *Nature* **434**:45–49. DOI: https://doi.org/10.1038/nature03289, PMID: 15744293

Shiono T, Kubota Y, Kusumoto B. 2021. Area-based conservation planning in Japan: The importance of OECMs in the post-2020 Global Biodiversity Framework. *Global Ecology and Conservation* **30**:e01783. DOI: https://doi.org/10.1016/j.gecco.2021.e01783

Soga M, Gaston KJ. 2023. Nature benefit hypothesis: direct experiences of nature predict self-reported pro-biodiversity behaviors. *Conservation Letters* **16**:e12945. DOI: https://doi.org/10.1111/conl.12945

Steen VA, Elphick CS, Tingley MW. 2019. An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions* **25**:1857–1869. DOI: https://doi.org/10.1111/ddi.12985

Stockwell DRB, Peterson AT. 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling* **148**:1–13. DOI: https://doi.org/10.1016/S0304-3800(01)00388-X

Tilman D, Reich PB, Knops JMH. 2006. Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature* **441**:629–632. DOI: https://doi.org/10.1038/nature04742, PMID: 16738658

TNFD. 2023. Taskforce on nature-related financial disclosures (TNFD) recommendations version 1.0. version 1.0. TNFD Recommendations.

Udy K, Fritsch M, Meyer KM, Grass I, Hanß S, Hartig F, Kneib T, Kreft H, Kukunda CB, Pe'er G, Reininghaus H, Tietjen B, Tscharntke T, van Waveren C, Wiegand K. 2021. Environmental heterogeneity predicts global species richness patterns better than area. *Global Ecology and Biogeography* **30**:842–851. DOI: https://doi.org/10.1111/geb.13261

Urban MC, Bocedi G, Hendry AP, Mihoub JB, Pe'er G, Singer A, Bridle JR, Crozier LG, De Meester L, Godsoe W, Gonzalez A, Hellmann JJ, Holt RD, Huth A, Johst K, Krug CB, Leadley PW, Palmer SCF, Pantel JH, Schmitz A, et al. 2016. Improving the forecast for biodiversity under climate change. *Science* **353**:aad8466. DOI: https://doi.org/10.1126/science.aad8466

Ushio M, Hsieh CH, Masuda R, Deyle ER, Ye H, Chang CW, Sugihara G, Kondoh M. 2018. Fluctuating interaction network and time-varying stability of a natural fish community. *Nature* **554**:360–363. DOI: https://doi.org/10.1038/nature25504, PMID: 29414940

Valavi R, Guillera-Arroita G, Lahoz-Monfort JJ, Elith J. 2022. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs* **92**:e01486. DOI: https://doi.org/10.1002/ecm.1486

Visser ME, Gienapp P. 2019. Evolutionary and demographic consequences of phenological mismatches. *Nature Ecology & Evolution* **3**:879–885. DOI: https://doi.org/10.1038/s41559-019-0880-8, PMID: 31011176

Wallace RD, Bargeron CT. 2014. Identifying invasive species in real time: early detection and distribution mapping system (Eddmaps) and other mapping tools. Ziska LH (Ed). *Invasive Species and Global Climate Change*. CABI. p. 219–231. DOI: https://doi.org/10.1079/9781780641645.0000

Ward G, Hastie T, Barry S, Elith J, Leathwick JR. 2009. Presence-only data and the EM algorithm. *Biometrics* **65**:554–563. DOI: https://doi.org/10.1111/j.1541-0420.2008.01116.x, PMID: 18759851

Wepfer PH, Guénard B, Economo EP. 2016. Influences of climate and historical land connectivity on ant beta diversity in East Asia. *Journal of Biogeography* **43**:2311–2321. DOI: https://doi.org/10.1111/jbi.12762

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* **14**:763–773. DOI: https://doi.org/10.1111/j.1472-4642.2008.00482.x

Wood C, Sullivan B, Iliff M, Fink D, Kelling S. 2011. eBird: engaging birders in science and conservation. *PLOS Biology* **9**:e1001220. DOI: https://doi.org/10.1371/journal.pbio.1001220, PMID: 22205876

Yamasaki T. 2017. Biogeographic pattern of Japanese birds: a cluster analysis of faunal similarity and a review of phylogenetic evidence in. Motokawa M, Kajihara H (Eds). *Species Diversity of Animals in Japan, Diversity and Commonality in Animals*. Tokyo: Springer Japan. p. 117–134. DOI: https://doi.org/10.1007/978-4-431-56432-4

Zapponi L, Cini A, Bardiani M, Hardersen S, Maura M, Maurizi E, Redolfi De Zan L, Audisio P, Bologna MA, Carpaneto GM, Roversi PF, Sabbatini Peverieri G, Mason F, Campanaro A. 2017. Citizen science data as an efficient tool for mapping protected saproxylic beetles. *Biological Conservation* **208**:139–145. DOI: https://doi.org/10.1016/j.biocon.2016.04.035, PMID: 28393617

Zhang W, Sheldon BC, Grenyer R, Gaston KJ. 2021. Habitat change and biased sampling influence estimation of diversity trends. *Current Biology* **31**:3656–3662. DOI: https://doi.org/10.1016/j.cub.2021.05.066, PMID: 34171303

Appendix 1

Determine the best blend of Traditional survey and *Biome* data

Methods

In this investigation, we aimed to determine the optimal proportion of *Biome* data within the training dataset of SDMs in order to enhance the accuracy of SDMs. To conduct this assessment, we initially selected a subset of species for which sufficient test data was available (as detailed below).

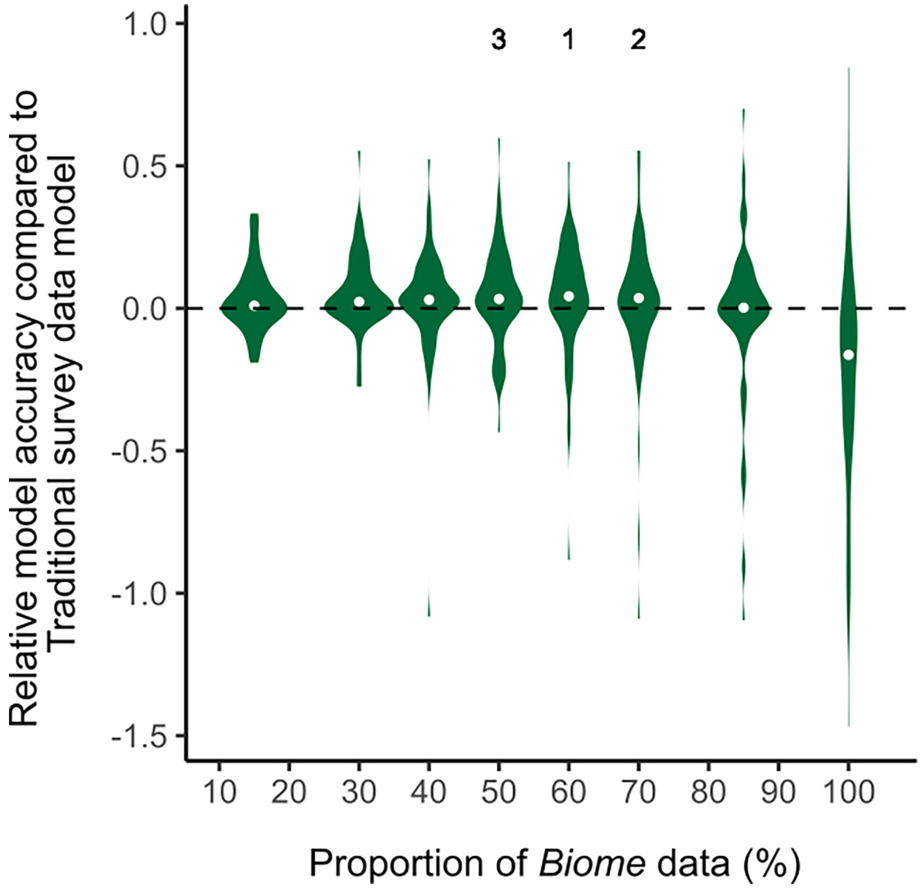
For each of the selected species, we generated training datasets by combining Traditional survey data with *Biome* data at different proportions: 15, 30, 40, 50, 60, 70, 85, and 100%, referred to as *Biome*-blended datasets.

To compare the accuracy of SDMs, we created and evaluated models using both the Traditional survey dataset and each of the *Biome*-blended datasets. SDMs were created by following the methodology employed in the main analysis. To ensure equitable comparison, we equalised the amount of data in each pair of blended and Traditional survey datasets. This equalisation was achieved by randomly downsampling the larger dataset to match the size of the smaller one.

We assessed the accuracy of the models using the BI, which follows the same methodology as employed in the main analysis. In this specific investigation, we did not control the proportion of *Biome* data within the test data. We selected a set of species for which the test dataset contained at least 50 locations and randomly chose 20 species from each of the seed plants, insects, and birds (see ***Supplementary file 3***).

Results

The analysis revealed that the relative model accuracy becomes high positive values when training data comprises 50–70% of *Biome* data (***Appendix 1—figure 1***). This indicates that SDM accuracy is substantially enhanced when the training data incorporates 50–70% of *Biome* data. The relative model accuracy remained positive in the 15–70% *Biome*-blended datasets, but decreased to negative values in the 85 and 100% *Biome*-blended datasets (***Appendix 1—figure 1***). This suggests that blending *Biome* data generally enhances the accuracy of SDMs, but it is important to include at least 30% Traditional survey data to maintain accuracy. Based on the high performance observed and simplicity, we selected the 50% *Biome*-blended dataset as the Biome + Traditional data for comparing model accuracy with the Traditional survey data in the main text.



Appendix 1—figure 1. The violin plots of relative model accuracy between species distribution models (SDMs) using Biome-blended data and Traditional survey data. The median values are shown as grey dots. The positive relative model accuracy indicates that SDMs that used Biome data outperformed models that used Traditional survey data.