Boosting biodiversity monitoring using smartphone-driven, rapidly accumulating citizen data

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

[1]*Biome Inc., 134 Chudoji Minamimachi, Shimogyo-ku, Kyoto 600-8813, Japan*

[2]*Department of Ocean Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR*

[3]*Toyohashi Museum of Natural History, 1-238 Oana, Oiwa, Toyohashi, Aichi 441-3147, Japan*

*Correspondence: fujiki@biome.co.jp

Abstract

Kunming-Montreal Global Biodiversity Framework increased the demand for biodiversity distribution data. To gather species observation from the public, we introduced a mobile application called '*Biome*' in Japan. By employing species identification algorithms and gamification elements, *Biome* has gathered >5M observations since its launch in 2019. However, cloud-sourced data often exhibit spatial and taxonomic biases. Species distribution models (SDMs) enable infer species distribution while accommodating such bias. We investigated *Biome* data's quality and how incorporating *Biome* data influences the performance of SDMs. Species identification accuracy of *Biome data* exceeds 95% for birds, reptiles, mammals, and amphibians, but seed plants, molluscs, and fishes scored below 90%. The distributions of 132 terrestrial plants and animals across Japan were modeled, and their accuracy was improved by incorporating *Biome* data into traditional survey data. For endangered species, traditional survey data required >2,000 records to build accurate models (Boyce index ≥ 0.9), though only ca.300 records were required when *Biome* was blended. The unique data distributions may explain this improvement: *Biome* data covers urban-natural gradients uniformly, while traditional data is biased towards natural areas. Combining multiple data sources offers insights into species distributions across Japan, aiding protected area designation and ecosystem service assessment.

KEY WORDS:

Citizen science, Habitat suitability model, Maxent, Big data, Taskforce on Nature-related Financial Disclosures: TNFD, Other Effective area-based Conservation Measures: OECMs, Natural capital

Introduction

Kunming-Montreal Global Biodiversity Framework (GBF) by the United Nations envisions reversing the nature loss by 2030. As direct means for nature conservation, GBF targeted making 30% of Earth's land and ocean area as protected areas by 2030 (i.e. 30by30). As an indirect but influential way, 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 will be guided by Taskforce on Nature-related Financial Disclosures [1]. 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. 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.

Since the emergence of digital devices and the internet, citizens have been able to share their observations with communities through various media, such as images and video/audio recordings. These citizen-sourced data have contributed significantly to accumulating ecosystem information, including phenology and species occurrence [2–4]. 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 [3,4]. Individuals submit records to these platforms for various reasons, including a desire to contribute to science and engage with cutting-edge technologies [5,6]. By making the process more enjoyable (i.e. gamification), we can potentially gather even more biological data from the public [7,8]. Yet, the collection process of citizen-sourced data is usually not well-designed (e.g., spatially biased "presence-only" data) [9,10] and its interpretation is challenging without proper statistical modeling. Thus, although much effort has been invested in developing effective monitoring and modeling methods for biodiversity assessment, current approaches can be further improved by incorporating (i) more enjoyable citizen-based survey platform using mobile applications and (ii) employing an advanced statistical modeling framework in estimating species distribution.

To fuel citizens' engagement in biodiversity surveys and environmental education, we launched the mobile application '*Biome*' in 2019 in Japan [11]. 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 (Fig 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 [11,12]. 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 citizen science [7,13]. Consequently, *Biome* has accumulated data rapidly. Since its launch, 5.8 million records have been collected through the app. This is more than four times greater than the number of records accumulated by GBIF (Global Biodiversity Information Facility) 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 analyzing species occurrence records.

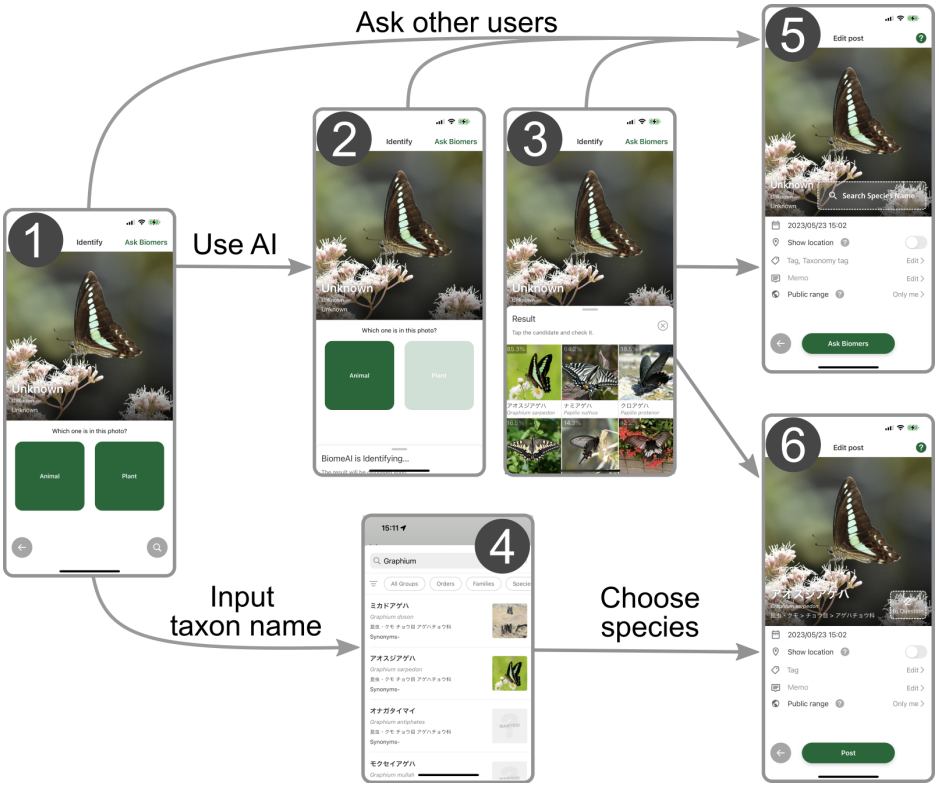

Fig. 1 | Workflow of submitting records to Biome. (1) Users can upload images that were taken by the smartphone camera or import existing images from the storage, including those imported from external devices. (2) Users select whether the image is about animals or plants to activate the species identification AI. (3) The AI analyzes 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.

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 [14,15]. These models play a crucial role in conservation and restoration planning by helping predict how changes in land use and climate impact species distributions [16,17]. While species presence/absence data—which needs extensive surveys by experts—is limited, presence-only data—which can be obtained from citizens' observations—is much more available. Maxent is one of the most popular SDM methods, which can estimate species distribution from presence-only data by maximizing the entropy of the probability distribution while satisfying constraints based on the available information [15,18]. Since Maxent only requires occurrence records, it is well-suited for empowering citizen-based observations to predict species distributions. Also, while citizen science data often suffer from spatially-biased sampling efforts (i.e., sampling tends to concentrate in densely populated or touristic areas [19,20]), SDMs such as Maxent can account for such spatial biases by considering the spatial distribution of sampling efforts when selecting pseudo-absence (background) locations [21,22]. When sampling efforts are adequately controlled, adding citizen science data improves the accuracy of SDMs [10,23,24]. This implies that SDMs may be substantially improved by utilizing rapidly accumulating *Biome*'s species occurrence records if we adequately control the sampling efforts.

Here, we show the quality of citizen-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 modeled 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 citizen science 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 (Fig. 2A). The amount of occurrence records submitted to *Biome* has increased across the years (Fig. 2B). On average in 2022, users submitted 5,407 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 summarize all environmental variables. The two datasets demonstrated divergent distribution patterns along PC1 (Fig. 2C). This component, accounting for 6.1% of the total variation, is primarily influenced by land use, topography, and climate (supplementary material S1). 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 substantially biased towards natural areas (Fig. 2C). The majority of records are attributed to insects (31.2%) and to seed plants (41.8%), which are relatively accessible and can be easily photographed using smartphones (Fig. 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 utilized 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.

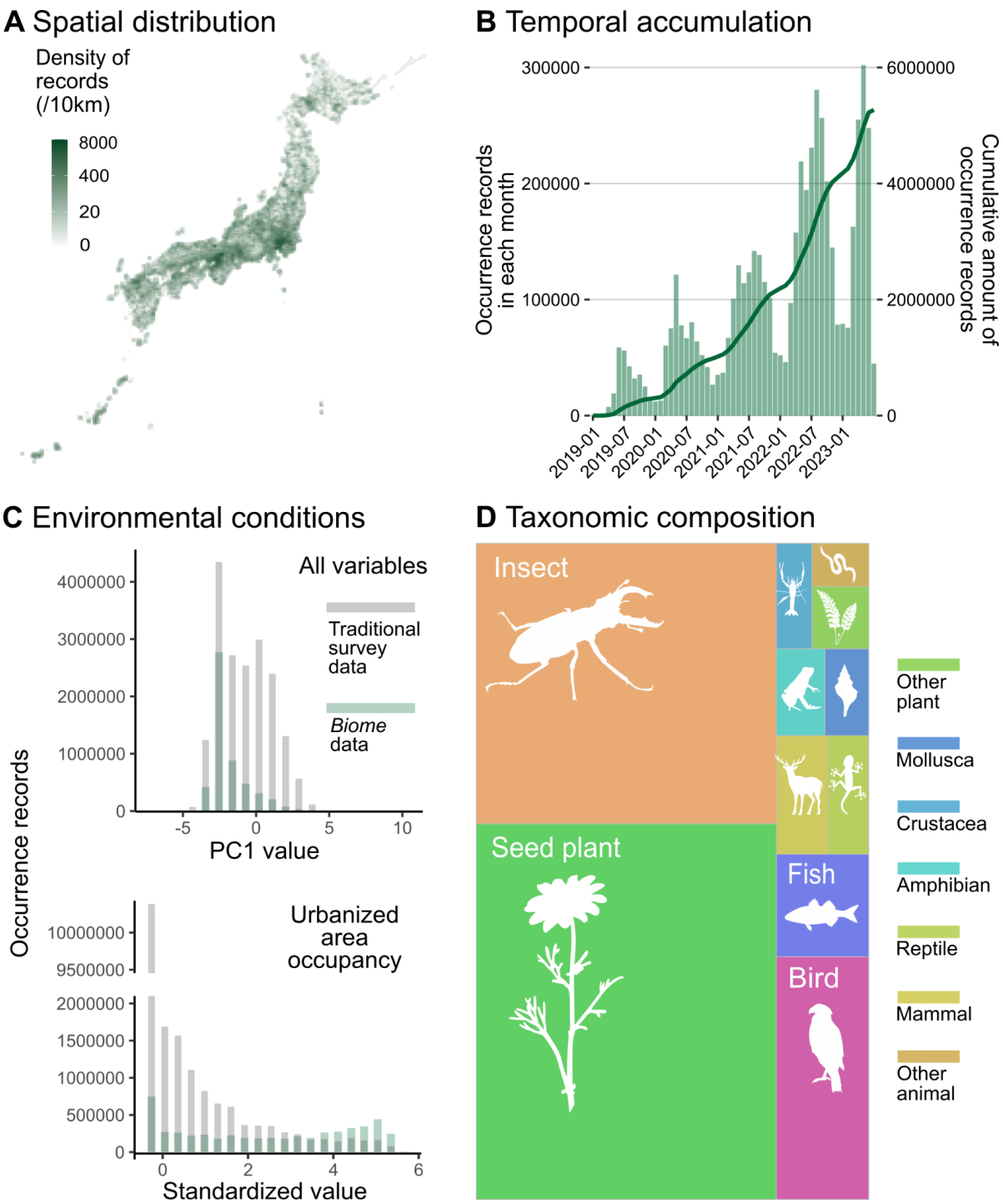

Fig. 2 | Description of data accumulated by *Biome*. Data distributions are shown based on all records submitted to *Biome* by 7 July 2023 (*N*=5,275,457). A Spatial distribution of records across Japan. B Accumulation of records through time. The barplot represents the number of records each month and the line shows the cumulative amount of records. C Distributions of records along with PC1 of all environmental variables and standardized 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

The performance of species distribution models

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 (Fig. 3). Our analysis revealed that although the intercept of the Boyce Index (BI, model accuracy metric that ranges between -1 to 1) did not differ between the two datasets (β=0.02±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, comparing to models solely relying on Traditional survey data (β=0.02±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 2,129±4,157 records to achieve the same accuracy. This was also true in endangered species (included in Japanese national or prefectural red lists); although 2,336±3,718 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).

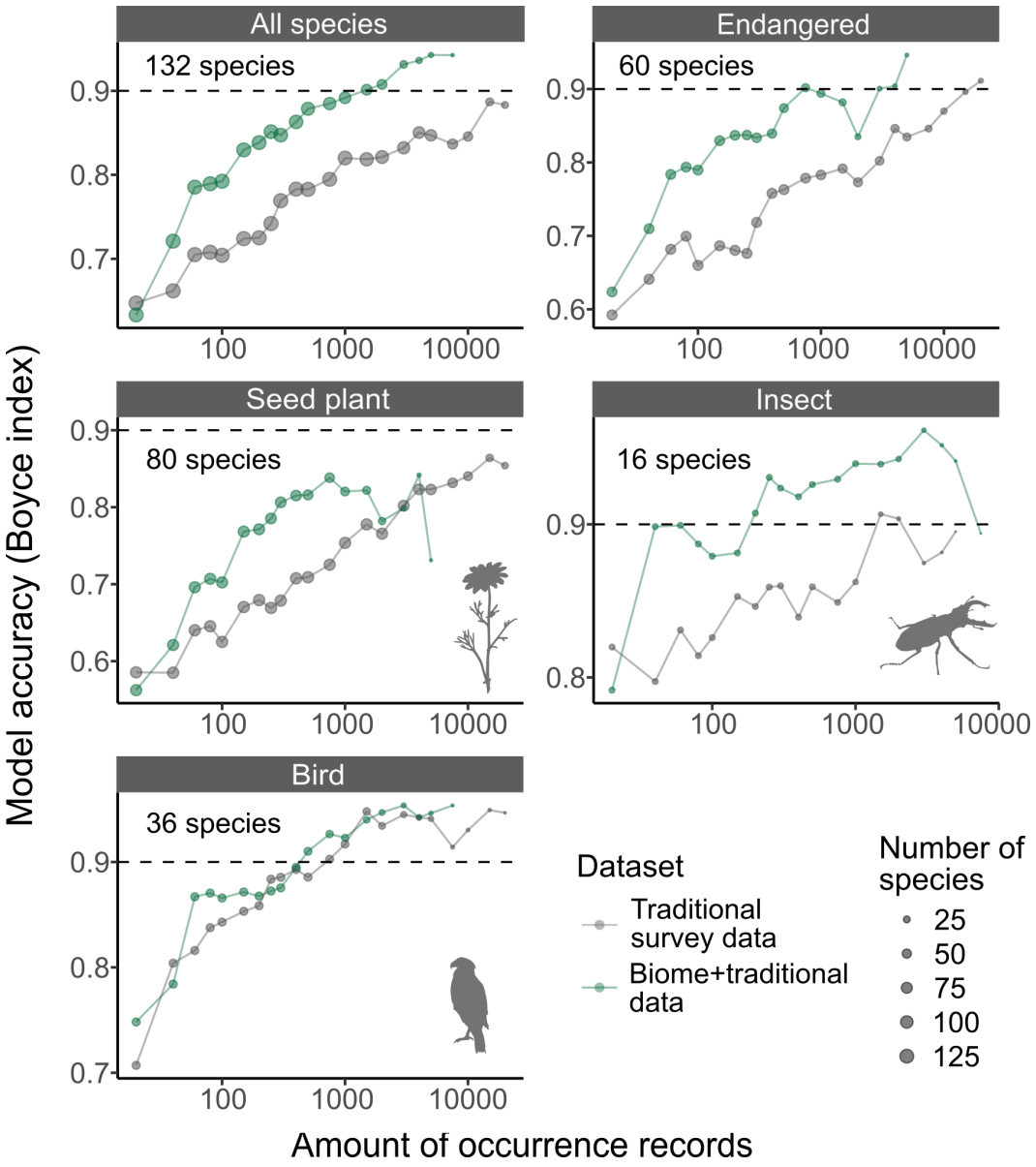

Fig. 3 | The accuracy of species distribution models. Accuracy of SDMs using Traditional survey data (grey dots and lines) and *Biome*+Traditional data (i.e. 50% of *Biome* data: green). Each SDM was performed with a specific dataset, species, and the amount of records. For each species and amount of 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.

Discussion

Biome: The amount and quality of submitted data

Since its launch in 2019, the app *Biome* accumulates species occurrence data rapidly (Fig. 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' behavior on the app (supplementary material S2). The implementation of automatic data filtering techniques is expected to enhance the quality of the data: in the case of insects and birds, which encompass numerous 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. The app's posting flow should be revised to encourage users to label their records when documenting non-wild individuals.

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: [25]). During the invasive plants survey in the US, the reports by non-professional volunteers were 72% correct [26]. 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 citizen science app, the data quality of *Biome* is decent. Yet, rare species generally showed lower identification accuracy, which would require identification by experts.

Species distribution modeling

The inclusion of *Biome* data resulted in improved accuracy of SDMs (Fig. 3). The most accurate model predictions were obtained when the training data consisted of 50-70% *Biome* data (supplementary material S3), highlighting the necessity of incorporating both

traditional surveys and citizen observations for a comprehensive understanding of species distributions [23,27,28].

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 (Fig. 2C). A balanced distribution along with the natural-urbal gradient is noteworthy because citizen science data is typically biased towards human population centers [19,20]. 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 nine 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 (Fig. 3) [29,30]. 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 (supplementary material S3) 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 citizen science data on SDMs has primarily been investigated using birds, with a limited focus on plants [9]. In our investigation, we observed that the incorporating *Biome* data improved SDM accuracy for seed plants and insects, while the impact on birds remained unclear (Fig. 3). This ambiguity is likely because citizen science 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 (Fig. 3) and Traditional survey data alone (Fig. S4). 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 citizen science data and its impact on SDMs across different taxa, it is important to recognize that data quality can vary among taxa.

Importantly, SDMs for endangered species, which often suffer from data deficit [30,31], became accurate in a much fewer amount of records by blending *Biome* data (Fig. 3). Specifically, a threshold of >0.9 Boyce index could be reached with only around 300 records when using *Biome* data, whereas over 6 times of data is required when using Traditional survey data only. This finding highlights the importance of citizen science data not only for monitoring the dynamics of endangered species [4,32] but also for modeling purposes. Considering the rapid accumulation of *Biome* data, *Biome* data would make a significant contribution to the more effective distribution modeling 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 (Fig. 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., unfavorable conditions for *Biome*+Traditional data SDMs), the accuracy of SDMs derived from *Biome*+Traditional and Traditional survey data did not differ (supplementary material S4). 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 [15,33]. 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 citizens, we can now estimate distributions of many terrestrial species across the Japanese archipelago. Estimated distributions will be useful in selecting new protected areas or areas with OECMs (Other Effective area-based Conservation Measures: allowing a wider range of landuse 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 [34] or OECMs [35]. Further, estimated distributions can be used as input for spatial conservation prioritization tools (e.g. Marxan [36]).

The rapid accumulation of data from various locations throughout Japan enables early detection of range expansions in invasive species (Sakai et al., in prep) and changes in the distribution patterns of native species [37]. Citizen science is increasingly recognized as a powerful tool for detecting distribution changes [38,39]. This wealth of rapidly accumulating data also opens up opportunities for conducting time-series analysis, which may contribute to forecasting ecosystem dynamics such as population dynamics, stability, and phenology [40,41].

For financial disclosures, companies will assess how their activities rely on ecosystem services and their opportunities for protecting/recovering nature [1]. By incorporating taxon-specific ecosystem services, multifaceted ecosystem services can be preliminarily screened. 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 [16,17], 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.

Supporting natural experiences for a wide range of people is also expected to contribute to changing people's minds towards nature. Through experiencing nature, people become familiar with it and subsequently make pro-nature decisions [42]. We believe that citizen science can significantly contribute to creating a sustainable society by fostering nature-positive awareness in society and providing data tools that enable effective action.

Methods

Occurrence record accumulation through 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 September 13, 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 [12]. The social networking function provides opportunities for communication among users, including non-experts [11]. 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 (supplementary material S2). This filtered *Biome* data is used in the subsequent investigations.

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 (see also Fig. S5 for the flowchart of selecting records to be checked). We defined rare species as those with fewer 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 seed plant and insect species which account for the majority of *Biome* data (Fig. 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 documented wild individuals were manually reidentified by experts with taxonomic knowledge (Fig. S5). 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 if the identification was correct at genus and family levels.

Species distribution models

Modeling

We modeled distributions of terrestrial seed plants and animals at a scale of 1 x 1 km grid cell. To model species distributions from presence-only data, we used Maxent [18] via ENMeval 2.0 package [43] on R 4.1.3 [44]. As predictor variables, continuous environmental variables—land use, climate, landform, vegetation and geology (supplementary material S6)—were transformed into linear, quadratic and hinge feature classes to illustrate nonlinear associations between environments and species occurrence [45]. The regularization multiplier was set at 2.5.

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 (see Table S2 for a list of datasets). For the species analysed (Table S9), traditional survey data contains a negligible portion of citizen science data (5.5%) because GBIF contains citizen science data from iNaturalist and eBird. 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 (supplementary material S3).

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) [21,22]. 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 (supplementary material S7).

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 minimize spatial autocorrelation between training and test data, we set a spatial block for splitting data [46,47]. As the spatial block, we chose the central Japan region (latitude, 33.7˚–37.7˚ N; longitude, 136.2˚–137.6˚ E: Fig. S8) which covers various environments—alpine to coastal lowlands, metropolis to highly intact areas. Boyce index (BI) was used to measure model performance because it was designed to evaluate presence-only SDMs [48]. 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 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. Since *Biome* data might include misidentifications or records of non-wild individuals, we derived test data from *Biome* data only in cases where multiple *Biome* users had submitted records of the same species at the identical location (within a 1km grid cell, see supplementary material S8). 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. To reliably calculate BI, at least 50 occurrences should be needed in test data [29]. Thus, we used 132 species that have more than 50 occurrences in test data for calculating BI (Table S9).

Examining influences of blending Biome data on SDM accuracy

Given that the accuracy of SDMs is affected by the amount and quality of data [29,30,46], 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 analyzed 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 modeling 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 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 generalized 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 [49] and lmerTest [50], respectively.

Conflict of interest

KA and YN are employed by Biome Inc., of which the CEO is SF and the CTO is TG. SF and TG are inventors of the species-identification-AI-algorithm JPN patents 6590417 and US patents 11048969. MU and HN declare that they have no competing interests. All authors will not financially benefit directly from the publication of this paper.

Author contributions

SF, TG and KA conceived the study; YN performed the analyses with the orientation by MU and KA; HN checked species identifications; KA led the writing of manuscript with substantial inputs from MU. All authors approved the manuscript.

Acknowledgements

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; Dalelan Anderson for comments on the manuscript. We appreciate the invaluable contributions of *Biome* users.

References

1. TNFD. The TNFD Nature-Related Risk and Opportunity Management and Disclosure Framework—Beta v0.3. 2022. Available: https://framework.tnfd.global/wp-content/uploads/2022/11/TNFD_Executive_Summary_v0-3_B.pdf

2. Kobori H, Dickinson JL, Washitani I, Sakurai R, Amano T, Komatsu N, et al. Citizen science: a new approach to advance ecology, education, and conservation. Ecol Res. 2016;31: 1–19. doi:10.1007/s11284-015-1314-y

3. Wood C, Sullivan B, Iliff M, Fink D, Kelling S. eBird: Engaging Birders in Science and Conservation. PLOS Biol. 2011;9: e1001220. doi:10.1371/journal.pbio.1001220

4. Chandler M, See L, Copas K, Bonde AMZ, López BC, Danielsen F, et al. Contribution of citizen science towards international biodiversity monitoring. Biol Conserv. 2017;213: 280–294. doi:10.1016/j.biocon.2016.09.004

5. Kaplan Mintz K, Arazy O, Malkinson D. Multiple forms of engagement and motivation in ecological citizen science. Environ Educ Res. 2023;29: 27–44. doi:10.1080/13504622.2022.2120186

6. Herodotou C, Ismail N, I. Benavides Lahnstein A, Aristeidou M, Young AN, Johnson RF, et al. Young people in iNaturalist: a blended learning framework for biodiversity monitoring. Int J Sci Educ Part B. 2023;0: 1–28. doi:10.1080/21548455.2023.2217472

7. Bowser A, Hansen D, He Y, Boston C, Reid M, Gunnell L, et al. Using gamification to inspire new citizen science volunteers. Proceedings of the First International Conference on Gameful Design, Research, and Applications. New York, NY, USA: Association for Computing Machinery; 2013. pp. 18–25. doi:10.1145/2583008.2583011

8. Ponti M, Hillman T, Stankovic I. Science and Gamification: The Odd Couple? Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play. New York, NY, USA: Association for Computing Machinery; 2015. pp. 679–684. doi:10.1145/2793107.2810293

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

10. Steen VA, Elphick CS, Tingley MW. An evaluation of stringent filtering to improve species distribution models from citizen science data. Divers Distrib. 2019;25:

1857–1869. doi:10.1111/ddi.12985

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

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

13. Groom Q, Pernat N, Adriaens T, de Groot M, Jelaska SD, Marčiulynienė D, et al. Species interactions: next-level citizen science. Ecography. 2021;44: 1781–1789. doi:10.1111/ecog.05790

14. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006;190: 231–259. doi:10.1016/j.ecolmodel.2005.03.026

15. Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. Divers Distrib. 2011;17: 43–57. doi:10.1111/j.1472-4642.2010.00725.x

16. Porfirio LL, Harris RMB, Lefroy EC, Hugh S, Gould SF, Lee G, et al. Improving the Use of Species Distribution Models in Conservation Planning and Management under Climate Change. PLOS ONE. 2014;9: e113749. doi:10.1371/journal.pone.0113749

17. Urban MC, Bocedi G, Hendry AP, Mihoub J-B, Pe'er G, Singer A, et al. Improving the forecast for biodiversity under climate change. Science. 2016;353: aad8466. doi:10.1126/science.aad8466

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

19. Reddy S, Dávalos LM. Geographical sampling bias and its implications for conservation priorities in Africa. J Biogeogr. 2003;30: 1719–1727. doi:10.1046/j.1365-2699.2003.00946.x

20. Kendal D, Egerer M, Byrne JA, Jones PJ, Marsh P, Threlfall CG, et al. City-size bias in knowledge on the effects of urban nature on people and biodiversity. Environ Res Lett. 2020;15: 124035. doi:10.1088/1748-9326/abc5e4

21. Milanesi P, Mori E, Menchetti M. Observer-oriented approach improves species distribution models from citizen science data. Ecol Evol. 2020;10: 12104–12114. doi:10.1002/ece3.6832

22. Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl. 2009;19: 181–197. doi:10.1890/07-2153.1

23. Robinson OJ, Ruiz-Gutierrez Viviana, Reynolds MD, Golet GH, Strimas-Mackey M, Fink D. Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. Divers Distrib. 2020;26: 976–986. doi:10.1111/ddi.13068

24. Johnston A, Fink D, Hochachka WM, Kelling S. Estimates of observer expertise improve species distributions from citizen science data. Methods Ecol Evol. 2018;9: 88–97. doi:10.1111/2041-210X.12838

25. Hart AG, Bosley H, Hooper C, Perry J, Sellors-Moore J, Moore O, et al. Assessing the accuracy of free automated plant identification applications. People Nat. n/a. doi:10.1002/pan3.10460

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

27. Miller DAW, Pacifici K, Sanderlin JS, Reich BJ. The recent past and promising future for data integration methods to estimate species' distributions. Methods Ecol Evol. 2019;10: 22–37. doi:10.1111/2041-210X.13110

28. Pacifici K, Reich BJ, Miller DAW, Gardner B, Stauffer G, Singh S, et al. Integrating multiple data sources in species distribution modeling: a framework for data fusion*. Ecology. 2017;98: 840–850. doi:10.1002/ecy.1710

29. Stockwell DRB, Peterson AT. Effects of sample size on accuracy of species distribution models. Ecol Model. 2002;148: 1–13. doi:10.1016/S0304-3800(01)00388-X

30. Erickson KD, Smith AB. 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;2023: e06500. doi:10.1111/ecog.06500

31. Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, et al. Effects of sample size on the performance of species distribution models. Divers Distrib. 2008;14: 763–773. doi:10.1111/j.1472-4642.2008.00482.x

32. Zapponi L, Cini A, Bardiani M, Hardersen S, Maura M, Maurizi E, et al. Citizen science data as an efficient tool for mapping protected saproxylic beetles. Biol Conserv. 2017;208: 139–145. doi:10.1016/j.biocon.2016.04.035

33. Ward G, Hastie T, Barry S, Elith J, Leathwick JR. Presence-Only Data and the EM Algorithm. Biometrics. 2009;65: 554–563. doi:10.1111/j.1541-0420.2008.01116.x

34. Jones KR, Watson JEM, Possingham HP, Klein CJ. Incorporating climate change into spatial conservation prioritisation: A review. Biol Conserv. 2016;194: 121–130. doi:10.1016/j.biocon.2015.12.008

35. Shiono T, Kubota Y, Kusumoto B. Area-based conservation planning in Japan: The

importance of OECMs in the post-2020 Global Biodiversity Framework. Glob Ecol Conserv. 2021;30: e01783. doi:10.1016/j.gecco.2021.e01783

36. Ball IR, Possingham HP, Watts ME. Marxan and relatives: Software for spatial conservation prioritization. Spatial conservation prioritisation: Quantitative methods and computational tools. Oxford University Press; 2009. p. 320.

37. Biome Inc. The report of Climate Change Biosurvey in 2022. 2023. Available: https://ccbio.jp/

38. Wallace RD, Bargeron CT. Identifying invasive species in real time: early detection and distribution mapping system (EDDMapS) and other mapping tools. Invasive Species Glob Clim Change. 2014; 219–231. doi:10.1079/9781780641645.0219

39. Larson ER, Graham BM, Achury R, Coon JJ, Daniels MK, Gambrell DK, et al. From eDNA to citizen science: emerging tools for the early detection of invasive species. Front Ecol Environ. 2020;18: 194–202. doi:10.1002/fee.2162

40. Taylor SD, Meiners JM, Riemer K, Orr MC, White EP. Comparison of large-scale citizen science data and long-term study data for phenology modeling. Ecology. 2019;100: e02568. doi:10.1002/ecy.2568

41. Ushio M, Hsieh CH, Masuda R, Deyle ER, Ye H, Chang CW, et al. Fluctuating interaction network and time-varying stability of a natural fish community. Nature. 2018;554: 360–363. doi:10.1038/nature25504

42. Soga M, Gaston KJ. Nature benefit hypothesis: Direct experiences of nature predict self-reported pro-biodiversity behaviors. Conserv Lett. 2023;n/a: e12945. doi:10.1111/conl.12945

43. Kass JM, Muscarella R, Galante PJ, Bohl CL, Pinilla-Buitrago GE, Boria RA, et al. ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. Methods Ecol Evol. 2021;12: 1602–1608. doi:10.1111/2041-210X.13628

44. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing , Vienna, Austria. Vienna, Austria; 2021. Available: https://www.R-project.org

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

46. Araújo MB, Anderson RP, Márcia Barbosa A, Beale CM, Dormann CF, Early R, et al. Standards for distribution models in biodiversity assessments. Sci Adv. 2019;5: eaat4858. doi:10.1126/sciadv.aat4858

47. Santini L, Benítez-López A, Maiorano L, Čengić M, Huijbregts MAJ. Assessing the reliability of species distribution projections in climate change research. Divers Distrib. 2021;27: 1035–1050. doi:10.1111/ddi.13252

48. Hirzel AH, Le Lay G, Helfer V, Randin C, Guisan A. Evaluating the ability of habitat suitability models to predict species presences. Ecol Model. 2006;199: 142–152. doi:10.1016/j.ecolmodel.2006.05.017

49. Bates DM, Maechler M, Bolker B, Walker S. lme4: linear mixed-effects models using S4 classes. J Stat Softw. 2015;67: 1–48.

50. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: Tests in linear mixed effects models. J Stat Softw. 2017;82. doi:10.18637/jss.v082.i13