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

Ecosystem services, which derive in part from biological diversity, are a fundamental support for human society. However, human activities are causing harm to biodiversity, ultimately endangering these critical ecosystem services. Halting nature loss and mitigating these impacts necessitates comprehensive biodiversity distribution data, a requirement for implementing the Kunming-Montreal Global Biodiversity Framework. To efficiently collect species observations from the public, we launched the '*Biome*' mobile application in Japan. By employing species identification algorithms and gamification elements, the app has gathered >6M observations since its launch in 2019. However, community-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 the data influences the performance of SDMs. Species identification accuracy 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 our 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 the two data sources were 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. Providing a platform to accumulate community-sourced distribution data and improving data processing protocol will contribute to not only conserving natural ecosystems but also detecting species distribution changes and testing ecological theories.

KEY WORDS:

Community science, 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

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 & Biggs, 2005). 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. 30by30). As 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 Taskforce on Nature-related Financial Disclosures (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 urbanization 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 & Bargeron, 2014), exploring large-scale geographic variations in traits (Atsumi & 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 utilization 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 & Ackerly, 2009; Tilman et al., 2006; Ushio et al.,

2018), capitalizing on the spatio-temporally 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., 2023; 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 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 community-based survey platform using mobile applications and (ii) employing an advanced statistical modeling 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 & 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 (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 (Fujiki & 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 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.

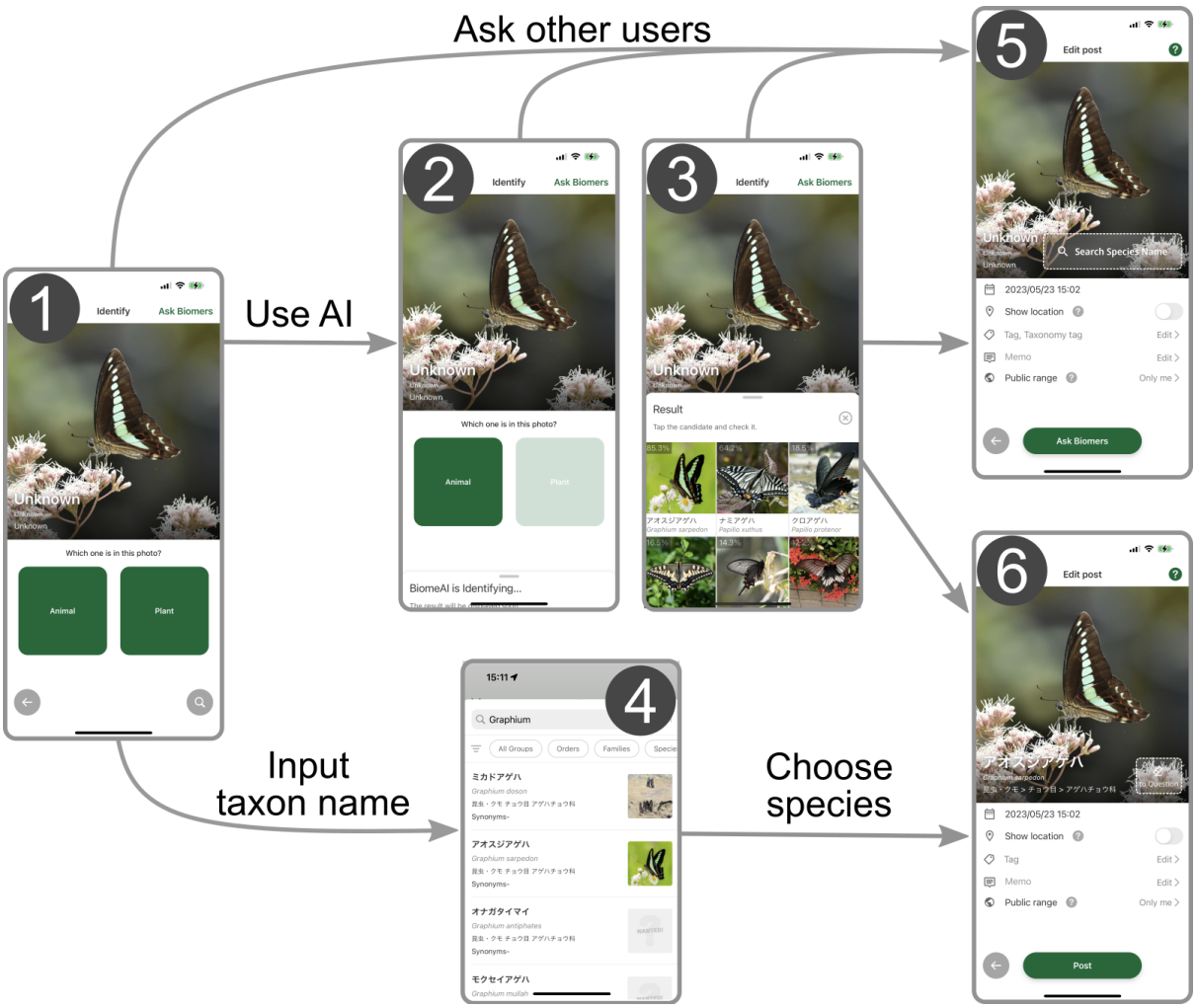

Fig 1. Workflow of submitting records to *Biome*. (1) Users can upload images that were taken by the smartphone camera or import existing images from the storage, including those imported from external devices. (2) Users select whether the image is about animals or plants to activate the species identification 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 (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 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 (Elith et al., 2011; Phillips & 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 & 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 utilizing 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 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 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 (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 (S1 File). 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).

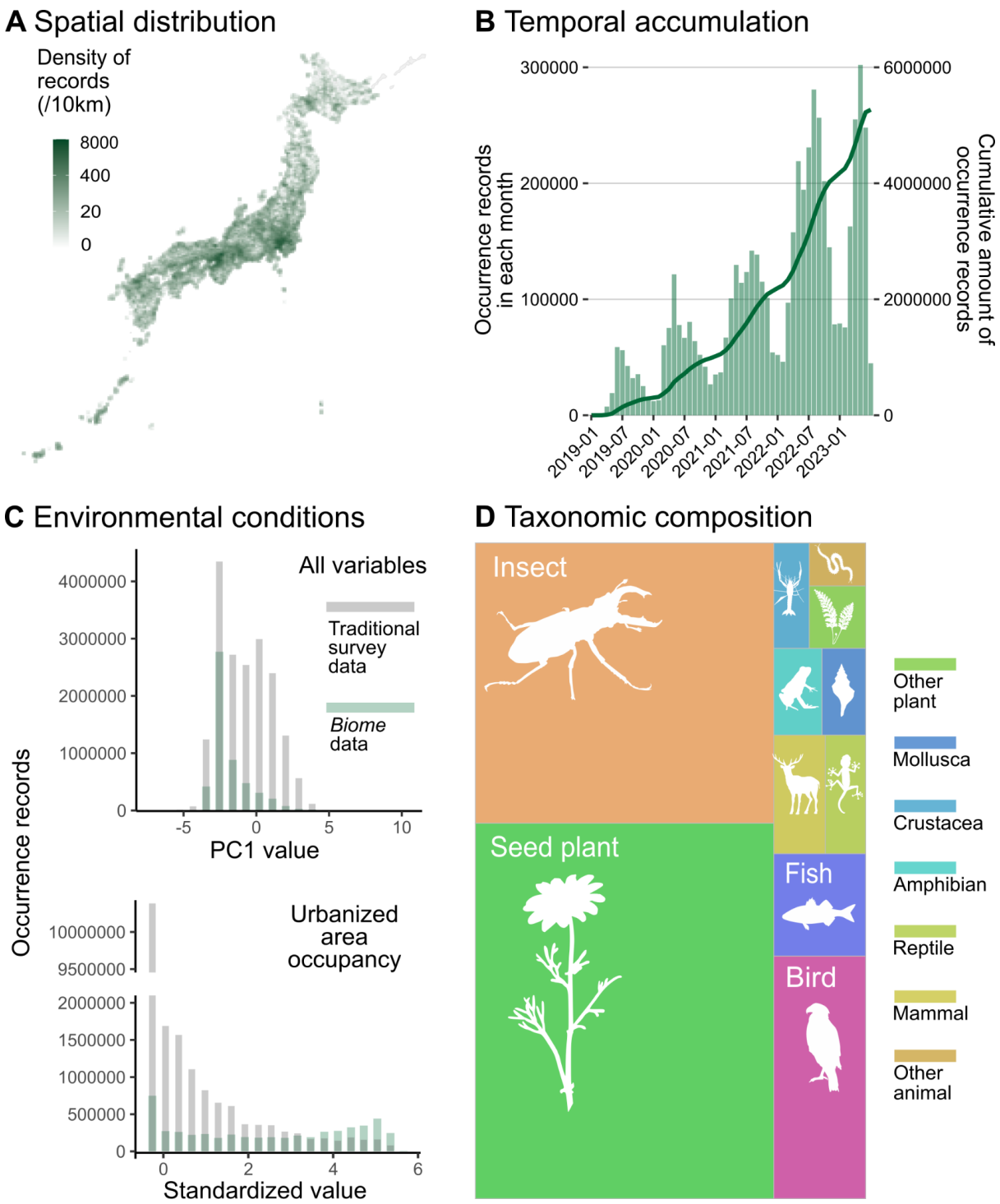

Fig 2. Description of data accumulated by *Biome*. Data distributions are shown based on all records submitted to *Biome* by 7 July 2023 (*N*=5,275,457). **A** Spatial distribution of records across Japan. **B** Accumulation of records through time. The barplot represents the number of records each month 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.

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.

Table 1. Data quality of *Biome*.

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 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.

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).

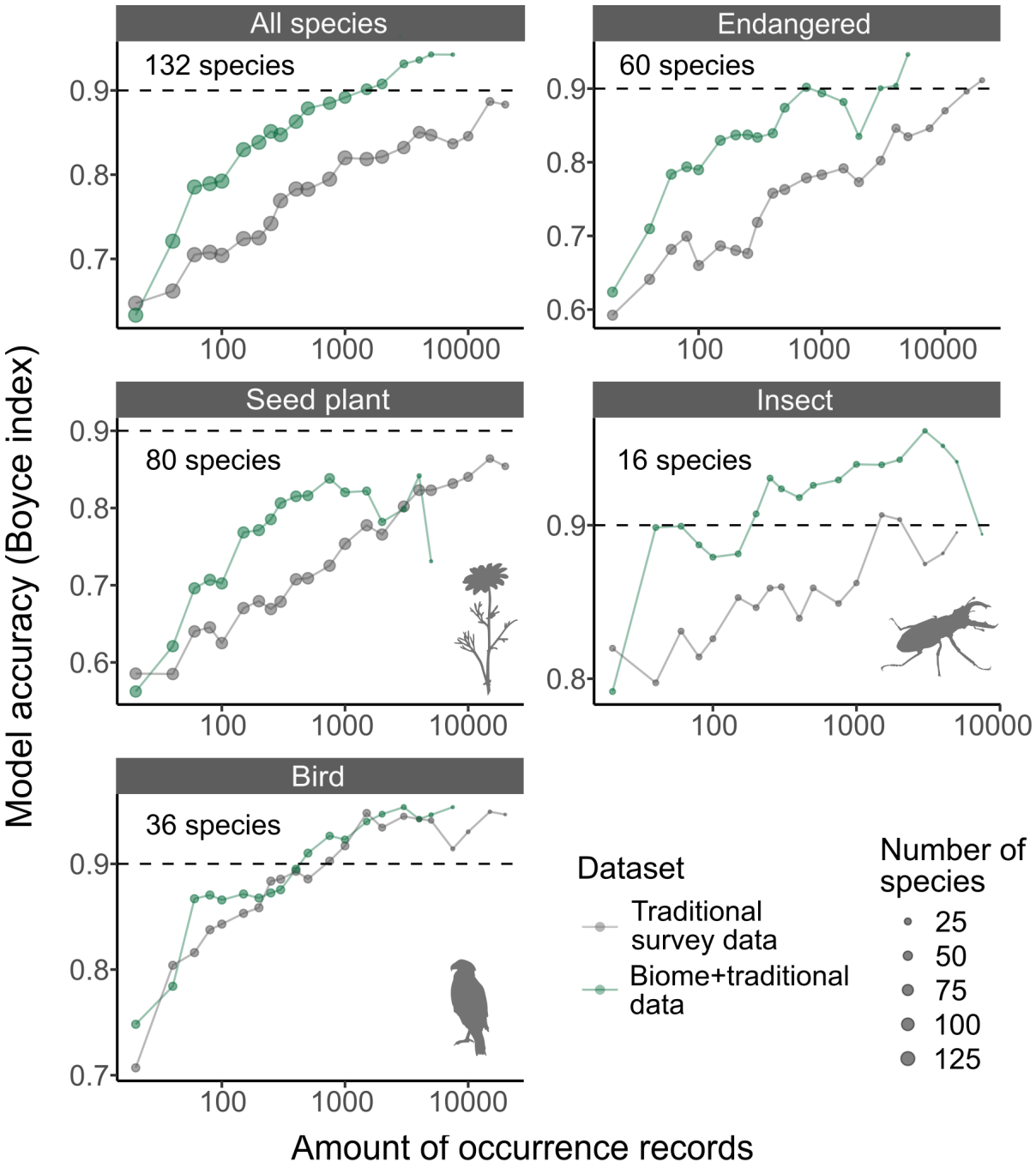

Fig 3. The accuracy of species distribution models. Accuracy of SDMs using Traditional survey data (grey dots and lines) and *Biome*+Traditional data (i.e. 50% of *Biome* data: green). Each SDM was performed with a specific dataset, species, and the amount of records. For each species and amount of records, 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 (S2 File). 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: (Hart et al., 2023)). During the invasive plants survey in the US, 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 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 (S3 File), 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 (Fig 2C). A balanced distribution along with the natural-urbal gradient is noteworthy because citizen science data is typically biased towards human population centers (Kendal et al., 2020; Reddy & 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 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) (Erickson & Smith, 2023; Stockwell & 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 (S3 File) 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 (Fig 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 (Fig 3) and Traditional survey data alone (S4 File). 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 recognize that data quality can vary among taxa.

Importantly, SDMs for endangered species, which often suffer from data deficit (Erickson & Smith, 2023; Wisz et al., 2008), 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 community-sourced data not only for monitoring the dynamics of endangered species (Chandler et al., 2017; Zapponi et al., 2017) 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 (S4 File). 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 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 (Jones et al., 2016) or OECMs (Shiono et al., 2021). Further, estimated distributions can be used as input for spatial conservation prioritization tools (e.g. Marxan (Ball et al., 2009)).

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 prep). 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 (Fig. 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 & Zohner, 2018; Visser & Gienapp, 2019). Moreover, *Biome* data is well-suited for assessing the effects of urbanization on ecosystems since it comprehensively spans both urban and natural habitats (Fig. 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 spatio-temporal 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. 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 & 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. Through experiencing nature, people become familiar with it and subsequently make pro-nature decisions (Soga & Gaston, 2023). We believe that community 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 (Koide et al., 2023). The social networking function provides opportunities for communication among users, including non-experts (Fujiki & 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 (S2 File). 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 S5 Fig 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 (S5 Fig). 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 (Phillips & Dudík, 2008) via ENMeval 2.0 package (Kass et al., 2021) on R 4.1.3 (R Core Team, 2021). As predictor variables, continuous environmental variables—land use, climate, landform, vegetation and geology (S6 File)—were transformed into linear, quadratic and hinge feature classes to illustrate nonlinear associations between environments and species occurrence (Phillips et al., 2017). 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 S2 File for a list of datasets and citation). For the species analysed (S9 Table), traditional survey data contains a negligible portion of community-sourced data (5.5%) because GBIF contains community-sourced 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 (S3 File).

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 (S7 File).

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 (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: 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 (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 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, S8 File). 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 (S9 Table).

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 & Smith, 2023; Stockwell & 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 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 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017), respectively.

Acknowledgments

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

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

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*(4), 567–578. https://doi.org/10.1007/s11284-017-1466-z

Ball, I. R., Possingham, H. P., & Watts, M. E. (2009). Marxan and relatives: Software for spatial conservation prioritization. In *Spatial conservation prioritisation: Quantitative methods and computational tools* (p. 320). Oxford University Press.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using S4 classes. *Journal of Statistical Software*, *67*(1), 1–48.

Biome Inc. (2023). *The report of Climate Change Biosurvey in 2022*. https://ccbio.jp/

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). bioclim: The first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, *20*(1), 1–9. 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. *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, 18–25. https://doi.org/10.1145/2583008.2583011

Box, E. O. (1981). *Macroclimate and Plant Forms* (Vol. 1). Springer Netherlands. https://doi.org/10.1007/978-94-009-8680-0

Bury, T. M., Sujith, R. I., Pavithran, I., Scheffer, M., Lenton, T. M., Anand, M., & Bauch, C. T. (2021). Deep learning for early warning signals of tipping points. *Proceedings of the*

National Academy of Sciences of the United States of America, *118*(39). https://doi.org/10.1073/pnas.2106140118

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)*. The Secretariat of the Convention on Biological Diversity. https://www.cbd.int/article/2021-singapore-index

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

Cornwell, W. K., & Ackerly, D. D. (2009). Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California. *Ecological Monographs*, *79*(1), 109–126. https://doi.org/10.1890/07-1134.1

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011). Assessing citizen science data quality: An invasive species case study. *Conservation Letters*, *4*(6), 433–442. https://doi.org/10.1111/j.1755-263X.2011.00196.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Erickson, K. D., & Smith, A. B. (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*(6), e06500. https://doi.org/10.1111/ecog.06500

Feldman, M. J., Imbeau, L., Marchand, P., Mazerolle, M. J., Darveau, M., & Fenton, N. J. (2021). Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PLOS ONE*, *16*(3), e0234587. https://doi.org/10.1371/journal.pone.0234587

Fink, D., Johnston, A., Strimas-Mackey, M., Auer, T., Hochachka, W. M., Ligocki, S., Oldham Jaromczyk, L., Robinson, O., Wood, C., Kelling, S., & Rodewald, A. D. (2023). A Double machine learning trend model for citizen science data. *Methods in Ecology and Evolution*, *14*(9), 2435–2448. https://doi.org/10.1111/2041-210X.14186

Fuccillo Battle, K., Duhon, A., Vispo, C. R., Crimmins, T. M., Rosenstiel, T. N., Armstrong-Davies, L. L., & de Rivera, C. E. (2022). Citizen science across two centuries reveals phenological change among plant species and functional groups in the Northeastern US. *Journal of Ecology*, *110*(8), 1757–1774. 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. *Japanese Journal of Ecology*, *71*(2), 85–90. https://doi.org/10.18960/seitai.71.2_85

Gonzalez, A., Chase, J. M., & O'Connor, M. I. (2023). A framework for the detection and attribution of biodiversity change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1881), 20220182. https://doi.org/10.1098/rstb.2022.0182

Groom, Q., Pernat, N., Adriaens, T., de Groot, M., Jelaska, S. D., Marčiulynienė, D., Martinou, A. F., Skuhrovec, J., Tricarico, E., Wit, E. C., & Roy, H. E. (2021). Species interactions: Next-level citizen science. *Ecography*, *44*(12), 1781–1789. https://doi.org/10.1111/ecog.05790

Hart, A. G., Bosley, H., Hooper, C., Perry, J., Sellors-Moore, J., Moore, O., & Goodenough, A. E. (2023). Assessing the accuracy of free automated plant identification applications. *People and Nature*, *5*(3), 929–937. https://doi.org/10.1002/pan3.10460

Herodotou, C., Ismail, N., I. Benavides Lahnstein, A., Aristeidou, M., Young, A. N., Johnson, R. F., Higgins, L. M., Ghadiri Khanaposhtani, M., Robinson, L. D., & Ballard, H. L. (2023). Young people in iNaturalist: A blended learning framework for biodiversity monitoring. *International Journal of Science Education, Part B*, *0*(0), 1–28. https://doi.org/10.1080/21548455.2023.2217472

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of

habitat suitability models to predict species presences. *Ecological Modelling*, *199*(2), 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Hutchinson, E. G. (1957). Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, *22*.

IPBES. (2019). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. IPBES secretariat. https://zenodo.org/records/6417333

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, *9*(1), 88–97. https://doi.org/10.1111/2041-210X.12838

Jones, K. R., Watson, J. E. M., Possingham, H. P., & Klein, C. J. (2016). Incorporating climate change into spatial conservation prioritisation: A review. *Biological Conservation*, *194*, 121–130. 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*(1), 27–44. https://doi.org/10.1080/13504622.2022.2120186

Kass, J. M., Muscarella, R., Galante, P. J., Bohl, C. L., Pinilla-Buitrago, G. E., Boria, R. A., Soley-Guardia, M., & Anderson, R. P. (2021). ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. *Methods in Ecology and Evolution*, *12*(9), 1602–1608. https://doi.org/10.1111/2041-210X.13628

Kendal, D., Egerer, M., Byrne, J. A., Jones, P. J., Marsh, P., Threlfall, C. G., Allegretto, G., Kaplan, H., Nguyen, H. K. D., Pearson, S., Wright, A., & Flies, E. J. (2020). City-size bias in knowledge on the effects of urban nature on people and biodiversity. *Environmental Research Letters*, *15*(12), 124035. https://doi.org/10.1088/1748-9326/abc5e4

Keough, H. L., & Blahna, D. J. (2006). Achieving Integrative, Collaborative Ecosystem Management. *Conservation Biology*, *20*(5), 1373–1382.

https://doi.org/10.1111/j.1523-1739.2006.00445.x

Kindt, R. (2023). TreeGOER: A database with globally observed environmental ranges for 48,129 tree species. *Global Change Biology*. https://doi.org/10.1111/gcb.16914

Klinger, Y. P., Eckstein, R. L., & Kleinebecker, T. (2023). iPhenology: Using open-access citizen science photos to track phenology at continental scale. *Methods in Ecology and Evolution*, *14*(6), 1424–1431. https://doi.org/10.1111/2041-210X.14114

Kobori, H., Dickinson, J. L., Washitani, I., Sakurai, R., Amano, T., Komatsu, N., Kitamura, W., Takagawa, S., Koyama, K., Ogawara, T., & Miller-Rushing, A. J. (2016). Citizen science: A new approach to advance ecology, education, and conservation. *Ecological Research*, *31*(1), 1–19. 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. *Japanese Journal of Conservation Ecology*. https://doi.org/10.18960/hozen.2217

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Larson, E. R., Graham, B. M., Achury, R., Coon, J. J., Daniels, M. K., Gambrell, D. K., Jonasen, K. L., King, G. D., LaRacuente, N., Perrin-Stowe, T. I., Reed, E. M., Rice, C. J., Ruzi, S. A., Thairu, M. W., Wilson, J. C., & Suarez, A. V. (2020). From eDNA to citizen science: Emerging tools for the early detection of invasive species. *Frontiers in Ecology and the Environment*, *18*(4), 194–202. https://doi.org/10.1002/fee.2162

Laubmeier, A. N., Cazelles, B., Cuddington, K., Erickson, K. D., Fortin, M. J., Ogle, K., Wikle, C. K., Zhu, K., & Zipkin, E. F. (2020). Ecological Dynamics: Integrating Empirical, Statistical, and Analytical Methods. *Trends in Ecology and Evolution*, *35*(12), 1090–1099. https://doi.org/10.1016/j.tree.2020.08.006

Leighton, G. R. M., Hugo, P. S., 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*(9), 1060–1070. 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*(9), Article 9. https://doi.org/10.1038/s41559-023-02113-w

Loh, J., Green, R. E., 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: Biological Sciences*, *360*(1454), 289–295. 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*(21), 12104–12114. https://doi.org/10.1002/ece3.6832

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. 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. https://doi.org/10.3897/mbmg.6.80444

Mori, A. S., Suzuki, K. F., 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: Biological Sciences*, *378*(1881), 20220192. https://doi.org/10.1098/rstb.2022.0192

Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., De Palma, A., Ferrier, S., Hill, S. L. L., Hoskins, A. J., Lysenko, I., Phillips, H. R. P., Burton, V. J., Chng, C. W. T., Emerson,

S., Gao, D., Hale, G. P., Hutton, J., Jung, M., Sanchez-Ortiz, K., Simmons, B. I., … Purvis, A. (2016). Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science*, *353*(6296), 291–288. https://doi.org/10.1126/science.aaf2201

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., & Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion*. *Ecology*, *98*(3), 840–850. https://doi.org/10.1002/ecy.1710

Pennekamp, F., Iles, A. C., Garland, J., Brennan, G., Brose, U., Gaedke, U., Jacob, U., Kratina, P., Matthews, B., Munch, S., Novak, M., Palamara, G. M., Rall, B. C., Rosenbaum, B., Tabi, A., Ward, C., Williams, R., Ye, H., & Petchey, O. L. (2019). The intrinsic predictability of ecological time series and its potential to guide forecasting. *Ecological Monographs*, *89*(2), e01359. https://doi.org/10.1002/ecm.1359

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography*, *40*(7), 887–893. https://doi.org/10.1111/ecog.03049

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., 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*(1), 181–197. https://doi.org/10.1890/07-2153.1

Pocock, M. J. O., Chandler, M., Bonney, R., Thornhill, I., Albin, A., August, T., Bachman, S., Brown, P. M. J., Cunha, D. G. F., Grez, A., Jackson, C., Peters, M., Rabarijaon, N. R.,

Roy, H. E., Zaviezo, T., & Danielsen, F. (2018). Chapter Six—A Vision for Global Biodiversity Monitoring With Citizen Science. In D. A. Bohan, A. J. Dumbrell, G. Woodward, & M. Jackson (Eds.), *Advances in Ecological Research* (Vol. 59, pp. 169–223). Academic Press. https://doi.org/10.1016/bs.aecr.2018.06.003

Pocock, M. J. O., Tweddle, J. C., Savage, J., Robinson, L. D., & Roy, H. E. (2017). The diversity and evolution of ecological and environmental citizen science. *PLoS ONE*, *12*(4), 1–17. https://doi.org/10.1371/journal.pone.0172579

Ponti, M., Hillman, T., & Stankovic, I. (2015). Science and Gamification: The Odd Couple? *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 679–684. https://doi.org/10.1145/2793107.2810293

Porfirio, L. L., Harris, R. M. B., Lefroy, E. C., Hugh, S., Gould, S. F., Lee, G., Bindoff, N. L., & Mackey, B. (2014). Improving the Use of Species Distribution Models in Conservation Planning and Management under Climate Change. *PLOS ONE*, *9*(11), e113749. https://doi.org/10.1371/journal.pone.0113749

R Core Team. (2021). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing , Vienna, Austria.* [Computer software]. https://www.R-project.org

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*(11), 1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x

Renner, S. S., & Zohner, C. M. (2018). Climate Change and Phenological Mismatch in Trophic Interactions Among Plants, Insects, and Vertebrates. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 165–182. https://doi.org/10.1146/annurev-ecolsys-110617-062535

Robinson, O. J., Ruiz-Gutierrez, Viviana., Reynolds, M. D., Golet, G. H., 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*(8), 976–986. https://doi.org/10.1111/ddi.13068

Roy, H. E., Pauchard, A., Stoett, P., Renard Truong, T., Bacher, S., Galil, B. S., Hulme, P. E., Ikeda, T., Sankaran, K. V., McGeoch, M. A., Meyerson, L. A., Nuñez, M. A., Ordonez, A., Rahlao, S. J., Schwindt, E., Seebens, H., Sheppard, A. W., & Vandvik, V. (2023). *IPBES Invasive Alien Species Assessment: Summary for Policymakers*. Zenodo. https://doi.org/10.5281/zenodo.8314303

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., & Huijbregts, M. A. J. (2021). Assessing the reliability of species distribution projections in climate change research. *Diversity and Distributions*, *27*(6), 1035–1050. https://doi.org/10.1111/ddi.13252

Scholes, R. J., & Biggs, R. (2005). A biodiversity intactness index. *Nature*, *434*(7029), 45–49. https://doi.org/10.1038/nature03289

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*(July), e01783. https://doi.org/10.1016/j.gecco.2021.e01783

Soga, M., & Gaston, K. J. (2023). Nature benefit hypothesis: Direct experiences of nature predict self-reported pro-biodiversity behaviors. *Conservation Letters*, *n/a*(n/a), e12945. https://doi.org/10.1111/conl.12945

Steen, V. A., Elphick, C. S., & Tingley, M. W. (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions*, *25*(12), 1857–1869. https://doi.org/10.1111/ddi.12985

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*(1), 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Tilman, D., Reich, P. B., & Knops, J. M. H. (2006). Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature*, *441*(7093), Article 7093. https://doi.org/10.1038/nature04742

TNFD. (2023). *Taskforce on Nature-related Financial Disclosures (TNFD) Recommendations*

version 1.0. https://tnfd.global/publication/recommendations-of-the-taskforce-on-nature-related-financial-disclosures/

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J.-B., Pe'er, G., Singer, A., Bridle, J. R., Crozier, L. G., De Meester, L., Godsoe, W., Gonzalez, A., Hellmann, J. J., Holt, R. D., Huth, A., Johst, K., Krug, C. B., Leadley, P. W., Palmer, S. C. F., Pantel, J. H., … Travis, J. M. J. (2016). Improving the forecast for biodiversity under climate change. *Science*, *353*(6304), aad8466. https://doi.org/10.1126/science.aad8466

Ushio, M., Hsieh, C. H., Masuda, R., Deyle, E. R., Ye, H., Chang, C. W., Sugihara, G., & Kondoh, M. (2018). Fluctuating interaction network and time-varying stability of a natural fish community. *Nature*, *554*(7692), 360–363. https://doi.org/10.1038/nature25504

Visser, M. E., & Gienapp, P. (2019). Evolutionary and demographic consequences of phenological mismatches. *Nature Ecology and Evolution*, *12*. https://doi.org/10.1038/s41559-019-0880-8

Wallace, R. D., & Bargeron, C. T. (2014). Identifying invasive species in real time: Early detection and distribution mapping system (EDDMapS) and other mapping tools. *Invasive Species and Global Climate Change*, 219–231. https://doi.org/10.1079/9781780641645.0219

Ward, G., Hastie, T., Barry, S., Elith, J., & Leathwick, J. R. (2009). Presence-Only Data and the EM Algorithm. *Biometrics*, *65*(2), 554–563. https://doi.org/10.1111/j.1541-0420.2008.01116.x

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & Group, N. P. S. D. W. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. 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*(12), e1001220.

https://doi.org/10.1371/journal.pbio.1001220

Zapponi, L., Cini, A., Bardiani, M., Hardersen, S., Maura, M., Maurizi, E., Redolfi De Zan, L., Audisio, P., Bologna, M. A., Carpaneto, G. M., Roversi, P. F., 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. https://doi.org/10.1016/j.biocon.2016.04.035

Zhang, W., Sheldon, B. C., Grenyer, R., & Gaston, K. J. (2021). Habitat change and biased sampling influence estimation of diversity trends. *Current Biology*, *31*(16), 3656-3662.e3. https://doi.org/10.1016/j.cub.2021.05.066

Supporting information

S1 File. Distributions of occurrence records along with environmental variables.

S2 File. Occurrence records compilation.

S3 File. Determine the best blend of Traditional survey and *Biome* data.

S4 File. Comparing SDM accuracy using only Traditional survey test data.

S5 Fig. The workflow of checking records from the *Biome*. The records passed auto-filters were assessed (see S2 File).

S6 File. Predictor variables for SDMs.

S7 File. Selecting pseudo-absence grid cells for SDM.

S8 File. Preparing SDM test data.

S9 Table. List of species for constructed species distribution models.