DivShift: Exploring Domain-Specific Distribution Shift in Large-Scale, Volunteer-Collected Biodiversity Datasets

Elena Sierra[1,2,3*], **Lauren E. Gillespie**[1,3,4,5*], **Salim Soltani**[2], **Moises Exposito-Alonso**[5,6], **Teja Kattenborn**[2]

[1]Stanford University Department of Computer Science
[2] Chair of Sensor-based Geoinformatics (geosense), University of Freiburg
[3]Carnegie Science Plant Biology
[4]Federal University of Minas Gerais Department of Vegetation Biology
[5]University of California, Berkeley Department of Integrative Biology
[6]Howard Hughes Medical Institute
esierra@cs.stanford.edu, gillespl@cs.stanford.edu

Abstract

Large-scale, volunteer-collected datasets of community-identified natural world imagery like iNaturalist have enabled marked performance gains for fine-grained visual classification of species using machine learning methods. However, such data—sometimes referred to as citizen science data—are opportunistic and lack a structured sampling strategy. This volunteer-collected biodiversity data contains geographic, temporal, taxonomic, observers, and sociopolitical biases that can have significant effects on biodiversity model performance, but whose impacts are unclear for fine-grained species recognition performance. Here we introduce Diversity Shift (DivShift), a framework for quantifying the effects of domain-specific distribution shifts on machine learning model performance. To diagnose the performance effects of biases specific to volunteer-collected biodiversity data, we also introduce DivShift - North American West Coast (DivShift-NAWC), a curated dataset of almost 7.5 million iNaturalist images across the western coast of North America partitioned across five types of expert-verified bias. We compare species recognition performance across these bias partitions using a diverse variety of species- and ecosystem-focused accuracy metrics. We observe that these biases confound model performance less than expected from the underlying label distribution shift, and that more data leads to better model performance but the magnitude of these improvements are bias-specific. These findings imply that while the structure within natural world images provides generalization improvements for biodiversity monitoring tasks, the biases present in volunteer-collected biodiversity data can also affect model performance; thus these models should be used with caution in downstream biodiversity monitoring tasks.

Code — github.com/moiexpositoalonsolab/DivShift
Dataset —
huggingface.co/datasets/elenagsierra/DivShift-NAWC
Extended version — arxiv.org/abs/2410.19816

*These authors contributed equally.

Introduction

Monitoring biodiversity is vital for understanding the state of the natural world, and frequent and accurate monitoring via automated tools is crucial for guiding decisions to protect the world's ecosystems. Building machine learning tools for this automated monitoring requires large volumes of natural world imagery. In recent years, participatory science applications that enable public volunteers to observe, share, and help identify species in their natural environments have seen a surge in popularity.

These scientific efforts on the part of the general public now mean that large-scale biodiversity image datasets are readily available with the number of observations rapidly approaching the scale of internet-scale image datasets (Schuhmann et al. 2022). With these finely labeled volunteer-collected datasets, computer vision models have shown impressive improvement in a variety of biodiversity monitoring-related machine learning tasks, including species recognition, species distribution modeling, novel species identification, and visual question answering (Van Horn et al. 2021; Huynh et al. 2024; Teng et al. 2024; Sastry et al. 2024; Garcin et al. 2021; Goëau, Bonnet, and Joly 2023; Gillespie, Ruffley, and Exposito-Alonso 2024; Stevens et al. 2024). However, the volume of these opportunistic volunteer records comes at a cost: as these observations become easier for the public to collect, sampling becomes unstructured, and injects a variety of biases into these data (Arazy and Malkinson 2021; Pernat et al. 2021; Geldmann et al. 2016; Isaac and Pocock 2015; Di Cecco et al. 2021; Boakes et al. 2010; Dimson and Gillespie 2023). These biases mean these volunteer-collected data do not reflect the state of the world's biodiversity in many aspects and present challenges for the general uptake of these unstructured, opportunistic data for biodiversity monitoring (Backstrom et al. 2024; Cooper 2014; Callaghan et al. 2021; Kishimoto and Kobori 2021; Johnston, Matechou, and Dennis 2023; Botts, Erasmus, and Alexander 2011; Deacon, Govender, and Samways 2023; Wolf et al. 2022).

To help quantify the effects of these biases on model performance, we introduce Diversity Shift (DivShift), a frame-

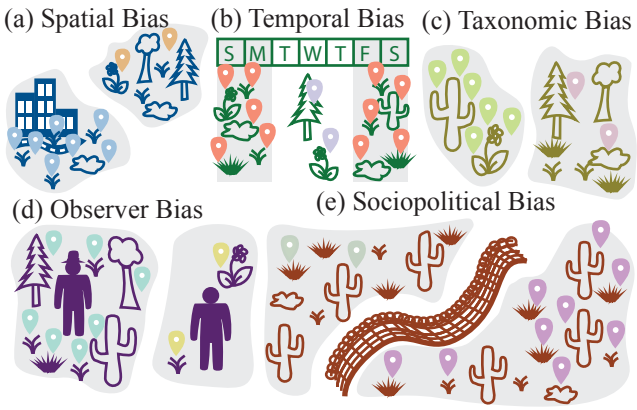

Figure 1: Biases present in biodiversity data include **(a)** spatial bias, **(b)** temporal bias, **(c)** taxonomic bias, **(d)** observer behavior bias, and **(e)** sociopolitical bias.

work for linking domain shift-driven computer vision model performance disparities to biases manifest in large-scale, volunteer-collected biodiversity datasets. We specifically focus on known biases in these data present across space, time, taxonomy, observers, and sociopolitical boundaries (Fig. 1). We also introduce a new public biodiversity imagery dataset DivShift-North American West Coast (DivShift-NAWC), a dataset of nearly 7.5 million observations of over 7,500 plant species across the North American West Coast designed to help quantify these disparities in a controlled case study. Performance varies both positively and negatively under these five different domain shifts. Synthesizing these quantitative results with previous work, we suggest recommendations for downstream use of computer vision models trained on these volunteer-collected biodiversity data.

Related Works

Large-Scale Natural World Imagery Datasets

Large-scale natural world imagery datasets for training computer vision models for biodiversity monitoring tasks span a variety of modalities, including handheld phone images, high-quality archival and herbaria images, long-distance camera imagery, terrestrial camera traps, ocean sonar cameras, google street view imagery and remote sensing imagery (Van Horn et al. 2021; Huynh et al. 2024; Sastry et al. 2024; Garcin et al. 2021; Goëau, Bonnet, and Joly 2023; Van Horn et al. 2018; Stevens et al. 2024; de Lutio et al. 2022; Wang et al. 2023; Kshitiz et al. 2024; Beery et al. 2021; Swanson et al. 2015; Kay et al. 2022; Beery et al. 2022; Lee et al. 2024; Cole et al. 2020; Gillespie, Ruffley, and Exposito-Alonso 2024; Teng et al. 2024; Huynh et al. 2024; Weinstein et al. 2021; Soltani et al. 2024). The app iNaturalist—where users can upload photos of species in their natural environments, identify them, and help identify other observations—has especially seen significant and sustained growth year over year, and now reaches over 40 million observations with nearly 300,000 species observed annually (Di Cecco et al. 2021; Backstrom et al. 2024; Dimson and Gillespie 2023; iNaturalist 2023).

Biases in Volunteer-Collected Biodiversity Datasets

Collections of large-scale volunteer datasets are subject to social and ecological filters, which inject many types of bias into biodiversity datasets (Carlen et al. 2024; Isaac and Pocock 2015; Di Cecco et al. 2021; Isaac et al. 2014). In this work, we focus on five kinds of biases common to volunteer-collected biodiversity datasets: spatial, temporal, taxonomic, observer, and sociopolitical (Fig. 1). **Spatial bias** includes observer preferences to sampling easy-to-access green spaces in urban or touristic areas (Gratzer and Brodschneider 2021; McGoff et al. 2017; Backstrom et al. 2024; Dimson and Gillespie 2023). **Temporal bias** includes a skew towards more observations on weekends when observers are free from work and during seasons with pleasant weather and attractive appearances of plants (Sweet, Rödl, and Weisser 2022; Sánchez-Clavijo et al. 2021; Crimmins et al. 2021; iNaturalist 2023; Courter et al. 2013; Cooper 2014). **Taxonomic bias** includes observer preference for identifying larger, exotic, or charismatic species and species that are easy to identify (Aristeidou et al. 2021; Unger et al. 2021; Ward 2014; Mair and Ruete 2016; McMullin and Allen 2022; Hochmair et al. 2020; Boakes et al. 2016; Deacon, Govender, and Samways 2023; Callaghan et al. 2021; Stoudt, Goldstein, and de Valpine 2022). **Observer bias** manifests as a small but dedicated group of users that tend to observe more species in more diverse habitats (Van Eupen et al. 2021; Milanesi, Mori, and Menchetti 2020; Boakes et al. 2016; Rosenblatt et al. 2022). Lastly, **sociopolitical bias** in who has access to the resources, time, and areas to collect biodiversity observations includes a skew towards whiter and wealthier regions (Blake, Rhanor, and Pajic 2020; Ellis-Soto, Chapman, and Locke 2023; Mahmoudi et al. 2022; Chen et al. 2022; Burgess et al. 2017; Cooper et al. 2023; Soleri et al. 2016; Pandya 2012; Mac Domhnaill, Lyons, and Nolan 2020; Pateman, Dyke, and West 2021). While these biases are well-documented, their performance effects on fine-grained visual classification of species is not well understood.

The DivShift Framework

In order to quantify the performance effects of bias present in volunteer-collected biodiversity datasets, we propose Diversity Shift (DivShift), a new framework that casts these domain-specific biases as distribution shifts (Fig. 2). The DivShift framework quantifies the effect of bias by measuring the in-domain versus out-of-domain model performance of any two partitions of a dataset and further compares these changes to the underlying label distribution shift present across these two partitions.

Given any finite labeled dataset D consisting of pairs of inputs x and labels y, we first define partition P_A as any subset of D such that $P_A \subset D$. We similarly define a second partition P_B such that $P_B \subset D$ and $P_B \bigcap P_A = \emptyset$. These partitions are then each further split into two sub-partitions P_{Atrain} and P_{Atest} where again $P_{Atest} \bigcap P_{Atrain} = \emptyset$ (Fig. 2a). If the sampling process $\overset{S}{\sim}$ for P_A and P_B is identical,

(a) DivShift Framework (b) Measuring DivShift with JSD

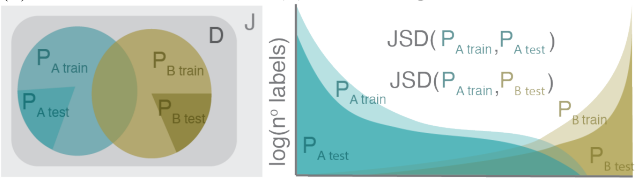

Figure 2: The Diversity Shift (DivShift) Framework **(a)** quantifies impacts of domain-specific biases by first partitioning data into partitions P_A and P_B using expert-verified types of bias. Bias impacts are then quantified by measuring the accuracy of models trained on P_{Atrain} using P_{Atest} and P_{Btest} which is further compared to **(b)** the distribution shift between labels in P_{Atrain} to labels in P_{Atest} and P_{Btest} using the Jensen-Shannon Distance (JSD).

P_A and P_B are considered *in-domain*. However, when the sampling processes for P_A and P_B are biased (e.g. more observers selectively uploading a few species in a certain area) then $P_A \overset{S_a}{\sim} J(x, y)$ and $P_B \overset{S_b}{\sim} J(x, y)$ will be out-of-distribution relative to each other, even if the underlying joint distribution $J(x, y)$—or the true representative biodiversity in an area—is the same. Therefore, any model trained on P_{Atrain} will exhibit a changed performance on P_{Btest} relative to P_{Atest} under these conditions.

To quantify this performance change, we first assume that the distribution of labels y in P_A and P_B can be used to estimate their joint distributions and summarily use these labels to estimate the underlying distribution shift between these partitions. Namely, we measure the Jensen-Shannon Divergence (JSD) between $P_{Atrain}(y)$ and $P_{Btest}(y)$, specifically using a base 2 log to ensure the distance is bound between 1 and 0 where 0 is perfectly aligned and 1 is perfectly misaligned distribution (Endres and Schindelin 2003) (Fig. 2b). We choose JSD for quantifying distribution shift as it is a bounded symmetric metric, which allows the comparison between changes in the magnitude of the JSD across partitions to changes in model performance. Furthermore, JSD is a well-established metric in the ecology literature for comparing biodiversity across sites.

While each P_{train} and P_{test} pair are uniformly sub-partitioned and sampled from the same distribution, the datasets are finite and the random sampling process is not truly random, meaning the label distributions between P_{train} and P_{test} will not perfectly match. To account for this noise in which observations are ultimately selected for each P_{train}, for each partition, the JSD between each paired P_{train} and P_{test} is subtracted from the estimates of distribution shift to other partitions. Specifically, the DivShift framework first measures the performance decrease between models trained on P_{Atrain} and tested on P_{Btest} and compares those decreases to the JSD between $P_{Atrain}(y)$ and $P_{Btest}(y)$ adjusted by the JSD between $P_{Atrain}(y)$ and $P_{Atest}(y)$.

For any set of partitions where the underlying JSD is smaller than the difference in models' test accuracy across the partitions, we consider that to be a *strongly biased* partition, implying that the distribution shift between $P_A(x, y)$ and $P_B(x, y)$ is even greater than the shift between $P_A(y)$ and $P_B(y)$. Conversely, partitions where the JSD is greater than the difference in model accuracy can be considered to be *weakly biased* partitions, implying that the distribution shift between $P_A(x, y)$ and $P_B(x, y)$ is smaller than the shift between $P_A(y)$ and $P_B(y)$.

While the magnitude of model performance change across partitions is informative for comparing to the underlying label distribution shift, to drill down on the importance of the sign of performance changes across partitions, the DivShift framework also measures the performance changes between models trained on P_{Atrain} and P_{Btrain} tested on both P_{Atest} and P_{Btest}. Whether a model performs better or worse on its out-of-distribution test set partition depends on the nature of the biased samplers $\overset{S_a}{\sim}$ and $\overset{S_b}{\sim}$; or in other words, some biases in the data generation process may be more helpful than others for estimating the underlying distribution $J(x, y)$. Specifically, when a model trained on P_{Atrain} has a higher out-of-partition accuracy on P_{Btest} than the model trained on P_{Btrain}, then the model trained on P_{Atrain} is a *strong generalizer* with respect to the model trained on P_{Btrain}, which is *overfitted*. This implies that some structure in the joint distribution $P_A(x, y)$ captures useful information about $P_B(x, y)$ that P_{Btrain} potentially lacks.

DivShift-NAWC Case Study

To prototype the DivShift framework, we introduce the DivShift–North American West Coast (DivShift-NAWC) dataset. DivShift-NAWC consists of ∼7.3 million images from ∼3.9 million research-grade and in need of ID iNaturalist observations across the west coast of North America (Fig. 3). The states in DivShift-NAWC cover seven of the world's nine terrestrial biomes, and include some of the coldest (Denali peak), hottest, driest (Death Valley), and wettest (Olympic peninsula) places on Earth. These states further include a high variation in socioeconomic status (2022 CA GDP: USD$3,600 bil.; 2022 BS GDP: USD$13 bil.) and data availability (CA: 6.22 obs/km^2; AK: 0.01 obs/km^2).

Distribution Shifts

While there are many taxonomies for classifying the kinds of bias present in biodiversity data (Carlen et al. 2024; Isaac and Pocock 2015), for this work, we partition DivShift-NAWC based on five main types of bias: spatial, temporal, taxonomic, quality, and sociopolitical (Fig. 1). For each partition, we randomly split 80% of the images into train and 20% into test data.

To measure the underlying data partition distribution shifts, we filtered each paired partition to only species shared between the partitions, and used Scipy's JSD function with a log base of 2 to calculate the distance between the label distributions (Fig. 2b). To account for noise in the sampling process, for the JSD calculation we randomly resample the 80% train 20% test splits five times for each partition and re-

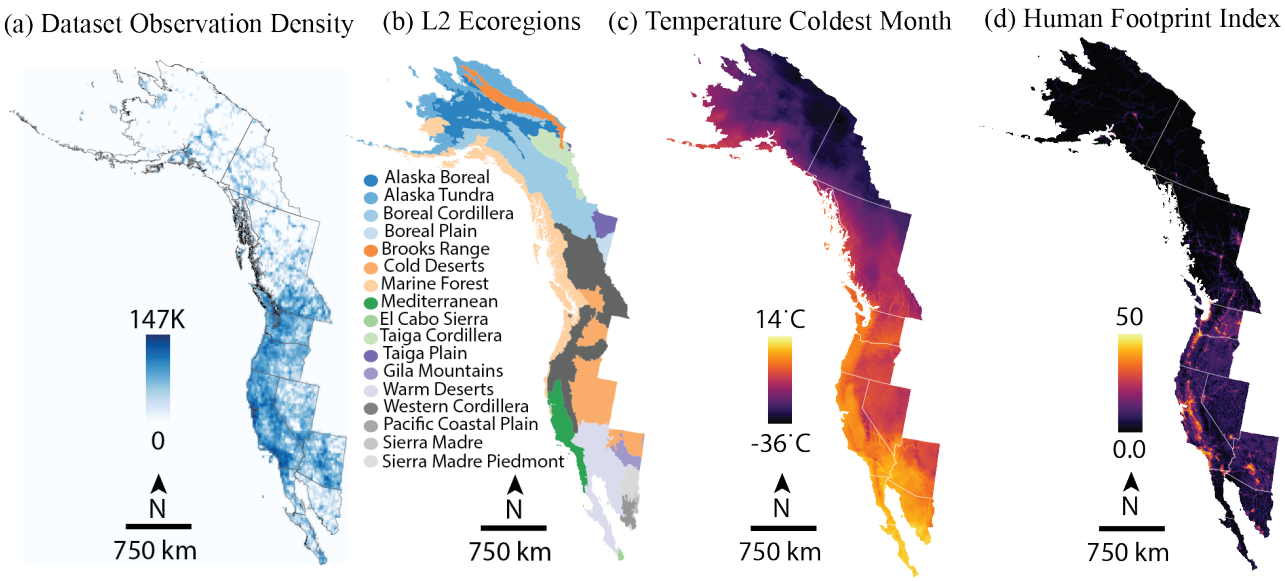

Figure 3: **Overview of DivShift–North American West Coast Dataset (DivShift-NAWC). (a)** Density plot of the DivShift-NAWC's iNaturalist observations (Naturalist 2024). Observations are skewed to U.S. and coastal states. **(b)** DivShift-NAWC spans a diverse set of habitats and ecosystems (Omernik 1987), **(c)** along with climates (WorldClim 2024). **(d)** DivShift-NAWC observations are concentrated in human-modified areas (Mu et al. 2022).

port the mean and standard deviation. To measure the impact on machine learning model performance, we train models on both partitions and test on said partition plus its complement, comparing the differences in performance.

Spatial Partition: Human Footprint Human-driven land use change is widespread across the planet, but there still exist large tracts of undisturbed habitat especially in the polar regions (Fig. 3d). However, these wilder regions are also harder to reach, making it difficult for volunteers to collect imagery there (Fig. 1a) and skewing volunteer-collected biodiversity data towards human-modified habitats (Fig. 4a). Using the Global Human Footprint Index (HFI) (Mu et al. 2022), we partition DivShift-NAWC into wilderness (HFI $\leq$ 1) and highly modified observations (HFI $\geq$ 4). Interestingly, over 90% of the 7.3 million images in DivShift-NAWC are from highly human-modified regions while only $\sim$6% are from minimally-modified wilderness (Table 1), as compared to $\sim$48% of all landmass in the DivShift-NAWC states being wilderness versus $\sim$37% being highly-modified.

Temporal Partition: City Nature Challenge The City Nature Challenge happens every year during the last weekend in April. This challenge creates a large spike in observations (Fig. 4b) (Di Cecco et al. 2021; iNaturalist 2023) and leads to altered observer behavior (Fig. 1b), as volunteers are encouraged to maximize the number of observations and unique species they observe within the week. While the majority of iNaturalist photos are taken outside of this challenge, a higher proportion of observations from the City Nature Challenge are labeled. Indeed, the Challenge captures more than half of the species from the entire DivShift-NAWC dataset despite having less than 6% of the total observations (Table 1), implying that observer behavior patterns shift significantly during the challenge. To test the benefits and drawbacks of this altered user behavior on model performance, we partition DivShift-NAWC so all observations taken during official City Nature Challenge (CNC) dates for the four years of study comprise one partition, while observations from all other weeks comprise the other.

Taxonomic Partition: Long-Tailed Versus Balanced While most species are rare and few species are common (Enquist et al. 2019), volunteer observations tend to be especially skewed towards charismatic or interesting species (Fig. 1c). To minimize these long-tail performance effects, many computer vision datasets built from volunteer-collected biodiversity image collections tend to re-balance the number of examples per-species to be more uniform (Van Horn et al. 2021; Deng et al. 2009). However, some of these more commonly observed species are also more ecologically abundant, and thus have a larger diversity of phenotypes, leading to a larger intra-class variability. This artificial balancing decision thus means that additional observations for these abundant species are excluded, potentially harming performance for these common species.

To explore the ramifications of this choice, we compare the performance of models trained on two train partitions with the same test set. Namely, after removing species with fewer than 25 observations, we consider a long-tailed partition where all observations are kept for common species. We also consider a balanced partition, where we only train on up to 300 randomly-selected images for common classes, ultimately discarding observations for 2,375 of the 7,607

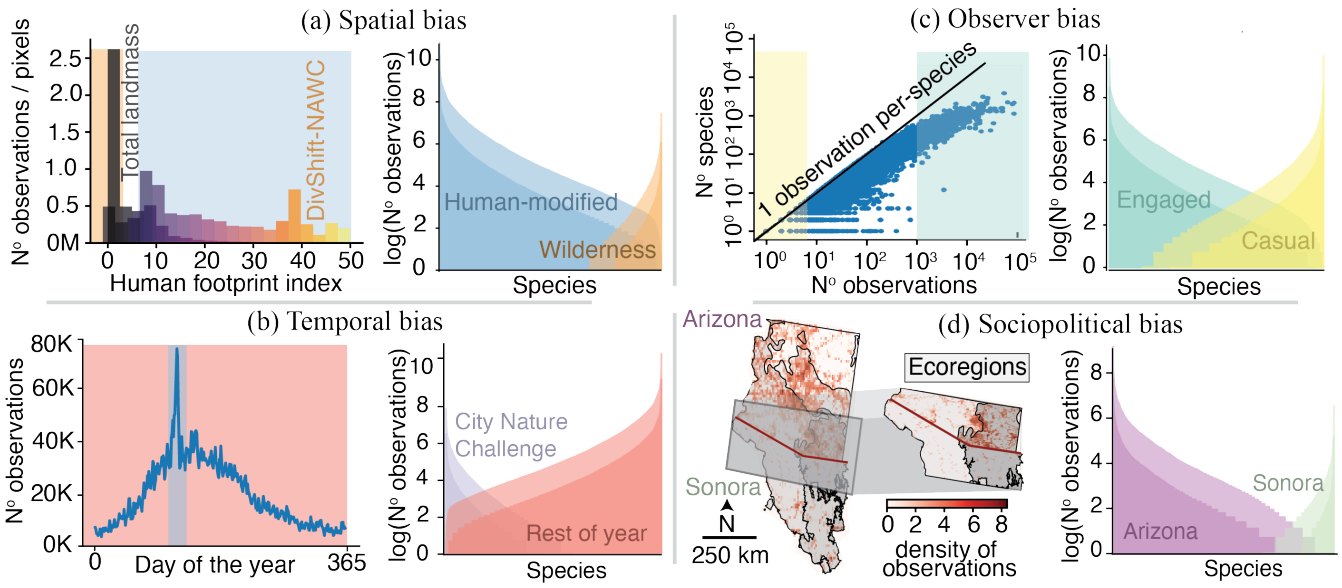

Figure 4: Biases in the DivShift-NAWC dataset. **(a)** Human footprint index (Mu et al. 2022) across human-modified and wilderness areas. **(b)** Observations per-day, with City Nature Challenge spike highlighted. **(c).** Observations per-observer with casual/engaged lines highlighted. **(d)** Density of observations in shared ecoregions across Arizona-Sonora border.

species in this partition.

Quality Partition: Observer Engagement Since iNaturalist observations are collected by volunteers with differing amounts of enthusiasm, time, and resources (Mac Domhnaill, Lyons, and Nolan 2020; Blake, Rhanor, and Pajic 2020), observer engagement varies widely between observers (Fig. 1d). Given that observers who use the app frequently and collect more data tend to observe a wider diversity of species in more diverse habitats (Fig. 4c) (Di Cecco et al. 2021), we also partition DivShift-NAWC by user engagement, with the casual partition consisting of all observations from observers with fewer than 50 total research-grade observations, and the engaged partition as observations from observers with more than 1,000 research-grade observations (Di Cecco et al. 2021).

Sociopolitical Partition: State Boundaries Where certain plant species can grow are demarcated by ecological boundaries (Fig. 3b). Similarly, volunteer observation trends are demarcated by political boundaries which may not necessarily reflect ecological ones (Figs. 1e). For example, while the Sonora and Mojave deserts extend beyond the borders of California and Arizona, the stark effects of political boundaries can be seen in the difference in abundance of observations between the U.S. and Mexico, especially across the Arizona–Sonora border, which bifurcates two similar ecosystems (Fig. 4d).

To test whether predictive accuracy of models trained in observation-rich geographies can extend across these at times ecologically arbitrary political boundaries to observation-poor regions, we compare model performance trained with observations from two states with a large number of observations (British Columbia and California) and test them on nearby states with varying levels of volunteer-collected biodiversity data availability (Alaska, Washington, Oregon, Yukon, and California for British Columbia; British Columbia, Washington, Oregon, Arizona, Nevada, Baja California, and Baja California Sur for California).

Baseline Partitions Lastly, we compare the absolute accuracy of these partitions to a variety of classic partitioning schemes from natural world imagery datasets. Specifically, we recreated the filtering and partitioning schema of the iNat2021 benchmarking dataset (Van Horn et al. 2021). We also recreated the iNat2021 mini train partition by randomly sub-sampling exactly 50 images per-species from the train set. Additionally, we tested spatial stratification, partitioning the study area into a 50 x 50 km grid and randomly assigning 20% of the grid boxes and the DivShift-NAWC images that fell within these boxes to test and the rest to train (Cole et al. 2020; Huynh et al. 2024). We also recreated the Imagenet train / test partitioning strategy, (Deng et al. 2009) and tested a naive train / test partition where 20% of observations and all of their corresponding images are selected as test images while the rest are used for training.

Model Training and Testing

To quantify bias effects with the DivShift-NAWC dataset, we focus on the specific task of fine-grained species recognition (identifying species from their images), which is an important biodiversity monitoring task for automating species detection. We only use research-grade iNaturalist observations as these images have been community-verified that the species identification is correct. As the goal of this work is to test distribution shift effects across partitions of volunteer-collected data as opposed to maximizing predictive perfor-

Partition	Images	Research Grade Obs.		Spec.
DivShift-NAWC Dataset	7.348M	4.726M	3.905M	7,607
Baselines				
iNat21	3.554M	3.554M	1.937M	1,852
iNat21 mini	0.185M	0.185M	0.109M	1,852
ImageNet	1.614M	1.614M	0.858M	1,260
Spatial Stratified	7.348M	4.726M	3.905M	7,607
Taxonomic Bias				
Long-tail	4.725M	4.725M	2.527M	7,607
Balanced	1.992M	1.992M	1.007M	7,607
Temporal Bias				
City Nature (CNC)	0.362M	0.245M	0.220M	3,929
Not City Nature	6.986M	4.480M	3.685M	7,604
Observer Bias				
Engaged	3.476M	2.324M	1.697M	7,361
Casual	1.113M	0.660M	0.756M	5,706
Spatial Bias				
Modified	6.642M	4.280M	3.536M	7,513
Wilderness	0.141M	0.083M	0.068M	2,395
Sociopolitical Bias				
Alaska (AK)	0.099M	0.064M	0.057M	875
Arizona (AZ)	0.497M	0.313M	0.272M	2,191
Baja California (BN)	0.142M	0.098M	0.090M	1,466
Baja California Sur (BS)	0.046M	0.033M	0.022M	716
British Columbia (BC)	1.080M	0.691M	0.622M	2,329
California (CA)	4.039M	2.558M	2.115M	4,654
Nevada (NV)	0.259M	0.177M	0.121M	1,860
Oregon (OR)	0.604M	0.399M	0.300M	2,711
Sonora (SO)	0.018M	0.010M	0.010M	673
Washington (WA)	0.529M	0.357M	0.279M	2,393
Yukon (YK)	0.034M	0.026M	0.018M	746

Table 1: DivShift-NAWC Data Representation by Partition. Obs=Observations. Spec=Species. Research-grade images are verified by at least two iNaturalist community members.

Train Partition	Test Partition	JSD Diff $(\mu \pm \sigma) \times 100$	Top1-Img Diff	Top1-Spec Diff
Spatial Bias (Human Footprint)				
Wild	Modified	49.29 $\pm$0.15	-35.3	-15.4
Modified	Wild	71.87 $\pm$0.30	-11.3	+6.9
Temporal Bias (City Nature Challenge (CNC))				
CNC	Not CNC	19.53 $\pm$0.08	-17.9	-10.0
Not CNC	CNC	36.39 $\pm$0.14	+1.9	+9.2
Quality Bias (Observer Engagement)				
Casual	Engaged	26.70 $\pm$0.09	-23.1	-12.5
Engaged	Casual	33.20 $\pm$0.08	+4.9	+3.8
Sociopolitical Bias (State Borders)				
CA	BC	79.04 $\pm$0.08	-29.1	-15.2
CA	WA	77.08 $\pm$0.09	-26.1	-11.6
CA	OR	70.27 $\pm$0.06	-24.6	-0.9
CA	AZ	74.83 $\pm$0.09	-23.0	-6.5
CA	NV	75.64 $\pm$0.06	-14.2	-0.2
CA	BN	59.66 $\pm$0.13	-10.3	+5.9
CA	BS	87.52 $\pm$0.15	-39.9	-10.1
CA	SO	86.77 $\pm$0.39	-26.1	+2.9
BC	AK	65.28 $\pm$0.26	-20.5	-3.7
BC	YK	75.61 $\pm$0.17	-32.4	-10.8
BC	WA	28.36 $\pm$0.18	-8.9	-5.4
BC	OR	38.72 $\pm$0.13	-16.7	-7.0
BC	CA	62.38 $\pm$0.12	-29.1	-8.5

Table 2: Comparison of label distribution shift to performance shift across bias partitions on DivShift-NAWC. The difference in JSD is calculated as $JSD(P_{Atrain}, P_{Btest}) - JSD(P_{Atrain}, P_{Atest})$ and the difference in observation performance (in %) is the model's out-of-distribution's test set performance minus the test set performance on the in-distribution train partition. JSD=Jensen-Shannon Divergence, Diff=Difference, Img=Image, Spec=Species.

mance, for each partition we train a small computer vision model for a limited number of epochs with the same hyperparameter configuration for each model.

Specifically, for each partition we train a ResNet-18 initialized with ImageNet pre-trained weights for 10 epochs with a batch size of 64, an SGD optimizer, single-label cross-entropy loss, and a learning rate of 0.064. Image augmentations were limited to resizing each image to at least 256 x 256 pixels and center cropping to 224 pixels, then normalizing the image with Imagenet mean and standard deviation. For testing, we employ early stopping using Top-1 species accuracy, and for all partitions we test only with images from species present in the split the model was trained on. To demonstrate how architecture and size choices improve absolute accuracy, we also ablate the model architecture and size, training a large ResNet50 and a base-size vision transformer (ViT) on the casual partition of the dataset.

Models' accuracies were measured using eight diverse accuracy metrics common to the machine community, such as Top-1 per-image (Top1-Img), per-species accuracy (Top1-Spec), accuracy aggregated by rarity (Beery et al. 2022)(Top1-FAR, CAR, RAR), and accuracy broken down by ecoregion (Top1-Eco) (Huynh et al. 2024). We also introduce a new rarity-weighted loss function that emphasizes performance on classes that are rare within a partition (Top1-Wgt), and accuracy broken down by land use category (Top1-LUC), emphasizing performance across both human-modified and untouched habitats.

Results and Discussion

Generally speaking, comparing the dataset label distribution shift measured by the JSD (Table 2, JSD Diff) versus the model performance shift (Table 2, Top1-Img Diff), when looking at strong versus weak biases per partition, we see encouragingly that all bias partitions are weakly biased as overall model performance changes for in- vs. out-of-distribution are smaller than the JSD across labels. For the sociopolitical bias partition, despite having the largest in- vs. out-of-domain model performance drops (Table 2, Sociopolitical Bias, Top1-Img), these drops in performance are much smaller than the drop in JSD between the partitions (Table 2, Sociopolitical Bias, JSD Diff). This implies that computer vision model performance losses across geographies tend to be less pronounced than underlying data availability differences (Beery, Van Horn, and Perona 2018).

Train-Test	Wgt	FAR	CAR	RAR	LUC	Eco
Spatial Bias (Human Footprint)						
Wild-Wild	16.1	34.8	30.0	18.1	55.5	52.8
Modified-Wild	**43.2**	**66.5**	**44.0**	10.8	**61.9**	**59.5**
Modified-Modified	15.6	69.3	48.8	21.3	70.2	70.1
Wild-Modified	8.5	19.2	12.4	5.2	17.9	18.0
Temporal Bias (City Nature Challenge (CNC))						
CNC-CNC	8.7	32.5	13.8	5.2	43.4	49.0
Not CNC-CNC	**37.0**	**65.8**	**44.7**	**18.6**	**70.8**	**75.8**
Not CNC-Not CNC	17.6	69.0	49.2	22.7	71.1	69.5
CNC-Not CNC	2.2	21.8	6.1	1.4	27.0	26.2
Taxonomic Bias (Balanced vs Long Tailed)						
Long-Long	17.8	70.9	50.4	24.7	70.1	72.0
Balanced-Long	**19.2**	55.0	**52.5**	**26.6**	52.0	51.3
Quality Bias (Observer Engagement)						
Casual-Casual	11.4	51.6	22.3	8.7	61.0	63.8
Engaged-Casual	**24.6**	**62.5**	**36.7**	**11.1**	**65.9**	**70.6**
Engaged-Engaged	15.0	63.2	41.2	18.4	63.0	62.1
Casual-Engaged	4.7	39.9	11.2	1.8	41.8	41.3
Sociopolitical Bias (State Borders)						
CA-CA	14.6	63.9	44.6	21.0	72.0	-
CA-BC	20.6	42.5	17.0	4.6	41.1	-
CA-WA	24.2	47.6	21.2	3.9	45.7	-
CA-OR	26.6	49.9	27.6	7.6	47.2	-
CA-AZ	29.3	51.5	25.3	8.7	50.1	-
CA-NV	39.0	56.1	35.3	19.9	56.1	-
CA-BN	36.2	62.3	37.7	11.3	60.1	-
CA-BS	32.2	49.7	23.3	4.2	32.8	-
CA-SO	38.3	50.3	22.0	11.4	51.6	-
BC-BC	11.4	57.5	35.0	13.6	70.5	-
BC-AK	21.4	55.6	23.4	4.4	47.8	-
BC-YK	17.9	49.8	23.3	2.3	34.9	-
BC-WA	12.5	50.8	23.1	3.6	59.3	-
BC-OR	15.3	46.3	20.0	4.4	50.6	-
BC-CA	18.2	43.5	17.2	1.7	40.7	-

Table 3: Top-1 accuracy results (in %) on DivShift-NAWC across bias partitions. Wgt=Weighted, FAR=Frequent Average Recall, CAR=Common Average Recall, RAR=Rare Average Recall, LUC=Land Use Category. Top-1 Eco is excluded for sociopolitical bias due to the lack of consistency of overlapping ecoregions between states.

Wilderness Vs. Modified Habitats For the spatial split, training on observations from less-disturbed habitats (Table 3, Wild-Modified) leads to worse performance than training in areas of high human activity (Table 3, Modified-Wild), likely due to the significant difference in the number of observations between these partitions, with modified regions having more than 3.5 million more unique observations than wilderness regions (Table 1, Spatial Bias). Indeed, the modified-trained model is a strong generalizer to wilderness regions, showing a higher out-of-distribution accuracy on wilderness observations for all but the rarest species in the DivShift-NAWC dataset (bolded entries, Table 3, Modified-Wild). The wilderness-trained model's low transferability even for rare species implies it has overfit and the wilderness partition is strongly biased, lacking sufficient data to build well-generalized models. Interestingly, the modified-trained model's out-of-distribution rarity-weighted accuracy for rare species within the wilderness partition (Wgt, Table 3, Modified-Wild) is much higher than its accuracy on species rare across the entire DivShift-NAWC dataset (RAR, Table 3, Modified-Wild). This implies that many species that are rare in wilderness regions are commonly dispersed in highly modified regions, and highlights that what is considered rare can very across geographies.

City Nature Challenge For the temporal split, we also see uniformly worse performance training on observations from the City Nature Challenge and testing on observations from outside the Challenge (Table 3, CNC-Not CNC), likely a consequence of significant differences in partition size (Table 1, Temporal Bias). Meanwhile, training on observations from outside the Challenge strongly generalizes and leads to improved performance on observations from the Challenge in all cases (bolded entries, Table 3, Not CNC-CNC), while the Challenge-trained model is especially overfitted for rare species. This implies that iNaturalist users are more drawn to increase their species count as opposed to their observation count during this bioblitz, leading to too many species and not enough observations to effectively train models from the City Nature Challenge (~ 62 observations per-species CNC, ~ 589 observations per-species Not CNC).

Long-Tailed Vs. Balanced For the taxonomic partition, we see that as expected, sub-sampling the most frequent classes improves accuracy per-class for all but the most common species (Table 3, Balanced-Long, FAR). Indeed, balancing the training set leads to strong generalization for rare species (bolded entries, Table 3, Balanced-Long), but conversely using the maximal training data available leads to the best common species performance (Table 3, Long-Long, FAR), highlighting the inherent tension between maximizing rare vs. common species performance.

Casual Vs. Engaged Observers The observation quality split shows the most marked performance differences across partitions, with the model trained on observations from engaged users showing substantially better and strongly generalized performance (bolded entries, Table 3, Engaged-Casual) over models trained with observations from casual users (Table 3, Casual-Casual). Given that this bias partition has the most similar number of observations between the two partitions, this implies that the overall lower performance of the casual-trained model stems from lower image quality for observations taken by casual users as opposed to insufficient training data volumes.

State Boundaries For the sociopolitical boundaries, we see substantial distribution shifts between states, with a general correlation between distance in space and distance in distribution (Table 2, Sociopolitical Bias, JSD Diff). Similarly, model accuracy drops off with larger distances between states and fewer shared ecoregions. However, when controlling for distance, states with a much higher data density tend to have a much lower performance (Table 3, Sociopolitical Bias, CA-BC, BC-CA) than similarly distant states with a low data density (Table 3, Sociopolitical Bias,

Baseline	JSD $(\mu \pm \sigma) \times 100$	Top1- Img %	Top1- Spec %	Top1- Wgt %	Top1- FAR %	Top1- CAR %	Top1- RAR %	Top1- LUC %	Top1- Eco %
Spatial	31.23	66.3	37.6	21.4	65.8	42.6	15.6	65.1	66.1
ImageNet	7.27	69.5	69.5	69.5	69.5	-	-	69.9	70.4
iNat2021	40.20	68.0	68.0	68.0	71.8	52.5	-	69.1	64.7
iNat2021 Mini	0.00	33.7	33.7	33.7	34.0	32.3	-	31.0	29.3
Random	8.09	70.6	40.5	20.5	69.8	48.4	21.3	70.9	69.8

Table 4: Results for baseline partitions (Huynh et al. 2024; Deng et al. 2009; Van Horn et al. 2021). Some CAR and RAR values are missing due to lack of common and rare species in baseline partition.

CA-BS, BC-AK), implying that the data volume of low density states may be insufficient to reliably test model transfer performance. These results imply that while predictive power decreases across boundaries, there is still some transferability across geography in the North American West Coast when data density is sufficient.

Baselines For the baseline partitions, we find that in general performance is on par with that of the bias partitions trained on their in-domain test sets, like the engaged and long-tailed partitions (Table 4). While the Random partition has high per-image accuracy, its much lower rarity-weighted accuracy implies that much of these gains may be concentrated in just the most common species, a common critique of this sampling approach. The iNat2021 Mini split has the lowest absolute but also most consistent performance across accuracy metrics, because sub-sampling a small number of images for every dense-enough class likely captures a less-biased sub-sampling in expectation. Lastly, the spatial split has frequency-binned accuracies comparable to other baseline splits but a significantly lower species- and rarity-weighted accuracy, implying in aggregate that spatial block sampling can capture dataset-wide trends, but sometimes at the cost of high train-test variance (JSD: 31.23, Table 4).

Model Architecture and Size Ablation Comparing performance across the casual split of the quality bias partition for a larger ResNet architecture and a transformer-based architecture (Table 5), we see both ablations outperform the ResNet-18 model across all accuracy metrics. Using a larger ResNet model led to modest performance improvements between $\sim$2 and $\sim$12% depending on the metric. Using a more modern vision transformer architecture led to significantly larger performance improvements, between $\sim$15 and $\sim$20% depending on the metric. Importantly, the differences between in-domain and out-domain per-image accuracy stay relatively consistent between ablations (Table 2, Casual-Engaged, Top1-Img Diff; Table 5, Img % Diff), highlighting the effectiveness of the DivShift framework to measure performance shifts independent of modeling choice.

Metric	ResNet50	ResNet50 Diff	ViT	ViT Diff
Img %	71.1	-22.7	79.0	-23.1
Spec %	35.2	-14.1	47.4	-21.0
Wgt %	15.6	-8.1	28.0	-17.6
FAR %	59.1	-11.4	70.2	-13.9
CAR %	28.4	-12.9	42.3	-20.7
RAR %	13.0	-9.4	24.0	-19.0
LUC %	74.0	-26.6	78.5	-21.5
Eco %	65.6	-16.4	81.1	-24.4

Table 5: Top-1 in-distribution and out-of-distribution difference accuracy results for ablated model architecture and size. Models trained on Casual Observers and tested on Casual Observers and Engaged Observers.

Recommendations for Downstream Modeling

Spatial Bias Takeaways: Wilderness regions simply lack sufficient volunteer-collected biodiversity data to train effective models (Table 2, Human Footprint), thus downstream modeling efforts targeted for undisturbed regions and their species will likely require additional data collection. **Temporal Bias Takeaways:** Normal iNaturalist user behavior leads to denser training data than from the City Nature Challenge data collection campaign alone (Table 3, CNC), thus modelers should complement models trained on bioblitz observations with data taken from the rest of the year when possible. **Taxonomic Bias Takeaways:** Using more data even if long-tailed improves common species performance but reduces rare species performance, leading to an inevitable rare vs. common trade-off (Table 3, Balanced vs Long-Tailed). Thus, training data sub-sampling should be chosen with downstream biodiversity monitoring use cases in mind (e.g. maximal accuracy for detection of common invasive species, vs. endangered species recognition). **Quality Bias Takeaways:** Observations from more engaged observers are of resounding higher quality (Table 3, User Engagement), thus modelers should consider discarding observations from observers with < 50 observations. **Sociopolitical Bias Takeaways:** Accuracy across geographies tends to degrade with larger distances but is obscured when data density is low (Table 2, State Borders), thus modelers working in data-sparse regions should take care to validate models with expert-collected data when possible.

Limitations and Future Work

Our findings represent the first comprehensive effort to quantify and document the downstream effects of bias in biodiversity data on computer vision model species recognition performance. Despite documenting the effects of five unique partitions, there are yet even more kinds of biases not tested here, and further complex interactions and intersections between these biases should be explored (Carlen et al. 2024; Bowler et al. 2022). So long as those biases enable the partitioning of biodiversity datasets, the flexibility of the

DivShift framework should allow for the targeted testing of these additional and intersectional biases.

While our framework can provide quantitative estimates of underlying distribution shift, it still lacks a mechanism to causally attribute performance changes to a given bias, an important direction for future work. Similarly, we only evaluate common supervised approaches to contrast performance across shifts. In future work, we envision expanding DivShift to the unsupervised and long-tailed learning settings to benchmark more modern machine learning techniques for dealing with distribution shift.

Furthermore, relying on label distribution shift across iNaturalist images may not be the most biologically-plausible way to measure ecological shifts, and alone may not capture all correlations between features and labels influenced by environmental or other factors. In the future, we aim to measure underlying distribution shift within environmental space by comparing climate hulls across the bias partitions, and within image space by comparing shifts within image features via image embedding manifold analysis.

Lastly, by performing train/test splitting per-image instead of per-observation, within partitions there is the risk that different images from the same observation end up in both the train and test split, inflating in-distribution test performance and reducing model generalizability. The DivShift framework can directly test for these effects when models show high in-distribution but low out-of-distribution performance. Furthermore, the DivShift framework is agnostic to train/test splitting choices and future versions of the DivShift-NAWC dataset will include both per-observation and per-image splitting options.

Conclusion

Here we present DivShift, a framework for quantifying bias-induced distribution shift across biodiversity datasets and introduce DivShift-NAWC, a new large-scale natural world imagery dataset designed to benchmark distribution shift effects on computer vision model performance for biodiversity monitoring tasks like fine-grained species recognition. This framework and dataset enable the rigorous testing of problems known to the conservation biology community in a machine learning setting to help enable the building of more robust, accurate biodiversity monitoring tools from large-scale volunteer datasets.

Acknowledgments

We first thank all the participants who contributed observations on iNaturalist. We further thank Noah Goodman and Gabriel Poesia for their comments and discussion, and we further thank CoCoLab for donating compute for this project. We also thank the TomKat Center for Sustainable Energy, The Fulbright Brasil Commission, Carnegie Science, the Howard Hughes Medical Institute, and the University of California, Berkeley for funding support for this research. This research was also funded by the NSF Graduate Research Fellowship DGE-1656518 (L.G.), the TomKat Graduate Fellowship for Translational Research (L.G.), and the Krupp Internship Program for Stanford Students in Germany (E.S.). T.K. and S.S. acknowledge funding from the German Research Foundation (DFG) for the projects BigPlantSens (project no. 444524904) and PANOPS (project no. 504978936). Lastly, M.E.-A. is supported by the Office of the Director of the National Institutes of Health's Early Investigator Award (1DP5OD029506-01), the U.S. Department of Energy, Office of Biological and Environmental Research (DE-SC0021286), and by the U.S. National Science Foundation's DBI Biology Integration Institute WALII (Water and Life Interface Institute, 2213983). Compute for this project was performed on the Calc cluster at Carnegie Science and the Stanford SC Compute Cluster.

References

Arazy, O.; and Malkinson, D. 2021. A framework of observer-based biases in citizen science biodiversity monitoring: Semi-structuring unstructured biodiversity monitoring protocols. *Frontiers in Ecology and Evolution*, 9: 693602.

Aristeidou, M.; Herodotou, C.; Ballard, H. L.; Young, A. N.; Miller, A. E.; Higgins, L.; and Johnson, R. F. 2021. Exploring the participation of young citizen scientists in scientific research: The case of iNaturalist. *Plos one*, 16(1): e0245682.

Backstrom, L. J.; Callaghan, C. T.; Worthington, H.; Fuller, R. A.; and Johnston, A. 2024. Estimating sampling biases in citizen science datasets. *Ibis*.

Beery, S.; Agarwal, A.; Cole, E.; and Birodkar, V. 2021. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*.

Beery, S.; Van Horn, G.; and Perona, P. 2018. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, 456–473.

Beery, S.; Wu, G.; Edwards, T.; Pavetic, F.; Majewski, B.; Mukherjee, S.; Chan, S.; Morgan, J.; Rathod, V.; and Huang, J. 2022. The auto arborist dataset: a large-scale benchmark for multiview urban forest monitoring under domain shift. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21294–21307.

Blake, C.; Rhanor, A.; and Pajic, C. 2020. The demographics of citizen science participation and its implications for data quality and environmental justice.

Boakes, E. H.; Gliozzo, G.; Seymour, V.; Harvey, M.; Smith, C.; Roy, D. B.; and Haklay, M. 2016. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Scientific Reports*, 6(1): 33051. Publisher: Nature Publishing Group.

Boakes, E. H.; McGowan, P. J.; Fuller, R. A.; Chang-qing, D.; Clark, N. E.; O'Connor, K.; and Mace, G. M. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS biology*, 8(6): e1000385.

Botts, E. A.; Erasmus, B. F.; and Alexander, G. J. 2011. Geographic sampling bias in the South African Frog Atlas Project: implications for conservation planning. *Biodiversity and Conservation*, 20: 119–139.

Bowler, D. E.; Callaghan, C. T.; Bhandari, N.; Henle, K.; Benjamin Barth, M.; Koppitz, C.; Klenke, R.; Winter, M.; Jansen, F.; Bruelheide, H.; et al. 2022. Temporal trends in the spatial bias of species occurrence records. *Ecography*, 2022(8): e06219.

Burgess, H. K.; DeBey, L.; Froehlich, H.; Schmidt, N.; Theobald, E. J.; Ettinger, A. K.; HilleRisLambers, J.; Tewksbury, J.; and Parrish, J. K. 2017. The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation*, 208: 113–120.

Callaghan, C. T.; Poore, A. G.; Hofmann, M.; Roberts, C. J.; and Pereira, H. M. 2021. Large-bodied birds are over-represented in unstructured citizen science data. *Scientific reports*, 11(1): 19073.

Carlen, E. J.; Estien, C. O.; Caspi, T.; Perkins, D.; Goldstein, B. R.; Kreling, S. E.; Hentati, Y.; Williams, T. D.; Stanton, L. A.; Des Roches, S.; et al. 2024. A framework for contextualizing social-ecological biases in contributory science data. *People and Nature*, 6(2): 377–390.

Chen, B.; Wu, S.; Song, Y.; Webster, C.; Xu, B.; and Gong, P. 2022. Contrasting inequality in human exposure to greenspace between cities of Global North and Global South. *Nature Communications*, 13(1): 4636.

Cole, E.; Deneu, B.; Lorieul, T.; Servajean, M.; Botella, C.; Morris, D.; Jojic, N.; Bonnet, P.; and Joly, A. 2020. The geolifeclef 2020 dataset. *arXiv preprint arXiv:2004.04192*.

Cooper, C.; Martin, V.; Wilson, O.; and Rasmussen, L. 2023. Equitable data governance models for the participatory sciences. *Community Science*, 2(2): e2022CSJ000025.

Cooper, C. B. 2014. Is there a weekend bias in clutch-initiation dates from citizen science? Implications for studies of avian breeding phenology. *International journal of biometeorology*, 58: 1415–1419.

Courter, J. R.; Johnson, R. J.; Stuyck, C. M.; Lang, B. A.; and Kaiser, E. W. 2013. Weekend bias in Citizen Science data reporting: implications for phenology studies. *International journal of biometeorology*, 57: 715–720.

Crimmins, T. M.; Posthumus, E.; Schaffer, S.; and Prudic, K. L. 2021. COVID-19 impacts on participation in large scale biodiversity-themed community science projects in the United States. *Biological Conservation*, 256: 109017.

de Lutio, R.; Park, J. Y.; Watson, K. A.; D'Aronco, S.; Wegner, J. D.; Wieringa, J. J.; Tulig, M.; Pyle, R. L.; Gallaher, T. J.; Brown, G.; et al. 2022. The herbarium 2021 half–earth challenge dataset and machine learning competition. *Frontiers in Plant Science*, 12: 787127.

Deacon, C.; Govender, S.; and Samways, M. J. 2023. Overcoming biases and identifying opportunities for citizen science to contribute more to global macroinvertebrate conservation. *Biodiversity and Conservation*, 32(6): 1789–1806.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Di Cecco, G. J.; Barve, V.; Belitz, M. W.; Stucky, B. J.; Guralnick, R. P.; and Hurlbert, A. H. 2021. Observing the observers: How participants contribute data to iNaturalist and implications for biodiversity science. *BioScience*, 71(11): 1179–1188.

Dimson, M.; and Gillespie, T. W. 2023. Who, where, when: Observer behavior influences spatial and temporal patterns of iNaturalist participation. *Applied Geography*, 153: 102916.

Ellis-Soto, D.; Chapman, M.; and Locke, D. H. 2023. Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States. *Nature Human Behaviour*, 7(11): 1869–1877.

Endres, D. M.; and Schindelin, J. E. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7): 1858–1860.

Enquist, B. J.; Feng, X.; Boyle, B.; Maitner, B.; Newman, E. A.; Jørgensen, P. M.; Roehrdanz, P. R.; Thiers, B. M.; Burger, J. R.; Corlett, R. T.; et al. 2019. The commonness of rarity: Global and future distribution of rarity across land plants. *Science advances*, 5(11): eaaz0414.

Garcin, C.; Joly, A.; Bonnet, P.; Affouard, A.; Lombardo, J.-C.; Chouet, M.; Servajean, M.; Lorieul, T.; and Salmon, J. 2021. Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution. In *NeurIPS Datasets and Benchmarks*.

Geldmann, J.; Heilmann-Clausen, J.; Holm, T. E.; Levinsky, I.; Markussen, B.; Olsen, K.; Rahbek, C.; and Tøttrup, A. P. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22(11): 1139–1149.

Gillespie, L. E.; Ruffley, M.; and Exposito-Alonso, M. 2024. Deep learning models map rapid plant species changes from citizen science and remote sensing data. *Proceedings of the National Academy of Sciences*.

Goëau, H.; Bonnet, P.; and Joly, A. 2023. Overview of PlantCLEF 2023: image-based plant identification at global scale. In *24th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF-WN 2023*, volume 3497, 1972–1981.

Gratzer, K.; and Brodschneider, R. 2021. How and why beekeepers participate in the INSIGNIA citizen science honey bee environmental monitoring project. *Environmental Science and Pollution Research*, 28(28): 37995–38006.

Hochmair, H. H.; Scheffrahn, R. H.; Basille, M.; and Boone, M. 2020. Evaluating the data quality of iNaturalist termite records. *PLoS One*, 15(5): e0226534.

Huynh, A. V.; Gillespie, L. E.; Lopez-Saucedo, J.; Tang, C.; Sikand, R.; and Expósito-Alonso, M. 2024. Contrastive ground-level image and remote sensing pre-training improves representation learning for natural world imagery. In *Proceedings of the 18th European Conference on Computer Vision ECCV 2024*.

iNaturalist. 2023. Year In Review 2023. https://www.inaturalist.org/stats/2023. Accessed: 2024-07-30.

Isaac, N. J.; and Pocock, M. J. 2015. Bias and information in biological records. *Biological Journal of the Linnean Society*, 115(3): 522–531.

Isaac, N. J.; van Strien, A. J.; August, T. A.; de Zeeuw, M. P.; and Roy, D. B. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, 5(10): 1052–1060.

Johnston, A.; Matechou, E.; and Dennis, E. B. 2023. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, 14(1): 103–116.

Kay, J.; Kulits, P.; Stathatos, S.; Deng, S.; Young, E.; Beery, S.; Van Horn, G.; and Perona, P. 2022. The Caltech Fish Counting Dataset: A Benchmark for Multiple-Object Tracking and Counting. In *European Conference on Computer Vision (ECCV)*.

Kishimoto, K.; and Kobori, H. 2021. COVID-19 pandemic drives changes in participation in citizen science project "City Nature Challenge" in Tokyo. *Biological Conservation*, 255: 109001.

Kshitiz, .; Shreshtha, S.; Dutta, B.; Dosi, M.; Vatsa, M.; Singh, R.; Anand, S.; Sarkar, S.; and Parihar, S. M. 2024. BirdCollect: A Comprehensive Benchmark for Analyzing Dense Bird Flock Attributes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21879–21887.

Lee, J. J.; Li, B.; Beery, S.; Huang, J.; Fei, S.; Yeh, R. A.; and Benes, B. 2024. Tree-D Fusion: Simulation-Ready Tree Dataset from Single Images with Diffusion Priors. arXiv:2407.10330.

Mac Domhnaill, C.; Lyons, S.; and Nolan, A. 2020. The citizens in citizen science: demographic, socioeconomic, and health characteristics of biodiversity recorders in Ireland. *Citizen Science: Theory and Practice*, 5(1).

Mahmoudi, D.; Hawn, C. L.; Henry, E. H.; Perkins, D. J.; Cooper, C. B.; and Wilson, S. M. 2022. Mapping for whom? Communities of color and the citizen science gap. *ACME: An International Journal for Critical Geographies*, 21(4): 372–388.

Mair, L.; and Ruete, A. 2016. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PloS one*, 11(1): e0147796.

McGoff, E.; Dunn, F.; Cachazo, L. M.; Williams, P.; Biggs, J.; Nicolet, P.; and Ewald, N. C. 2017. Finding clean water habitats in urban landscapes: professional researcher vs citizen science approaches. *Science of the Total Environment*, 581: 105–116.

McMullin, R. T.; and Allen, J. L. 2022. An assessment of data accuracy and best practice recommendations for observations of lichens and other taxonomically difficult taxa on iNaturalist. *Botany*, 100(6): 491–497.

Milanesi, P.; Mori, E.; and Menchetti, M. 2020. Observer-oriented approach improves species distribution models from citizen science data. *Ecology and Evolution*, 10(21): 12104–12114.

Mu, H.; Li, X.; Wen, Y.; Huang, J.; Du, P.; Su, W.; Miao, S.; and Geng, M. 2022. A global record of annual terrestrial Human Footprint dataset from 2000 to 2018. *Scientific Data*, 9(1): 176.

Naturalist. 2024. Naturalist. https://www.inaturalist.org. Accessed: 2024-07-30.

Omernik, J. 1987. Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, 77: 118–125.

Pandya, R. E. 2012. A framework for engaging diverse communities in citizen science in the US. *Frontiers in Ecology and the Environment*, 10(6): 314–317.

Pateman, R. M.; Dyke, A.; and West, S. E. 2021. The diversity of participants in environmental citizen science. *Citizen Science: Theory and Practice*.

Pernat, N.; Kampen, H.; Jeschke, J. M.; and Werner, D. 2021. Citizen science versus professional data collection: Comparison of approaches to mosquito monitoring in Germany. *Journal of Applied Ecology*, 58(2): 214–223.

Rosenblatt, C. J.; Dayer, A. A.; Duberstein, J. N.; Phillips, T. B.; Harshaw, H. W.; Fulton, D. C.; Cole, N. W.; Raedeke, A. H.; Rutter, J. D.; and Wood, C. L. 2022. Highly specialized recreationists contribute the most to the citizen science project eBird. *Ornithological Applications*, 124(2): duac008.

Sánchez-Clavijo, L. M.; Martínez-Callejas, S. J.; Acevedo-Charry, O.; Diaz-Pulido, A.; Gómez-Valencia, B.; Ocampo-Peñuela, N.; Ocampo, D.; Olaya-Rodríguez, M. H.; Rey-Velasco, J. C.; Soto-Vargas, C.; et al. 2021. Differential reporting of biodiversity in two citizen science platforms during COVID-19 lockdown in Colombia. *Biological Conservation*, 256: 109077.

Sastry, S.; Khanal, S.; Dhakal, A.; Huang, D.; and Jacobs, N. 2024. Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7136–7145.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Soleri, D.; Long, J. W.; Ramirez-Andreotta, M. D.; Eitemiller, R.; and Pandya, R. 2016. Finding pathways to more equitable and meaningful public-scientist partnerships. *Citizen Science: Theory and Practice*, 1(1): 9–9.

Soltani, S.; Ferlian, O.; Eisenhauer, N.; Feilhauer, H.; and Kattenborn, T. 2024. From simple labels to semantic image segmentation: leveraging citizen science plant photographs for tree species mapping in drone imagery. *Biogeosciences*, 21(11): 2909–2935.

Stevens, S.; Wu, J.; Thompson, M. J.; Campolongo, E. G.; Song, C. H.; Carlyn, D. E.; Dong, L.; Dahdul, W. M.; Stewart, C.; Berger-Wolf, T.; et al. 2024. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19412–19424.

Stoudt, S.; Goldstein, B. R.; and de Valpine, P. 2022. Identifying engaging bird species and traits with community science observations. *Proceedings of the National Academy of Sciences*, 119(16): e2110156119.

Swanson, A.; Kosmala, M.; Lintott, C.; Simpson, R.; Smith, A.; and Packer, C. 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific data*, 2(1): 1–14.

Sweet, F. S.; Rödl, T.; and Weisser, W. W. 2022. COVID-19 lockdown measures impacted citizen science hedgehog observation numbers in Bavaria, Germany. *Ecology and Evolution*, 12(6): e8989.

Teng, M.; Elmustafa, A.; Akera, B.; Bengio, Y.; Radi, H.; Larochelle, H.; and Rolnick, D. 2024. Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data. *Advances in Neural Information Processing Systems*, 36.

Unger, S.; Rollins, M.; Tietz, A.; and Dumais, H. 2021. iNaturalist as an engaging tool for identifying organisms in outdoor activities. *Journal of Biological Education*, 55(5): 537–547.

Van Eupen, C.; Maes, D.; Herremans, M.; Swinnen, K. R.; Somers, B.; and Luca, S. 2021. The impact of data quality filtering of opportunistic citizen science data on species distribution model performance. *Ecological Modelling*, 444: 109453.

Van Horn, G.; Cole, E.; Beery, S.; Wilber, K.; Belongie, S.; and Mac Aodha, O. 2021. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12884–12893.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wang, H.; Lu, H.; Guo, H.; Jian, H.; Gan, C.; and Liu, W. 2023. Bird-Count: a multi-modality benchmark and system for bird population counting in the wild. *Multimedia Tools and Applications*, 82(29): 45293–45315.

Ward, D. F. 2014. Understanding sampling and taxonomic biases recorded by citizen scientists. *Journal of insect conservation*, 18: 753–756.

Weinstein, B. G.; Marconi, S.; Bohlman, S. A.; Zare, A.; Singh, A.; Graves, S. J.; and White, E. P. 2021. A remote sensing derived data set of 100 million individual tree crowns for the National Ecological Observatory Network. *Elife*, 10: e62922.

Wolf, S.; Mahecha, M. D.; Sabatini, F. M.; Wirth, C.; Bruelheide, H.; Kattge, J.; Moreno Martínez, Á.; Mora, K.; and Kattenborn, T. 2022. Citizen science plant observations encode global trait patterns. *Nature Ecology & Evolution*, 6(12): 1850–1859.

WorldClim. 2024. WorldClim Dataset. https://worldclim.org/. Accessed: 2024-07-29.