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

Elena Sierra[1,2,3*] **Lauren E. Gillespie**[1,3,4,5*] **Salim Soltani**[2]
Moises Exposito-Alonso[3,5,6] **Teja Kattenborn**[2]
[1]Stanford University [2]Universität Freiburg [3]Carnegie Science
[4]Federal University of Minas Gerais [5]University of California Berkeley
[6]Howard Hughes Medical Institute
{esierra,gillespl}@cs.stanford.edu
{salim.soltani,teja.kattenborn}@geosense.uni-freiburg.de
moiexpositoalonso@berkeley.edu

Abstract

Climate change is negatively impacting the world's biodiversity. To build automated systems to monitor these negative biodiversity impacts, large-scale, volunteer-collected datasets like iNaturalist are built from community-identified, natural imagery. However, such volunteer-based data are opportunistic and lack a structured sampling strategy, resulting in geographic, temporal, observation quality, and socioeconomic, biases that stymie uptake of these models for downstream biodiversity monitoring tasks. Here we introduce DivShift North American West Coast (DivShift-NAWC), a curated dataset of almost 8 million iNaturalist plant images across the western coast of North America, for exploring the effects of these biases on deep learning model performance. We compare model performance across four known biases and observe that they indeed confound model performance. We suggest practical strategies for curating datasets to train deep learning models for monitoring climate change's impacts on the world's biodiversity.

1 Introduction

The world's biodiversity is under threat from climate and land-use change [96, 90]. Biodiversity helps ecosystems combat climate change by improving carbon sequestration [137] and novel climate adaptation [105]. Therefore, monitoring the world's biodiversity via automated tools [100] is critical to mitigate climate change's effects on the natural world. Building machine learning tools for this automated monitoring requires large volumes of natural world imagery [53]. Participatory science applications—where users can upload and identify photos of species in their natural environments—have surged in popularity [37, 3, 39] (see Appdx. 5.1.1 for participatory science overview). These applications now provide sufficient finely-labeled imagery data to build large-scale biodiversity datasets [125, 60, 122, 107, 50, 55] and train deep learning models for automated biodiversity monitoring tasks such as species recognition [125, 50, 55, 60], species distribution modeling [122, 60, 107, 53], novel species identification [115], and visual question answering [125].

However, as these observations become easier for the public to make, sampling becomes more unstructured [2, 98, 52] and injects biases into these data [63, 37, 11, 39]. Therefore, volunteer-collected biodiversity datasets often do not fully reflect the world's biodiversity [3, 27, 20, 70], presenting challenges for the general uptake of these data for biodiversity monitoring [66, 12, 34]. To help quantify the effects of these biases on model performance, here we introduce a new public

[*]Denotes equal contribution to this work

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

Figure 1: Biodiversity data biases include **spatial** bias, e.g. more observations in urban compared to wild areas; **temporal** bias, e.g. more observations on weekends compared to weekdays; **observation quality** bias, e.g. more diverse observations made by highly engaged observers than infrequent users; and **sociopolitical** bias, e.g. disparities between observation density due to political as opposed to ecological boundaries.

biodiversity imagery dataset DivShift-North American West Coast (DivShift-NAWC), comprised of 8 million observations and 10,000 plant species across the North American west coast. DivShift-NAWC spans wide climatic, ecological, and sociopolitical gradients, enabling the targeted testing of downstream performance effects of spatial, temporal, observation quality, and sociopolitical biases present in volunteer-collected biodiversity datasets.

2 Related Works

Biases in volunteer-collected biodiversity datasets. Collection of large-scale volunteer datasets are subject to social and ecological filters [21, 63], which inject many types of bias into biodiversity datasets built from these collections [21, 38, 64] (see Appdx. 5.1.2 for further definition). In this work, we focus on four kinds of bias common to volunteer-collected biodiversity datasets: spatial, temporal, observation quality, and sociopolitical (Fig. 1). **Spatial bias** includes observer preferences to sampling easy-to-access greenspaces in urban areas [56, 83, 3, 38, 39, 21] (see Appdx. 5.1.3). **Temporal bias** [3, 39, 119, 106, 30, 61] includes a skew towards more observations on weekends when observers are free from work [38, 29, 27] (see Appdx. 5.1.3). **Observation quality bias** [39, 124, 85, 10] manifests as a small but dedicated group of users that tend to observe more species in more diverse habitats [38, 104] (see Appdx. 5.1.3). Lastly, **sociopolitical bias** in who has access to the resources, time, and areas to collect biodiversity observations [9, 41, 80, 34, 24, 21, 16, 26, 111, 93] includes a skew towards whiter, wealthier, and older observers [79, 97] (see Appdx. 5.1.3).

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 [125, 60, 107, 50, 55, 126], high-quality archival and herbaria images [115, 33], long-distance camera imagery [131, 71], terrestrial camera traps [6, 118], ocean sonar cameras [69], google street view imagery [7, 73] and remote sensing imagery [25, 53, 122, 60, 132]. With these datasets, efforts to minimize the domain-specific bias of biodiversity data range from using context to differentially select training examples [94], mixing high-quality, expert-curated images in with volunteer-collected images [115], evenly sampling images by class [125], and spatial stratification [60, 25, 53]. However, none of these datasets nor techniques explicitly address how each type of bias present in these datasets affect downstream model performance.

3 DivShift Framework and DivShift-NAWC Dataset

To explore the downstream effects of bias in volunteer-collected biodiversity datasets, we first propose the bioDiversity Shift (DivShift) framework, then introduce the DivShift North American West Coast (DivShift-NAWC) dataset.

DivShift Framework: The bioDiversity Shift (DivShift) framework casts biases present in volunteer-collected biodiversity data as *distribution shifts*. Specifically, the dataset of individuals actually uploaded by observers, D, is the result of the biased sampling process $\overset{S_a}{\sim}$ from the true distribution of biodiversity, J, leading to strong skews in D across space, time, taxon, observer, and sociopolitical boundaries. To explore these differences, we partition D into P_A and P_B by some known bias in biodiversity data, (e.g. for spatial bias P_A being observations from modified cities and P_B being observations from undisturbed wilderness). We then further sub-partition observations in P_A and

P_B randomly into 80% train P_{Atrain} and 20% P_{Atest} test as normally done in machine learning datasets [35, 125, 7], then measure the underlying effect of this distribution shift via the Jensen-Shannon Distance (JSD) between P_{Atrain} and P_{Btest}. Finally, we compare the JSD to the change in performance between deep learning models trained on P_{Atrain} and tested on P_{Atest} to the same models tested on P_{Btest}. If a model's performance decreases when tested on P_{Btest}, we consider that to be a negative bias, and a positive bias if it increases. Moreso, if the change in a model's performance is less than the underlying JSD between the partitions P_{Atrain} and P_{Btest}, it is considered a *weak bias*, while if it is greater, it is considered a *strong bias*.

DivShift North American West Coast Dataset: To test these distribution effects, we use the new DivShift - North American West Coast (DivShift-NAWC) dataset. DivShift-NAWC consists of 7.3 million iNaturalist images from the west coast of North America between 2019 and 2023, which spans three countries, eleven states, and eleven ecosystems (Table A1, Fig. 2). Biodiversity data by nature is long-tailed [43], and so like all other popular natural world imagery datasets [126, 125, 7, 60], DivShift-NAWC is heavily long-tailed.

We focus on spatial, temporal, observation quality, and sociopolitical bias (Fig. 1). For **spatial bias**, we use the Human Footprint Index (HFI) [87] where observations in wilderness are HFI <= 1 and modified are HFI >= 4 (Appdx. 5.3.2). For **temporal bias** we compare observations from the City Nature Challenge (CNC) [37, 61] to those not (Appdx. 5.3.3). For **observation quality bias**, compare engaged observers (> 1,000 observations) to casual observers (< 50 observations) [38] (Appdx. 5.3.4). Finally for **sociopolitical bias**, we compare two high-resource states (> 1 mil. images, California-C) and British Columbia-BC) to two low-resource states ($< 50,000$ images, Sonora-SO and Yukon-YT) [102] (Appdx. 5.3.5). We also report accuracies on baseline train / test partitions as commonly executed in other natural world imagery datasets [125, 25, 60, 35] (Appdx. 5.3.6).

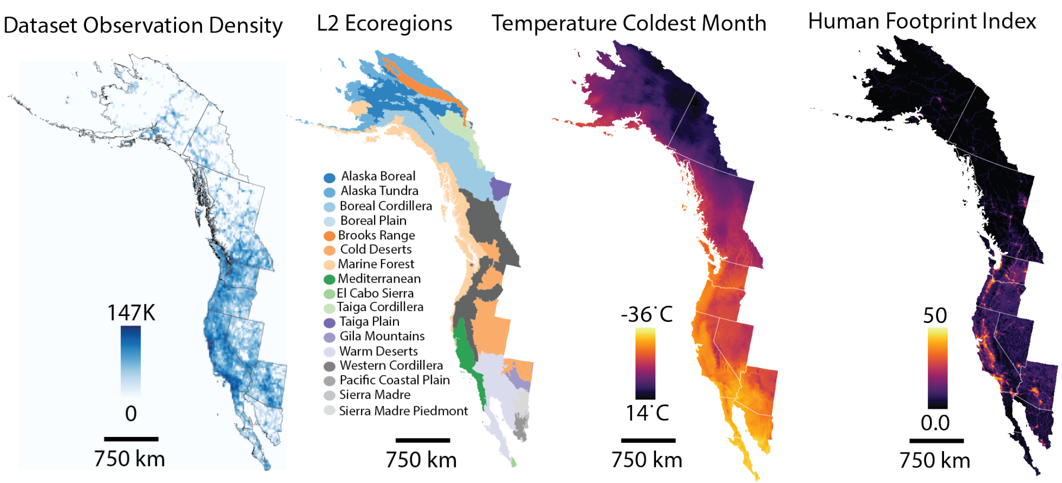

Figure 2: **Overview of DivShift-North American West Coast Dataset. a.** Density plot of the dataset's iNaturalist observations [89]. Observations are skewed to U.S. and coastal states. **b.** DivShift-NAWC spans a diverse set of habitats and ecosystems [92], **c.** along with climates [136]. **d.** DivShift-NAWC observations are concentrated in human-modified areas [87].

Model Training and Evaluation Metrics: To measure the underlying data partition distribution shifts, we filter each bias partition (e.g. modified cities vs. wilderness) to only species shared between the partitions, and calculated the Jensen-Shannon Distance (JSD) between the label distributions (Appdx. 5.5). For each bias partition, we split images into either train (80%) or test (20%) unless otherwise specified, and train a ResNet18 for 10 epochs on a given train partition (see Appdx. 5.6 for training details). For each model, we report Top1 accuracy per-observation (Top1-Obs), Top1 accuracy per-species (Top1-Spec), a new Top1 accuracy metric weighted by rarity (Top1-Wgt), and Top1 accuracy by land use category (Top1-LUC) (see Appdx. 5.7 for metric definitions).

4 Results and Discussion

For all bias partitions, surprisingly all biases are weak, with performance drop greater than the underlying JSD of the partition (Tables 1a, 1b). For the **spatial partition**, there is a larger distribution shift from wild to modified observations than vice-versa (Table 1b). However, training on observations from less-disturbed habitats leads to worse performance than training in areas of high human activity (Table 1a). For the **temporal partition**, the distribution shift is largely symmetric (Table 1b), but training on City Nature leads universally worse performance on observations from outside the Challenge, while training with observations from outside the Challenge leads to better performance except when accounting for land use (Table 1a). For the **observation quality partition** we see that the distribution shift between both the engaged and casual observer partitions are surprisingly symmetric (Table 1b). However, the model trained using images from engaged observers showed increased accuracy across the board, while the model trained on casual observations saw a universal decrease (Table 1a). For the **sociopolitical bias**, the distribution shift is greater for CA to SO than BC to YT (Table 1b), yet surprisingly the decrease in performance is less pronounced for CA to SO than BC to YT, except when correcting for land use type (Table 1a). Across the four splits, model accuracies (Table A3) are well within the range of the baseline splits (Table A2), implying that the bias splits are not dramatically skewed compared to previous approaches.

Train-Test Diff.	**Top1-Obs**	**Top1-Spec**	**Top1-Wgt**	**Top1-LUC**
Spatial Bias (Human Footprint)				
Wild-Modified diff	-0.353	-0.154	-0.076	-0.376
Modified-Wild diff	-0.113	0.068	0.322	-.111
Temporal Bias (City Nature Challenge)				
CNC-Not CNC diff	-0.179	-0.100	-0.065	-0.164
Not CNC-CNC diff	0.007	0.091	0.200	-0.009
Observer Quality Bias				
Casual-Engaged diff	-0.231	-0.125	-0.067	-0.192
Engaged-Casual diff	0.049	0.038	0.097	0.29
Sociopolitical Bias				
CA-SO diff	-0.248	0.014	0.294	-0.336
BC-YT diff	-0.309	-0.095	0.069	-0.284

(a)

Train - Test	**JSD**
Spatial Bias (Human Footprint)	
Wild - Modified	0.643
Modified - Wild	0.618
Temporal Bias (City Nature)	
CNC - Not CNC	0.307
Not CNC - CNC	0.296
Observation Quality Bias	
Casual - Engaged	0.376
Engaged - Casual	0.372
Socio-Political Bias	
CA - SO	0.571
BC - YT	0.518

(b)

Table 1: (1a) Performance differences for all partitions (absolute values in Table A3). Blue values indicate positive bias, red values indicate negative bias. Obs = Observation, Spec=Species, Wgt=Weighted, LUC=Land Use Category, CNC = City Nature Challenge, diff = difference. (1b) Jensen-Shannon differences between partitions. JSD = Jensen-Shannon Difference.

Climate Change Impact: Our findings on the DivShift-NAWC dataset suggest four recommendations for training deep learning models on voluminous but noisy volunteer-collected biodiversity datasets. First, observations from urban areas provide useful training signal even for wild areas for species found in both kinds of habitats. Second, using many noisy observations with a random sampling pattern is better than using fewer observations with a more structured sampling pattern. Third, using data from more engaged observers is generally better than less-engaged ones. Fourth, even though sociopolitical boundaries can have a significant effect on performance, training on diverse and extremely well-sampled high-resource regions like CA can still provide some predictive benefit for low-resource regions even if geographically far away.

Conclusion Here we introduce DivShift-NAWC, a new large-scale natural world imagery dataset designed to benchmark distribution shift effects on computer vision models for biodiversity monitoring. 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 for counteracting climate change's effects on the world's biodiversity.

Acknowledgments and Disclosure of Funding

We first and foremost thank all of the participants who volunteered and 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) within the framework of BigPlantSens (Assessing the Synergies of Big Data and Deep Learning for the Remote Sensing of Plant Species; project no. 444524904) and PANOPS (Revealing Earth's plant functional diversity with citizen science; project no. 504978936). This research was further funded by the Office of the Director of the National Institutes of Health's Early Investigator Award (1DP5OD029506-01, M.E.-A.). Compute for this project was performed on the Calc cluster at Carnegie Science and the Stanford SC Compute Cluster.

References

[1] Eréndira Aceves-Bueno, Adeyemi S Adeleye, Darcy Bradley, W Tyler Brandt, Patrick Callery, Marina Feraud, Kendra L Garner, Rebecca Gentry, Yuxiong Huang, Ian McCullough, et al. Citizen science as an approach for overcoming insufficient monitoring and inadequate stakeholder buy-in in adaptive management: criteria and evidence. *Ecosystems*, 18:493–506, 2015.

[2] Ofer Arazy and Dan Malkinson. A framework of observer-based biases in citizen science biodiversity monitoring: Semi-structuring unstructured biodiversity monitoring protocols. *Frontiers in Ecology and Evolution*, 9:693602, 2021.

[3] Louis J Backstrom, Corey T Callaghan, Hannah Worthington, Richard A Fuller, and Alison Johnston. Estimating sampling biases in citizen science datasets. *Ibis*, 2024.

[4] Vijay V Barve, Laura Brenskelle, Daijiang Li, Brian J Stucky, Narayani V Barve, Maggie M Hantak, Bryan S McLean, Daniel J Paluh, Jessica A Oswald, Michael W Belitz, et al. Methods for broad-scale plant phenology assessments using citizen scientists' photographs. *Applications in Plant Sciences*, 8(1):e11315, 2020.

[5] Nature Kids BC. Naturekids' inaturalist project. `https://naturekidsbc.ca/inaturalist-project/`, 2024. Accessed: 2024-08-12.

[6] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

[7] Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. 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*, pages 21294–21307, 2022.

[8] Michael W Belitz, Elise A Larsen, Leslie Ries, and Robert P Guralnick. The accuracy of phenology estimators for use with sparsely sampled presence-only observations. *Methods in Ecology and Evolution*, 11(10):1273–1285, 2020.

[9] Charlie Blake, Allison Rhanor, and Cody Pajic. The demographics of citizen science participation and its implications for data quality and environmental justice. *Citizen Science: Theory and Practice*, 2020.

[10] Elizabeth H. Boakes, Gianfranco Gliozzo, Valentine Seymour, Martin Harvey, Chloë Smith, David B. Roy, and Muki Haklay. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Scientific Reports*, 6(1):33051, September 2016. Publisher: Nature Publishing Group.

[11] Elizabeth H Boakes, Philip JK McGowan, Richard A Fuller, Ding Chang-qing, Natalie E Clark, Kim O'Connor, and Georgina M Mace. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS biology*, 8(6):e1000385, 2010.

[12] Emily A Botts, Barend FN Erasmus, and Graham J Alexander. Geographic sampling bias in the south african frog atlas project: implications for conservation planning. *Biodiversity and Conservation*, 20:119–139, 2011.

[13] Diana E Bowler, Corey T Callaghan, Netra Bhandari, Klaus Henle, M Benjamin Barth, Christian Koppitz, Reinhard Klenke, Marten Winter, Florian Jansen, Helge Bruelheide, et al. Temporal trends in the spatial bias of species occurrence records. *Ecography*, 2022(8):e06219, 2022.

[14] Robin J Boyd, Gary D Powney, Fiona Burns, Alain Danet, François Duchenne, Matthew J Grainger, Susan G Jarvis, Gabrielle Martin, Erlend B Nilsen, Emmanuelle Porcher, et al. Robitt: A tool for assessing the risk-of-bias in studies of temporal trends in ecology. *Methods in Ecology and Evolution*, 13(7):1497–1507, 2022.

[15] Eleanor D Brown and Byron K Williams. The potential for citizen science to produce reliable and useful information in ecology. *Conservation Biology*, 33(3):561–569, 2019.

[16] Hillary K Burgess, LB DeBey, HE Froehlich, Natalaie Schmidt, Elli J Theobald, Ailene K Ettinger, Janneke HilleRisLambers, Joshua Tewksbury, and Julia K Parrish. The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation*, 208:113–120, 2017.

[17] Calflora. Calflora: Information on california plants for education, research and conservation. `https://www.calflora.org/`, 2024. Accessed: 2024-08-01.

[18] Corey T Callaghan, Richard E Major, Mitchell B Lyons, John M Martin, John H Wilshire, Richard T Kingsford, and William K Cornwell. Using citizen science data to define and track restoration targets in urban areas. *Journal of Applied Ecology*, 56(8):1998–2006, 2019.

[19] Corey T. Callaghan, Ian Ozeroff, Colleen Hitchcock, and Mark Chandler. Capitalizing on opportunistic citizen science data to monitor urban biodiversity: A multi-taxa framework. *Biological Conservation*, 251:108753, November 2020.

[20] Corey T Callaghan, Alistair GB Poore, Max Hofmann, Christopher J Roberts, and Henrique M Pereira. Large-bodied birds are over-represented in unstructured citizen science data. *Scientific reports*, 11(1):19073, 2021.

[21] Elizabeth J Carlen, Cesar O Estien, Tal Caspi, Deja Perkins, Benjamin R Goldstein, Samantha ES Kreling, Yasmine Hentati, Tyus D Williams, Lauren A Stanton, Simone Des Roches, et al. A framework for contextualizing social-ecological biases in contributory science data. *People and Nature*, 6(2):377–390, 2024.

[22] Curtis Champion, Alistair J Hobday, Sean R Tracey, and Gretta T Pecl. Rapid shifts in distribution and high-latitude persistence of oceanographic habitat revealed using citizen science data from a climate change hotspot. *Global Change Biology*, 24(11):5440–5453, 2018.

[23] Mark Chandler, Linda See, Kyle Copas, Astrid MZ Bonde, Bernat Claramunt López, Finn Danielsen, Jan Kristoffer Legind, Siro Masinde, Abraham J Miller-Rushing, Greg Newman, et al. Contribution of citizen science towards international biodiversity monitoring. *Biological conservation*, 213:280–294, 2017.

[24] Bin Chen, Shengbiao Wu, Yimeng Song, Chris Webster, Bing Xu, and Peng Gong. Contrasting inequality in human exposure to greenspace between cities of global north and global south. *Nature Communications*, 13(1):4636, 2022.

[25] Elijah Cole, Benjamin Deneu, Titouan Lorieul, Maximilien Servajean, Christophe Botella, Dan Morris, Nebojsa Jojic, Pierre Bonnet, and Alexis Joly. The geolifeclef 2020 dataset. *arXiv preprint arXiv:2004.04192*, 2020.

[26] Caren Cooper, Vincent Martin, Omega Wilson, and Lisa Rasmussen. Equitable data governance models for the participatory sciences. *Community Science*, 2(2):e2022CSJ000025, 2023.

[27] Caren B Cooper. 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, 2014.

[28] Myriah L Cornwell and Lisa M Campbell. Co-producing conservation and knowledge: Citizen-based sea turtle monitoring in north carolina, usa. *Social Studies of Science*, 42(1):101–120, 2012.

[29] Jason R Courter, Ron J Johnson, Claire M Stuyck, Brian A Lang, and Evan W Kaiser. Weekend bias in citizen science data reporting: implications for phenology studies. *International journal of biometeorology*, 57:715–720, 2013.

[30] Theresa M. Crimmins, Erin Posthumus, Sara Schaffer, and Kathleen L. Prudic. COVID-19 impacts on participation in large scale biodiversity-themed community science projects in the United States. *Biological Conservation*, 256:109017, April 2021.

[31] Martin Dallimer and Niels Strange. Why socio-political borders and boundaries matter in conservation. *Trends in Ecology & Evolution*, 30(3):132–139, 2015.

[32] Karis A Daniel and Leslie G Underhill. Temporal dimensions of data quality in bird atlases: The case of the second southern african bird atlas project. *Citizen Science: Theory and Practice*, 8(1), 2023.

[33] Riccardo de Lutio, John Y Park, Kimberly A Watson, Stefano D'Aronco, Jan D Wegner, Jan J Wieringa, Melissa Tulig, Richard L Pyle, Timothy J Gallaher, Gillian Brown, et al. The herbarium 2021 half–earth challenge dataset and machine learning competition. *Frontiers in Plant Science*, 12:787127, 2022.

[34] Charl Deacon, Suvania Govender, and Michael J Samways. Overcoming biases and identifying opportunities for citizen science to contribute more to global macroinvertebrate conservation. *Biodiversity and Conservation*, 32(6):1789–1806, 2023.

[35] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[36] Vincent Devictor, Robert J Whittaker, and Coralie Beltrame. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and distributions*, 16(3):354–362, 2010.

[37] Grace J Di Cecco, Vijay Barve, Michael W Belitz, Brian J Stucky, Robert P Guralnick, and Allen H Hurlbert. Observing the observers: How participants contribute data to inaturalist and implications for biodiversity science. *BioScience*, 71(11):1179–1188, 2021.

[38] Grace J Di Cecco, Vijay Barve, Michael W Belitz, Brian J Stucky, Robert P Guralnick, and Allen H Hurlbert. Observing the observers: How participants contribute data to inaturalist and implications for biodiversity science. *BioScience*, 71(11):1179–1188, 2021.

[39] Monica Dimson and Thomas W. Gillespie. Who, where, when: Observer behavior influences spatial and temporal patterns of iNaturalist participation. *Applied Geography*, 153:102916, 2023.

[40] Jonathan P Drury, Morgan Barnes, Ann E Finneran, Maddie Harris, and Gregory F Grether. Continent-scale phenotype mapping using citizen scientists' photographs. *Ecography*, 42(8):1436–1445, 2019.

[41] Diego Ellis-Soto, Melissa Chapman, and Dexter H Locke. Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the united states. *Nature Human Behaviour*, 7(11):1869–1877, 2023.

[42] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[43] Brian J Enquist, Xiao Feng, Brad Boyle, Brian Maitner, Erica A Newman, Peter Møller Jørgensen, Patrick R Roehrdanz, Barbara M Thiers, Joseph R Burger, Richard T Corlett, et al. The commonness of rarity: Global and future distribution of rarity across land plants. *Science advances*, 5(11):eaaz0414, 2019.

[44] The Global Biodiversity Information Facility. Gbif: The global biodiversity information facility (2024) what is gbif? `https://www.gbif.org/what-is-gbif`, 2024. Accessed: 2024-07-30.

[45] Daniel Fink, Alison Johnston, Matt Strimas-Mackey, Tom Auer, Wesley M Hochachka, Shawn Ligocki, Lauren Oldham Jaromczyk, Orin Robinson, Chris Wood, Steve Kelling, et al. A double machine learning trend model for citizen science data. *Methods in Ecology and Evolution*, 14(9):2435–2448, 2023.

[46] Amélie Fontaine, Anouk Simard, Nicolas Brunet, and Kyle H Elliott. Scientific contributions of citizen science applied to rare or threatened animals. *Conservation Biology*, 36(6):e13976, 2022.

[47] ML Forister, CA Halsch, CC Nice, JA Fordyce, TE Dilts, JC Oliver, KL Prudic, AM Shapiro, JK Wilson, and J Glassberg. Fewer butterflies seen by community scientists across the warming and drying landscapes of the american west. *Science*, 371(6533):1042–1045, 2021.

[48] BC Parks Foundation. inaturalist. `https://bcparksfoundation.ca/projects/wildlife-forever/bc-parks-inaturalist-project/`, 2024. Accessed: 2024-08-11.

[49] Travis Gallo and Damon Waitt. Creating a successful citizen science model to detect and report invasive species. *BioScience*, 61(6):459–465, 2011.

[50] Camille Garcin, Alexis Joly, Pierre Bonnet, Antoine Affouard, Jean-Christophe Lombardo, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, and Joseph Salmon. Plantnet-300k: a plant image dataset with high label ambiguity and a long-tailed distribution. In *NeurIPS Datasets and Benchmarks*, 2021.

[51] Milan Gazdic and Quentin Groom. inaturalist is an unexploited source of plant-insect interaction data. *Biodiversity Information Science and Standards*, (41), 2019.

[52] Jonas Geldmann, Jacob Heilmann-Clausen, Thomas E. Holm, Irina Levinsky, Bo Markussen, Kent Olsen, Carsten Rahbek, and Anders P. Tøttrup. What determines spatial bias in citizen science? exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22(11):1139–1149, 2016.

[53] Lauren E. Gillespie, Megan Ruffley, and Moises Exposito-Alonso. Deep learning models map rapid plant species changes from citizen science and remote sensing data. *Proceedings of the National Academy of Sciences*, 2024.

[54] Krishna S Girish and Umesh Srinivasan. Community science data provide evidence for upward elevational range shifts by eastern himalayan birds. *Biotropica*, 54(6):1457–1465, 2022.

[55] Hervé Goëau, Pierre Bonnet, and Alexis Joly. 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, pages 1972–1981, 2023.

[56] Kristina Gratzer and Robert Brodschneider. How and why beekeepers participate in the insignia citizen science honey bee environmental monitoring project. *Environmental Science and Pollution Research*, 28(28):37995–38006, 2021.

[57] TL Hawthorne, V Elmore, A Strong, P Bennett-Martin, J Finnie, J Parkman, T Harris, J Singh, L Edwards, and J Reed. Mapping non-native invasive species and accessibility in an urban forest: A case study of participatory mapping and citizen science in atlanta, georgia. *Applied Geography*, 56:187–198, 2015.

[58] Thomas Hiller and Danny Haelewaters. A case of silent invasion: Citizen science confirms the presence of harmonia axyridis (coleoptera, coccinellidae) in central america. *PloS one*, 14(7):e0220082, 2019.

[59] Joshua J Horns, Frederick R Adler, and Çağan H Şekercioğlu. Using opportunistic citizen science data to estimate avian population trends. *Biological conservation*, 221:151–159, 2018.

[60] Andy V. Huynh, Lauren E. Gillespie, Jael Lopez-Saucedo, Claire Tang, Rohan Sikand, and Moisés Expósito-Alonso. 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*, 2024.

[61] iNaturalist. Year in review 2023. `https://www.inaturalist.org/stats/2023`, 2023. Accessed: 2024-07-30.

[62] iNaturalist. inaturalist open data. `https://github.com/inaturalist/inaturalist-open-data`, 2024. Accessed: 2024-07-09.

[63] Nick JB Isaac and Michael JO Pocock. Bias and information in biological records. *Biological Journal of the Linnean Society*, 115(3):522–531, 2015.

[64] Nick JB Isaac, Arco J van Strien, Tom A August, Marnix P de Zeeuw, and David B Roy. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, 5(10):1052–1060, 2014.

[65] Clemens Jacobs and Alexander Zipf. Completeness of citizen science biodiversity data from a volunteered geographic information perspective. *Geo-spatial Information Science*, 20(1):3–13, 2017.

[66] Alison Johnston, Eleni Matechou, and Emily B Dennis. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, 14(1):103–116, 2023.

[67] Alison Johnston, Nick Moran, Andy Musgrove, Daniel Fink, and Stephen R Baillie. Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, 422:108927, 2020.

[68] Johannes Kamp, Steffen Oppel, Henning Heldbjerg, Timme Nyegaard, and Paul F Donald. Unstructured citizen science data fail to detect long-term population declines of common birds in denmark. *Diversity and Distributions*, 22(10):1024–1035, 2016.

[69] Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The caltech fish counting dataset: A benchmark for multiple-object tracking and counting. In *European Conference on Computer Vision (ECCV)*, 2022.

[70] Keidai Kishimoto and Hiromi Kobori. Covid-19 pandemic drives changes in participation in citizen science project "city nature challenge" in tokyo. *Biological Conservation*, 255:109001, 2021.

[71] . Kshitiz, Sonu Shreshtha, Bikash Dutta, Muskan Dosi, Mayank Vatsa, Richa Singh, Saket Anand, Sudeep Sarkar, and Sevaram Mali Parihar. Birdcollect: A comprehensive benchmark for analyzing dense bird flock attributes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21879–21887, 2024.

[72] Frank A La Sorte, Jeremy M Cohen, and Walter Jetz. Data coverage, biases, and trends in a global citizen-science resource for monitoring avian diversity. *Diversity and Distributions*, page e13863, 2024.

[73] Jae Joong Lee, Bosheng Li, Sara Beery, Jonathan Huang, Songlin Fei, Raymond A. Yeh, and Bedrich Benes. Tree-d fusion: Simulation-ready tree dataset from single images with diffusion priors, 2024.

[74] Maiju Lehtiniemi, Okko Outinen, and Riikka Puntila-Dodd. Citizen science provides added value in the monitoring for coastal non-indigenous species. *Journal of Environmental Management*, 267:110608, 2020.

[75] Misha Leong and Michelle Trautwein. A citizen science approach to evaluating us cities for biotic homogenization. *PeerJ*, 7:e6879, 2019.

[76] Meixi Lin, Ariel Levi Simons, Ryan J Harrigan, Emily E Curd, Fabian D Schneider, Dannise V Ruiz-Ramos, Zack Gold, Melisa G Osborne, Sabrina Shirazi, Teia M Schweizer, et al. Landscape analyses using edna metabarcoding and earth observation predict community biodiversity in california. *Ecological Applications*, 31(6):e02379, 2021.

[77] Bianca Lopez, Emily Minor, and Andrew Crooks. Insights into human-wildlife interactions in cities from bird sightings recorded online. *Landscape and Urban Planning*, 196:103742, 2020.

[78] Scott R Loss, Sara S Loss, Tom Will, and Peter P Marra. Linking place-based citizen science with large-scale conservation research: a case study of bird-building collisions and the role of professional scientists. *Biological Conservation*, 184:439–445, 2015.

[79] Ciarán Mac Domhnaill, Seán Lyons, and Anne Nolan. The citizens in citizen science: demographic, socioeconomic, and health characteristics of biodiversity recorders in ireland. *Citizen Science: Theory and Practice*, 5(1), 2020.

[80] Dillon Mahmoudi, Chris L Hawn, Erica H Henry, Deja J Perkins, Caren B Cooper, and Sacoby M Wilson. Mapping for whom? communities of color and the citizen science gap. *ACME: An International Journal for Critical Geographies*, 21(4):372–388, 2022.

[81] Oscar H. Marín-Gómez, Claudia Rodríguez Flores, and María del Coro Arizmendi. Assessing ecological interactions in urban areas using citizen science data: Insights from hummingbird–plant meta-networks in a tropical megacity. *Urban Forestry & Urban Greening*, 74:127658, August 2022.

[82] Sabrina McCormick. After the cap: risk assessment, citizen science and disaster recovery. *Ecology and society*, 17(4), 2012.

[83] Elaine McGoff, Francesca Dunn, Luis Moliner Cachazo, Penny Williams, Jeremy Biggs, Pascale Nicolet, and Naomi C Ewald. Finding clean water habitats in urban landscapes: professional researcher vs citizen science approaches. *Science of the Total Environment*, 581:105–116, 2017.

[84] Duncan C McKinley, Abe J Miller-Rushing, Heidi L Ballard, Rick Bonney, Hutch Brown, Susan C Cook-Patton, Daniel M Evans, Rebecca A French, Julia K Parrish, Tina B Phillips, et al. Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological conservation*, 208:15–28, 2017.

[85] Pietro Milanesi, Emiliano Mori, and Mattia Menchetti. Observer-oriented approach improves species distribution models from citizen science data. *Ecology and Evolution*, 10(21):12104–12114, 2020.

[86] Nicolas Moulin. When citizen science highlights alien invasive species in france: the case of indochina mantis, hierodula patellifera (insecta, mantodea, mantidae). *Biodiversity Data Journal*, 8, 2020.

[87] Haowei Mu, Xuecao Li, Yanan Wen, Jianxi Huang, Peijun Du, Wei Su, Shuangxi Miao, and Mengqing Geng. A global record of annual terrestrial human footprint dataset from 2000 to 2018. *Scientific Data*, 9(1):176, 2022.

[88] Freddy Nachtergaele, Harrij van Velthuizen, Luc Verelst, Dave Wiberg, Matieu Henry, Frederica Chiozza, Yusuf Yigini, Ece Aksoy, Niels Batjes, Enoch Boateng, et al. *Harmonized world soil database version 2.0*. FAO, 2023.

[89] Naturalist. Naturalist. `https://www.inaturalist.org`, 2024. Accessed: 2024-07-30.

[90] Tim Newbold, Lawrence N Hudson, Andrew P Arnell, Sara Contu, Adriana De Palma, Simon Ferrier, Samantha LL Hill, Andrew J Hoskins, Igor Lysenko, Helen RP Phillips, et al. Has land use pushed terrestrial biodiversity beyond the planetary boundary? a global assessment. *Science*, 353(6296):288–291, 2016.

[91] Thomas Neyens, Peter J Diggle, Christel Faes, Natalie Beenaerts, Tom Artois, and Emanuele Giorgi. Mapping species richness using opportunistic samples: a case study on ground-floor bryophyte species richness in the belgian province of limburg. *Scientific Reports*, 9(1):19122, 2019.

[92] J.M. Omernik. Ecoregions of the conterminous united states. *Annals of the Association of American Geographers*, 77:118–125, 1987.

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

[94] Omiros Pantazis, Gabriel J Brostow, Kate E Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 10583–10592, 2021.

[95] Daniel S Park, Erica A Newman, and Ian K Breckheimer. Scale gaps in landscape phenology: challenges and opportunities. *Trends in Ecology & Evolution*, 36(8):709–721, 2021.

[96] Camille Parmesan and Gary Yohe. A globally coherent fingerprint of climate change impacts across natural systems. *nature*, 421(6918):37–42, 2003.

[97] Rachel Mary Pateman, Alison Dyke, and Sarah Elizabeth West. The diversity of participants in environmental citizen science. *Citizen Science: Theory and Practice*, 2021.

[98] Nadja Pernat, Helge Kampen, Jonathan M Jeschke, and Doreen Werner. Citizen science versus professional data collection: Comparison of approaches to mosquito monitoring in germany. *Journal of Applied Ecology*, 58(2):214–223, 2021.

[99] Michael JO Pocock, Mark Chandler, Rick Bonney, Ian Thornhill, Anna Albin, Tom August, Steven Bachman, Peter MJ Brown, Davi Gasparini Fernandes Cunha, Audrey Grez, et al. A vision for global biodiversity monitoring with citizen science. In *Advances in ecological research*, volume 59, pages 169–223. Elsevier, 2018.

[100] Christophe F Randin, Michael B Ashcroft, Janine Bolliger, Jeannine Cavender-Bares, Nicholas C Coops, Stefan Dullinger, Thomas Dirnböck, Sandra Eckert, Erle Ellis, Néstor Fernández, et al. Monitoring biodiversity in the anthropocene using remote sensing in species distribution models. *Remote sensing of environment*, 239:111626, 2020.

[101] Giovanni Rapacciuolo, Alison Young, and Rebecca Johnson. Deriving indicators of biodiversity change from unstructured community-contributed data. *Oikos*, 130(8):1225–1239, 2021.

[102] Christopher G Reddick, Roger Enriquez, Richard J Harris, and Bonita Sharma. Determinants of broadband access and affordability: An analysis of a community survey on the digital divide. *Cities*, 106:102904, 2020.

[103] Orin J Robinson, Viviana Ruiz-Gutierrez, Daniel Fink, Robert J Meese, Marcel Holyoak, and Evan G Cooch. Using citizen science data in integrated population models to inform conservation. *Biological Conservation*, 227:361–368, 2018.

[104] Connor J Rosenblatt, Ashley A Dayer, Jennifer N Duberstein, Tina B Phillips, Howard W Harshaw, David C Fulton, Nicholas W Cole, Andrew H Raedeke, Jonathan D Rutter, and Christopher L Wood. Highly specialized recreationists contribute the most to the citizen science project ebird. *Ornithological Applications*, 124(2):duac008, 2022.

[105] Boris Sakschewski, Werner Von Bloh, Alice Boit, Lourens Poorter, Marielos Peña-Claros, Jens Heinke, Jasmin Joshi, and Kirsten Thonicke. Resilience of amazon forests emerges from plant trait diversity. *Nature climate change*, 6(11):1032–1036, 2016.

[106] Lina María Sánchez-Clavijo, Sindy Jineth Martínez-Callejas, Orlando Acevedo-Charry, Angélica Diaz-Pulido, Bibiana Gómez-Valencia, Natalia Ocampo-Peñuela, David Ocampo, María Helena Olaya-Rodríguez, Juan Carlos Rey-Velasco, Carolina Soto-Vargas, et al. Differential reporting of biodiversity in two citizen science platforms during covid-19 lockdown in colombia. *Biological Conservation*, 256:109077, 2021.

[107] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, and Nathan Jacobs. 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*, pages 7136–7145, 2024.

[108] Stephen DA Smith and Matt J Nimbs. Citizen scientists record significant range extensions for tropical sea slug species in subtropical eastern australia. *Diversity*, 14(4):244, 2022.

[109] California Native Plant Society. Cnps field protocols and guidelines. `https://www.cnps.org/plant-science/field-protocols-guidelines`, 2022. Accessed: 2024-08-02.

[110] National Audubon Society. The christmas bird count historical results. `http://www.christmasbirdcount.org`, 2020. Accessed: 2024-08-03.

[111] Daniela Soleri, Jonathan W Long, Mónica D Ramirez-Andreotta, Rose Eitemiller, and Rajul Pandya. Finding pathways to more equitable and meaningful public-scientist partnerships. *Citizen Science: Theory and Practice*, 1(1):9–9, 2016.

[112] Valerie A Steen, Chris S Elphick, and Morgan W Tingley. An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions*, 25(12):1857–1869, 2019.

[113] Valerie A Steen, Morgan W Tingley, Peter WC Paton, and Chris S Elphick. Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods in Ecology and Evolution*, 12(2):216–226, 2021.

[114] Ryan Steiner, Gerald Niemi, Frank Nicoletti, Mary Jane Evans, Edmund Zlonis, and Matthew A Etterson. Changes in survey effort can influence conclusions about migration phenology. *Journal of Raptor Research*, 56(2):171–179, 2022.

[115] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024.

[116] Brian L Sullivan, Tina Phillips, Ashley A Dayer, Christopher L Wood, Andrew Farnsworth, Marshall J Iliff, Ian J Davies, Andrea Wiggins, Daniel Fink, Wesley M Hochachka, et al. Using open access observational data for conservation action: A case study for birds. *Biological Conservation*, 208:5–14, 2017.

[117] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009.

[118] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1):1–14, 2015.

[119] Fabio ST Sweet, Thomas Rödl, and Wolfgang W Weisser. Covid-19 lockdown measures impacted citizen science hedgehog observation numbers in bavaria, germany. *Ecology and Evolution*, 12(6):e8989, 2022.

[120] North American Land Change Monitoring System, Canada Centre for Remote Sensing (CCRS), U.S. Geological Survey (USGS), Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), Comisión Nacional Forestal (CONAFOR), and Instituto Nacional de Estadística y Geografía (INEGI). 2020 land cover of north america at 30 meters.

[121] Becky Tang, James S Clark, and Alan E Gelfand. Modeling spatially biased citizen science effort through the ebird database. *Environmental and Ecological Statistics*, 28(3):609–630, 2021.

[122] Mélisande Teng, Amna Elmustafa, Benjamin Akera, Yoshua Bengio, Hager Radi, Hugo Larochelle, and David Rolnick. Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data. *Advances in Neural Information Processing Systems*, 36, 2024.

[123] Ellinore J Theobald, Ailene K Ettinger, Hillary K Burgess, Lauren B DeBey, Natalie R Schmidt, Halley E Froehlich, Christian Wagner, Janneke HilleRisLambers, Joshua Tewksbury, Melanie A Harsch, et al. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, 181:236–244, 2015.

[124] Camille Van Eupen, Dirk Maes, Marc Herremans, Kristijn RR Swinnen, Ben Somers, and Stijn Luca. The impact of data quality filtering of opportunistic citizen science data on species distribution model performance. *Ecological Modelling*, 444:109453, 2021.

[125] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021.

[126] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[127] Arco J Van Strien, Chris AM Van Swaay, and Tim Termaat. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, 50(6):1450–1458, 2013.

[128] Nature Vancouver. inaturalist ecowalk: Learn how to use inaturalist. `https://naturevancouver.ca/events/inaturalist-ecowalk-learn-how-to-use-inaturalist/`, 2022. Accessed: 2024-08-12.

[129] Jay M Ver Hoef, Devin Johnson, Robyn Angliss, and Matt Higham. Species density models from opportunistic citizen science data. *Methods in Ecology and Evolution*, 12(10):1911–1925, 2021.

[130] Jacob Walker and Philip D Taylor. Using ebird data to model population change of migratory bird species. *Avian Conservation & Ecology*, 12(1), 2017.

[131] Hongchang Wang, Huaxiang Lu, Huimin Guo, Haifang Jian, Chuang Gan, and Wu Liu. Bird-count: a multi-modality benchmark and system for bird population counting in the wild. *Multimedia Tools and Applications*, 82(29):45293–45315, 2023.

[132] Ben G Weinstein, Sergio Marconi, Stephanie A Bohlman, Alina Zare, Aditya Singh, Sarah J Graves, and Ethan P White. A remote sensing derived data set of 100 million individual tree crowns for the national ecological observatory network. *Elife*, 10:e62922, 2021.

[133] Victoria Werenkraut, Florencia Baudino, and Helen E Roy. Citizen science reveals the distribution of the invasive harlequin ladybird (harmonia axyridis pallas) in argentina. *Biological Invasions*, 22(10):2915–2921, 2020.

[134] Thomas Wesener. An integrative and citizen science based approach to the rediscovery and redescription of the only known high-altitude endemic pill millipede, glomeris aurita koch (diplopoda, glomerida). *PeerJ*, 6:e5569, 2018.

[135] Joseph S Wilson, Aaron D Pan, David Emmanuel M General, and Jonathan B Koch. More eyes on the prize: an observation of a very rare, threatened species of philippine bumble bee, bombus irisanensis, on inaturalist and the importance of citizen science in conservation biology. *Journal of Insect Conservation*, 24:727–729, 2020.

[136] WorldClim. Worldclim dataset. `https://worldclim.org/`, 2024. Accessed: 2024-07-29.

[137] Yi Yang, David Tilman, George Furey, and Clarence Lehman. Soil carbon sequestration accelerated by restoration of grassland biodiversity. *Nature communications*, 10(1):718, 2019.

[138] Alexander Zizka, Alexandre Antonelli, and Daniele Silvestro. Sampbias, a method for quantifying geographic sampling biases in species distribution data. *Ecography*, 44(1):25–32, 2021.

5 Appendix

5.1 Extended Related Works

5.1.1 Participatory Science for Biodiversity Monitoring

There are a plethora of types of participatory science collection strategies for biodiversity monitoring (also known as citizen science and community science). Briefly, these include observation platforms like the Global Biodiversity Information Facility [44] which allow researchers and registered members of the public to upload geolocated and timestamped observations for both individual species observations and community checklists, including specifically for plants [17] and birds [110]; specific easy-to-use apps for expediting similar types of collections like iNaturalist [89], Pl@ntNet [50], and eBird [117] that allow users to upload geolocated and timestamped photos of individuals or checklists in real-time, identify them, and share them to publicly; and more structured and specialized checklists like relevés [109], targeted collection campaigns focused on specific taxa [52], and eDNA soil collection campaigns [76].

Community engagement projects built around these strategies have in turn enabled a wide array of novel and impactful biodiversity monitoring breakthroughs such as monitoring species and habitats [28, 46, 74, 57, 18], tracking invasive species spread [133, 86, 58, 49], detecting new populations of species [135, 108], rediscovering cryptic species [134], quantifying and monitoring species richness [91, 19], quantifying anthropogenic biodiversity changes [75, 47, 54, 101, 22], understanding species interactions [51, 77, 81], characterizing within-species diversity and behavior [40, 4], estimating species' population sizes [129, 59, 130, 127], tracking ecological disaster recovery efforts [82], and aiding conservation decisions [116, 103, 78]. These projects are now considered to be an essential tool for reaching conservation goals across the world [15, 36, 23, 1, 123, 99, 84]. From these data collections and sampling strategies, we focus specifically on iNaturalist as it is the largest data collection that has linked images for almost every observation (except for some bird observations identified by an audio recording of their call) [61].

5.1.2 Definitions of Bias in Biodiversity Data

The taxonomy, attribution of, and even fundamental definitions of bias in biodiversity data is an active area of study [21, 63, 64, 38]. Isaac et. al. classified define bias as a property of the observation sampling process, specifically as "variation in recorder activity" [64], and acknowledged four forms of bias: non-biological variation in number of observations over time, non-biological variation across space, variation in observation collection effort per-visit, and variation in detectability of organisms [64, 63]. Meanwhile, Di Cecco et. al. partitions biases into spatial, temporal, taxonomic, and user activity level bias [38]. Lastly, Carlen et. al. defines bias as "an uneven or disproportionate representation of a particular subject or variable within the larger group" [21], and further categorizes biases that affect observers (referred to as "filters", namely participation, detection, sampling, and preference) and the downstream biases resultant in biodiversity data (such as spatial and temporal) [21]. Carlen et. al. explicitly highlight how sociopolitical biases (referred to as "unconscious bias") strongly affect the participation filter [21], and importantly Carlen et. al. acknowledges that the there are further intersectional interactions between these biases [21]. For the purposes of this work, we adopt the definitions of spatial, temporal, and observation quality bias from Di Cecco et. al. and additionally include effects of the participation filter from Carlen et. al. as a fourth bias, named sociopolitical bias.

5.1.3 Biases in Volunteer-Collected Biodiversity Datasets

Drivers of spatial bias include participants sampling closer to home [56, 83], differential preferences for protected versus urban spaces [3, 72, 38, 39], and access to greenspaces [21]. Without proper care, these spatial biases can in turn affect inferences about demographic changes [11, 3], biodiversity changes [101], and the utility of these data for conservation planning [12]. Various methods have been proposed and tested to mitigate these effects [138, 65], mainly for species distribution modeling [124, 113, 112, 121, 67].

Drivers of temporal bias include the year-over-year rise in popularity of participatory science [37, 3, 39], relative ease of observing on weekends versus the workweek [38, 29, 27], the COVID-19 pandemic [119, 106, 30], and the City Nature Challenge [38, 61]. These temporal biases can make it difficult to accurately assess bird migration patterns [114], changes in species distributions [13, 32], population declines [68], and flowering time [95, 8] from these data. Methods do exist to mitigate these effects [14], mainly for estimating demographic changes over time [3, 45].

Observation quality—defined here as how representative a collection of observations are of the underlying biodiversity of an area—are driven in part by who is observing, thus *observer quality* can strongly influence the observed biodiversity. For bird surveys, a small but highly-specialized subset of observers contribute the most [104], and more generally more active users tend to observe more species in more diverse habitats [38]. Observer behavior and observation quality also differ based on whether observers are local residents or visitors [39]. Filtering observations from the most active users or their most active days [124, 85, 10] is the main approach for mitigating these effects currently.

Lastly, sociopolitical factors influence who observes where, including a skew towards whiter, wealthier, older, and more educated observers [79, 97] and fewer observations in areas and communities of environmental justice concern [9], fewer observers in historically redlined districts or communities of color in the U.S. [41, 80], fewer observations in lower GDP countries [34], differential access to green spaces [24], and conservation and land management policy differences across political boundaries [31]. However, little work has been done to account for these differences outside of calls for more broad structural reform of participatory science [21, 16, 26, 111, 93]. The various biases in volunteer-collected biodiversity datasets mean that the potential of participatory science has yet to be fully realized [12], and in this work we specifically we focus on how they negatively impact computer vision model performance.

5.2 The DivShift Framework

The bioDiversity Shift (DivShift) framework casts the effect of performance changes due to biases present in volunteer-collected biodiversity datasets as a problem of distribution shift.

Briefly, 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 subpartitions P_{Atrain} and P_{Atest} where again $P_{Atest} \bigcap P_{Atrain} = \emptyset$. The size of these subpartitions is arbitrary, but a standard protocol in computer vision is to assign 80% of P_A to P_{atrain} and the remaining 20% to P_{atest}, which we follow here. Each of these individual subpartitions can then be seen as a finite sample from the joint distribution over inputs and labels, e.g. $P_A \overset{S}{\sim} J(x, y)$. However, as a finite sample, the partitions P_{Atrain} and P_{Atest} won't be identical, and thus there exists a ceiling for models trained to estimate the distribution P_{Atrain} when tested on P_{Atest} driven by this divergence between these two finite distributions. If the sampling process $\overset{S}{\sim}$ for P_A and P_B is identical, the same ceiling will summarily hold for P_B and they are considered *in-domain*, but when the sampling process is biased, $\overset{S_b}{\sim}$, for example due to selective behavior by observers in what species, 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 $J(x, y)$ is the same. These partitions P_A and P_B will be out-of-distribution and any model trained on P_{Atrain} will exhibit decreased performance relative to P_{Btrain} below the in-domain ceiling. Furthermore, if a model trained on P_{Atrain} is well-calibrated, then these performance decreases will mirror the distribution divergence between P_A and P_B. Models that are poorly calibrated can then err one of two ways, either being overfitting manifested and accuracy decreases more than the expected distribution shift, or being more generalist and accuracy decreases are less bad than expected. Whether a model performs better or worse than the expectation 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 distribution $J(x, y)$. When our distribution which for the purposes of this work represents the general underlying distribution of biodiversity across the planet.

Whether these biases are harmful or helpful is the key question we aim to answer in this work, at least in the setting where $J(x, y)$ represents the distribution of biodiversity across the planet, $\overset{S_b}{\sim}$ represents volunteer collectors, and D represents the iNaturlist dataset. To quantify these biases, we first assume that the distribution of labels y in P_A and P_B can be used to estimate their joint distributions, and use these labels to estimate the underlying distribution shift between these partitions.

Split	Images	Obs.	Labeled	Species
Spatial Bias				
Modified	6.642M	3.536M	64.70%	7,513
Wilderness	0.141M	0.068M	63.79%	2,395
Temporal Bias				
Out of City Nature	6.986M	3.685M	64.52%	7,604
In City Nature	0.362M	0.220M	67.98%	3,929
Observation Quality Bias				
Engaged	3.476M	1.697M	69.54%	7,361
Casual	1.113M	0.756M	56.94%	5,706
Sociopolitical Bias				
US-AK	0.099M	0.057M	66.37%	875
CAN-YT	0.034M	0.018M	77.81%	746
CAN-BC	1.080M	0.622M	67.39%	2,329
US-WA	0.529M	0.279M	67.02%	2,393
US-OR	0.604M	0.300M	65.62%	2,711
US-CA	4.039M	2.115M	63.00%	4,654
US-NV	0.259M	0.121M	70.18%	1,860
US-AZ	0.497M	0.272M	63.02%	2,191
MX-SO	0.018M	0.010M	57.84%	673
MX-BJ	0.142M	0.090M	68.97%	1,466
MX-BJS	0.046M	0.022M	73.68%	716
Baselines				
iNat21	3.554M	1.937M	100.0%	1,852
iNat21 mini	0.185M	0.109M	100.0%	1,852
ImageNet	1.614M	0.858M	100.0%	1,260
Spatial Stratified	7.348M	3.905M	64.71%	7,607

Table A1: DivShift-NAWC Characteristics by Data Partition. M = Million, Obs. = Observation, US = United States, AK = Alaska, CAN = Canada, YT = Yukon Territories, BC = British Columbia, WA = Washington, OR = Oregon, CA = California, NV = Nevada, AZ = Arizona, MX = Mexico, SO = Sonora, BJ = Baja California, BJS = Baja California Sur.

Namely, we measure the Jensen-Shannon Distance (JSD) between $P_{Atrain}(y)$, and both $P_{Btrain(y)}$, 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 disaligned [42]. Assuming there is no distribution shift (JSD= 0) between P_{Atrain} and P_{Atest}, we then measure the performance decrease between models trained on P_{Atrain} and tested on P_{Atest} to those trained on P_{Atrain} and tested on P_{Btest} and compare those decreases to the JSD between $P_{Atrain}(y)$ and $P_{Btest}(y)$. For any set of partitions where the underlying JSD is smaller than the difference in accuracy for models tested 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 $P_B(y)$. These strong and weak biases are then further categorized into negative and positive bias. Positive bias resulting in a performance increase implies that some structure in the joint distribution $P_A(x, y)$ captures additionally useful information about $P_B(x, y)$ Meanwhile, negative bias resulting in a performance decrease implies $P_A(x, y)$ lacks critical information about the distribution of $P_B(x, y)$.

5.3 Building the DivShift-NAWC Dataset

Here we describe the processes used to generate the DivShift-NAWC dataset, and further explain choices on why and how bias splits were chosen.

5.3.1 Dataset Download and Observation Cleaning

Observations were downloaded from the iNaturalist Open Data repository [62]. Only research-grade or observations in need of ID were kept. Observations were further filtered to those with a positional accuracy of under 120 m to ensure that spatial associations with geographic variables like climate and habitat type were accurate. Spatial and temporal biases can be taxa-specific [27, 4], thus given that many plant communities have been undersampled in the past [38], we chose to only work with observations of plants, specifically vascular plants (tracheophyta). After filtering to vascular plants, we only kept observations from the years 2019-2023 that fell within the administrative boundaries of the states of Alaska, Yukon, British Columbia, Washington, Oregon, California, Nevada, Arizona, Baja California, Baja California Sur, and Sonora. We further rolled up subspecies, varieties, and phenotypes to the species level to ensure a more uniform intra-class diversity. Lastly, we removed any species not observed in at least two years and only kept species with at least 15 observations. This left us with 3.9 million unique observations, of which 64% are research grade and labeled with 7,607 unique species (Table A1). For each of these observations, we downloaded all available photos per-observation from the iNaturalist Open Data Repository [62], leaving us with 7.3 million unique images of plants (Table A1).

iNaturalist data provides crucial and useful information about each image, such as the latitude, longitude, date, and observer [61]. Using this information, for each observation we added more geologically-relevant data for each image, specifically L2 and L3 ecoregion, 19 current-day WorldClim bioclim variables, land use type, soil type, and Human Footprint data [92, 136, 120, 88, 87].

5.3.2 Spatial Partition: Human Footprint

Human influence on biodiversity is widespread across the planet, especially near urban areas, with the most undisturbed areas focused in the polar regions. However, these wilder regions are also harder to reach, making it difficult for volunteers to collect imagery there. Using the Global Human Footprint Index (HFI) [87], we partition DivShift-NAWC into wilderness (HFI <= 1) and highly modified observations (HFI >= 4). Interestingly, over 6.6 million of the 7.3 million images in DivShift-NAWC are from highly human-modified regions while only about 141,000 are from minimally-modified wilderness (Table A1).

5.3.3 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 [37, 61] and altered observer behavior, as volunteers are aiming to maximize their number of observations and unique species. While the majority of iNaturalist photos are taken outside of this challenge, a higher proportion of observations from the the City Nature challenge are labeled, and 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 A1). For this partition, we consider all observations taken during official City Nature Challenge (CNC) dates for the four years of study as one train / test partition, and all other observations as the other.

5.3.4 Quality Partition: Observer Engagement

Since iNaturalist observations are collected by volunteers with differing amounts of enthusiasm, time, and resources [79, 9], observer engagement varies widely between observers. Following observations from [38] that observers with more observations tend to observe a wider diversity of species in more diverse habitats, we also partition DivShift-NAWC into partitions based on user engagement, with the casual partition consisting of all observations from observers with fewer than 50 total valid observations, and the engaged partition as observations from observers with more than 1,000 total observations [37].

5.3.5 Sociopolitical Partition: Administrative Boundaries

Where certain plant species can grow are demarcated by ecological boundaries, and similarly volunteers observation trends are demarcated by political boundaries. iNaturalist is based in California, the state with half of the images in DivShift-NAWC. It similarly has a larger human population and a higher proportion of people with access to smartphones and expendable cellular data to collect biodiversity images [102]. British Columbia meanwhile has implemented many programs encouraging the use of iNaturalist [5, 128, 48]. However, the reach of these community-based programs often are cut off by political boundaries. For example, this stark effect can be seen in the abundance of observations between the U.S. and Mexico, especially between the border of Arizona and Sonora, which are the same ecosystem yet Sonora has 3.6% of the observations that Arizona has (Table A1). Similar differences can be observed between southern British Columbia, Yukon, and Alaska. To test the effects of these ecologically arbitrary political boundaries, we partition each state, then train models separately on the British Columbia and California train partitions, then test these models on all nearby states (Alaska, Washington, Oregon, Yukon, and California for British Columbia-trained models, and British Columbia, Washington, Oregon, Arizona, Nevada, Baja California, and Baja California Sur for California-trained models).

5.3.6 Baseline Partitions

Lastly, we compare the absolute accuracy of these various 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 [125] by only keeping species with at least 50 observations from 10 unique observers and species with at least 60 overall observations, selecting up to 310 observations one observer at a time, then further randomly partitioning these random 60-310 observations into 10 validation images and 50-300 test images. For the test set, we used observations from September 25th of 2019 to September 25th of 2020 as test data and iteratively sampling one observation per observer until 50 images are reached (for an even 50 test observations per-species). We also recreated the iNat2021mini train partition by randomly subsampling exactly 50 images per-species from the train set. We also considered a spatial stratification partitioning strategy [25, 60], where we partitioned the study area into a 50 x 50 km grid. The DivShift-NAWC images were then split into train and test depending on what labeled grid cell they fell between. We also recreated the Imagenet train / test partitioning strategy [35], namely only keeping species with at least 850 observations, then for each species randomly selecting images such that 100 images for each species are test images, 50 are validation, and the remaining images are train up to a threshold of 1,300 images. Lastly, we consider a naive 80 / 20 train / test split partition.

5.4 Dataset Licensing and Reuse

Images and observations available through the iNaturalist Open Data program include data with Creative Commons licenses range from CC-BY-NC, CC-BY-NC-SA, CC-BY-ND, CC0, CC-BY-SA, CC-BY, to CC-BY-NC-ND. These images may be reused for non-commercial purposes and by associativity, the DivShift-NAWC dataset is therefore free and open for research purposes and will be made publicly available along with the associated code to build the dataset and train the models. Individual images can be reproduced with proper attribution given per-image, depending on a photo's given license. License information is provided in the DivShift-NAWC dataset under the column titled "license".

5.5 Measuring Distribution Shift with Jensen-Shannon Distance

For each four of the four bias partitions (namely spatial, temporal, sociopolitical, observation quality), we measured the Jensen-Shannon Distance (JSD) between the train set of one partition (e.g. for the spatial bias partition, observations in wilderness areas) to both the train and the test partition of the second partition (e.g. for the spatial bias partition, separately the train and test observations in modified areas). Of the available statistical distance metrics, we chose to report JSD as it has many desirable properties, namely that is a symmetric metric (e.g. the distance from P_A to P_B is the same as from from P_B to P_A) and the metric is bounded from 0 to 1 when using a log base of 2, meaning its range can be mapped to the range of differences in accuracies for models trained on these data. JSD was calculated using scipy's distance module's "jensenshannon" function with a log base of 2. JSD was calculated only for classes shared between the corresponding splits.

Model	Top-1-Obs	Top-1-Eco	Top-t-LUC
Spatial	0.663	0.661	0.651
Imagenet	0.695	0.699	0.704
iNat2021	0.68	0.691	0.647
Random	0.706	0.709	0.698
iNatMini	0.337	0.31	0.293

Table A2: Baseline DivShift-NAWC partition performance.

5.6 Model Training and Testing

As the goal of this work is to test distribution shift effects across partitions of these volunteer-collected data as opposed to maximizing predictive performance, for each partition, we train a small computer vision model for a limited number of epochs with the same hyperparameter configuration.

For each of the four bias and additional baseline partitions, we train a ResNet-18 initialized with ImageNet pretrained 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 observation accuracy, and for all partitions we test only with images from species present in the split the model was trained on.

All experiments were performed on a machine running RHEL Rocky Linux version 9.4 with 96 Intel Skylake CPUs, 1.5 terabytes of RAM and one 40 GiB NVIDIA Tesla A100 GPU. All training and testing code for the models was implemented in Python 3.11 and Pytorch version 2.3.1.

Train-Test	Top1-Obs	Top1-Spec	Top1-Wgt	Top1-LUC
Spatial Bias (Human Footprint)				
Wild-Wild	0.540	0.291	0.161	0.555
Modified-Modified	0.708	0.396	0.156	0.702
Temporal Bias (City Nature Challenge)				
CNC-CNC	0.445	0.208	0.087	0.434
Not CNC-Not CNC	0.696	0.401	0.176	0.690
Observer Quality Bias				
Casual-Casual	0.646	0.290	0.114	0.610
Engaged-Engaged	0.633	0.354	0.150	0.630
Socio-Political Bias				
CA-CA	0.721	0.424	0.146	0.720
BC-BC	0.689	0.359	0.114	0.705

Table A3: (Performance values for all partitions. Obs = Observation, Spec=Species, Wgt=Weighted, LUC=Land Use Category, CA = California, SO = Sonora, BC = British Columbia, YT = Yukon Territories, CNC = City Nature Challenge, diff = difference.

5.7 Evaluation Metrics

We report standard Top-1 accuracy, referred to here as Top-1 per-observation accuracy. As the DivShift-NAWC is extremely long-tailed, overweighting the contribution of more common classes to Top-1 per-observation accuracy, we also report Top-1 accuracy averaged per-class (sometimes referred to as macro Top-1 accuracy or average recall), which we refer to here as Top-1 per-species accuracy. This metric considers the Top-1 accuracy per-class independently of classes' frequency. We also introduce a new rarity-weighted Top-1 accuracy that upweights the relative importance of rarer classes and downweights more common ones:

$$\frac{1}{\sum_{i=1}^{C} \frac{1}{Sum(y_i)}} \cdot \sum_{i=1}^{C} \frac{Acc(y_i, K)}{Sum(y_i)^2}$$

where y are the model predictions per-class and per-observation, $Acc(y_i, K)$ is the Top-K accuracy for observations of class i, and $Sum(y_i)$ is the number of observations of class i. This metric can be thought of as an inverse of Top-K per-observation accuracy, where rarer classes are upweighted and more common classes are downweighted. As rarer species summarily have fewer data, these classes are naturally more difficult to correctly classify, thus model accuracies with this metric will be significantly depressed compared to other top-K metrics. Lastly, to modulate the effects of spatial biases, we also calculate per-land use category top-1 accuracy [38]. These accuracies are identical to species Top-K accuracy, except instead of calculating the accuracy per-label class and then averaging across classes, we calculate the Top-K accuracy for all images that fall within a given land-use category, then average those accuracies across the categories.