Challenges in Applying Audio Classification Models to Datasets Containing Crucial Biodiversity Information

Anonymous Authors[1]

Abstract

The acoustic signature of a natural soundscape can reveal consequences of climate change on biodiversity. Hardware costs, human labor time, and expertise dedicated to labeling audio are impediments to conducting acoustic surveys across a representative portion of an ecosystem. These barriers are quickly eroding away with the advent of low-cost, easy to use, open source hardware and the expansion of the machine learning field providing pre-trained neural networks to test on retrieved acoustic data. One consistent challenge in passive acoustic monitoring (PAM) is a lack of reliability from neural networks on audio recordings collected in the field that contain crucial biodiversity information that otherwise show promising results from publicly available training and test sets. To demonstrate this challenge, we tested a hybrid recurrent neural network (RNN) and convolutional neural network (CNN) binary classifier trained for bird presence/absence on two Peruvian bird audiosets. The RNN achieved an area under the receiver operating characteristics (AUROC) of 95% on a dataset collected from Xeno-canto and Google's AudioSet ontology in contrast to 65% across a stratified random sample of field recordings collected from the Madre de Dios region of the Peruvian Amazon. In an attempt to alleviate this discrepancy, we applied various audio data augmentation techniques in the network's training process which led to an AUROC of 77% across the field recordings.

1. Introduction

Anthropogenic activities that lead to catastrophes, such as wildfires and deforestation, cascade into challenges in maintaining biodiversity in an ecosystem (Ward et al., 2020) . The intersectionality between biodiversity loss and climate change is becoming increasingly apparent leading to an intergovernmental multidisciplinary workshop on the subject matter (Otto-Portner et al., 2021). To properly understand the ramifications of anthropogenic activity on wildlife populations, reliable and large-scale tools must be developed to monitor biodiversity across various ecosystems.

Historically, field biologists surveyed wildlife populations through techniques that are challenging to scale up such as trapping individual specimens and monitoring feeding sites (Lopez-Baucells et al., 2016; Welsh Jr. & Ollivier, 1998). A growing method amongst biologists and ecologists involves deploying remote camera trap arrays to monitor the population densities of large fauna over a large area (Tobler et al., 2018; Norouzzadeh et al., 2018; Willi et al., 2019) . New breakthroughs by researchers in the field of automated image classification driven by neural networks have made these camera trap arrays more practical by driving down the amount of resources required to label and extract relevant biodiversity information from the images collected (Tabak et al., 2019; He et al., 2015).

Many indicator species such as insects, birds, amphibians, and bats can reveal consequences of climate change on ecosystems(Borges, 2007; Kim, 1993; Medellín et al., 2000; Woodford & Meyer, 2003) . These species are oftentimes too small or mobile for stationary camera trap arrays to measure to any statistical significance. Passive acoustic monitoring with low-cost open source audio recorders fills this niche, as it enables detection of species such as cicadas that are small and noisy (Hill et al., 2018). Audiosets from these surveys are oftentimes impractical for human labeling from a temporal standpoint. This challenge naturally leads to the use of machine learning.

Many techniques derived from image classification translate into the audio domain once the sounds have been converted into spectrogram images (Kahl et al., 2021; Colonna et al., 2016). One such neural network we have chosen to test was designed for audio event detection with low-resource training sets (Morfi & Stowell, 2018). This model is a hybrid RNN-CNN model that consists of a 2d-convolutional block that computes features from the audio that has been

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

converted into a mel spectrogram, a recursive block that computes features at each time step from the features of neighboring time steps, a time-distributed dense block that's layers are applied independently of one another on each of the time step features, and a max-pooling layer that pools the predictions across all time steps to generate a global label for a given sequence. We leveraged a Github repository called Microfaune that encapsulates the neural network with ease-of-use features such as pre-trained weights for the network.

In this paper, we compare Microfaune's bird presence/absence capabilities across audio recordings of Peruvian birds taken from the crowd-sourced bird vocalization database Xeno-canto combined with bird absent recordings taken from the Google AudioSet ontology (Gemmeke et al., 2017; Vellinga & Planqué, 2015) to field recordings collected from the Peruvian Amazon. This will aid in determining what sort of challenges are to be expected by scientists considering deploying neural networks on PAM field data. We also demonstrate the efficacy of audio data augmentation techniques in the training of neural networks (Ko et al., 2015) to improve a model's generalizability across field recordings labeled for bird audio.

2. Methodology

2.1. Deployment

We collected field audio recordings in two logging concessions (Forestal Otorongo and MADERACRE) in Madre de Dios, Peru, a biodiversity hotspot in southeastern Peru (BROTTO et al., 2010). These logging concessions are located in lowland Amazonian moist forest and are sustainably managed under a Forest Stewardship Council (FSC) certification. From June to September 2019, we deployed 35 Audiomoth devices along logging roads or inside unlogged forest (6) . The Audiomoth devices were attached to tree trunks at a height of approximately 2 meters (5) and were set to record 1 minute every 10 minutes at a 384 kilohertz sampling rate. In total, 31 devices successfully recorded for approximately 1 month generating nearly 1500 hours of audio.

To generate a test set from the field recordings, a smaller stratified random sample was constructed by collecting a random clip from each hour of the day from each Audiomoth device. This technique left us with a representative subset of the field recordings amounting to approximately 12 hours of audio. The stratified clips from 16 devices were split up into 3 second segments (20 clips per recording) amounting to a total of 7120 3 second clips. These audio clips were then labeled for bird presence/absence resulting in a 3113/4007 split between the two classes.

To generate a test set from internet audio data, we scraped Xeno-canto for a list of approximately 1000 species given to us by an ornithologist familiar with Madre de Dios bird species. From these variable-length audio clips, we selected approximately 50 species we determined to be high priority due to their abundance of available recordings and distinct calls. We randomly selected 50 recordings from each of these species. To make the model more robust with a wider variety of species we randomly selected 2-3 clips from each species in the list provided marked as "A" quality on Xeno-canto. We combined these two Xeno-canto datasets together amounting to 4774 bird-present recordings. To balance the bird-present recordings, we scraped the Google AudioSet ontology database for 4774 recordings from classes unlikely to contain bird vocalizations.

2.2. Training

For reproducibility purposes, we retrained Microfaune's built-in model weights as a baseline with the DCASE 2018 competition datasets "freefield1010" and "warblr10k". The freefield1010 dataset contains 7690 field recordings from around the world and the warblr10k contains 8000 crowd-sourced smartphone audio recordings from the United Kingdom. These audio recordings were broken down into 10 second segments and divided into an 80/20 random split between training and validation, respectively.

To create a new set of model weights that leverages audio data augmentation, we used the same process as the baseline model with the addition of alternative versions of freefield1010 and warblr10k in the training process. These augmented alternative versions included increasing the speed by 10%, decreasing the speed by 10%, injecting gaussian noise with a mean of 0 and standard deviation of 0.005, and injecting gaussian noise with a mean of 0 and standard deviation of 0.1.

2.3. Testing

To vet the trained models on the test sets, we used Microfaune's built-in audio clip global score functionality that is equivalent to taking the maximum score from the model's multiple temporal predictions. We treat this as the model's prediction on the probability of at least one bird vocalization existing within an audio clip. All of the audio was normalized to have a sampling rate less than or equal to 44.1 kilohertz. All stereo recordings were converted to mono. Both the ideal Xeno-canto/Google AudioSet and our hand-labeled field recordings were given global score predictions across both the baseline and data augmented models.

Table 1. Summary of ROC Curves

Metric	Baseline XC Data	Baseline field data	Augmentation XC data	Augmentation field data
AUCROC TP/FP	.95	.65	.98	.77
AUCROC Precision/Recall Bird-present (Class 1)	.96	.64	.98	.71
AUCROC Precision/Recall Bird-absent (Class 0)	.94	.66	.98	.78

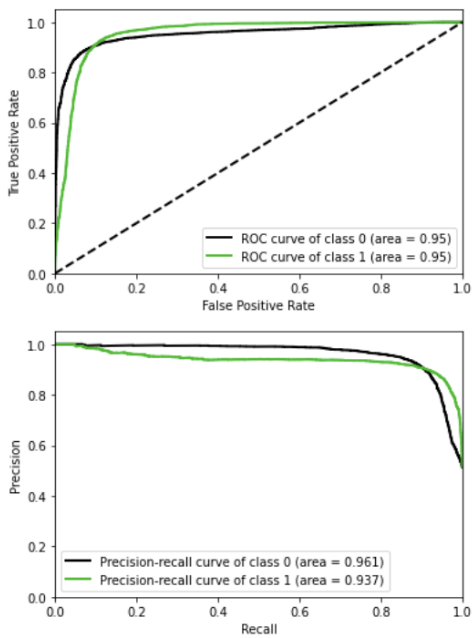

Figure 1. Xeno-canto w/ baseline

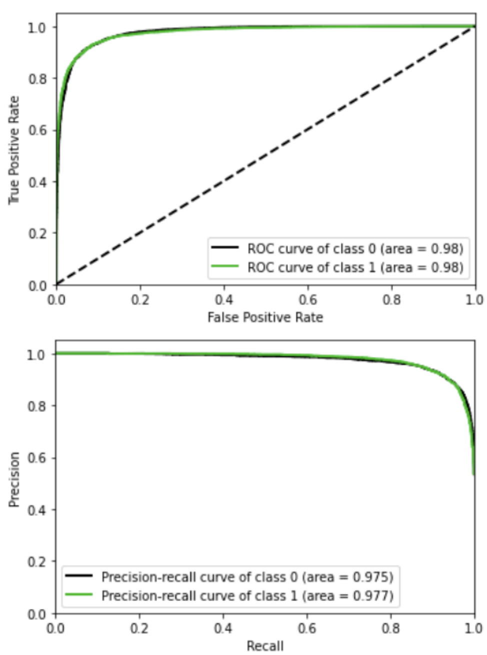

Figure 2. Xeno-canto w/ data augmentation

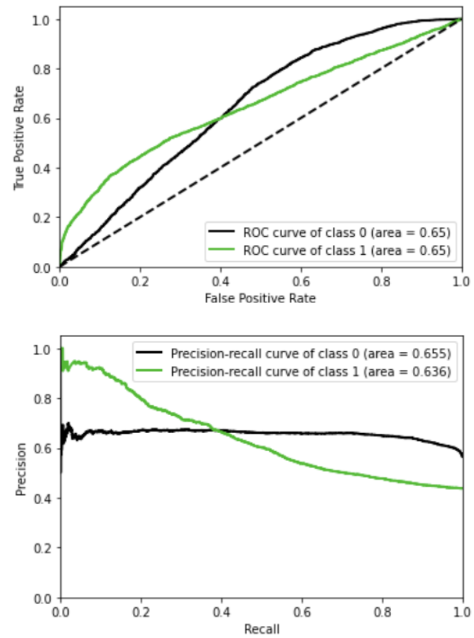

Figure 3. Peru field recordings w/ baseline

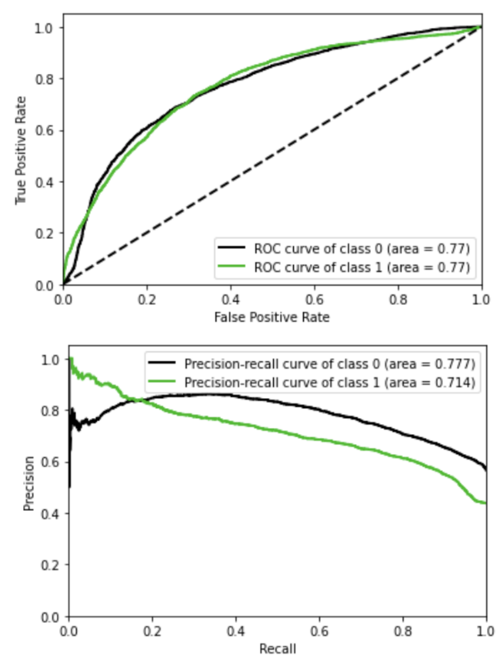

Figure 4. Peru field recordings w/ data augmentation

3. Results

To statistically compare the hand labels to the global scores, we used ROC curves (1, 2, 3, 4) that are common tools for measuring binary classifiers (Davis & Goadrich, 2006) . Using Scikit-learn we examined the tradeoffs of increasing the global score threshold for classifying an audio clip as a bird-present (Class 1) true positive(1). We focused on the relationships defined by the AUROC between true positive and false positive rates as well as the tradeoff between precision and recall (5).

4. Conclusion

These results demonstrate how individuals interested in acquiring biodiversity related information from their field audio can be led on by promising results from neural networks on ideal test sets that show metrics above 90% but may not smoothly translate onto their recordings. This is evident by the large differences in our ROC Curves. We observed a 30% difference between the AUROC's of the true positive/false positive curves of the Xeno-canto dataset and field recordings. We also observed a 32% difference between the AUROC's of the bird-present precision-recall curves of the Xeno-canto dataset and field recordings. Data augmentation with speed modulation and gaussian noise injection appears to be a very simple method to reduce the discrepancy between these two datasets as the difference between the AUROC's of true positive/false positive and bird-present precision-recall curves come out to be 21% and 27% respectively.

5. Discussion

There are many potential factors that could be driving the discrepancy between datasets scraped from the internet and field recordings. The most clear factor comes in the form of potential false positives in field recordings from various fauna such as insects, frogs, and monkeys that are challenging to distinguish from bird vocalizations. In the future we hope to implement a referee labeling process where each audio clip is labeled by two humans and a third human makes the final decision on any labeling discrepancy.

Many machine learning techniques exist that have the potential to improve the performance of neural networks on field recordings that we are interested in trying. One technique known as active learning involves integrating data acquisition and the training process together through unsupervised machine learning methods to assess which clips yield the greatest representation of the underlying dataset (Dasgupta, 2011) . This technique has been used to drastically reduce the amount of data required to achieve the same results on camera trap models (Norouzzadeh et al., 2019) . Another technique referred to as transfer learning involves taking lower layers of a pre-trained neural network and training the final layers on user training sets that bias the model towards said dataset. This has the benefit of reducing computational time and the amount of labeled data dedicated to training (Pan & Yang, 2010) . These techniques are worth pursuing as the combined power of PAM and neural networks has the potential to be invaluable in measuring biodiversity loss driven by the climate crisis.

Additional Materials

Audiomoth Deployment Photos

Figure 5. Example Audiomoth Attachment

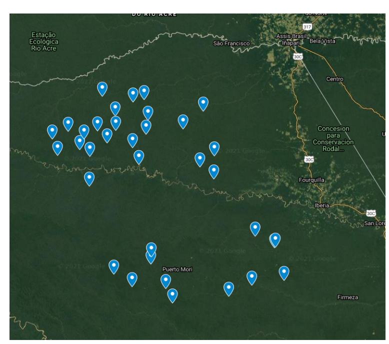

Figure 6. Madre de Dios Deployment Map

Precision and Recall Statistical Metrics

$$Precision = \frac{(TruePositiveCount)}{(TruePositiveCount) + (FalsePositiveCount)}$$

$$Recall = \frac{(TruePositiveCount)}{(TruePositiveCount) + (FalseNegativeCount)}$$

Project Github Repository (private for now for review process)

https://tinyurl.com/4e4k9nfk

References

Borges, S. H. Bird assemblages in secondary forests developing after slash-and-burn agriculture in the brazilian amazon. *Journal of Tropical Ecology*, 23(4):469–477, 2007. ISSN 02664674, 14697831. URL http://www.jstor.org/stable/4499120.

BROTTO, L., MuRRAY, J., PETTENELLA, D., SECCO, L., and MASIERO, M. 4.2 biodiversity in the peruvian amazon. *Biodiversity conservation in certified forests*, pp. 112, 2010.

Colonna, J., Peet, T., Ferreira, C. A., Jorge, A. M., Gomes, E. F., and Gama, J. a. Automatic classification of anuran sounds using convolutional neural networks. In *Proceedings of the Ninth International C* Conference on Computer Science Software Engineering*, C3S2E '16, pp. 73–78, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340755. doi: 10.1145/2948992.2949016. URL https://doi.org/10.1145/2948992.2949016.

Dasgupta, S. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

Hill, A. P., Prince, P., Piña Covarrubias, E., Doncaster, C. P., Snaddon, J. L., and Rogers, A. Audiomoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods in Ecology and Evolution*, 9(5):1199–1211, 2018. doi: 10.1111/2041-210X.12955. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12955.

Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2021.101236. URL https://www.sciencedirect.com/science/article/pii/S1574954121000273.

Kim, K. C. Biodiversity, conservation and inventory: why insects matter. *Biodiversity and Conservation*, 2(3):191–214, Jun 1993. doi: 10.1007/bf00056668.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. Audio augmentation for speech recognition. In *INTERSPEECH*, 2015.

Lopez-Baucells, A., Rocha, R., Bobrowiec, P., Bernard, E., Palmeirim, J., and Meyer, C. *Field Guide to Amazonian Bats*. 09 2016. ISBN 978-85-211-0158-1. doi: 10.13140/RG.2.2.23475.84003.

Medellín, R. A., Equihua, M., and Amin, M. A. Bat diversity and abundance as indicators of disturbance in neotropical rainforests. *Conservation Biology*, 14(6):1666–1675, 2000. doi: https://doi.org/10.1111/j.1523-1739.2000.99068.x. URL https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/j.1523-1739.2000.99068.x.

Morfi, V. and Stowell, D. Deep learning for audio event detection and tagging on low-resource datasets. *Applied Sciences*, 8:1397, 08 2018. doi: 10.3390/app8081397.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018.

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., and Clune, J. A deep active learning system for species identification and counting in camera trap images, 2019.

Otto-Portner, H., Scholes, B., Agard, J., Archer, E., Arneth, A., Bai, X., Barnes, D., Burrows, M., Chan, L., Cheung, W. L. W., Diamond, S., Donatti, C., Duarte, C., Eisenhauer, N., Foden, W., Gasalla, M., Handa, C., Hickler, T., Hoegh-Guldberg, O., Ichii, K., Jacob, U., Insarov, G., Kiessling, W., Leadley, P., Leemans, R., Levin, L., Lim, M., Maharaj, S., Managi, S., Marquet, P., Mcelwee, P., Midgley, G., Oberdorff, T., Obura, D., Osman Elasha, B., Pandit, R., Pascual, U., Pires, A. P. F., Popp, A., Reyes-García, V., Sankaran, M., Settele, J., Shin, Y.-J., Sintayehu, D. W., Smith, P., Steiner, N., Strassburg, B., Sukumar, R., Trisos, C., Val, A. L., Wu, J., Aldrian, E., Parmesan, C., Pichs-Madruga, R., Roberts, D. C., Rogers, A. D., Díaz, S., Fischer, M., Hashimoto, S., Lavorel, S., Wu, N., and Ngo, H. Full scientific outcome of the IPBES-IPCC co- sponsored workshop report on biodiversity and climate change, June 2021. URL https://doi.org/10.5281/zenodo.4659159. Suggested citation: Pörtner, H.O., Scholes, R.J., Agard, J., Archer, E., Arneth, A., Bai, X., Barnes, D., Burrows, M., Chan, L., Cheung, W.L., Diamond, S., Donatti, C., Duarte, C., Eisenhauer, N., Foden, W., Gasalla, M., Handa, C., Hickler, T., Hoegh-Guldberg, O., Ichii, K., Jacob, U., Insarov, G., Kiessling, W., Leadley, P., Leemans, R., Levin, L., Lim, M., Maharaj, S., Managi, S., Marquet, P., McElwee, P., Midgley, G., Oberdorff, T., Obura, D., Osman, E., Pandit, R., Pascual, U., Pires, A. P. F., Popp, A., Reyes-García, V., Sankaran, M., Settele, J., Shin, Y. J., Sintayehu, D. W., Smith, P., Steiner, N., Strassburg, B., Sukumar, R., Trisos, C., Val, A.L., Wu, J., Aldrian, E., Parmesan, C., Pichs-Madruga, R., Roberts, D.C., Rogers, A.D., Díaz, S., Fischer, M., Hashimoto, S., Lavorel, S., Wu, N., Ngo, H.T. 2021. Full scientific outcome of the IPBES-IPCC co-sponsored workshop report on biodiversity and climate change; IPBES secretariat, Bonn, Germany, DOI:10.5281/zenodo.4659158.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1345–1359, 2010.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., Vercauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., Lukacs, P. M., Moeller, A. K., Mandeville, E. G., Clune, J., and Miller, R. S. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019. doi: https://doi.org/10.1111/2041-210X.13120. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13120.

Tobler, M. W., Garcia Anleu, R., Carrillo-Percastegui, S. E., Ponce Santizo, G., Polisar, J., Zuñiga Hartley, A., and Goldstein, I. Do responsibly managed logging concessions adequately protect jaguars and other large and medium-sized mammals? two case studies from guatemala and peru. *Biological Conservation*, 220:245 – 253, 2018. ISSN 0006-3207. doi: https://doi.org/10.1016/j.biocon.2018.02.015. URL http://www.sciencedirect.com/science/article/pii/S000632071731563X.

Vellinga, W.-P. and Planqué, R. The xeno-canto collection and its relation to sound recognition and classification. In *CLEF (Working Notes)*, 2015.

Ward, M., Tulloch, A. I., Radford, J. Q., Williams, B. A., Reside, A. E., Macdonald, S. L., Mayfield, H. J., Maron, M., Possingham, H. P., Vine, S. J., et al. Impact of 2019–2020 mega-fires on australian fauna habitat. *Nature Ecology & Evolution*, 4(10): 1321–1326, 2020.

Welsh Jr., H. H. and Ollivier, L. M. Stream amphibians as indicators of ecosystem stress:a case study from california's redwoods. *Ecological Applications*, 8(4):1118–1132, 1998. doi: https://doi.org/10.1890/1051-0761(1998)008[1118:SAAIOE]2.0.CO;2. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/1051-0761%281998%29008%5B1118%3ASAAIOE%5D2.0.CO%3B2.

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., and Fortson, L. Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1):80–91, 2019.

Woodford, J. E. and Meyer, M. W. Impact of lakeshore development on green frog abundance. *Biological Conservation*, 110 (2):277–284, 2003.