2019

# Tools for landscape-scale automated acoustic monitoring to characterize wildlife occurrence dynamics

Cathleen Michelle Balantic
*University of Vermont*

# TOOLS FOR LANDSCAPE-SCALE AUTOMATED ACOUSTIC MONITORING TO CHARACTERIZE WILDLIFE OCCURRENCE DYNAMICS

A Dissertation Presented

by

Cathleen Michelle Balantic

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfilment of the Requirements
For the Degree of Doctor of Philosophy
Specializing in Natural Resources

January, 2019

Defense Date: September 21, 2018
Dissertation Examination Committee:

Therese Donovan, Ph.D., Advisor
Donna Rizzo, Ph.D., Chairperson
James Murdoch, Ph.D.
Allan Strong, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

# ABSTRACT

In a world confronting climate change and rapidly shifting land uses, effective methods for monitoring natural resources are critical to support scientifically-informed management decisions. By taking audio recordings of the environment, scientists can acquire presence-absence data to characterize populations of sound-producing wildlife over time and across vast spatial scales. Remote acoustic monitoring presents new challenges, however: monitoring programs are often constrained in the total time they can record, automated detection algorithms typically produce a prohibitive number of detection mistakes, and there is no streamlined framework for moving from raw acoustic data to models of wildlife occurrence dynamics. In partnership with a proof-of-concept field study in the U.S Bureau of Land Management's Riverside East Solar Energy Zone in southern California, this dissertation introduces a new R software package, AMMonitor, alongside a novel body of work: 1) temporally-adaptive acoustic sampling to maximize the detection probabilities of target species despite recording constraints, 2) values-driven statistical learning tools for template-based automated detection of target species, and 3) methods supporting the construction of dynamic species occurrence models from automated acoustic detection data. Unifying these methods with streamlined data management, the AMMonitor software package supports the tracking of species occurrence, colonization, and extinction patterns through time, introducing the potential to perform adaptive management at landscape scales.

# CITATIONS

Material from this dissertation has been submitted for publication to Bioacoustics on October, 1, 2018 in the following form:

Balantic, C.M, Donovan, T.D.. (2018). Statistical learning mitigation of false positive detections in automated acoustic wildlife monitoring. Bioacoustics.

Material from this dissertation has been submitted for publication to Methods in Ecology and Evolution on October, 3, 2018 in the following form:

Balantic, C.M, Donovan, T.D.. (2018). Temporally-adaptive acoustic sampling to maximize detection across a suite of focal wildlife species. Methods in Ecology and Evolution.

Material from this dissertation has been submitted for publication to Ecological Applications on October, 3, 2018 in the following form:

Balantic, C.M, Donovan, T.D.. (2018). Dynamic wildlife occupancy models using automated acoustic monitoring data. Ecological Applications.

# ACKNOWLEDGEMENTS

Thank you to Terri Donovan, my visionary advisor, for helping me become a better scientist and person. Thank you also to my committee members: Jed Murdoch, Allan Strong, and Donna Rizzo, whose diverse backgrounds and feedback have provided meaningful insights throughout this experience. Thank you to Jon Katz for invaluable assistance with the AMMonitor package, for patiently teaching me how to solder, and for helping me turn smartphones into reliable automated scientific research units. Thank you to Mark Massar at the U.S. Bureau of Land Management for supporting all of our field trips into the Solar Energy Zone, and generously sharing expertise, knowledge of the land, and good company along the way. Thank you to the students and faculty in the IGERT Smart Grid and Complex Systems groups who pushed me to expand my comfort zone, and to all of my friends and colleagues in IGERT and RSENR. Above all, thank you to my family for your boundless support and unconditional love through this process. Thank you especially to Robert Caulk, for volunteering to help that out-of-place Natural Resources student code up an automatic birdsong detector in Donna's ANN class years ago (the rest is history). Finally, thank you to the Black-tailed Gnatcatcher who sang near Phone #8 at 9am on November, 13th, 2015 in the Solar Energy Zone. Following a tumultuous year of academic self-doubt, you were the first animal we remotely captured on our homemade recording units, and I wouldn't have made it this far without you.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1: INTRODUCTION AND COMPREHENSIVE LITERATURE REVIEW

## 1.1. Introduction

Amid climate change and rapidly shifting land uses, effective methods for monitoring natural resources priorities are critical to support scientifically-informed resource management decisions (Pollock et al. 2002). Acoustic monitoring of wildlife provides one such method: by taking audio recordings of the environment, scientists can acquire data to characterize status and trends of populations of sound-producing wildlife over time and across vast spatial scales (Furnas & Callas 2015, Cerquiera & Aide 2016). Many animals – from birds, to amphibians, to mammals, to insects – use vocalizations or other sounds to defend territory, attract mates, and communicate (Kroodsma 1996, Catchpole & Slater 2008, Suthers et al. 2016). The scientific community has long used animal sounds to gain insights into the whereabouts, abundance, and behavior of wildlife species and communities, amid a changing climate and changing land uses (Robbins et al. 1986). Typically, auditory information is gathered by way of researchers present on site to listen for species of interest (Ralph et al. 1995, Rosenstock et al. 2002). Though such field study is vital to the growing body of ecological knowledge, it also has drawbacks, such as detection mistakes by the human observer (Campbell & Francis 2011, Swiston & Mennill 2011), modification of animal behavior in the presence of human observers (Gutzwiller & Markum 1997, Bye et al. 2001, Gaynor et al. 2018), and the logistical effort required to be physically present to detect wildlife when they are available for observation (Moore & McCarthy 2016).

To circumvent these challenges, some research settings have traded in-person field study for autonomous recording units (ARUs), in which audio recorders are deployed for some length of time at some number of study sites to capture recordings of the environment, which can later be analyzed for species of interest (Shonfield & Bayne 2017). The practice of using ARUs to monitor wildlife species has grown immensely in the past decade, with monitoring projects across species from birds (Furnas & Callas 2015), to bats (Zamora-Gutierrez et al. 2016), elephants (Wrege et al. 2012), wolves (Root-Gutteridge et al. 2014), primates (Heinicke et al. 2015), amphibians (Brauer et al. 2016), insects (Newson et al. 2017), and marine mammals (Bioacoustics Research Program 2018).

Critically, recorded audio data confers the capacity to verify and analyze species identifications and/or vocalization behavioral patterns *a posteriori* (Hobson et al. 2002), and uses of this information windfall have been diverse. Recordings from ARUs have been interrogated to characterize wildlife occurrence patterns (Furnas & Callas 2015, Cerquiera & Aide 2016), avian density (Dawson & Efford 2009), biodiversity and ecological communities (Gage et al. 2001, Sueur & Farina 2015), animal behavior patterns (e.g. Mennill & Vehrencamp 2008), and environmental soundscape characteristics, which provide an alternative to species-focused acoustic monitoring (Pijanowski et al. 2011). Audio data have also been used to identify individual wolf packs by the frequency of their howls (Root-Gutteridge et al. 2014), recognize and mitigate poaching incidents (Astaras et al. 2017), avoid marine shipping strikes of North Atlantic

Right Whales (*Eubalaena glacialis*) (Bioacoustics Research Program 2018), classify the cryptic nocturnal flight calls of migratory birds (Farnsworth et al. 2005, Salamon et al. 2016), and monitor wildlife and soundscapes in remote or unsafe areas, such as the Chernobyl Exclusion Zone (Stowell et al. 2018).

Although applications of environmental audio data are numerous, this dissertation focuses specifically on the use of ARU-acquired audio data in occurrence-based wildlife population models, and the utility of ARUs for this purpose is emerging alongside two-overarching challenges. Firstly, fundamental questions remain with regard to the ARU hardware itself: how much data can be recorded, when and how often may recordings be taken, and how can the resulting audio data be efficiently collected by researchers for subsequent processing? Existing ARU choices vary in data storage volume, capacity to schedule recordings, and options for collecting the data. Regardless of the hardware used, if recordings are not taken under suitable conditions, target monitoring species may not be acoustically captured even if the species is truly present, resulting in problematic false negatives at the site level. Secondly, once large volumes of audio data have been acquired, humans often cannot search through audio recordings in a timely manner, making automated detection systems necessary for expedient data processing and analysis. The performance of these nascent automated systems, however, varies widely depending on the expertise of the user, acoustic characteristics of target species sounds, and soundscape circumstances of the study area. Automated detection systems can frequently miss individual sound events emitted by a target species (false negatives) or mistakenly detect non-target sounds (false positives) (Marques et al. 2012).

Beyond acoustic monitoring, these false negatives and false positives represent a ubiquitous challenge in occurrence-based population modeling, where the goal is to characterize presence and absence (rather than abundance) of a species across a landscape. Dynamic occurrence models (e.g. Mackenzie et al. 2003), in which researchers strive to gain insight not only on an initial occurrence pattern, but on local extinction and colonization patterns over time, are especially valuable, but particularly susceptible to the vagaries of false negatives and false positives (McClintock et al. 2010, Miller et al. 2015, Ruiz-Guitierrez et al. 2016). If the challenges posed by false negatives and false positives in acoustic monitoring can be adequately addressed, this method will be better equipped to produce dynamic species occurrence models that can inform adaptive management, in which learning over time can be used systematically for prediction to aid land management decisions (Williams et al. 2009).

## 1.2. The challenge of taking recordings in a way that minimizes false negatives

One of the first hurdles a remote acoustic monitoring program encounters is determining when and how much to record, decisions which are often constrained by the hardware used for recording. There are a growing number of hardware solutions for gathering acoustic data, with wide variation in costs, recording quality, convenience, and ease of use. The vast majority of ARU options take audio recordings and store them directly on a memory card on the device, obliging researchers to physically collect data from the device on a regular basis: these include the Song Meter SM4 ($849 USD per unit) (Wildlife Acoustics 2018), the AudioMoth ($43 USD per unit) (Hill et al. 2018), the

Swift Recorder (price varies) (Bioacoustics Research Program 2018), or the fully open-source and customizable Raspberry Pi-based Solo, which records continuously ($199 USD per unit) (Whytock & Christie 2017).

Some devices may be programmed to record at specified intervals if desired; certain individual brands of storage card-based ARU can capture up to 400 hours of audio data (Song Meter SM4: Wildlife Acoustics 2018), though total run-time depends on the recording sampling rate, with up to 1,065 hours of recording possible at a sample rate of 8 kHz (Swift: Bioacoustics Research Program 2018). In some cases, devices are limited more by their power source than by storage restrictions, recording independently for up to three weeks before requiring a recharge (Swift Recorder: Bioacoustics Research Program 2018), while still others are capable of recording continuously for nearly six weeks at a time depending on the audio sampling rate used (Solo: Whytock & Christie 2017).

An alternative to on-device storage is for the ARU to use a cellular or wireless network to transmit recordings from the device to a storage server in near real-time, a method which has been deployed to remotely monitor seabirds (McKown 2012), and in a variety of projects undertaken by the Remote Environmental Assessment Laboratory at Michigan State University (Gage et al. 2015). The monitoring devices used in those cases were constructed specifically for the projects in question and are not commercially available. Currently, the only commercially available hardware option for transmission of acoustic files over a WiFi or cellular network is the ARBIMON permanent Acoustic Monitoring Station ($4000 USD before WiFi or cellular data plan) (Aide et al. 2013, Sieve Analytics 2018).

ARUs may be limited in how much data they can record, and when they can record it, particularly if constrained by transmission over a cellular network (McKown et al. 2012, Aide et al. 2013, Gage et al. 2015) or if there is minimal power and/or storage capacity on the device within a desired monitoring period. Recording constraints may cause sampling to occur under suboptimal conditions, with ARUs potentially failing to record a present species when it is actually active (Sidie-Slettedahl et al. 2015). Thus, recording constraints can lead to “site-level false negatives”: if recordings are routinely taken under conditions inappropriate for detecting a species of interest, or if the probability of detecting a present species is less than one, researchers may mistakenly conclude that a species is absent where it is actually present (Mackenzie et al. 2002, Pollock et al. 2002).

False negatives in wildlife monitoring are often either the consequence of a suboptimal sampling method or of sampling at a time when the animal was not active or available for detection (Thompson 2004). Such circumstances are often rooted in logistical or budgetary constraints that limit sampling power across time and space. For wildlife monitoring, literature around allocation of sampling resources typically focuses on the optimization of observation effort for field studies, or on the most efficient allocation of sampling effort in the spatial dimension (Thompson et al. 1998, 2004, Turk and Borkowski 2005, Moore & McCarthy 2016). For example, a “removal design” typically refers to a framework in which sites are surveyed multiple times until the target species is detected, after which remaining sampling efforts can be directed to sites where a species has not yet been detected (Mackenzie & Royle 2005).

In the context of wildlife monitoring, adaptive sampling means that information from prior surveys is explicitly incorporated into future sampling efforts in order to improve the chances of detecting a target species given that it is present (Thompson & Seber 1994), but the concept of adaptive sampling is not limited to wildlife monitoring. Incorporation of prior information into future sampling efforts amid sampling constraints, in order to maximize the utility of available sampling resources, is a challenge present across domains and applications concerned with monitoring and prediction (Bucher 1988, Bishop et al. 2001).

The problem of sampling optimization is commonly encountered in the implementation of wireless sensor networks (WSNs), which are composed of a number of sensors deployed to monitor environmental or other conditions at different spatial locations (Anastasi et al. 2009). ARUs might be integrated into a WSN framework, and like ARUs, WSNs often confront constraints on memory, data communication, data transmission, and power (Anastasi et al. 2009). Spatio-temporal adaptive sampling techniques have been used in WSNs to detect military base intruders (Raghunathan et al. 2006), monitor snowpack for avalanche prediction (Alippi et al. 2012), and monitor water level sensor nodes to warn of floods (Zhou & De Roure 2007). They have also been used for wildlife applications, such as tracking the social behavior of badgers (Dyo et al. 2012), characterizing the occupancy status of Leach's Storm Petrel nests (Mainwaring et al. 2002), long-term censusing of birds in spruce-fir habitat, and others (Porter et al. 2005). The implementation of temporal adaptive sampling in WSNs range has been achieved via Kalman filters (Jain & Change 2004), hidden Markov models, reinforcement learning (Dyo et al. 2012), Bayesian frameworks (Xu et al. 2011), and more (Anastasi et al. 2009). The resource allocation routines of WSNs provide inspiration for an analogous pursuit of sampling resource optimization by ARUs.

Though the notion of adaptive sampling for wildlife across space has attracted much attention in the literature (Smith et al. 1995, Thompson et al. 2004), the idea of adaptive sampling for wildlife through time has invited comparatively little study, though this is likely to change with the rise of long-term remote monitoring opportunities offered by acoustic monitoring and camera trapping. In most commercially available ARU hardware, recordings are taken and then stored locally on a memory card on the device, requiring periodic field trips to collect the audio data (Whytock & Christie 2017, Bioacoustics Research Program 2018, Hill et al. 2018, Wildlife Acoustics 2018). Another emergent paradigm is for the ARU to use the cellular or WiFi network to transmit recordings to a server in near-real time, eliminating the need for frequent field trips into areas where human disturbance may impact the quality of monitoring data (McKown 2012, Aide et al. 2013, Gage et al. 2015). The latter case of network-linked ARUs provides the opportunity to conduct temporally adaptive allocation of constrained sampling resources, due to the availability of two-way communication between the server and the ARU, wherein fresh recording schedules might be dispatched to ARU monitoring locations on a daily basis, contingent on information from the day before and weather information predicted for the next day (Balantic & Donovan *in prep*).

Many animals vocalize under very specific conditions, often based on seasonal changes and weather (Hayes & Huntly 2005, Frick et al. 2012). Sampling resources for characterizing wildlife populations thus may be allocated intelligently to avoid wasteful recording at times when target species will not be active. Additionally, if researchers have multiple target species they want to detect, it behooves them to sample at times when the highest proportion of target species will be active concurrently. This idea has been explored in the context of wildlife community interactions monitored with camera trap data, where researchers optimally only sample (take photos) on occasions with temporal overlap of multiple focal species, instead of wasting the camera's power resources on photos likely to contain only a single species (Frey et al. 2017).

## 1.3. The challenge of analyzing recordings, given that automated detection systems are mistake-prone

Despite constraints on total recording time, ARUs can typically capture a markedly larger volume of observation time than the traditional 3-10 minute point counts common to avian field studies (Ralph et al. 1995). This information increase confers a double-edged sword: ARUs, and the audio recordings they produce, can very quickly engender a big data problem, wherein signals of interest must be efficiently located within large bodies of data. Real-time listening by a researcher is often prohibitively time-consuming, and thus, automated acoustic monitoring approaches are necessary, wherein computer algorithms are employed to automatically detect signals of interest (Stowell et al. 2016, Shonfield & Bayne 2017).

The notion of mature computer-automated acoustic detection of wildlife is an illusion, however: existing automated detection systems can be so rife with mistakes that the consequences of automated systems may outweigh any benefits, unless the detection process includes some manual validation by a human researcher (Acevedo et al. 2009, Hutto & Stutzman 2009, Buxton & Jones 2012, Duan et al. 2013). For example, automated detection systems may flag prohibitively high numbers of detections wherein a target species is not actually vocalizing (“event-level false positive”) or may fail to detect a vocalizing species when it is producing the target sound (“event-level false negative”) (Acevedo et al. 2009, Marques et al. 2012). If not adequately addressed, both of these mistakes may compromise scientific inference and undermine monitoring and management objectives.

Several software options exist for the purpose of automatically detecting wildlife signals out of audio recordings. The objectives of an automated detection system can vary, ranging from requiring perfect detection and time-stamped data for every single species calling event for use in behavior-based research, to information that facilitates abundance-based research, to simple presence/absence data for use in occurrence models (Stowell et al. 2016). A widespread paradigm in automated acoustic wildlife detection is to focus on the most difficult of these, individual event detection, which is commonly conveyed as time vs. frequency event boxes in a spectrogram. Detection mistakes are a challenge due to the nature of environmental audio recordings: in addition to sounds produced by a species of monitoring interest, the acoustically captured soundscape may include sounds from nontarget species, wind, rain, and anthropogenic noise, all of which can either produce event-level detections where the target species was not actually calling, or which can mask sounds produced by the target species (Towsey et al. 2012, Aide et al. 2013, Potamitis 2014).

Widely commercial software options include Raven Pro, which uses band-limited energy detection, and Wildlife Acoustics Kaleidoscope program, which uses a hidden Markov model paired with clustering and classification algorithms to detect and classify events. At the time of this writing, Raven Pro costs $100 per year and Wildlife Acoustics Kaleidoscope costs $299-$399 per year depending on license type (Bioacoustics Research Program 2018, Wildlife Acoustics 2018). Alternatively, the ARBIMON Bioacoustics analysis software platform provides cloud-based automated detection functions and data storage at varying cost rates (5 minutes recording per day * 365 days per year * 20 devices * 0.03 cents per minute = $1095 per year), and uses template-based detection paired with a random forest classifier (Aide et al. 2013, Corrada-Bravo et al. 2016).

Many free programs exist as well, such the R packages *seewave* (for general soundwave analysis) (Sueur et al. 2008), *soundecology* (to compute soundscape ecology indices) (Villanueva-Rivera & Pijanowski 2018), *warbleR* (functions for acquiring and analyzing avian vocalizations) (Araya-Salas & Vidaurre 2017), and *monitoR* (template-based acoustic event detection) (Hafner & Katz 2018). Of these R packages, only *monitoR* supports automated species detection meant to be used without continuous manual checking; it does so via binary template matching or spectrogram cross-correlation templates. Other free recently released software includes the hidden Markov model-based MatlabHTK (Ranjard et al. 2017), and Tadarida, which uses random forests paired with discriminant analysis (Bas et al. 2017).

Several other automated detection methods have been detailed in the scientific literature (Knight et al. 2017, Shonfield & Bayne 2017), but most implementations were designed for specific research purposes, and lack code or software platforms that explicitly make the methodologies accessible for widespread use. Common methods include hidden Markov models (e.g. Wildlife Acoustics), support vector machines (Acevedo et al. 2009, Fagerlund 2007), classification trees (Adams et al. 2010), wavelets (Priyadarshani et al. 2016), dynamic time warping (Anderson et al. 1996), convolutional neural networks (Knight et al. 2017), ensemble approaches that make use of multiple layers or combinations therein (Stowell et al. 2016), and deep learning methods (Goeau et al. 2016). Towsey et al. (2012) emphasize that for single-species detection, the method employed might vary based on sound characteristics of that specific species (as well as the soundscape environment it inhabits), and recommend solutions such as energy-based segmentation, spectral peak tracking, detection of amplitude modulation within small frequency ranges, and spectrogram template matching (as in Katz et al. 2016), depending on the circumstances.

There is a distinction between the algorithm used to automatically detect events (these might be algorithms like a random forest, a convolutional neural network, or a support vector machine) and the features used by the algorithm to decide whether a sound is in fact a signal from a target species. These features are properties of the sound that can be used by the automated detection system to assess whether a sound is a target signal. Commonly used features include the zero-crossing rate, signal energy, or features constructed from a short-time Fourier transform, such as linear prediction cepstrum coefficients (LPCC), mel frequency cepstral coefficients (MFCC) (Levy et al. 2003), and other statistical properties of the frequency spectrum (Sueur et al. 2008).

Regardless of the particular algorithm and set of acoustic features employed by an automated detection system, the human user of that system constitutes perhaps the most important factor of all. Automated detection systems typically involve a training period, wherein the human user provides the algorithm with examples of sounds from a target monitoring species (Bishop 2006). The algorithm then produces a model for detecting those target sounds. If this trained model performs adequately during a subsequent testing period, wherein its detection and classification performance is evaluated against target sounds not encountered during training, then the human user may decide that the model is ready for use in an active research for the automated detection of target monitoring species.

The quality of example data provided during the training phase is critical, and if the system is given inappropriate examples during training and testing, it will likely perform poorly in a real monitoring program (Friedman et al. 2001). It is desirable for animal sound training data to come from the same environment from which it will be collected (Katz et al. 2016, Shonfield & Bayne 2017), and it is likely best if a recognizer is trained on recordings that have been collected with the same type of microphone and audio sampling rate that will be used for the study. Ideally, during the training phase, the user is able to provide examples of target monitoring sounds that capture every single fragment of variation the automated detection system will confront when it is deployed in an active monitoring context. In practice – as is true for many scientific domains broadly concerned with predictive modeling – achieving this level of training data quality is impossible, due to the relative rarity of examples that can encompass all possible scenarios an automated detection system might encounter (Friedman et al. 2001, Bishop 2006). To illustrate, imagine that a research program wants to automatically detect the vocalizations of a species of songbird. In order to train an automated detection system to successfully detect every single vocalization in every single recording taken by the monitoring program, without detecting *any* false positives, that automated detection system will first require examples of every type of sound this species makes, with the target species vocalizing very close to the recorder, vocalizing far away from the recorder, vocalizing while other animals are vocalizing, and vocalizing against a backdrop of wind, vehicle noise, and any other non-target sounds present in the research setting.

Many such training examples are difficult or impossible to capture *a priori*, given that the impacts of climate and land use change on wildlife are ongoing. Wildlife species may modify their vocalization behavior based on landscape changes, a phenomenon observed in populations of White-crowned Sparrows (*Zonotrichia leucophrys*) who increased the pitch of their songs to cope with increasing urban development in California (Luther & Baptista 2010). Additionally, some species of birds and whales are known to learn their songs from neighbors or family members, and consequently modify their vocalizations on a generational basis (Kroodsma et al. 1996, Noad et al. 2000). Thus, what constitutes suitable training data for an automated detection system may shift over time, giving this domain the property of non-stationarity (Sugiyama & Kawanabe 2012).

Human curation of training data thus has a substantial impact on any automated acoustic detection system's performance (Buxton & Jones 2012, Duan et al. 2013). Automated detectors may do well in the circumstances in which they were trained, only to dip in performance in new environments (Wrege et al. 2017). Depending on the life histories of target species, it is conceivable that researchers might unwittingly train an automated detection model on only a scant few vocalizing individuals in a given monitoring season. In the next season, when population turnover culls some individuals that contributed to the model, and introduces new individuals that present new variation in target vocalizations, the automated detection system may struggle. Systems may therefore benefit from seasonal updates to the training data to accommodate these changes, though no literature has explicitly implemented nor documented the outcome of such an action.

In conclusion, the vast variety in animal sounds offers complex challenges, making certain tools and approaches appropriate only in certain contexts. No one-size-fits-all automated detection approach meets the needs of every research problem (Acevedo et al. 2009, Towsey et al. 2012).

## 1.4. The challenge of modeling wildlife occurrence patterns with acoustic monitoring data

Though it is increasingly easy to collect reams of acoustic data within study areas of interest, and emerging methods can assist with expedient data processing, less is known about how to use these data to create population models with actual utility for researchers and land managers. The ecological research community emphasizes that long-term studies should be prioritized, because they carry more weight in policy and decision-making (Hughes et al. 2018), and are more likely to have conservation and land management utility compared with static, single-season approaches (Dugger et al. 2015, Nichols et al. 2015).

Automated acoustic monitoring using ARUs may be able to support this type of research, most imminently by way of translating audio data streams with automated species detections into dynamic occupancy models (sensu Mackenzie et al. 2003). In contrast to abundance-based population models, an occupancy model requires only presence-absence data for species of interest. Because the probability of detecting a species, given that it is present, is often less than 1 (i.e., false negatives are possibility), the standard occupancy model requires estimation of a nuisance parameter, p, the probability of detecting a species given that it is present (Mackenzie et al. 2002). Numerous extensions to the basic single-season occupancy model have emerged in the past decade, including dynamic (multiple season models) (Mackenzie et al. 2003), and models that accommodate heterogeneity (Royle 2006), false positives (Miller et al. 2011, 2013), both false positives and heterogeneity (Ferguson et al. 2015), and more (Bailey et al. 2014). Occupancy models can be fit using covariates that predict occupancy, and model selection techniques and goodness-of-fit methods can be used to assess and select models, from which inference may be gained in order to guide land management decisions. The free software packages PRESENCE (Hines 2018) and *unmarked* (R package, Fiske & Chandler 2011) support streamlined fitting of these models.

Prior to the work described in chapter 3 of this dissertation, no method had yet been introduced for the development of dynamic occupancy models from automated acoustic monitoring data. Such a framework may be able to support adaptive management paradigms across landscapes, amid climate change and land use change. Adaptive management is a form of management that involves learning over time to inform future management decisions, though it can be difficult to implement systematically (Williams et al. 2009). One successful example of adaptive management is the U.S. Fish and Wildlife Agency's adaptive management program for American waterfowl populations, which uses a mixture of management objectives, monitoring, and model prediction under alternative management scenarios in order to achieve optimal management decisions by way of reductions in uncertainty around population responses to land management over time (Nichols et al. 2007, Johnson et al. 2015). There are not yet examples of dynamic occupancy models used for adaptive management purposes from automated acoustic monitoring data. Camera trap data provides an analog as it is also high volume and often requires some automation to detect focal species, but no cases of using camera trap data for adaptive management exist yet either (Nichols et al. 2011).

## 1.5. Conclusions

Automated acoustic monitoring of wildlife is a rapidly emerging field with several challenges yet to address. The simple act of taking recordings at study areas can present difficult logistical constraints and yield inadequate information about the presence or absence of a target monitoring species. Once recordings have been acquired, overwhelming quantities of audio data, in combination with processing by mistake-prone automated detection systems, can further undermine attempts to characterize the occurrence of focal species through time. The following body of work describes novel methodology for alleviating these challenges.

## 1.6. References

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J. & Aide, T.M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics, 4, 206–214. doi: https://doi.org/10.1016/j.ecoinf.2009.06.005

Adams M.D., Law B.S., & Gibson M.S. (2010). Reliable automation of bat call identification for Eastern New South Wales, Australia, using classification trees and anascheme software. Acta Chiropterologica, 12(1):231–245.

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. PeerJ 1:e103. doi: https://doi.org/10.7717/peerj.103

Alippi, C., Anastasi, G., Galperti, C., Mancini, F., & Roveri, M. (2007). Adaptive sampling for energy conservation in wireless sensor networks for snow monitoring applications. In Mobile Adhoc and Sensor Systems, 2007. MASS 2007. IEEE International Conference on (pp. 1-6). IEEE.

Anastasi, G., Conti, M., Di Francesco, M., & Passarella, A. (2009). Energy conservation in wireless sensor networks: A survey. Ad hoc networks, 7(3), 537-568.

Anderson, S. E., Dave, A. S.  & Margoliash, D. (1996). Template-based automatic recognition of birdsong syllables from continuous recordings. J Acoustical Soc America, 100 (2) Part 1, pp. 1209–1219.

Araya-Salas, M. & Smith-Vidaurre, G. (2017). warbleR: an r package to streamline analysis of animal acoustic signals. Methods Ecol Evol. 8, 184-191.

Astaras, C., Linder, J.M., Wrege, P., Diotoh Orume, R., Macdonald, D.W. (2017). Passive acoustic monitoring as a law enforcement tool for Afrotropical rainforests. Frontiers in Ecology and the Environment, 15(5).

Bailey, L.L., MacKenzie, D.I. & Nichols, J.D. (2014) Advances and applications of occupancy models (E. Cooch, Ed.). Methods in Ecology and Evolution, 5, 1269–1279. DOI:  10.1111/2041-210X.12100

Bas, Y., Bas, D. & Julien, J.-F. (2017). Tadarida: A Toolbox for Animal Detection on Acoustic Recordings. Journal of Open Research Software. 5(1), p.6. doi: http://doi.org/10.5334/jors.154

Bioacoustics Research Program. (2018). Raven Pro 1.5: Interactive Sound Analysis Software [Computer Software]. URL http://www.birds.cornell.edu/raven.

Bishop, C. H., Etherton, B. J., & Majumdar, S. J. (2001). Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. Monthly weather review, 129(3), 420-436.

Bishop, C.M. (2006). Pattern recognition and machine learning. Springer. ISBN: 0387310738 9780387310732.

Brauer, C., T. Donovan, R. Mickey, J. Katz, & Mitchell, B. (2016). A comparison of acoustic monitoring methods for common anurans of the northeastern United States. Wildlife Society Bulletin 40:140-149. doi: 10.1002/wsb.619

Bucher, C. G. (1988). Adaptive sampling—an iterative fast Monte Carlo procedure. Structural safety, 5(2), 119-126.

Buxton, R. T., & Jones, I.L. (2012). Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration. Journal of Field Ornithology 83:47-60. http://dx.doi.org/10.1111/j.1557-9263.2011.00355.x

Bye, S. L., Robel, R. J., & Kemp, K. E. (2001). Effects of human presence on vocalizations of grassland birds in Kansas. The Prairie Naturalist, 33(4):249–256.

Campbell, M., & Francis, C.M. (2011). Using stereo-microphones to evaluate observer variation in North American Breeding Bird Survey point counts. Auk 128:303-312. http://dx.doi.org/10.1525/auk.2011.10005

Catchpole, C.K. & Slater, P.J.B. (2008). Bird Song: Biological Themes and Variations, 2nd Ed. Cambridge University Press, Cambridge, UK.

Cerquiera, M. & Aide, T.M. (2016). Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling. *Methods in Ecology and Evolution*, **7**, 1340-1348. DOI: 10.1111/2041-210X.12599

Corrada-Bravo, C.J.C., Berrios, R.A, & Aide, T.M. (2017). Species-specific audio detection: a comparison of three template-based detection algorithms using random forests. PeerJ Computer Science 3:e113. doi: https://doi.org/10.7717/peerj-cs.113

Dawson, D. K., & M. G. Efford. (2009). Bird population density estimated from acoustic signals. Journal of Applied Ecology 46:1201-1209. http://dx.doi.org/10.1111/j.1365-2664.2009.01731.x

Duan, S., Zhang, J., Roe, P., Wimmer, J., Dong, X., Truskinger, A., & Towsey, M. (2013) Timed Probabilistic Automaton : a bridge between Raven and Song Scope for automatic species recognition. In Muñoz-Avila, Hector & Stracuzzi, David J. (Eds.). Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference, AAAI, Bellevue, Washington, USA, pp. 1519-1524.

Dugger, K. M., Forsman, E. D., Franklin, A. B., Davis, R. J., White, G. C., Schwarz, C. J., et al., & Doherty Jr, P. F. (2015). The effects of habitat, climate, and Barred Owls on long-term demography of Northern Spotted Owls. The Condor, 118(1), 57-116.

Dyo, V., Ellwood, S. A., MacDonald, D. W., Markham, A., Trigoni, N., Wohlers, R., Mascolo, C., Pasztor, B., Scellato, S., & Yosef, K. (2012). WILDSENSING: Design and deployment of a sustainable sensor network for wildlife monitoring. ACM Transactions on Sensor Networks 8 (4), Article 29. doi: 10.1145/2240116.2240118

Fagerlund S. (2007). Bird species recognition using support vector machines. EURASIP Journal on Applied Signal Processing 2007(1):1–8.

Farnsworth, A. (2005). Flight calls and their value for future ornithological studies and conservation research. The Auk, 122(3), 733-746.

Ferguson, P. F., Conroy, M. J., & Hepinstall-Cymerman, J. (2015). Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys. Methods in Ecology and Evolution, 6(12), 1395-1406.

Fiske, I., & Chandler, R. (2011). unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. Journal of Statistical Software, 43(10), 1-23. URL http://www.jstatsoft.org/v43/i10/.

Frey, S., Fisher, J.T., Cole Burton, A., & Volpe, J.P. (2017). Investigating animal activity patterns and temporal niche partitioning using camera-trap data: challenges and opportunities. Remote Sensing in Ecology and Conservation, 3(3).

Frick, W. F., Stepanian, P. M., Kelly, J. F., Howard, K. W., Kuster, C. M., Kunz, T. H., & Chilson, P. B. (2012). Climate and weather impact timing of emergence of bats. PLoS One, 7(8), e42737.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York, NY, USA: Springer series in statistics.

Furnas, B. J., & Callas, R.L. (2015). Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. Journal of Wildlife Management 79:325-337. DOI: http://dx.doi.org/10.1002/jwmg.821

Gage, S.H., Napoletano, B., & Cooper, M., (2001). Assessment of ecosystem biodiversity by acoustic diversity indices. J. Acoust. Soc. Am. 109 (5), 2430.

Gage, S. H., W. Joo, E. P. Kasten, J. Fox, & S. Biswas. (2015). Acoustic observations in agricultural landscapes. Pages 360-377 in S. K. Hamilton, J. E. Doll, & G. P. Robertson, editors. The Ecology of Agricultural Landscapes: Long-Term Research on the Path to Sustainability. Oxford University Press, New York, New York, USA. ISBN: 978-0199773350.

Gaynor, K.M., Hojnowski, C.E., Carter, N.H., & Brashares, J.S. (2018). The influence of human disturbance on wildlife nocturnality. Science, 360 (6394) 1232-1235. DOI: 10.1126/science.aar7121

Goëau, H., Glotin, H., Vellinga, W.-P., Planqué, R. & Joly, A. (2016). LifeCLEF bird identification task 2016: The arrival of deep learning. In Working Notes of CLEF 2016-Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016, 440–449.

Gutzwiller, K. J., & Marcum, H. A. (1997). Bird reactions to observer clothing color: implications for distance-sampling techniques. Journal of Wildlife Management 61:935-947. http://dx.doi.org/10.2307/3802203

Hafner S. & Katz J. (2018). monitoR: Acoustic template detection in R. R package version 1.0.7, URL: http://www.uvm.edu/rsenr/vtcfwru/R/?Page=monitoR/monitoR.htm.

Hayes, A. R., & Huntly, N. J. (2005). Effects of wind on the behavior and call transmission of pikas (Ochotona princeps). Journal of mammalogy, 86(5), 974-981.

Heinicke, S., Kalan, A.K., Wanger, O.J., Mundry, R., Lukashevich, H., & Kuhl, H.S. (2015). Assessing the performance of a semi-automated acoustic monitoring system for primates. Methods in Ecology and Evolution, 6(7): 753-763. doi: 10.1111/2041-210X.12384

Hill, A. P., Prince, P., Piña Covarrubias, E., Doncaster, C. P., Snaddon, J. L., & Rogers, A. (2018). AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. Methods in Ecology and Evolution, 9(5), 1199-1211.

Hines, J. (2018). RPresence for PRESENCE: Software to estimate patch occupancy and related parameters. Version 12.10. https://www.mbr-pwrc.usgs.gov/software/presence.html

Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, & Van Wilgenburg, S.L. (2002). Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. Wildlife Society Bulletin 30:709-720.

Hughes, B.B., Beas-Luna, R., Barner, A.K., Brewitt, K., Brumbaugh, D.R., Cerny-Chipman, E.B., Close, S.L., Coblentz, K.E., de Nesnera, K.L., Drobnitch, S.T., et al. (2017). Long-term studies contribute disproportionately to ecology and policy. Bioscience, 67(3), 271-281. DOI: 10.1093/biosci/biw185

Hutto, R. L., & Stutzman, R.J. (2009). Humans versus autonomous recording units: a comparison of point-count results. Journal of Field Ornithology 80:387–398.

Jain, A., & Chang J.Y. (2004). Adaptive sampling for sensor networks, in: Proc. 1st international workshop on Data management for sensor networks (DMSN 2004), Toronto, Canada, August 30th, 2004, pp. 10–16.

Johnson, F. A., Boomer, G. S., Williams, B. K., Nichols, J. D., & Case, D. J. (2015). Multilevel learning in the adaptive management of waterfowl harvests: 20 years and counting. Wildlife Society Bulletin, 39(1), 9-19.

Katz, J., Hafner, S.D. & Donovan, T. (2016). Assessment of Error Rates in Acoustic Monitoring with the R package monitoR. Bioacoustics, 25, 177–196. doi: https://doi.org/10.1080/09524622.2015.1133320

Knight, E. C., Hannah, K.C., Foley, G., Scott, C., Mark Brigham, R., & Bayne, E. (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology 12(2):14. https://doi.org/10.5751/ACE-01114-120214

Kroodsma, D.E., & Miller, E.H. (eds). (1996). Ecology and evolution of acoustic communication in birds. Comstock Pub, University of Michigan. ISBN: 0801430496, 9780801430497

Leecaster, M. K., & Weisberg, S. B. (2001). Effect of sampling frequency on shoreline microbiology assessments. Marine pollution bulletin, 42(11), 1150-1154.

Lévy, C., Linarès, G., & Nocera, P. (2003). Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems. In Workshop on DSP in Mobile and Vehicular Systems, Nagoya, Japan.

Luther, D., & Baptista, L. (2010). Urban noise and the cultural evolution of bird songs. Proceedings of the Royal Society B: Biological Sciences 277: 469–473.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A. & Langtimm, C.A. (2002). Estimating site occupancy rates when detection probabilities are less than one. Ecology, 83, 2248–2255. doi: 10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology, 84, 2200–2207. DOI: 10.1890/02-3090

MacKenzie, D. I., & Royle, J. A. (2005). Designing occupancy studies: general advice and allocating survey effort. Journal of applied Ecology, 42(6), 1105-1114.

Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R., & Anderson, J. (2002). Wireless sensor networks for habitat monitoring. In Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications (pp. 88-97). Acm.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D. & Tyack, P. L. (2012). Estimating animal population density using passive acoustics. Biological Reviews.

McClintock, B.T., Bailey, L.L., Pollock, K.H. & Simons, T.R. (2010). Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, **91**, 2446–2454. DOI: https://doi.org/10.1890/09-1287.1

McKown, M.W. (2012). A wireless acoustic sensor network for monitoring wildlife in remote locations. The Journal of the Acoustical Society of America, 132, 2036. DOI: https://doi.org/10.1121/1.4755484

Mennill, D. J., & Vehrencamp, S.L. (2008). Context-dependent functions of avian duets revealed by microphone-array recordings and multispeaker playback. Current Biology 18:1314-1319. http://dx.doi.org/10.1016/j.cub.2008.07.073

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology, 92(7), 1422-1428.

Miller, D. A., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., & Mitchell, M. S. (2013). Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data. PLoS one, 8(6), e65808.

Miller, D.A.W., Bailey, L.L., Grant, E.H.C., McClintock, B.T., Weir, L.A. & Simons, T.R. (2015). Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known (O. Gimenez, Ed.). *Methods in Ecology and Evolution*, **6**, 557–565. DOI: 10.1111/2041-210X.12342

Moore, A.L. & McCarthy, M.A. (2016). Optimizing ecological survey effort over space and time. Methods in Ecology and Evolution, 7, 891–899. doi: 10.1111/2041-210X.12564

Newson, S.E., Bas, Y., Murray, A., & Gillings, S. (2017). Potential for coupling the monitoring of bush-crickets with established large-scale acoustic monitoring of bats. *Methods in Ecology and Evolution*, **8**, 1051-1062. DOI: 10.1111/2041-210X.12720

Nichols, J. D., Runge, M. C., Johnson, F. A., & Williams, B. K. (2007). Adaptive harvest management of North American waterfowl populations: a brief history and future prospects. Journal of Ornithology, 148(2), 343-349.

Nichols, J. D., Karanth, K. U., & O’Connell, A. F. (2011). Science, conservation, and camera traps. In Camera Traps in Animal Ecology (pp. 45-56). Springer, Tokyo.

Nichols, J.D., Yackulic, C.B., Reid, J., Hines, J.E., Davis, R, & Forsman, E. (2015). Dynamic occupancy modeling for conservation. Presented at Ecological Society of America Annual Meeting, Baltimore, MD, August 2015.

Noad, M. J., Cato, D. H., Bryden, M. M., Jenner, M. N., & Jenner, K. C. S. (2000). Cultural revolution in whale songs. Nature, 408(6812), 537.

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., et al. (2011). Soundscape ecology: the science of sound in the landscape. BioScience, 61(3), 203-216.

Pollock, K. H., Nichols, J. D., Simons, T. R., Farnsworth, G. L., Bailey, L. L., & Sauer, J. R. (2002). Large scale wildlife monitoring studies: statistical methods for design and analysis. Environmetrics, 13(2), 105-119.

Porter, J., Arzberger, P., Braun, H. W., Bryant, P., Gage, S., Hansen, T., et al. & Michener, W. (2005). Wireless sensor networks for ecology. AIBS Bulletin, 55(7), 561-572.

Potamitis, I., Ntalampiras, S., Jahn, O. & Riede, K. (2014). Automatic bird sound detection in long real-field recordings: applications and tools. Applied Acoustics 80:1-9. doi: https://doi.org/10.1016/j.apacoust.2014.01.001

Priyadarshani N., Marsland S., Castro I., & Punchihewa A. (2016). Birdsong Denoising Using Wavelets. PLoS ONE 11(1): e0146790. doi:10.1371/ journal.pone.0146790

Raghunathan, V., Ganeriwal, S., & Srivastava, M. (2006). Emerging techniques for long lived wireless sensor networks. IEEE Communications Magazine, 44(4), 108-114.

Ralph, C. J., Sauer, J. R., & Droege, S. (1995). Monitoring bird populations by point counts. Gen. Tech. Rep. PSW-GTR-149. Albany, CA: US Department of Agriculture, Forest Service, Pacific Southwest Research Station. 187 p, 149.

Ranjard, L., Reed, B.S., Landers, T.J., Raynar, M.J., Friesen, M.R., Sagar, R.L., & Dunphy, B.J. (2017). MatlabHTK: a simple interface for bioacoustics analyses using hidden Markov models. Methods in Ecology and Evolution 8(5): 615-621. doi: 10.1111/2041-210X.12688

Robbins, C. S., D. Bystrak, & P. H. Geissler. (1986). The breeding bird survey: its first fifteen years, 1965–1979. Resource Publication No. 156, U.S. Fish and Wildlife Service, Washington, DC, USA.

Root-Gutteridge, H., Bencsik, M., Chebli, M., Gentle, L. K., Terrell-Nield, C., Bourit, A., & Yarnell, R. W. (2014). Identifying individual wild Eastern grey wolves (Canis lupus lycaon) using fundamental frequency and amplitude of howls. Bioacoustics, 23(1), 55-66.

Rosenstock, S. S., D. R. Anderson, K. M. Giesen, T. Leukering, & M. F. Carter. (2002). Landbird counting techniques: current practices and an alternative. Auk 119:46-53. http://dx.doi.org/10.1642/0004-8038(2002)119[0046:LCTCPA]2.0.CO;2

Royle, J. A. (2006). Site occupancy models with heterogeneous detection probabilities. Biometrics, 62(1), 97-102.

Ruiz-Gutierrez, V., Hooten, M.B. & Campbell Grant, E.H. (2016). Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias (N. Yoccoz, Ed.). Methods in Ecology and Evolution, 7, 900–909. doi: 10.1111/2041-210X.12542

Salamon, J., J. Pablo Bello, A. Farnsworth, M. Robbins, H. Klinck, S. Keen, & Kelling, S. (2016). Towards the automatic classification of avian flight calls for bioacoustic monitoring. PLoS ONE. doi: 10.1371

Shonfield, J. & Bayne, E.M. (2017). Autonomous recording units in avian ecological research: current use and future applications. Avian Conservation and Ecology, 12(1): 14. DOI: 10.5751/ACE-00974-120114

Sidie-Slettedahl, A. M., K. C. Jensen, R. R. Johnson, T. W. Arnold, J. E. Austin, & Stafford, J.D. (2015). Evaluation of autonomous recording units for detecting 3 species of secretive marsh birds. Wildlife Society Bulletin 39:626-634. http://dx.doi.org/10.1002/wsb.569

Sieve Analytics. (2018). Products. Url: https://www.sieve-analytics.com/products

Smith, D. R., Conroy, M. J., & Brakhage, D. H. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. Biometrics, 777-788.

Stowell, D., Wood, M., Stylianou, Y. & Glotin, H. (2016). Bird detection in audio: a survey and a challenge. IEEE International Workshop on Machine Learning for Signal Processing, Salerno, Italy. doi: 10.1109/MLSP.2016.7738875

Stowell, D., Stylianou, Y., Wood, M., Pamuła, H., & Glotin, H. (2018). Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. arXiv preprint arXiv:1807.05812.

Sueur J., Aubin T., Simonis C. (2008). Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics, 18: 213-226

Sueur, J., & Farina, A. (2015). Ecoacoustics: the ecological investigation and interpretation of environmental sound. Biosemiotics, 8(3), 493-502.

Sugiyama, M. & Kawanabe, M. (2012). Machine learning in non-stationary environments: Introduction to covariate shift adaptation. MIT press.

Suthers, R.A., Fitch, W.W., Fay, R.R., & Popper, A.N. (eds) (2016). Vertebrate sound production and acoustic communication. Springer International Publishing, Switzerland. DOI: https://doi.org/10.1007/978-3-319-27721-9 ISBN: 978-3-319-27721-9

Swiston, K. A., & Mennill, D.J. (2009). Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-Billed, and putative Ivory-Billed Woodpeckers. Journal of Field Ornithology 80:42-50. http://dx.doi.org/10.1111/j.1557-9263.2009.00204.x

Thompson, S.K. & Seber, G.A.F. (1994). Detectability in Conventional and Adaptive Sampling. Biometrics, 50, 712. doi: 10.2307/2532785

Thompson, W.L., White, G.C., & Gowan, C. (1998). Monitoring Vertebrate Populations. Academic Press, San Diego. ISBN: 0126889600

Thompson, W. L. (2004). Sampling rare or elusive species: concepts and techniques for estimating population parameters. Island Press, Washington, D.C., USA. ISBN: 9781559634519

Towsey, M., Planitz, B., Nantes, A., Wimmer, J. & Roe, P. (2012). A toolbox for animal call recognition. Bioacoustics, 21, 107–125.doi: http://dx.doi.org/10.1080/09524622.2011.648753

Turk, P. & Borkowski, J.J. (2005). A review of Adaptive Cluster Sampling: 1990-2003. Environmental and Ecological Statistics, 12(1), 55-94. doi: https://doi.org/10.1007/s10651-005-6818-0

Villanueva-Rivera, L.J, & Pijanowski, B. (2018). Soundecology: Soundscape Ecology. R Package. URL: https://cran.r-project.org/web/packages/soundecology/index.html

Whytock, R. C., & Christie, J. (2017). Solo: an open source, customizable and inexpensive audio recorder for bioacoustic research. Methods in Ecology and Evolution, 8(3), 308-312.

Wildlife Acoustics. (2018). Kaleidescope [Computer Software]. URL http://www. wildlifeacoustics.com.

Williams, B.K., Szaro, R.C., & Shapiro, C.D. (2009). Adaptive Management: The U.S. Department of the Interior Technical Guide, 2nd Edition. Adaptive Management Working Group, U.S. Department of the Interior, Washington DC.

Wrege, P.H., Rowland, E.D., Bout, N., & Doukaga, M. (2012). Opening a larger window onto forest elephant ecology. African Journal of Ecology, 50 (2).

Wrege, P.H., Rowland, E.D., Keen, S., & Shiu, Y. (2017). Acoustic monitoring for conservation in tropical forests: examples from forest elephants. Methods in Ecology and Evolution, 8 (10).

Xu, Y., Choi, J., Dass, S., & Maiti, T. (2011). Bayesian prediction and adaptive sampling algorithms for mobile sensor networks. In American Control Conference (ACC), 2011 (pp. 4195-4200). IEEE.

Zamora-Gutierrez, V., Lopez-Gonzalez, C., MacSwiney Gonzalez, M.C., Fenton, B., Jones, G., Kalko, E. K. V., Puechmaille, S.J., Stathopoulos, V., Jones, K.E. (2016). Acoustic identification of Mexican bats based on taxonomic and ecological constraints on call design. Methods in Ecology and Evolution, 7(9).

Zhou, J & De Roure, D. (2007). FloodNet: coupling adaptive sampling with energy aware routing in a flood warning system. Journal of Computer Science and Technology, 22 (1), pp. 121-130.

# CHAPTER 2: TEMPORALLY-ADAPTIVE ACOUSTIC SAMPLING TO MAXIMIZE DETECTION ACROSS A SUITE OF FOCAL WILDLIFE SPECIES

Cathleen Balantic[1*]

Therese Donovan[2]

[1] Corresponding Author: cathleen.balantic@uvm.edu; Vermont Cooperative Fish and Wildlife Research Unit, University of Vermont, Burlington, VT 05405, USA

[2] U.S. Geological Survey, Vermont Cooperative Fish and Wildlife Research Unit, Rubenstein School of Environment and Natural Resources, University of Vermont, Burlington, VT 05405, USA

*Prepublication Disclaimer:*

*This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy*

## 2.1. Abstract

1. Acoustic recordings of the environment can produce species presence-absence data for characterizing populations of sound-producing wildlife over multiple spatial scales. If a species is present at a site but does not vocalize during a scheduled audio recording survey, researchers may incorrectly conclude that the species is absent (‘false negative’). The risk of false negatives is compounded when audio devices do not record continuously and must be manually scheduled to operate at pre-selected times of day, particularly when research programs target multiple focal species with vocal availability that varies across temporal conditions.
2. We developed a temporally-adaptive acoustic sampling algorithm to maximize detection probabilities for a suite of focal species amid sampling constraints. The algorithm combines user-supplied species vocalization models with site-specific weather forecasts to set an optimized sampling schedule for the following day. To test our algorithm, we simulated hourly vocalization probabilities for a suite of focal species in a hypothetical monitoring area for the year 2016. We conducted a factorial experiment that sampled from the 2016 acoustic environment to compare the probability of acoustic detection by a fixed (stationary) schedule vs. a temporally-adaptive optimized schedule under several sampling efforts and monitoring durations.
3. We found that over the course of a study season, the probability of acoustically capturing a focal species at least once via automated acoustic monitoring was

greater (and acoustic capture occurred earlier in the season) when using the temporally-adaptive optimized schedule as compared to a fixed schedule.

4. The advantages of a temporally-adaptive optimized acoustic sampling schedule are magnified when a study duration is short, sampling effort is low, and/or species vocal availability is minimal. This methodology thus offers new possibilities to the existing paradigms for adaptive wildlife sampling and acoustic monitoring, potentially allowing research programs to maximize sampling efforts amid constraints.

Key Words

Automated acoustic monitoring; bioacoustics; optimization; adaptive sampling; occupancy; detection probability; wildlife

## 2.2. Introduction

Automated remote acoustic monitoring of wildlife offers a means to characterize the distribution of sound-producing species – such as birds, amphibians, bats, and insects – across vast landscapes (Dawson & Efford 2009, Marques et al. 2013). Because acquiring species abundance data is often logistically impractical at large spatial scales, research programs may instead collect species presence-absence data, an endeavor with which automated remote acoustic monitoring is compatible (Furnas & Callas 2014, Cerquiera & Aide 2016). In a typical remote acoustic monitoring program, audio recording devices deployed at fixed locations take environmental recordings based on a schedule that has been manually input to the device. Commercially available recording units often store recordings directly on the device (e.g., Wildlife Acoustics 2016), which obligates the researcher to be physically present to retrieve data from a storage card. Alternatively, in an emerging paradigm, recordings units may expedite data access and analysis by sending files in near-real time to a server using a cellular or Wi-Fi network (McKown et al. 2012, ARBIMON: Aide et al. 2013, Gage et al. 2017, Balantic & Donovan in prep).

The difficulty with manually setting a recording schedule to survey wildlife is that truly present species do not always cooperate by vocalizing during the recording session. If a species is present but does not vocalize during scheduled recording periods, the species is logged as absent, resulting in a "false negative" (MacKenzie et al. 2002). Across time and space, deficient fixed recording schedules can fail to adequately describe a pattern of occupancy, which could potentially result in conservation management decisions that are at odds with management objectives. For example, if an amphibian species of interest only vocalizes after the first substantial rainfall of the season, as is the case for Couch's Spadefoot Toad (*Scaphiopus couchii*) in the Sonoran Desert (Mayhew 1965), and no recordings were scheduled at a time that captures this event, then researchers may conclude the species is likely absent. Resource managers may subsequently use this information to make land use decisions that unwittingly sabotage their own conservation goals. As such, low species detection probabilities motivate the development of sampling protocols that improve the chances of detecting a species given that it is present (MacKenzie et al. 2006).

The task of avoiding false negatives is magnified when large-scale acoustic monitoring regimes attempt to track multiple focal species that are available under varying conditions (Manley et al. 2004, McKown 2012). Focal species may have diverse behaviors and life histories, driving vocalization activity patterns that vary across time of day, time of year, and weather conditions. For example, a comprehensive monitoring program may be interested in tracking the occurrence patterns of breeding birds that vocalize on spring mornings with minimal rain and wind, seasonally available amphibians that only vocalize after fall monsoon rains, and nocturnally active species such as nightjars (Caprimulgidae family) or coyotes (Canidae family). Certain species within the focal set may be of special concern and therefore merit higher monitoring priority. Thus, remote acoustic monitoring programs targeting multiple species face the prospect of low detection probabilities for some or all targets if using a fixed, manually applied schedule for sampling.

Alongside detection challenges, acoustic monitoring programs often encounter constraints that restrict sampling efforts, prompting the need for guidance in the development of effective sampling schedules that avoid squandering key resources. Contingent on program circumstances, budget and logistical limitations may curb the total number of allowable audio samples, total amount of sampling time, and sample file sizes for storage or efficient transfer over a mobile or Wi-Fi network (Gage et al. 2015). Even if a Wi-Fi or cellular network is available to facilitate the real-time transmission of audio recordings (allowing researchers to avoid collecting recordings from on-site memory cards), some portion of the research budget is required to support the Wi-Fi or cellular data plan, which may limit the total recording time that can be taken and transmitted over the network. Additionally, if network signal is weak, it is prudent to limit recordings to short intervals of time (~1-2 minutes) to ensure efficient and reliable file transmission over the network, particularly if using high sampling rates (44.1+ kHz) and/or uncompressed file formats (e.g. .wav).

Addressing these emergent acoustic monitoring challenges is crucial for building an expedient acoustic monitoring framework. As human land use and climate change continue to influence wildlife ranges and populations, there is a need to characterize status and trends of species that have been poorly understood and described (Thompson 2004). Lacking a framework for optimizing acoustic sampling schedules amid constraints, landscape-scale bioacoustic monitoring programs may fail to take full advantage of their monitoring efforts, resulting in compromised scientific inference and sub-optimal conservation management decisions.

In this paper, we introduce a novel, temporally-adaptive acoustic monitoring methodology for recording devices that can communicate remotely via Wi-Fi or cellular network. Devices that can send recordings are inherently equipped to receive external instructions about when to record on the following day. Our method optimizes these instructions across time and monitoring locations by tracking *p** or *p(capture),* the probability of acoustically capturing (detecting) a target species at least once at any monitoring site at any time during the study (*sensu* Otis et al. 1978). By tracking *p** for each species at each site on a daily basis, the timing of future acoustic surveys is allowed to vary across sites as a function of information from previous surveys. Once *p** reaches a user-defined threshold for target species at a given site, those species are released from future monitoring priority, allowing the recording schedule to focus more heavily on species that remain below their target thresholds. Acoustic monitoring thus offers an opportunity to implement flexible, temporally-adaptive sampling schedules that adjust automatically to optimize detection probabilities across a suite of focal species.

## Objectives

The goal of this work was to develop and evaluate the utility of a temporally-adaptive, automated acoustic sampling algorithm, and assess its potential for maximizing detection of multiple focal species. Our objectives were to 1) Develop a temporally-adaptive automated acoustic sampling algorithm for acoustic wildlife monitoring subject to species prioritization and sampling constraints, 2) Simulate hourly vocalization probabilities for nine species across 133 sites in a hypothetical monitoring area for the year 2016, 3) Implement a 2 x 6 x 2 factorial experiment to compare the probability of acoustic detection across sites in the 2016 vocalization simulation under differing monitoring protocols: schedule type (n = 2 levels; fixed schedule versus optimized temporally-adaptive schedule), sampling effort ($S$ = 6 levels: 2, 5, 10, 20, 30, or 40 minutes sampling per day at each site), and monitoring duration ($D$ = 2 levels: full year [$d$ = 366 days for the 2016 leap year] versus bird breeding season only [$d$ = 31 days]).

## 2.3. Materials and Methods

### Objective 1: Develop an optimized adaptive sampling algorithm subject to species prioritization and sampling constraints

We engineered a temporally-adaptive sampling algorithm (**Fig. 2. 1**) designed to maximize detections across $K$ target species and $R$ study sites for $D$ days, conditional on presence. The sampling schedule's unit of temporal adaptation was one day (i.e., the schedule updated every 24 hours, and could not change mid-day). In this approach, audio samples were collected on day $d$. Each day, based on these samples and forecasted temporal data, an optimized recording schedule was determined for the next day ($d$ + 1).

Three fundamental user-defined inputs provided the functionality for schedule optimization (**Fig. 2. 1**):

1. **Species Vocalization Models:** First, we created logistic regression vocalization models that reflected our knowledge about each of the $K$ target species' vocalization patterns. We then used these models to predict the probability of vocalization ($p_v$) for each species at each monitoring site during any given hour of the day given existing weather and temporal conditions (User Input 1; **Fig. 2. 1a**).

2. **Species Monitoring Priority Weights:** For each species in the focal group, we assigned an initial weight that reflected its user-defined monitoring priority throughout the entire study period. Weights can be equal across focal species, or asymmetrical if a research program has varied species monitoring priorities and/or anticipates greater or lesser calling availability of certain species *a priori.* Furthermore, these weights may be constant across the *R* sites, or may vary if certain monitoring sites are prioritized above others (User Input 2; **Fig. 2. 1a**). We used equal monitoring priority weights for each species at each site. The algorithm updated these weights on a daily basis as monitoring progressed.

3. **Species Acoustic Capture Thresholds:** Third, for each species and site combination, we chose a monitoring threshold that informed the allocation of samples at each site. We designated this user-defined monitoring threshold as $p*_{max}$, or the maximum cumulative probability of acoustic capture (Otis et al. 1978, White et al. 1982). For example, a $p*_{max}$ value of 0.95 for a given species at a given site indicated that monitoring should continue for this species at this site until the probability of detecting the species *at least one time* across the full monitoring period (D) met or exceeded 0.95 (i.e., monitor until $p* \geq p*_{max}$; User Input 3, **Fig. 2. 1a**).

These three key inputs drove the optimized schedule (**Fig. 2. 1**), and utilized functions within the R package *AMMonitor* (Balantic et al. in prep) on day *d* to produce the optimized recording schedule for each site on day *d* + 1. For each day *d* of monitoring, we used *AMMonitor*'s *temporalsGet()* function to obtain site-specific, hourly weather forecast data for the next day (*d* + 1). We combined this temporal data with the species vocalization models to predict each species' hourly probability of vocalization ($p_v$) at each site on day *d* + 1, hereafter "site-hour" (**Fig. 2. 1b-2. 1c**). Next, the *AMMonitor* function *scheduleOptim()* calculated a single overall score for each site-hour, computed as the dot-product of the species weights vector and the species vocalization probabilities vector (**Fig. 2. 1d**). On day *d* = 1, the weights vector consisted of the species monitoring weights assigned at the start of the monitoring program (**Fig. 2. 1a**). In later iterations, it was a vector updated based on the probability of acoustic capture ($p^*$) computed from previous sampling intervals (**Fig. 2. 1h**). The site-hour scores were then ranked for each site, identifying the optimal hour(s) for sampling within each site for day *d* + 1. The *scheduleOptim()* function then scheduled *S* one-minute samples, evenly spaced, into the highest scoring hour(s) for each site for day *d* + 1 (**Fig. 2. 1e**). The schedule was then sent to the recording unit, which collected audio samples as instructed the following day (**Fig. 2. 1f**). Based on the optimized recording schedules (which could vary from site to site) and the $p_v$ associated with that hour for each species, we then computed $p^*_d$ for each species at each site, where $p^*_d$ was defined as the probability of detecting the species at least once that day given the sampling schedule (**Fig. 2. 1g**). We recomputed the cumulative probability of acoustic capture across *all* previous days ($p^*$) for each species at each site at the end of each day (**Fig. 2. 1g**). The daily update of $p^*$ permitted priority weights of each species at each site to shrink or grow based on how likely it was that the species has already been adequately acoustically captured by previous sampling (**Fig. 2. 1h**). When $p^*$ equaled or exceeded our chosen $p^*_{max}$ threshold at a given site, the species' updated weight at that site dropped to zero, allowing remaining sampling to emphasize species for which acoustic capture remained inadequate. The algorithm repeated daily until the sampling period $D$ was complete or until all $p^* \geq p^*_{max}$ for each species at each site.

## Objective 2: Simulate hourly vocalization probabilities for nine species across 133 sites in a hypothetical monitoring area for the year 2016

### 2.1 Study Site

To test the utility of the algorithm, we simulated hourly vocalization probabilities for 9 species across 133 sites for 366 days (2016 was a leap year), and then sampled from this acoustic environment in Objective 3. Our focal study area in this work is the U.S. Bureau of Land Management's (U.S. BLM) Riverside East Solar Energy Zone, a 599 km$^2$ parcel allocated as a utility-scale solar renewable energy hub and located between Desert Center, CA and Blythe, CA. The Riverside East Solar Energy Zone (**Fig. 2. 2**) contains 133 sites actively monitored under an adaptive management protocol for vegetation indicators (U.S. Bureau of Land Management, 2016). We used these 133 sites as study locations for our simulation.

### 2.2 Study Species

Based on literature and the monitoring interests of U.S. BLM, we selected nine study species for this simulation: Black-tailed Gnatcatcher (*Polioptila melanura*), Common Poorwill (*Phalaenoptius nuttallii*), Couch's Spadefoot Toad (*Scaphiopus couchii*), Coyote (*Canis latrans*), Eurasian Collared-Dove (*Streptopelia decaocto*), Gambel’s Quail (*Callipepla gambellii*), Lesser Nighthawk (*Chordeiles acutipennis*), Phainopepla (*Phainopepla nitens*), and Verdin (*Auriparus flaviceps*). These species represented a mix of phylogenetic classes, diurnal and nocturnal vocalizers, early and late-year vocalizers, common and uncommon vocalizers, residents and non-residents, and species that are of conservation concern vs. invasive species (**Table 2. 1**).

## 2.3 Vocalization Models

We used the *AMMonitor* function *simGlmModel()* to create literature-based logistic regression models that predicted the probability of vocalizing at least once during a single hour of a given day for all nine target species ($p_v$), conditional on presence. This function produced a statistical model of class ‘*glm’* (generalized linear model) in R. Model covariates for any given species included date, hour of day, lunar phase, and proximity to sunrise and/or sunset, as well as weather conditions such as temperature, wind, and precipitation. In the interest of simplicity, we did not include any spatial (habitat) covariates.

To accommodate the circular nature of temporal predictive variables like day of year, hour of day, and lunar phase, we modeled sin and cosine-based coefficients. For example, we modeled hour of the day on a 24-hour scale as sin(2*pi*hour.of.day/24) and cosine(2*pi*hour.of.day/24). To provide finer control over the modeling outcome, we also modeled hour of the day on a 12-hour scale as sin(2*pi*hour.of.day/12) and cosine(2*pi*hour.of.day/12). To illustrate with a hypothetical example, the 0-intercept model *M* describes the vocalization process of Eurasian Collared-Dove (*Streptopelia decaocto*):

$$M = 0 + 1*\sin(2*pi*day.of.year/366) - 2*cosine(2*pi*hour.of.day/12) - 0.000005* time.to.sunrise^2 + 0.009*temperature - 0.000001*temperature^3 - 0.25*wind.speed$$

The probability of vocalizing at least once during a given hour on a given day ($p_v$) was subsequently obtained by applying the logit link function:

$$p_v = exp(M) / (1 + exp(M))$$

We developed logistic regression models that reflected our literature-based knowledge about vocalization activity for all nine focal species (**Table 2. 2**). All models used some combination of distance to sunrise/sunset and/or circular temporal variables (day of year, time of day) modeled with sin and cosine. We visualized the impacts of these variables on each species' vocalization probability in **Fig. 2.3**. Temperature and wind speed were included for all diurnal avian species (U.S. Geological Survey 2001). The nocturnal avian species models included variables for wind speed and cosine of the lunar phase because vocal availability may be improved on moonlit nights (Woods 2005). The coyote model also contained the cosine of the lunar phase because this species is often more vocally active at the new moon (Bender et al. 1996). The Couch's Spadefoot Toad model included rain accumulation within the past 24 hours (Mayhew 1965). Based on the literature, we made Couch's Spadefoot Toad, Coyote, and Lesser Nighthawk less vocally available and thus more difficult to detect (**Table 2. 2**).

### 2.4 Calculate $p_v$ for each site-hour for each species at each location

For each day of 2016, we acquired hourly weather data for all 133 study sites using the *AMMonitor* function *temporalsGet().* This function utilized the Dark Sky API

(Dark Sky 2017) to provide hourly data for precipitation intensity, precipitation probability, temperature, dew point, pressure, wind speed, cloud cover, ultraviolet index, visibility, and ozone, as well as the daily sunrise time, sunset time, and lunar phase associated with each monitoring site. The function appended variables such as the absolute value of time to sunrise or sunset, predicted rain accumulation in the previous 24 hours, day of year, and hour of day, and the aforementioned circular sin and cosine-based predictors. We supplied the finalized covariate dataset and the class *glm* vocalization models ($n$ = 9) to R's *predict()* function to generate the probability of vocalization ($p_v$) for each species, at each location, during each hour for the year 2016 in its entirety. This resulted in a dataset consisting of 9 species * 133 sites * 24 hours * 366 days = 10,514,448 $p_v$ records from which to sample in Objective 3.

## Objective 3: Apply both the optimized schedule and fixed (stationary) sampling schedule to the simulated environment, and compare performance of the optimized schedule and fixed schedule at different sampling efforts and study season lengths.

We implemented a 2 x 6 x 2 factorial experiment that subsampled the Objective 2 vocalization simulation. The experiment consisted of two scheduling treatments ($Tr$ = optimized or fixed) at six sampling effort levels ($S$ = 2, 5, 10, 20, 30, or 40 minutes per day of sampling) and under two study durations ($D$ = "Full Year (366 days)": the full 2016 year using all nine species, and "March Only (31 days)": a sole focus on the March 2016 breeding season, where most focal species were expected to be especially active, and where Couch's Spadefoot Toad was omitted since it was not expected to be active).

For the Full Year Optimization treatment, we applied our daily temporally-adaptive sampling protocol beginning on January 1, 2016 and ending on December 31, 2016. For the March Only Optimization treatment, the temporally-adaptive sampling protocol began on March 1, 2016, and concluded on March 31, 2016. In both cases, we set each initial **Species Monitoring Priority Weight** to be equal at each site (1 divided by the total number of focal species) (**Table 2. 3**). Additionally, we selected **Species Acoustic Capture Thresholds** ($p*_{max}$) of 0.95 for each species at each site.

For the fixed treatment, we created stationary schedules for each sampling effort (*S*) (**Table 2. 4**) in an effort to make them as competitive as possible with the optimized treatment at the same sampling effort. The $S$ = 2 minute sampling effort consisted of a one-minute sample in the morning (08:00:00), and a one-minute sample at night (23:00:00). At higher efforts, samples were generally clustered around the average sunrise and sunset times throughout the year, with recordings scheduled on an hourly and sub-hourly basis as sampling effort increased. The same fixed schedules were applied for both the Full Year and March Only study durations.

For the optimized treatment, the *scheduleOptim()* function allocated evenly-spaced samples to the highest scoring hour(s) in one minute increments, with a buffer of at least one minute between each sample. We settled upon this formulation as a consequence of real field testing within the Riverside East Solar Energy Zone, wherein we found that a), schedules with a high number of sampling occasions mitigated the risk of individual events not being received and logged by our audio recording devices, and b), smaller files produced by short recordings were more likely to be reliably dispatched over the cellular network. Thus, a maximum of 30 one-minute samples could be assigned to any single hour. For example, a sampling effort of $S$ = 30 one-minute samples would allot all 30 evenly-spaced samples, each one minute in length, with a one-minute buffer between each sample, into the highest scoring hour. For sampling efforts greater than 30 minutes (i.e., $S$ = 40), additional minutes spilled over into the second highest scoring hour.

For each species, under each sampling effort ($S$) and study duration ($D$), we used two metrics to compare the performance of the optimized and fixed treatments. First, we rendered $p*$ accumulation curves averaged across the 133 sites, and computed the total area under these curves (AUC), with AUC values closest to 1 being best. We also calculated the average date $p*_{max}$ was achieved for each species across sites (if at all), on the assumption that earlier achievement dates were more desirable.

## 2.4. Results

### Vocalization Simulation Results (Objective 2)

Driven by weather and temporal covariates, the simulated environment produced hourly probabilities of vocalization for each of the nine species at each site for the entire year of 2016. Summary statistics of monthly temperature, 24-hour rain accumulation, and wind speed demonstrated variation in weather covariates throughout the year, while sunrise and sunset times illustrated shifts in temporal covariates (**Table 2. 5**), all of which showed differences in conditions between the March Only and Full Year study durations.

The average probability of vocalization by species was summarized in **Fig. 2. 4** for both study durations, showing that breeding birds had a higher average vocalization probability during the March study duration as compared to the entire year, and also illustrating that three species – Couch's Spadefoot Toad (TOAD), Coyote, and Lesser Nighthawk (LENI) – were far less vocally available in general. Large standard deviations in **Fig. 2. 4** indicated the wide variation in overall vocalization probabilities across each hour of the year. (Note that **Fig. 2. 3** conveys in finer detail the influence of temporal and weather conditions on modeled vocalization probabilities.)

## Factorial Results (Objective 3)

Using the simulated environment for all species, the optimized treatment equaled or outperformed the fixed treatment on both metrics under all sampling efforts ($S$) and under both the Full Year and March Only durations ($D$), with only one exception (coyote $p*_{max}$ achievement at $S$ = 20, $D$ = Full Year).

In the optimized treatment, because we used equal initial monitoring priority weights for all species, gregarious species dominated the sampling allocation early on for both study durations. Species modeled to be more vocally available (**Fig. 2. 4**) initially had a greater effect on aggregate scores, causing sampling effort to be allotted in their favor early in the season. Once these species' weights began shrinking as their $p^*$ values increased, optimized sampling focus shifted to less vocally available species. If this phenomenon had been undesirable, we could have assigned higher monitoring priority weights *a priori* to species of special sampling concern.

Across species, AUC values produced by the $p^*$ accumulation curves for the optimized treatment equaled or exceeded those of the fixed treatment under all sampling efforts and for both study durations. At the extreme low end of sampling effort ($S$ = 2 minutes per day), the optimized treatment yielded AUC values that were typically at least 25% greater than those of the fixed treatment during the Full Year study (**Fig. 2. 5a**), ranging up to more than 50% greater for the March Only study (**Fig. 2. 5b**). Though the optimized AUC values are greater than the fixed AUC values in most cases, these differences became negligible for commonly available vocalizers during the Full Year study when sampling effort was high. For example, comparatively loquacious species like Black-tailed Gnatcatcher, Common Poorwill, Gambel's Quail, Eurasian Collared-Dove, Phainopepla, and Verdin attained relatively high AUC regardless of schedule type, provided that the study duration was sufficiently long and sampling effort was sufficiently high. Meanwhile, for the rarest vocalizers (e.g., Couch's Spadefoot Toad), the optimized treatment substantially outperformed the fixed treatment AUC even when sampling was high over the longer study duration.

Schedules only achieved $p^*_{max}$ values under certain conditions of sampling effort, study duration, and species vocal availability. For the full year study, where comparisons were possible, the optimized schedule reached $p^*_{max}$ earlier in the year than the fixed schedule for nearly all scenarios (**Fig. 2. 5a**). The sole departure from this pattern was presented by the coyote, for which $p^*_{max}$ was not obtained below a sampling effort of 20 minutes. At 20 minutes, both schedules attained $p^*_{max}$ for coyote, although the fixed schedule reached this value nine days earlier than the optimized schedule. In every other case, the opposite was true: across sites, for the full year study, the optimized schedule surpassed $p*_{max}$ anywhere from five to 180 days earlier than the fixed schedule depending on the species and sampling effort (average = 30 days earlier) (**Appendix A. 1**). Even at 40 samples, where the fixed schedule began to become more competitive, the optimized schedule still reached $p*_{max}$ an average of 14 days earlier than the fixed schedule for all species except for Couch's Spadefoot Toad, where no comparison was available because the optimized schedule achieved $p*_{max}$ and the fixed schedule did not (**Fig. 2. 5a**). In general, for both the fixed and optimized treatments, commonly available vocalizers (e.g., Eurasian Collared-Dove, Gambel's Quail, Verdin) exceeded $p*_{max}$ earlier in the season than less available vocalizers (e.g., Coyote, Lesser Nighthawk), and the least available species (Couch's Spadefoot Toad) only reached $p*_{max}$ with the optimized schedule. This outcome is consistent with simulated differences in average vocalization probability between species (**Fig. 2. 4**), given that we assigned equal initial monitoring priority weights to each species.

Under the abbreviated March sampling duration (where the toad was ommitted due to seasonal inactivity), the optimized schedule again proved superior on the $p*_{max}$ metric (**Fig. 2. 5b**). Only six out of the eight species hit $p*_{max}$ at all during the shorter sampling season. Often $p*_{max}$ was achieved only at higher sampling efforts, even for commonly available vocalizers such as Eurasian Collared-Dove, Gambel's Quail, and Verdin. In all cases, the fixed schedule lagged well behind the optimized schedule in attaining $p*_{max}$, if at all. For conditions under which a comparison was even possible, across all species and sampling efforts in the March Only study, the optimized schedule reached $p^*_{max}$ an average of 11 days earlier than the fixed schedule (**Appendix A. 1**).

## 2.5. Discussion

We demonstrated that a temporally-adaptive optimized sampling schedule can substantially outperform a fixed schedule in a simulation setting for maximizing the probability of detecting a suite of focal species, given presence. The advantage of the optimized schedule was magnified especially for the shorter study season and particularly at lower sampling efforts. The optimized schedule thus minimized the risk of encountering false negatives compared to the fixed schedule. In wildlife monitoring efforts striving to characterize distribution patterns of focal species, a temporally-adaptive sampling schedule may therefore improve capacity to definitively characterize true negatives within acoustic monitoring, wherein a species is truly not present on site.

This work contributes novel methodology to the adaptive sampling paradigm for monitoring wildlife. The bulk of research on adaptive sampling of wildlife is justifiably focused on sampling in the spatial dimension (e.g. Thompson et al. 1998, 2004; Turk and Borkowski 2005), while temporal adaptive sampling has not been explicitly explored in great depth (though see Dyo et al. 2012 and Charney et al. 2015). Recent work on the optimization of survey effort over space and time (Moore & McCarthy 2016), and when species detectability varies (Moore et al. 2014), focused on empirical field research, typically including a travel cost parameter that is fortunately irrelevant for spatially fixed automated acoustic recording units. Additionally, though the notion of time-sensitive sampling is present in wildlife surveys – for example, by surveying during seasonally-appropriate occasions for breeding amphibians, or on spring mornings during the dawn chorus for breeding birds – such sampling is not adaptive in nature unless information from prior surveys is incorporated into future sampling efforts (Thompson and Seber 1994, Charney et al. 2015).

Accordingly, the adaptive nature of this methodology enhances existing bioacoustic endeavors and introduces new possibilities. In terms of existing methodology, our adaptive sampling framework may be used to increase confidence in the local arrival and departure dates for migratory birds in a dynamic occupancy model framework (*sensu* MacKenzie et al. 2003). Though occupancy models already account for detection errors in the form of false negatives, the adaptive optimization framework described here can reduce the false negative rate to provide more confidence in detection probability estimates, potentially resulting in more precise occupancy estimates.

The optimization options developed here provide a framework for improved sampling granularity. First, in addition to local weather conditions, field-based implementations of the temporally-adaptive optimization scheme could incorporate real-time bird migration predictions which combine citizen science observations via the eBird database (Sullivan et al. 2009), flight calls of nocturnal migrants, and radar to detect ‘clouds’ of migrating birds (BirdCast: Cornell Lab of Ornithology 2017). Given brief study durations, sampling constraints, and multiple focal species with varied vocal availabilities, automated optimization of acoustic sampling may thus allow research programs to collect higher quality data with limited resources.

Second, though we used the simple daily site constraint option here, where an equal number *S* one-minute samples per day were taken at each site and distributed into the highest scoring hour(s), samples could also be distributed via the ‘max per hour’ argument within *AMMonitor*’s *scheduleOptim()* function. Using the ‘simple’ daily site constraint option, if daily sampling effort is ≤ 30 minutes, all of those minutes are distributed at equal intervals into the highest scoring hour of the day. If this is undesirable, the researcher may invoke the ‘max per hour’ option to specify a maximum number of samples that can be allocated into each hour. In an exploratory simulation using the March 2016 study duration, eight species, sampling efforts of *S* = 20, 30, 40 one-minute samples per day, and a maximum number of samples per hour of 10, we found that the ‘simple’ daily site constraint option outperformed the ‘max per hour’ option considerably (**Appendix A. 2**). Nevertheless, in a real-field scenario, researchers may elect to hedge their bets in this way depending on their confidence in the species vocalization models and accuracy of the weather forecast.

Third, optimization methods might sample during the highest scoring time increments independent of site. In this work, we forced all sites to take *S* one-minute samples daily, but future extensions could allocate all available sampling power within a given time period to the best ‘site-hours’ overall, perhaps across a one-, three-, or even five-day weather forecast. For example, if a study area is vast, and rain is forecasted for a subset of sites where Couch’s Spadefoot Toad is of high monitoring priority, available sampling power would be optimally distributed only to those site-hours with high predicted rain accumulation. Rainless site-hours, meanwhile, would be earmarked for no sampling during the forecast period of interest, minimizing wasteful sampling efforts if target species are only available under specific conditions.

Fourth, although this implementation optimized under an assumption of species presence, future extensions might set the adaptive schedule based on the joint probability of occupancy and vocalization. That is, our simulations set the optimized schedule based on the probability of calling, conditional on presence; we did not consider the factors that actually shape the presence or absence of species across the 133 sites, which was not necessary to test the optimization algorithm. However, site occupancy can be factored into the algorithm by redefining $p_v$ (currently, the conditional probability of vocalization given presence) as the joint probability of presence and vocalization. In this formulation, high presence probabilities produce a higher site-hour score, increasing the chances of sampling a given site-hour under the optimization scheme. In contrast, lower presence probabilities drive lower site-hour scores, resulting in a smaller chance that a site-hour will be selected for sampling.

Fifth, although this work focused on simulation results, in practice, researchers may incorporate additional considerations into a temporally-adaptive sampling scheme implemented in the field. Firstly, vocalization models producing species vocalization probabilities ($p_v$) may be generated such that they have confidence intervals that include upper and lower bounds. In practice, to accommodate model uncertainty, researchers may elect to use the upper bound, lower bound, or mean predicted $p_v$ values in the optimization scheme, depending on model confidence. Secondly, although we used equal initial priority monitoring weights at all sites for all species, in practice, researchers maybe set higher weights for species or sites of greater monitoring priority.

Finally, although our $p^*_{max}$ values were set to 0.95 for the simulation (i.e., sampling continued until there was a 95% chance the species was acoustically captured on our recording devices at least once), users may set this threshold to any value. For instance, we might relax the definition of $p^*_{max}$ as a probability bounded between zero and one, and set a $p^*_{max}$ value of 2.00 for a given species at a given site, which would indicate that monitoring should continue until we are quite confident that the species has been acoustically captured on at least two separate sampling occasions during the monitoring period. This arrangement could further safeguard against false negatives: first by providing an additional failsafe against recording at inopportune times, and second by adding preemptive cushion against false negatives that could occur as a consequence of using automated detection algorithms.

Although this work is simulation-based, we field-tested the mechanics of a temporally-adaptive sampling optimization protocol on N = 16 audio recorders by connecting each Android audio recording unit with a site-specific Google calendar account. We also developed a protocol linking the Android apps Easy Voice Recorder Pro (Digipom 2016) and Tasker (Tasker 2015) with the optimization protocol (Balantic et al. in prep). This combination allowed us to populate each device’s calendar with the optimized sampling schedule on a daily basis and collect acoustic recordings, providing a real field proof-of-concept for the simulation experiment detailed in this paper.  This protocol can be implemented in the field using the fully operational *AMMonitor* functions *scheduleOptim()* and *scheduleFixed()*, which can be combined to create daily optimized and/or fixed schedules that are automatically pushed to a remote recording unit's Google account and then synced automatically for the next day of acoustic sampling. Step-by-step instructions for linking the Android apps Easy Voice Recorder Pro (Digipom 2016) and Tasker (Tasker 2015) with *scheduleOptim()* are available in Balantic et al. in prep.

## 2.6. Acknowledgments

This project was supported by the U.S. Bureau of Land Management. Cathleen Balantic was supported by NSF IGERT grant 1144388. We thank Mark Massar (U.S. BLM) for providing the opportunity to work in the Riverside East Solar Energy Zone, Jonathan Katz for assistance with R programming for the *AMMonitor* package, and reviewers for their helpful comments. Use of trade names or products does not constitute endorsement by the U.S. Government. The Vermont Cooperative Fish and Wildlife Research Unit is jointly supported by the U.S. Geological Survey, University of Vermont, Vermont Department of Fish and Wildlife, and Wildlife Management Institute.

## 2.7. References

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. PeerJ 1:e103. doi: https://doi.org/10.7717/peerj.103

Balantic C. M., & Donovan, T.M. In prep. (Unpublished results) Statistical learning mitigation of false positive detections in automated acoustic wildlife monitoring.

Balantic C.M., Katz, J., & Donovan, T.M. In prep. (Unpublished results) AMMonitor R Package.

Bender, D.J., Bayne, E.M., & Brigham, R.M. (1996). Lunar condition influences coyote (*Canis latrans*) howling. American Midland Naturalist, 136(2), 413-417. doi: 10.2307/2426745

U.S. Bureau of Land Management. (2016). Riverside East Solar Energy Zone Long Term Monitoring Strategy: Final Report. Prepared by Environmental Science Division, Argonne National Laboratory, for the U.S. Department of the Interior Bureau of Land Management.

U.S. Geological Survey. (2001). North American Breeding Bird Survey Methodology: Methods and Requirements. Patuxent Wildlife Research Center, U.S. Department of the Interior, U.S. Geological Survey. URL https://www.pwrc.usgs.gov/bbs/participate/training/11.html

Cerqueira, M. C., & Aide, M.T. (2016). Improving distribution data of threatened species by combining acoustic monitoring and occupancy modeling. Methods in Ecology and Evolution. 7(11), 1340-1348. doi: 10.1111/2041-210X.12599

Charney, N.D., Kubel, J.E., Eiseman, & Eiseman, C.S. (2015). Temporally adaptive sampling: a case study in rare species survey design with marbled salamanders (*Ambystoma opacum*). PloS One, 10(3), e0120714. doi: https://doi.org/10.1371/journal.pone.0120714

CinixSoft. (2014). CinixSoft Remote Schedule Voice Recorder (v4.2.0). [Android App]. URL http://www.cinixsoft.com/

Cornell Lab of Ornithology. (2017). BirdCast: Bird migration forecasts in real-time. URL http://www.birdcast.info

Dark Sky. (2017). Dark Sky API [Application Programming Interface]. URL https://darksky.netDawson, D. K., & Efford, M.G. (2009). Bird population density estimated from acoustic signals. Journal of Applied Ecology 46:1201-1209. doi: http://dx.doi.org/10.1111/j.1365-2664.2009.01731.x

Digipom (2016). Easy Voice Recorder Pro [Android App]. URL http://www.digipom.com/portfolio-items/easy-voice-recorder/

Dyo, V., Ellwood, S. A., MacDonald, D. W., Markham, A., Trigoni, N., Wohlers, R., Mascolo, C., Pasztor, B., Scellato, S., & Yosef, K. (2012). WILDSENSING: Design and deployment of a sustainable sensor network for wildlife monitoring. ACM Transactions on Sensor Networks 8 (4), Article 29. doi: 10.1145/2240116.2240118

Furnas, B.J. & Callas, R.L. (2014). Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. Journal of Wildlife Management 79(2), 325-337. doi: 10.1002/jwmg.821

Gage, S. H., W. Joo, E. P. Kasten, J. Fox, & S. Biswas. (2015). Acoustic observations in agricultural landscapes. Pages 360-377 in S. K. Hamilton, J. E. Doll, & G. P. Robertson, editors. The Ecology of Agricultural Landscapes: Long-Term Research on the Path to Sustainability. Oxford University Press, New York, New York, USA. ISBN: 978-0199773350.

Gage, S.H. & Farina, A. (2017). Ecoacoustics Challenges. Pages 313-320 in A. Farina & S.H. Gage, editors. Ecoacoustics: The Ecological Role of Sounds. John Wiley & Sons. ISBN: 978-1-119-23069-4.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A. & Langtimm, C.A. (2002). Estimating site occupancy rates when detection probabilities are less than one. Ecology, 83, 2248–2255. doi: 10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G., & Franklin, A.B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology, 84(8), 2200-2207. doi:10.1890/02-3090.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006). Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Academic Press. **ISBN:** 9780120887668.Manley, P.N., Zielinski, W.J., Schlesinger, M.D., & Mori, S.R. (2004). Evaluation of a multiple-species approach to monitoring species at the ecoregional scale. Ecological Applications, 14(1) 296-310. doi: https://doi.org/10.1890/02-5249

Marques, T. A., L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, & Tyack, P.L. (2013). Estimating animal population density using passive acoustics. Biological Reviews 88:287-309. doi: http://dx.doi.org/10.1111/brv.12001

Mayhew, W.W. (1965). Adaptations of the amphibian, *Scaphiopus couchi*, to desert conditions. American Midland Naturalist, 74(1), 95-109. doi: 10.2307/2423123

McKown, M.W. (2012). A wireless acoustic sensor network for monitoring wildlife in remote locations. The Journal of the Acoustical Society of America, 132, 2036. doi: https://doi.org/10.1121/1.4755484

Moore, A., McCarthy, M. Parris, K., & Moore., J.L. (2014). The Optimal Number of Surveys when Detectability Varies. PLoS ONE, 9, e115345. doi: https://doi.org/10.1371/journal.pone.0115345

Moore, A.L. & McCarthy, M.A. (2016). Optimizing ecological survey effort over space and time. Methods in Ecology and Evolution, 7, 891–899. doi: 10.1111/2041-210X.12564

Otis, D. L., Burnham, K.P., White, G.C., & Anderson, D.R. (1978). Statistical inference from capture data on closed animal populations. Wildlife Monographs, 62, 1-135.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., & Kelling, S. (2009). eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142: 2282-2292. doi: https://doi.org/10.1016/j.biocon.2009.05.006

Tasker (2015). Tasker: Total Automation for Android (v4.8). [Android App]. URL http://tasker.dinglisch.net/

Thompson, S.K. & Seber, G.A.F. (1994). Detectability in Conventional and Adaptive Sampling. Biometrics, 50, 712. doi: 10.2307/2532785

Thompson, W.L., White, G.C., & Gowan, C. (1998). Monitoring Vertebrate Populations. Academic Press, San Diego. ISBN: 0126889600

Thompson, W. L. (2004). Sampling rare or elusive species: concepts and techniques for estimating population parameters. Island Press, Washington, D.C., USA. ISBN: 9781559634519

Turk, P. & Borkowski, J.J. (2005). A review of Adaptive Cluster Sampling: 1990-2003. Environmental and Ecological Statistics, 12(1), 55-94. doi: https://doi.org/10.1007/s10651-005-6818-0

White, G. C., Anderson, D. R., Burnham, K. P. & Otis, D. L. (1982). Capture-recapture and removal methods for sampling closed populations. LA-8787-NERP, Los Alamos National Laboratory, Los Alamos, NM. 235pp.

Wildlife Acoustics. (2016). Song Meter SM4 [Acoustic Recording Hardware]. URL https://www.wildlifeacoustics.com/products/song-meter-sm4

Woods, C. P., Csada, R.D., & Brigham, R.M. (2005). Common Poorwill (*Phalaenoptilus nuttallii*), The Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. doi: 10.2173/bna.32

## 2.8. Tables

Table 2. 1. Summary of nine focal species used for simulation

| Species | Species Code | Phylogenetic Class | Vocal availability throughout day | Vocal availability throughout year | Believed rarity of vocalizations, given presence | Resident or Migratory | Native vs. Invasive |
|---|---|---|---|---|---|---|---|
| Black-tailed Gnatcatcher | BTGN | Bird | Diurnal | Spring peak | Common | Resident | Native |
| Common Poorwill | COPO | Bird | Nocturnal | Spring peak | Common | Resident | Native |
| Couch’s Spadefoot Toad | TOAD | Amphibian | Nocturnal | Late summer/fall only | Rare | Resident | Native |
| Coyote | COYOTE | Mammal | Nocturnal | Peak at equinoxes | Uncommon | Resident | Native Invasive |
| Eurasian Collared-Dove | ECDO | Bird | Diurnal | Spring peak | Common | Resident | Invasive |
| Gambel’s Quail | GAQU | Bird | Diurnal | Spring peak | Common | Resident | Native |
| Lesser Nighthawk | LENI | Bird | Nocturnal | Spring peak | Uncommon | Migratory | Native |
| Phainopepla | PHAI | Bird | Diurnal | Spring peak | Common | Migratory | Native |
| Verdin | VERD | Bird | Diurnal | Spring peak | Common | Resident | Native |

Table 2. 2. Logistic regression models for nine focal species, each producing the hourly probability of vocalization.

| Species | Model |
|---|---|
| Black-tailed Gnatcatcher (BTGN) | -0.3 – 0.002*day.of.year + 1*sin(day.of.year) – 0.5*cosine($hour_{12}$) – 0.000007*time.to.sunrise$^{2}$ + 0.009*temperature – 0.000001*temperature$^{3}$ – 0.35*wind.speed |
| Common Poorwill (COPO) | -1.5 – 0.003*day.of.year – 0.5*cosine(day.of.year) + 0.6*sin(day.of.year) + 1*cosine($hour_{24}$) – 0.5*cosine($hour_{12}$) – 0.0005*time.to.sunrise – 0.0005*time.to.sunset – 0.1*wind.speed – 0.2*cosine(lunar.phase) |
| Couch’s Spadefoot Toad (TOAD) | -8 – 1*cosine(day.of.year) – 2*sin(day.of.year) + 3*cosine($hour.of.day_{24}$) + 5*rain accumulation in the past 24 hours |
| Coyote (COYOTE) | -3 – 0.5*cosine($day.of.year_{equinox}$) + 0.2* sin($day.of.year_{equinox}$) + 1* cosine($hour_{24}$) – 0.5* cosine($hour_{12}$) – 0.001* time.to.sunrise – 0.001* time.to.sunset + 0.2*cosine(lunar.phase) |
| Eurasian Collared-Dove (ECDO) | -1.4 + 1*sin(day.of.year) – 2*cosine($hour_{12}$) – 0.000005* time.to.sunrise$^{2}$ + 0.009*temperature – 0.000001*temperature$^{3}$ – 0.25*wind.speed |
| Gambel’s Quail (GAQU) | -1.2 - 0.002*day.of.year + 1.3*sin(day.of.year) – 2*cosine($hour.of.day_{12}$) – 0.000005*time.to.sunrise$^{2}$ + 0.009*temperature – 0.000001*temperature$^{3}$ – 0.25*wind.speed |
| Lesser Nighthawk (LENI) | -2 – 0.006*day.of.year – 0.4*cosine(day.of.year) + 0.7*sin(day.of.year) + 1*cosine($hour_{24}$) – 0.5*cosine($hour_{12}$) – 0.0005*time.to.sunrise – 0.0005*time.to.sunset – 0.25*wind.speed – 0.3*cosine(lunar.phase) |
| Phainopepla (PHAI) | -2.2 – 0.00001*day.of.year$^{2}$ + 0.7*cos(day.of.year) + 2.2*sin(day.of.year) – 2.5*cosine($hour_{12}$) – 0.000004*time.to.sunrise$^{2}$ + 0.009*temperature – 0.000001*temperature$^{3}$ – 0.25*wind.speed |
| Verdin (VERD) | -0.5 – 0.004*day.of.year + 1*sin(day.of.year) – 1.5*cosine($hour_{12}$) – 0.000007*time.to.sunrise$^{2}$ + 0.009*temperature – 0.000001*temperature$^{3}$ – 0.25*wind.speed |

Table 2. 3. Monitoring priority weights for focal species at 133 sites, used for the Full Year (**a**) and March Only (**b**) study durations.

**a.**

| | Site | | | | | | |
|---|---|---|---|---|---|---|---|
| Species | 1 | 2 | 3 | … | 131 | 132 | 133 |
| Black-tailed Gnatcatcher (BTGN) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| Common Poorwill (COPO) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| Couch’s Spadefoot Toad (TOAD) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| Coyote (COYOTE) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| Eurasian Collared-Dove (ECDO) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| Gambel’s Quail (GAQU) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| Lesser Nighthawk (LENI) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| Phainopepla (PHAI) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| Verdin (VERD) | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| *Sum* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

**b.**

| | Site | | | | | | |
|---|---|---|---|---|---|---|---|
| Species | 1 | 2 | 3 | … | 131 | 132 | 133 |
| Black-tailed Gnatcatcher (BTGN) | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| Common Poorwill (COPO) | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| Coyote (COYOTE) | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| Eurasian Collared-Dove (ECDO) | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| Gambel’s Quail (GAQU) | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| Lesser Nighthawk (LENI) | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| Phainopepla (PHAI) | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| Verdin (VERD) | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| *Sum* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 2. 4. Fixed sampling schedules used on the 24-hour clock at each sampling effort ($S$ = 2, 5, 10, 20, 30, or 40 minutes), applied to both the March Only and Full Year study durations.

| Number of Samples | Fixed Schedule |
|---|---|
| 2 | 08:00:00, 23:00:00 |
| 5 | 02:00:00, 05:00:00, 06:00:00, 08:00:00, 23:00:00 |
| 10 | 00:00:00, 01:00:00, 02:00:00, 06:00:00, 06:30:00, 07:00:00, 07:30:00, 08:00:00, 22:00:00, 23:00:00 |
| 20 | 00:00:00, 01:00:00, 02:00:00, 03:00:00, 04:00:00, 05:00:00, 05:30:00, 06:00:00, 06:30:00, 07:00:00, 07:30:00, 08:00:00, 18:00:00, 18:30:00, 19:00:00, 19:30:00, 22:00:00, 22:30:00, 23:00:00, 23:30:00 |
| 30 | 00:00:00, 01:00:00, 01:30:00, 02:00:00, 02:30:00, 03:00:00, 03:30:00, 04:00:00, 04:30:00, 05:00:00, 05:30:00, 06:00:00, 06:30:00, 07:00:00, 07:30:00, 08:00:00, 08:30:00, 09:00:00, 09:30:00, 10:00:00, 17:00:00, 17:30:00, 18:00:00, 18:30:00, 19:00:00, 19:30:00, 22:00:00, 22:30:00, 23:00:00, 23:30:00 |
| 40 | 00:00:00, 00:30:00, 01:00:00, 01:30:00, 02:00:00, 02:30:00, 03:00:00, 03:30:00, 04:00:00, 04:30:00, 05:00:00, 05:30:00, 05:45:00, 06:00:00, 06:15:00, 06:30:00, 06:45:00, 07:00:00, 07:15:00, 07:30:00, 07:45:00, 08:00:00, 08:15:00, 08:30:00, 08:45:00, 09:00:00, 09:30:00, 10:00:00, 17:00:00, 17:30:00, 18:00:00, 18:15:00, 18:30:00, 18:45:00, 19:00:00, 19:30:00, 22:00:00, 22:30:00, 23:00:00, 23:30:00 |

Table 2. 5. Summary statistics for weather and temporal covariates across the simulation study area in 2016. (Monthly summaries convey conditions in March as compared with the Full Year variation.)

| | Temperature (*C) | | | | 24-hr. Rain Accumulation (mm) | | | | Wind Speed (km/hr) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Month | Avg. | Min. | Max. | SD | Avg. | Min. | Max. | SD | Avg. | Min. | Max. | SD |
| January | 12.3 | 1.9 | 27.0 | 3.1 | 0.617 | 0.000 | 11.549 | 1.671 | 8.2 | 0.0 | 49.0 | 5.9 |
| February | 17.7 | 1.1 | 31.0 | 4.5 | 0.008 | 0.000 | 1.097 | 0.043 | 9.9 | 0.0 | 36.7 | 6.0 |
| March | 20.0 | 6.7 | 33.9 | 4.0 | 0.044 | 0.000 | 2.017 | 0.149 | 10.5 | 0.0 | 55.7 | 7.0 |
| April | 22.5 | 10.6 | 36.0 | 3.7 | 0.304 | 0.000 | 4.686 | 0.825 | 11.0 | 0.0 | 50.8 | 7.5 |
| May | 25.1 | 13.1 | 38.5 | 3.8 | 0.047 | 0.000 | 1.280 | 0.179 | 10.1 | 0.0 | 38.7 | 5.9 |
| June | 33.2 | 19.9 | 48.8 | 4.1 | 0.017 | 0.000 | 1.194 | 0.094 | 10.5 | 0.1 | 35.3 | 5.3 |
| July | 34.9 | 22.4 | 45.7 | 3.5 | 0.092 | 0.000 | 2.908 | 0.325 | 11.6 | 0.0 | 47.3 | 5.3 |
| August | 33.9 | 22.0 | 45.5 | 3.3 | 0.065 | 0.000 | 1.600 | 0.155 | 9.2 | 0.0 | 39.7 | 5.1 |
| September | 28.9 | 16.0 | 42.3 | 3.7 | 0.187 | 0.000 | 6.104 | 0.841 | 9.1 | 0.0 | 35.4 | 5.7 |
| October | 25.0 | 14.8 | 36.3 | 3.5 | 0.074 | 0.000 | 3.498 | 0.399 | 8.1 | 0.0 | 34.4 | 4.8 |
| November | 18.3 | 3.6 | 33.5 | 4.2 | 0.094 | 0.000 | 2.870 | 0.304 | 8.4 | 0.0 | 34.8 | 5.3 |
| December | 12.8 | 1.3 | 25.6 | 3.1 | 1.239 | 0.000 | 18.900 | 2.831 | 9.5 | 0.0 | 44.8 | 6.9 |

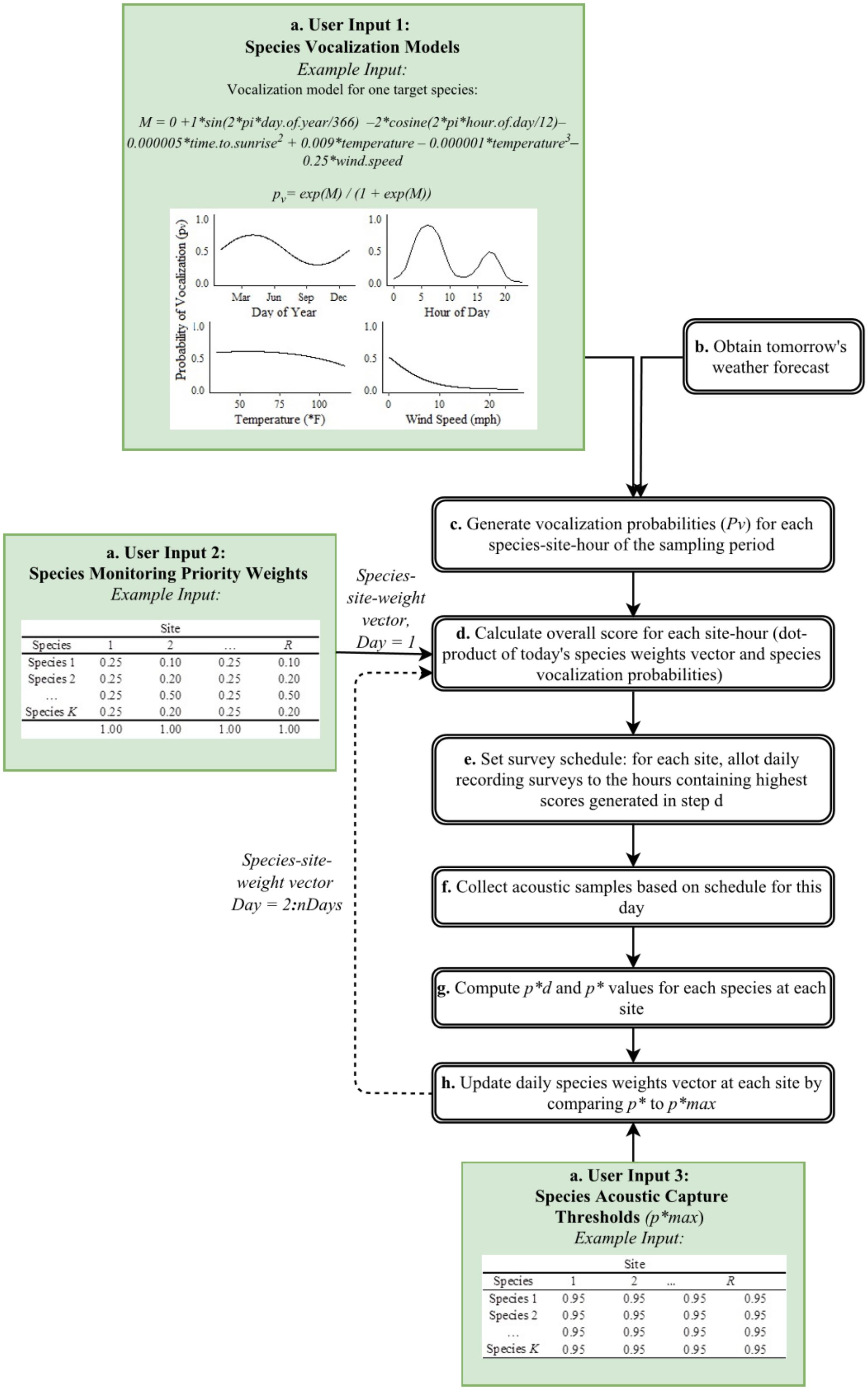
| | Sunrise Time | | | | Sunset Time | | | |
|---|---|---|---|---|---|---|---|---|
| Month | Avg. | Min. | Max. | SD (minutes) | Avg. | Min. | Max. | SD (minutes) |
| January | 6:44:36 | 6:37:05 | 6:48:47 | 2.5 | 16:56:21 | 16:41:32 | 17:12:21 | 8.3 |
| February | 6:25:40 | 6:08:47 | 6:40:09 | 8.4 | 17:24:54 | 17:09:49 | 17:39:11 | 7.8 |
| March* | 6:26:22 | 5:53:20 | 6:55:34 | 20.2 | 18:26:43 | 17:36:28 | 19:02:57 | 35.2 |
| April | 6:09:42 | 5:51:13 | 6:29:42 | 10.5 | 19:12:48 | 19:00:05 | 19:25:48 | 6.6 |
| May | 5:40:29 | 5:30:27 | 5:53:45 | 6.0 | 19:36:04 | 19:22:52 | 19:48:35 | 6.6 |
| June | 5:31:31 | 5:28:40 | 5:35:58 | 1.3 | 19:52:58 | 19:45:17 | 19:57:58 | 2.9 |
| July | 5:43:09 | 5:32:53 | 5:54:46 | 5.6 | 19:50:56 | 19:40:32 | 19:57:53 | 4.2 |
| August | 6:04:36 | 5:51:58 | 6:16:55 | 6.4 | 19:26:09 | 19:06:34 | 19:43:29 | 10.0 |
| September | 6:25:35 | 6:14:02 | 6:37:09 | 5.9 | 18:47:07 | 18:25:05 | 19:08:51 | 12.1 |
| October | 6:47:37 | 6:34:16 | 7:01:57 | 7.2 | 18:06:39 | 17:48:02 | 18:27:16 | 10.7 |
| November** | 6:24:31 | 6:03:46 | 7:06:33 | 18.2 | 16:49:13 | 16:31:24 | 17:50:38 | 26.1 |
| December | 6:38:59 | 6:26:55 | 6:48:04 | 5.4 | 16:36:07 | 16:31:08 | 16:44:50 | 3.3 |

* Begin DST

** End DST

## 2.9. Figures

Figure 2. 1. Objective 1 Workflow for an optimized temporally-adaptive sampling algorithm subject to species prioritization and sampling constraints.



**a. User Input 1:**
**Species Vocalization Models**
*Example Input:*
Vocalization model for one target species:

$$M = 0 + 1*\sin(2*pi*day.of.year/366) - 2*cosine(2*pi*hour.of.day/12) - 0.000005*time.to.sunrise^2 + 0.009*temperature - 0.000001*temperature^3 - 0.25*wind.speed$$

$$p_v = exp(M) / (1 + exp(M))$$

**b.** Obtain tomorrow's weather forecast

**c.** Generate vocalization probabilities (*Pv*) for each species-site-hour of the sampling period

**a. User Input 2:**
**Species Monitoring Priority Weights**
*Example Input:*

| | Site | | | |
|---|---|---|---|---|
| Species | 1 | 2 | … | *R* |
| Species 1 | 0.25 | 0.10 | 0.25 | 0.10 |
| Species 2 | 0.25 | 0.20 | 0.25 | 0.20 |
| … | 0.25 | 0.50 | 0.25 | 0.50 |
| Species *K* | 0.25 | 0.20 | 0.25 | 0.20 |
| | 1.00 | 1.00 | 1.00 | 1.00 |

*Species-site-weight vector, Day = 1*

**d.** Calculate overall score for each site-hour (dot-product of today's species weights vector and species vocalization probabilities)

**e.** Set survey schedule: for each site, allot daily recording surveys to the hours containing highest scores generated in step d

*Species-site-weight vector Day = 2:nDays*

**f.** Collect acoustic samples based on schedule for this day

**g.** Compute *p*d* and *p** values for each species at each site

**h.** Update daily species weights vector at each site by comparing *p** to *p*max*

**a. User Input 3:**
**Species Acoustic Capture Thresholds** (*p*max*)
*Example Input:*

| | Site | | | |
|---|---|---|---|---|
| Species | 1 | 2 | ... | *R* |
| Species 1 | 0.95 | 0.95 | 0.95 | 0.95 |
| Species 2 | 0.95 | 0.95 | 0.95 | 0.95 |
| … | 0.95 | 0.95 | 0.95 | 0.95 |
| Species *K* | 0.95 | 0.95 | 0.95 | 0.95 |

Figure 2. 2. Map of simulation study area: 133 sites were distributed across 599 km$^2$ located in southeastern California (USA) within the U.S. Bureau of Land Management's Riverside East Solar Energy Zone.

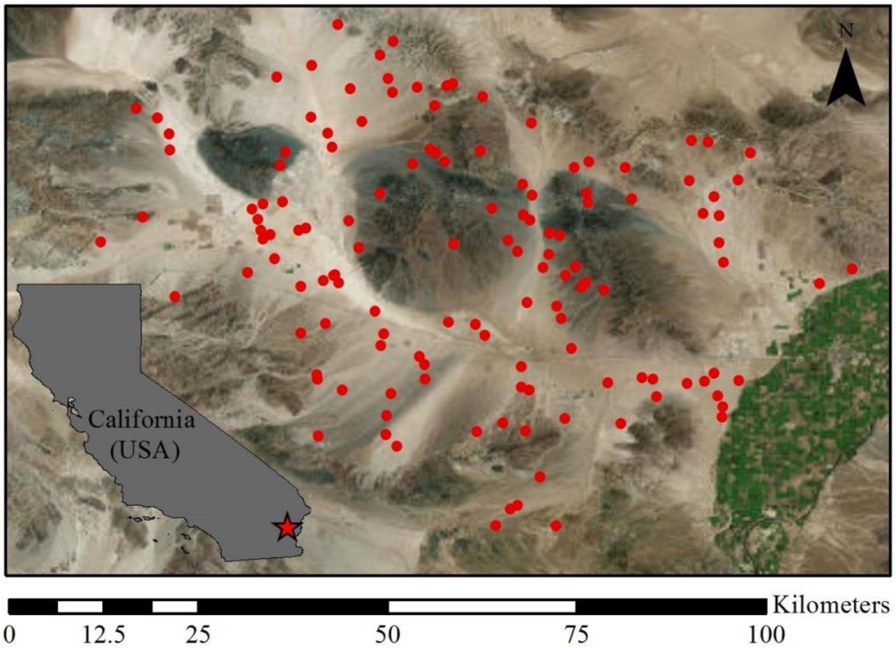


Figure 2. 3. Visual demonstration of species logistic regression vocalization models. Species codes and regression models are given in Table 2. 2. The probability of vocalization ($p_v$), given presence, is graphed as a function of key weather and temporal covariates to display vocalization characteristics across species. Because covariates are graphed separately, intercepts of zero are used for visual demonstration purposes.

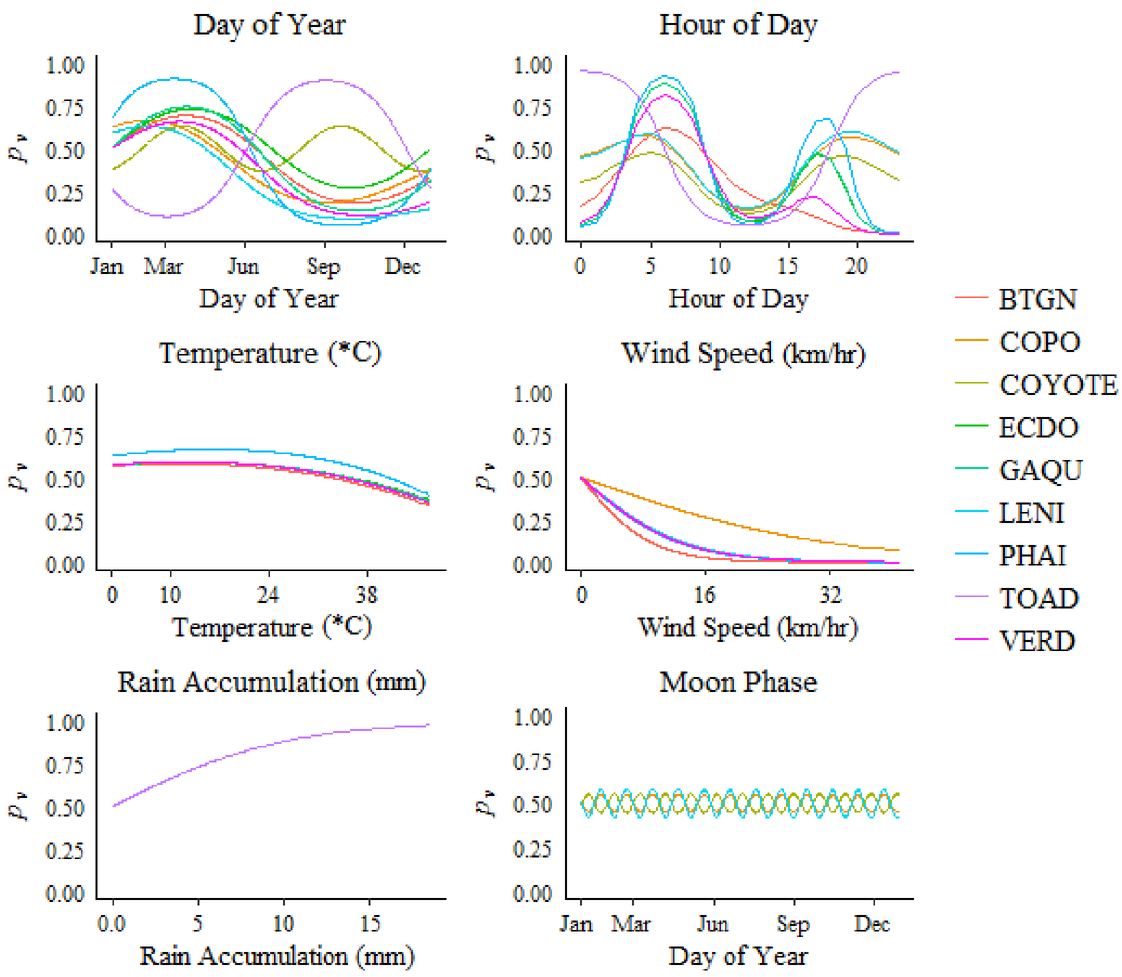


Figure 2. 4. Vocalization Simulation Results. Average probability of vocalization in a given hour across all hours and sites for each focal species during both the March Only and Full Year study durations. Species codes are provided in Table 2. 1. Standard deviation error bars reveal wide variation in vocalization probabilities contingent on weather and temporal conditions.

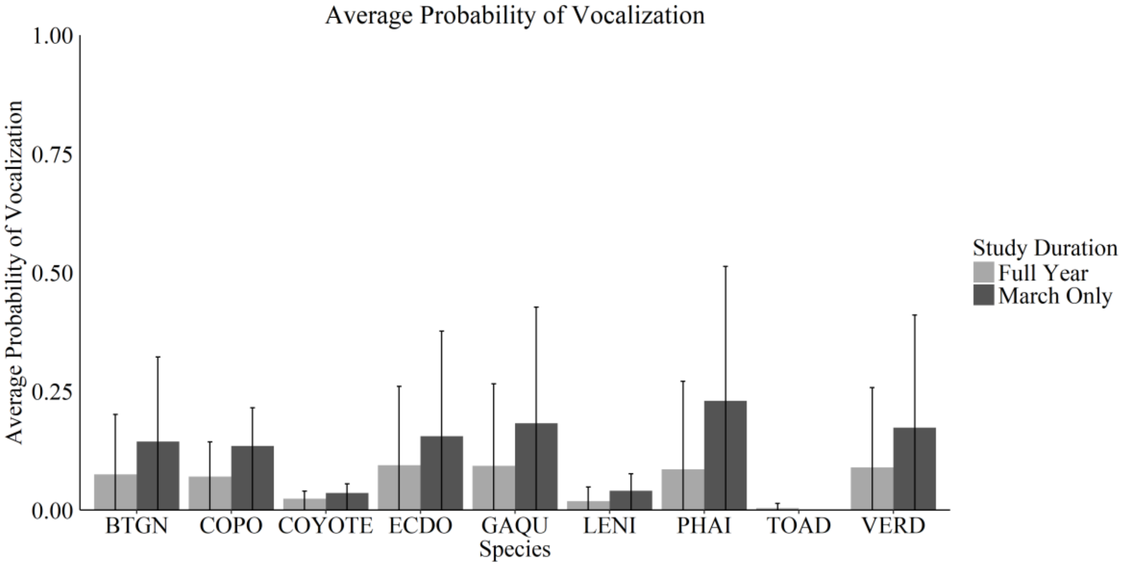
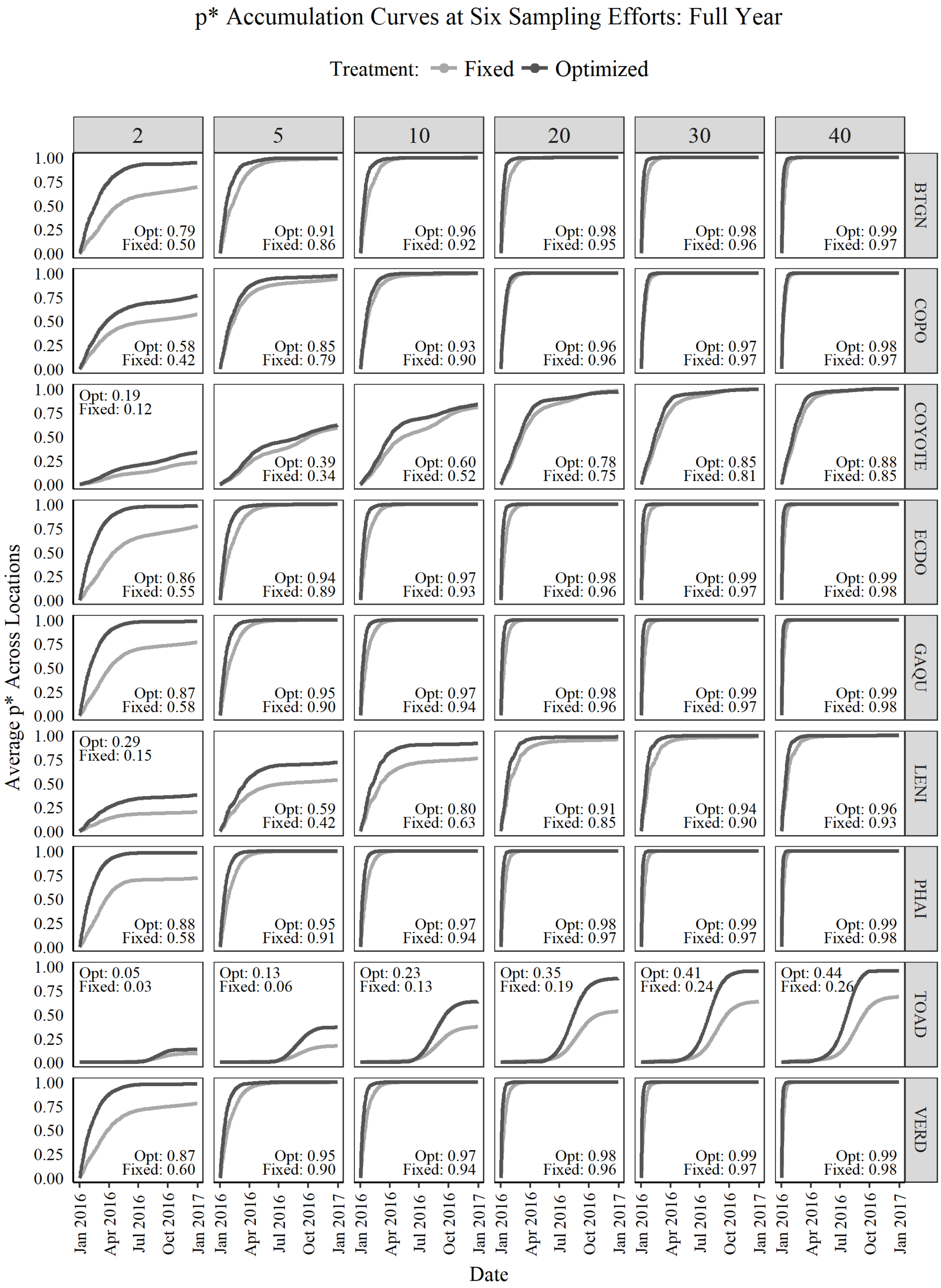
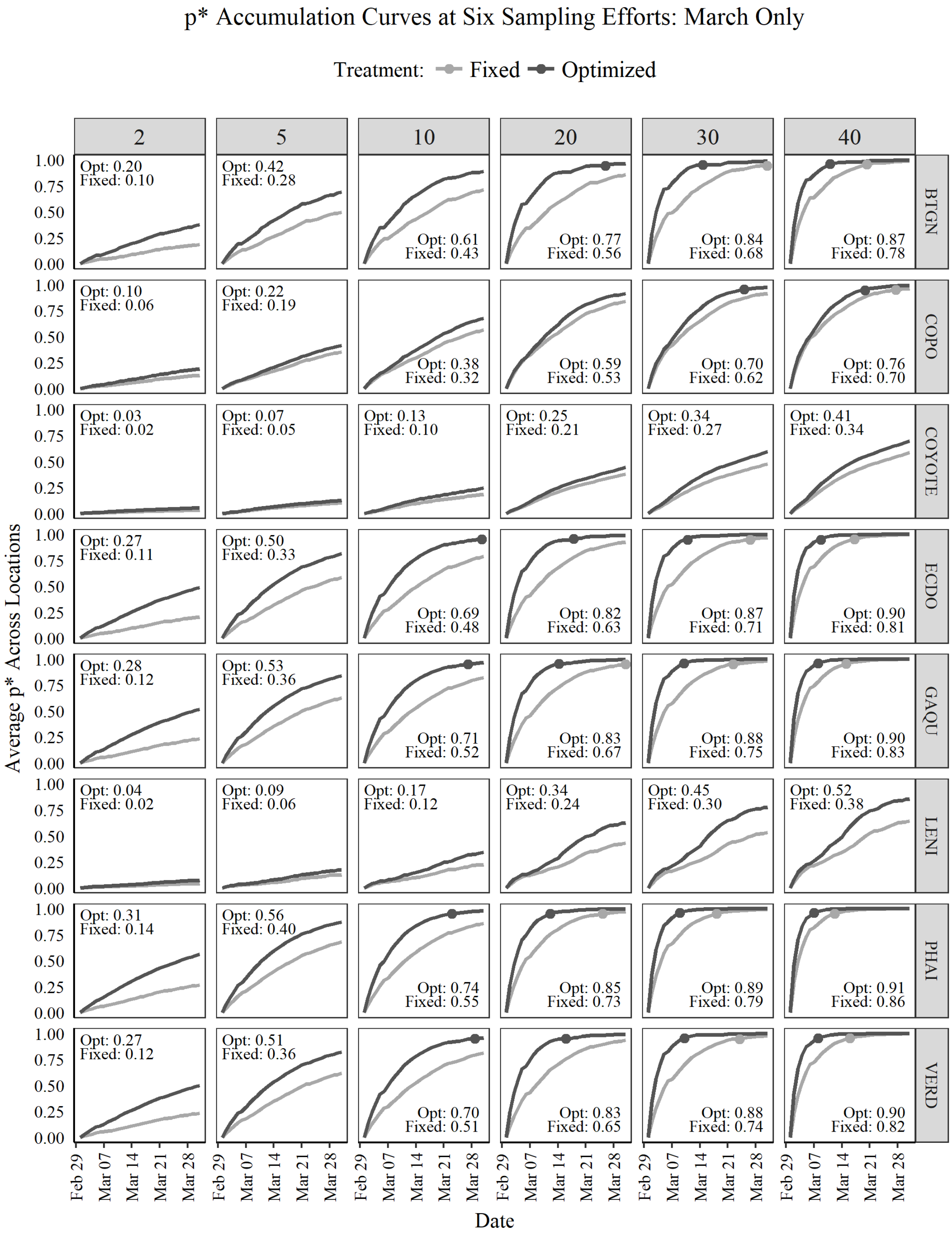


Figure 2. 5. Factorial Experiment Results. $p^*$ and $p^*_{max}$ charts are given for all focal species at six sampling efforts for the Full Year (**a**) and March Only (**b**) study durations. Species codes are provided in Table 2. 1. Lines track cumulative $p^*$ values for both the fixed and optimized schedule treatments. Total area under the cumulative $p^*$ curve (AUC) values are given for both treatments within each box. Where applicable, the date of first $p^*_{max}$ achievement is denoted by a single solid point on the line.

**a.**


p* Accumulation Curves at Six Sampling Efforts: Full Year
Treatment: Fixed Optimized
2
5
10
20
30
40
BTGN
Opt: 0.79 Fixed: 0.50
Opt: 0.91 Fixed: 0.86
Opt: 0.96 Fixed: 0.92
Opt: 0.98 Fixed: 0.95
Opt: 0.98 Fixed: 0.96
Opt: 0.99 Fixed: 0.97
COPO
Opt: 0.58 Fixed: 0.42
Opt: 0.85 Fixed: 0.79
Opt: 0.93 Fixed: 0.90
Opt: 0.96 Fixed: 0.96
Opt: 0.97 Fixed: 0.97
Opt: 0.98 Fixed: 0.97
COYOTE
Opt: 0.19 Fixed: 0.12
Opt: 0.39 Fixed: 0.34
Opt: 0.60 Fixed: 0.52
Opt: 0.78 Fixed: 0.75
Opt: 0.85 Fixed: 0.81
Opt: 0.88 Fixed: 0.85
ECDO
Opt: 0.86 Fixed: 0.55
Opt: 0.94 Fixed: 0.89
Opt: 0.97 Fixed: 0.93
Opt: 0.98 Fixed: 0.96
Opt: 0.99 Fixed: 0.97
Opt: 0.99 Fixed: 0.98
GAQU
Opt: 0.87 Fixed: 0.58
Opt: 0.95 Fixed: 0.90
Opt: 0.97 Fixed: 0.94
Opt: 0.98 Fixed: 0.96
Opt: 0.99 Fixed: 0.97
Opt: 0.99 Fixed: 0.98
LENI
Opt: 0.29 Fixed: 0.15
Opt: 0.59 Fixed: 0.42
Opt: 0.80 Fixed: 0.63
Opt: 0.91 Fixed: 0.85
Opt: 0.94 Fixed: 0.90
Opt: 0.96 Fixed: 0.93
PHAI
Opt: 0.88 Fixed: 0.58
Opt: 0.95 Fixed: 0.91
Opt: 0.97 Fixed: 0.94
Opt: 0.98 Fixed: 0.97
Opt: 0.99 Fixed: 0.97
Opt: 0.99 Fixed: 0.98
TOAD
Opt: 0.05 Fixed: 0.03
Opt: 0.13 Fixed: 0.06
Opt: 0.23 Fixed: 0.13
Opt: 0.35 Fixed: 0.19
Opt: 0.41 Fixed: 0.24
Opt: 0.44 Fixed: 0.26
VERD
Opt: 0.87 Fixed: 0.60
Opt: 0.95 Fixed: 0.90
Opt: 0.97 Fixed: 0.94
Opt: 0.98 Fixed: 0.96
Opt: 0.99 Fixed: 0.97
Opt: 0.99 Fixed: 0.98
Average p* Across Locations
0.00
0.25
0.50
0.75
1.00
Date
Jan 2016
Apr 2016
Jul 2016
Oct 2016
Jan 2017

**b.**


p* Accumulation Curves at Six Sampling Efforts: March Only
Treatment: Fixed Optimized
2
5
10
20
30
40
BTGN
COPO
COYOTE
ECDO
GAQU
LENI
PHAI
VERD
Opt: 0.20 Fixed: 0.10
Opt: 0.42 Fixed: 0.28
Opt: 0.61 Fixed: 0.43
Opt: 0.77 Fixed: 0.56
Opt: 0.84 Fixed: 0.68
Opt: 0.87 Fixed: 0.78
Opt: 0.10 Fixed: 0.06
Opt: 0.22 Fixed: 0.19
Opt: 0.38 Fixed: 0.32
Opt: 0.59 Fixed: 0.53
Opt: 0.70 Fixed: 0.62
Opt: 0.76 Fixed: 0.70
Opt: 0.03 Fixed: 0.02
Opt: 0.07 Fixed: 0.05
Opt: 0.13 Fixed: 0.10
Opt: 0.25 Fixed: 0.21
Opt: 0.34 Fixed: 0.27
Opt: 0.41 Fixed: 0.34
Opt: 0.27 Fixed: 0.11
Opt: 0.50 Fixed: 0.33
Opt: 0.69 Fixed: 0.48
Opt: 0.82 Fixed: 0.63
Opt: 0.87 Fixed: 0.71
Opt: 0.90 Fixed: 0.81
Opt: 0.28 Fixed: 0.12
Opt: 0.53 Fixed: 0.36
Opt: 0.71 Fixed: 0.52
Opt: 0.83 Fixed: 0.67
Opt: 0.88 Fixed: 0.75
Opt: 0.90 Fixed: 0.83
Opt: 0.04 Fixed: 0.02
Opt: 0.09 Fixed: 0.06
Opt: 0.17 Fixed: 0.12
Opt: 0.34 Fixed: 0.24
Opt: 0.45 Fixed: 0.30
Opt: 0.52 Fixed: 0.38
Opt: 0.31 Fixed: 0.14
Opt: 0.56 Fixed: 0.40
Opt: 0.74 Fixed: 0.55
Opt: 0.85 Fixed: 0.73
Opt: 0.89 Fixed: 0.79
Opt: 0.91 Fixed: 0.86
Opt: 0.27 Fixed: 0.12
Opt: 0.51 Fixed: 0.36
Opt: 0.70 Fixed: 0.51
Opt: 0.83 Fixed: 0.65
Opt: 0.88 Fixed: 0.74
Opt: 0.90 Fixed: 0.82
1.00
0.75
0.50
0.25
0.00
Average p* Across Locations
Feb 29
Mar 07
Mar 14
Mar 21
Mar 28
Date

# CHAPTER 3: STATISTICAL LEARNING MITIGATION OF FALSE POSITIVE DETECTIONS IN AUTOMATED ACOUSTIC WILDLIFE MONITORING

Cathleen Balantic[1*]

Therese Donovan[2]

[1] Corresponding Author: cathleen.balantic@uvm.edu; Vermont Cooperative Fish and Wildlife Research Unit, University of Vermont, Burlington, VT 05405, USA

[2] U.S. Geological Survey, Vermont Cooperative Fish and Wildlife Research Unit, Rubenstein School of Environment and Natural Resources, University of Vermont, Burlington, VT 05405, USA

*Prepublication Disclaimer:*

*This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy.*

## 3.1. Abstract

Audio sampling of the environment can provide long-term, landscape-scale presence-absence data to model populations of sound-producing wildlife. Automated detection systems allow researchers to avoid manually searching through large volumes of recordings, but often produce unacceptable false positive rates. We developed methods that allow researchers to improve template-based automated detection using a suite of statistical learning algorithms when false positive rates are problematic. To test our method, we acquired 675 hours of recordings in the Sonoran Desert, California USA between March 2016 and May 2017, and created spectrogram cross-correlation templates for three target avian species. We trained and tested five classification algorithms and four performance-weighted ensemble classifier methods on target signals and false alarms from March 2016, and then selected high-performing ensemble classifiers from the train/test phase to predict the class of new detections thereafter. For three target species, our ensemble classifiers were able to identify 98%, 85%, and 99% more false alarms than the baseline template detection system, and comparative positive predictive values improved from 6% to 75%, 87% to 97%, and 2% to 69%. We show that statistical learning approaches can be implemented to mitigate false detections acquired via template-based automated detection in automated acoustic wildlife monitoring.

Key Words

Automated acoustic monitoring; bioacoustics; false positives; machine learning; species identification; statistical learning

## 3.2. Introduction

Tracking wildlife population dynamics at regional scales requires methods that efficiently accumulate data on species of interest (Pollock et al. 2002). Automated acoustic monitoring of sound-producing wildlife offers one path for characterizing baseline species status and trends across vast landscapes, important within the context of climate change and rapidly shifting land uses. Because obtaining species abundance data is often inefficient, costly, and impractical, research at large spatial scales may instead collect species presence-absence data for use in occupancy models; remote acoustic monitoring is well-positioned to support such data collection because it affords the opportunity to identify presence or absence of species based on sounds captured on audio recordings (Furnas & Callas 2014, Cerquiera & Aide 2016).

Recent efforts have yielded tremendous growth in large-scale, long-term bioacoustic monitoring programs that accumulate vast amounts of acoustic data well beyond human capacity for efficient examination (Shonfield & Bayne 2017). Such large-scale data acquisition is accompanied by methodologies and software that enable semi-automated detection of sound-producing wildlife species from audio recordings. No approach for automated detection is perfect, and detection methods can vary based on research goals, soundscape characteristics, and acoustic features of a target species sound (Towsey et al. 2012; Stowell et al. 2016). Hidden Markov models (Agranat 2009, Aide et al. 2013, Potamitis et al. 2014, Ranjard et al. 2016, Wildlife Acoustics 2016), spectrogram cross correlation (Avisoft Bioacoustics e.K. 2016, Hafner & Katz 2018), binary point matching (Hafner & Katz 2018), band-limited energy detection (Figueroa

2012, Bioacoustics Research Program 2015) and convolutional neural networks (Knight et al. 2017) are common approaches. Probabilistic classification methods also show promise (Ovaskainen et al. 2018).

In any approach, a detected audio signal is either a true positive detection, which is a sound produced by the target species, or a false positive detection, which is not a signal from a target species. Throughout this paper we will use the convention of styling true positive detections as "target signals" and false positive detections as "false alarms". Regardless of the automated detection method employed, when acting without human assistance, computer-automated methods often produce an unacceptable number of false alarms, wherein non-target noise is detected and incorrectly assigned to a target species (Acevedo et al. 2009). False alarm rates from computer-automated methods may vary widely from project to project based on the prevalence of similar sounds from non-target sources in the soundscape, acoustic characteristics of sounds made by the target species (Towsey et al. 2012), the type of automated detection routine used (Corrado-Bravo et al. 2017), the available number of target sound examples upon which automated methods may be trained (Stowell et al. 2016), the quality of training data (Knight & Bayne 2018) and selection of score thresholds above which detections may occur (Knight et al. 2017).

We illustrate the process of acquiring both true target signals and false alarms using a spectrogram cross-correlation template as a screening mechanism to accumulate detections for a North American desert songbird, the Verdin *(Auriparus flaviceps)*. First, we render a spectrogram of an audio recording (**Fig. 3. 1a**), in which a Verdin vocalized three times, each with a characteristic three note whistle at about 4 kHz on the y-axis. We set time and frequency limits that define a cross correlation-based detection template for the song occurring at ~24 seconds within the example recording (**Fig. 3. 1b**). This template thus provides an acoustic pattern issued by a known target species, and can be used to scan many recordings in pursuit of Verdin vocalizations. The template is compared to an audio recording in a moving window analysis, in which a correlation between the template and audio file is obtained for each window (**Fig. 3. 2**). We then select a correlation detection threshold for the template, which is a user-specified detection threshold ranging from 0 (no correlation) to 1 (full correlation). Only peaks with scores above the chosen threshold are considered detections. This process facilitates rapid screening of acoustic data to acquire a set of detections, which are either true target signals or false alarms (**Table 3.1a**). Distinguishing target signals from false alarms is the focus of this paper. Tangentially, this process also results in false negatives at the level of actual vocalization occurrence, in which a species produced a sound but was not detected by the template matching process (**Table 3. 1a**; see Brauer et al. 2016 and Katz et al. 2016 for assessment of false negatives using a template-matching system).

Although humans may distinguish between true target signals and false alarms by visually examining the spectrogram or listening to the audio file, this approach is inefficient against the sheer volume of data collected in an acoustic monitoring program. Alternatively, after the template screening step has been performed, users may manually verify a small subset of detections as target signals and false alarms and use these to train a variety of classification algorithms that can predict whether template detections are true or false positives. Such an approach describes a form of statistical learning called supervised learning, in which a human labels a subset of data for the algorithm so that it can map existing data to known output classes (Bishop 2006). Once an algorithm has been trained on known data, it can be tested to predict the class – target signal or false alarm – of unknown data.

Two key components must be addressed to undertake supervised statistical learning. First, one must decide which acoustic features (predictive variables) of a detection can be used by the algorithm to predict the outcome (target signal or false alarm). The best predictive features may vary based on sounds produced by any given target species, as well as soundscape circumstances such as wind, rain, anthropogenic noise, or non-target species vocalizing within the same frequency range. An example of a set of acoustic features can be seen in **Fig. 3. 1b**: the colored shading in each pixel of the spectrogram represents the amplitude (or sound intensity) at that point, and each of the amplitude values in this spectrogram serves as an acoustic feature. Other potential predictive features include binned zero-crossing rates, time and frequency contours of the amplitude probability mass function, and summary statistics of the frequency spectrum (Sueur et al. 2008) (**Appendix B. 1**).

Once predictive features have been obtained, they are fed into a classification algorithm that will map predictive feature inputs to known output labels (target signal or false alarm). For example, the *k*-Nearest Neighbors classifier predicts the class of a new observation based on its feature similarity to some '*k*' number of observations within the training set (Cover & Hart 1967). Support Vector Machines seek an equation that optimally separates classes based on a high number of feature dimensions in geometric space (Boser et al. 1992), while Random Forests average a number of feature-based decision trees in order to make predictions (Breiman 2001). Regularized Logistic Regression uses penalized maximum likelihood to shrink the values of predictive feature coefficients, reducing variance so that the resulting model is better equipped to predict outside the range of data upon which it was trained (Zou & Hastie 2005). To improve classification, multiple algorithms may be combined into an ‘ensemble’ method to predict the class of a new detection.

Classification methods for discriminating between target signals and false alarms thus provide an opportunity in large-scale automated acoustic wildlife monitoring. Climate and land use change are forces that shift the occurrence of species across vast spatial scales, and monitoring these shifts at large scales is paramount for natural resource practitioners tasked with maintaining and sustaining species, populations, and ecosystems. Acoustic monitoring can produce vast amounts of data for this purpose, but existing automated detection algorithms often deliver high rates of false positives (Acevedo et al. 2009, Buxton and Jones 2012, Marques et al. 2012, Duan et al. 2013, Shonfield & Bayne 2017). Without accessible, straightforward, and generalizable methodologies for the mitigation of false positives in long-term data sets, occurrence-based bioacoustics research will continue to suffer the complications imposed by prohibitive numbers of detection errors, which often preempt poor model inference, ill-informed management decisions, and undesirable conservation outcomes (Royle & Link 2006, Miller et al. 2011, Ruiz-Gutierrez et al. 2016).

## Objectives

The aim of this study was to explore key functionality in the R package *AMMonitor* (Balantic et al. in prep*a*) for the semi-automatic removal of false positives to increase the quality of monitoring data. Our objectives were to 1) use spectrogram cross-correlation templates as an initial screening step to accumulate detections for focal species in a pilot acoustic monitoring program, 2) train and test statistical learning classification algorithms to distinguish between target signals and false alarms acquired during the template screening phase, 3) use a trained and tested classifier on new detections and compute overall classification performance in comparison to the template screening system.

## 3.3. Materials and Methods

## Objective 1: Use templates as a screening step to acquire focal species detections from field data

### *Acquire Acoustic Recordings*

We piloted an acoustic monitoring program in the Colorado-Sonoran Desert on public land managed under the auspices of the U.S. Bureau of Land Management (BLM). Autonomous recording units were installed at 16 sites within the BLM-managed Riverside East Solar Energy Zone, a 599 square-kilometer patch allocated as a utility-scale solar renewable energy hub. Because this work is a proof of concept with a focus on methodology rather than on ecological inference, study sites were selected nonrandomly near microphyll woodland habitat to record songbirds, and historic breeding pond locations with the intention of recording Couch's Spadefoot Toad (*Scaphiopus couchii*). Acoustic monitoring units were located at least 800 meters away from one another to maximize independence of acoustic events.

Each audio recording unit was a modified Android cellular phone (2015 2nd Generation Motorola Moto E model XT1527 with 5.0.2 Lollipop Android Operating System) contained within a weatherproof case and attached to an external 10-watt solar panel for power. Each unit was outfitted with an external omnidirectional electret condenser microphone (JLI-61A, JLI Electronics). All units were secured to U-posts elevated 1.83 meters aboveground. Units recorded in WAV format at a sampling rate of 44.1 kHz. The data collection period ran from March 2016 to May 2017. Units located in microphyll woodland habitats recorded every day for one minute at 6:00, 6:30, 7:00, 7:30, 8:00, 16:00, 16:30, 17:00, and 17:30 PST. Two units located next to historic toad breeding ponds also recorded for one minute each day at 5:30, 6:00, 6:30, 7:00, 21:00, 21:30, 22:00, 22:30, and 23:00 PST (n = 9 surveys per phone per day). We used the CinixSoft Remote Schedule Voice Recorder App (CinixSoft 2014) and Easy Voice Recorder Pro (Digipom 2016) to schedule recordings and remotely send them to our server using the cellular network. All units were in airplane mode while recording to prevent electromagnetic interference that occurs while in cellular data mode.

### *Create templates for target species*

As monitoring targets for this environment, we chose three avian species common to the region: Eurasian Collared-Dove (*Streptopelia decaocto*), Gambel's Quail

(*Callipepla gambelii*), and Verdin (*Auriparus flaviceps*). The canonical call from Eurasian Collared-Dove is a three note ‘advertising coo’ used for mate attraction and territory defense (Romagosa 2012), with calls occurring at frequencies around 0.5 kHz (**Fig. 3. 3a**). The Gambel’s Quail *kaa* or *cow* call emitted by males, whose principal function is to announce mating availability, is a single upside-down u-shaped note typically occurring between 1-3 kHz (Gee et al. 2013) (**Fig. 3. 3b**). The male Verdin’s ‘whistle song’ is a two to four note whistle occurring around 4-6 kHz; little documentation exists with regard to individual or geographic variation (Webster 1999) (**Fig. 3. 3c**).

We created one template for each species from song events chosen out of the recordings acquired in Objective 1 (**Fig. 3. 3 d-f**), using the *monitoR* R package function *makeCorTemplate()* with a window length of 512, zero overlap, and the Hanning window function. As suggested by Katz et al. (2016), we developed the templates and their accompanying score thresholds iteratively, testing them on recording data outside the recording from which the template was constructed before settling on finalized versions.

### *Accumulate detections and obtain associated predictive features*

Using the templates and accompanying score thresholds developed in Objective 2, we ran the *AMMonitor* function *scoreDetections()* to accumulate detections for all recordings from the rapid field prototype described in Objective 1. The *scoreDetections()* function employs Pearson’s correlation coefficient to score amplitude values of a moving frame against those in the template, and then isolates local maxima in the score vector to identify detection events (Katz et al. 2016). As in **Fig. 3. 2,** peaks with scores exceeding the score threshold were considered detections, which were either true target signals or false alarms.

Concomitant with the accumulation of detections, we used the *scoreDetections()* function to extract the raw amplitude matrix values associated with each detection, the correlation score, and a number of acoustic summary features acquired via the R package *seewave* (Sueur et al. 2008). These features included binned zero-crossing rates, time and frequency contours of the amplitude probability mass function for each time and frequency bin, quantiles calculated from the cumulative distribution functions of the time and frequency probability mass functions, and statistical properties of the frequency spectrum such as the spectral mean, standard deviation, standard error, median frequency, dominant (mode) frequency, frequency quartiles, centroid, skewness, kurtosis, flatness, and entropy (**Appendix B. 1)**.

## Objective 2: Train and test classification algorithms to distinguish between target signals and false alarms for each species

### *Manually label a subset of detections; split into training and testing data*

Once detections had been acquired via the template screening step in Objective 1, we manually verified all detections within the first month of field sampling (March 2016). We used the *AMMonitor* function *verify()* to assist with manual labeling of all detections as true and false positives for each target species. Each detection was labeled by the lead author primarily by visual identification on the spectrogram. We made additional effort to listen to visually ambiguous detections to confirm their class labels.

For Eurasian Collared-Dove (hereafter ECDO), we labelled detections as target signals if at least two notes were contained within the detection frame. For Gambel's Quail (hereafter GAQU), we labelled detections as target signals if they contained one frequency-modulated call signal. For Verdin (hereafter VERD), we labelled detections as target signals if at least one frequency-modulated whistle note was contained within the detection frame. Any other detections were labeled as false alarms. After verification, we split the labeled datasets into training (70%) and testing (30%) data, which is a common heuristic data split in statistical learning problems (Weinberger et al. 2006), and used the *createDataPartition()* function in the R package *caret* (Kuhn 2016) to obtain a balanced split of features, target signals, and false alarms.

### *Train and test statistical learning classifiers*

To construct models for each species and train them on the training data sets, we invoked the *AMMonitor* function *classifierTrain()*, which utilizes functions from the machine learning R package *caret* (Kuhn 2016). We trained our classifiers on raw data with no preprocessing (i.e., no scaling or transformation of the acoustic feature data). We used the method 'kknn' for kernelized *k*-nearest neighbors, which tunes to select an optimal *k,* 'svmLinear' for linear support vector machines, which entails the optimization of a cost parameter, 'svmRadial' for radial support vector machines, which optimizes both a cost parameter and the σ value of the radial basis function kernel, 'rf' for random forests, which involves tuning a parameter for the number of randomly selected predictor variables, and 'glmnet' for regularized logistic regression, which requires tuning a regularization parameter (λ) and a mixing parameter (α). Because a prohibitive entry point to the use of statistical learning methods is meticulously tuning algorithms to produce acceptable models, and because our aim was to generate extensible methodology accessible to researchers with little or no statistical learning experience, we used the default *caret* package tuning grids for all five models. Lastly, we applied 10-fold cross validation during the training phase to reduce model overfitting.

After the training phase, we tested the trained classifiers on the 30% of unseen data using the *AMMonitor* function *classifierTest()*. For every detection, each of the five classifiers yielded a probability that the detection was of the target signal class. While the logistic regression and random forest models output actual probabilities, the support vector machines and k-nearest neighbors fit a sigmoid function on their outputs to return probability-like values between 0 and 1 (Kuhn 2016). Detections with values of 0.5 or above were classified as target signals; those below were classified as false alarms.

### *Assess classifier performance on the test set*

Since labels for the test data were already known (target signal or false alarm), the training and testing procedure resulted in a confusion matrix summarizing the true classes of each detection and the classes to which they were assigned by each classifier (e.g., **Table 3. 1b**). We used the *AMMonitor* function *classifierAssess()* to calculate several measures of classifier performance (**Table 3. 1b)**. The literature contains rich debate over measures of classifier performance (Powers 2007), but the most useful evaluation measures depend on the research motivation behind using classification, as well as on the total number of observations and the balance of classes, which is why we sought a range of evaluation measurements.

For our study, we gave special merit to four performance metrics, all of which range in value from 0 to 1, with scores closest to 1 being most desirable (highlighted in **Table 3. 1b**). First, sensitivity (a.k.a. recall or true positive rate) is of particular interest because it denotes the proportion of target signals correctly identified by the classifier [TP / (TP + FN)] (**Table 3. 1b**). Second, specificity (or true negative rate) denotes the proportion of false alarms correctly identified as such, making them true negatives within the confusion matrix [TN / (TN + FP]. Third, positive predictive value (a.k.a. precision) expresses the proportion of *predicted* positive detections that are *actually* target signals [TP/(TP + FP]. Finally, the F1 score represents a weighted average of positive predictive value and sensitivity, quantifying the tradeoff between a desire for high positive predictive value and high sensitivity. The F1 score is calculated as 2 * Positive Predictive Value * Sensitivity / (Positive Predictive Value + Sensitivity). Maximizing all four of these metric scores was a primary goal in classifier evaluation.

We also constructed Receiver-Operating Characteristic (ROC) curves, which plot the true positive rate (sensitivity) against the false positive rate (1 – specificity). Many classification problems involve imbalanced datasets, in which the number of false positive cases greatly outweighs the number of true positive cases or vice versa. Class imbalances undermine performance metrics like accuracy and area under the ROC curve (AUC): a classifier may predict the majority class for most or all observations in the test set and still attain a high accuracy score, which is why measures beyond accuracy are necessary (Zhu & Davidson 2007). To account for this, we also constructed Precision-Recall Curves, which plot positive predictive value (a.k.a. precision) against sensitivity (a.k.a. recall) (Davis & Goadrich 2006). For both ROC and Precision-Recall curves, we defined AUC values matching or exceeding 0.80 as acceptable, and values matching or exceeding 0.90 as high performance, with values of 1 indicating perfect performance.

*Create and assess performance-weighted class probability ensemble methods*

After performance metrics were computed for each of the five classifiers individually, we used the *classifierAssess()* function to aggregate the results of the five classifiers. In statistical learning, such methods are known as 'ensembles,' in which classification occurs as a consequence of aggregation or integration across multiple distinct algorithms to improve predictive performance. The *classifierAssess()* function established four simple performance-weighted ensemble methods, weighting each classifier's probability that a detection was of class "target signal" by the classifier's test phase sensitivity, specificity, positive predictive value, or F1. Each performance-weighted method produced a single ensemble class probability of true detection (*Target Signal$_e$*), calculated as

$$P(Target\ Signal)_e = [\theta] \bullet [S]$$

where [θ] is a vector of length five consisting of the individual probability of a target signal for each of the five classifiers, and [S] is a length five vector of normalized performance scores that sums to 1.0. The [S] vector is computed by dividing each classifier's score on the metric of interest by the maximum score within the vector, resulting in a vector that represents how proportionally close each score is to the top score for that metric, which is then normalized to sum to 1 (**Appendix B. 2**). Thus, contributions of lower-scoring classifiers are diminished, while higher-scoring classifiers have stronger impact on ensemble class predictions for any given detection. We then computed the sensitivity, specificity, positive predictive value, and F1 of the ensemble results.

## Objective 3: Assess the performance of a trained and tested classifier on new detections

In releasing our trained and tested classifiers "into the wild" on new, incoming template detections, our goal was to use classification to eliminate as many false alarms as possible, while still retaining true target signals needed for meaningful estimations of species occurrence. For this reason, we chose to proceed using the ensemble method weighted by the F1 score as our predictive classifier on new data for all three species.

Using *AMMonitor's classifierPredict()* function, we invoked the ensemble method weighted by the F1 score to predict the class of all detections across the entire recording dataset that were *not* seen during the training and testing phase. Thus, the training and testing phase occurred on all data from March 2016, and the prediction phase occurred on all data spanning the 14 month period from April 2016 to May 2017. We then manually verified all detections in the prediction set, and computed metrics to evaluate whether our classification method improved upon the initial template screening step. We calculated the positive predictive value and F1 score for the template screening step, and calculated sensitivity, specificity, positive predictive value, and F1 score for the classifiers to compare performances of the two systems. We assumed that the template screening method had a sensitivity of 1 and specificity of 0 with regard to distinguishing target signals from false alarms.

## 3.4. Results

### Objective 1: Use templates as a screening step to acquire focal species detections from field data

We collected a total of 40,496 one-minute recordings from March 2016 to May 2017 across 16 cellphone-based audio recorders. An unknown number of recordings contained electromagnetic interference for reasons unknown, all of which were retained in the dataset. We created spectrogram cross-correlation templates for ECDO, GAQU, and VERD (**Fig. 3. 3 d- f**), and identified score thresholds of 0.43, 0.33, and 0.23, respectively. At these score thresholds, we collected a total of 4,427 detections for ECDO, 1,464 detections for GAQU, and 4,241 detections for VERD, resulting in a total of 10,132 detections.

### Objective 2: Train and test classification algorithms to distinguish between target signals and false alarms for each species

*Manually label a subset of detections; split into training and testing data*

There were 631 detections acquired from 54.3 hours of recordings from March 2016 at the selected score thresholds: 323 ECDO, 62 GAQU, and 246 VERD (**Table 3. 2**). It took approximately one hour to manually verify all March 2016 detections using our chosen verification standards. The ECDO and GAQU datasets were adequately balanced, with 135 true and 188 false for ECDO, and 34 true and 28 false for GAQU. The VERD dataset had a class imbalance with 49 true and 197 false (**Table 3. 2)**. A visual summary of verifications is contained in **Fig. 3. 4**, wherein spectrograms for verified detections were averaged across the amplitude values to show the mean target signal and mean false alarm.

*Train and test statistical learning classifiers*

Despite the low total number of GAQU detections and considerable class imbalance for VERD, all classification models converged during the training phase and were functional for testing and assessment. It took a total of 3.5 minutes to train and test the models for all three species. Because the *k*-nearest neighbors and support vector machines algorithms do not provide readily interpretable output with regard to predictive power of acoustic features, here, we only report feature selection results from the regularized logistic regression and random forest models.

Features summarizing statistical properties of the frequency spectrum served as the strongest predictors for distinguishing between target signals and false alarms. For ECDO, both the regularized logistic regression and random forest models identified spectral mean, spectral centroid, and spectral mode as the top predictors, with no other variables providing predictive value. For GAQU, both the regularized logistic regression and random forest models identified spectral kurtosis as the top predictor, with spectral skewness adding a lesser contribution. Correlation score, several binned zero-crossing rates, several time and frequency contours, and a number of individual amplitude values also supplied marginal predictive capacity. For VERD, the regularized logistic regression model identified spectral entropy as the top predictor, followed by spectral flatness and spectral kurtosis. The random forest model identified spectral skewness as the most important predictor, with spectral kurtosis and correlation score supplying some predictive impact. A number of time and frequency contours and binned zero crossing rates also offered minor predictive value.

*Assess classifier performance on the test set; create and assess performance-weighted class probability ensemble methods*

Performances across the various metrics, classification approaches, and templates varied (**Table 3. 3**). All five classifiers performed well on the ECDO data, reporting perfect sensitivity (1.00) with values ranging from 0.93 to 0.99 for specificity, positive predictive value, and F1. For the GAQU models, regularized logistic regression, random forest, and *k*-nearest neighbors each achieved sensitivities of 0.90, specificities and positive predictive values of 1.00, and F1 scores of 0.95, while the linear support vector machine had a lower sensitivity (0.80) and thus a lower F1 score (0.89). Though the radial support vector machine had perfect sensitivity (1.00), it failed to identify any false alarms (specificity = 0.00) and thus produced poor positive predictive value (0.56) and a sub-optimal F1 score (0.71). The large class imbalance in the VERD data, with many false alarms and few target signals, resulted in a radial support vector machine model adept at identifying false alarms (specificity = 1.00) but incapable of identifying target signals (sensitivity = 0.00), consequently producing NA results for positive predictive value and F1 score. Indeed, for VERD, all five classifiers were effective at identifying false alarms, as indicated by specificities ranging from 0.95 to 1.00, but weaker at identifying target signals, with sensitivities ranging from 0 to 0.86. The regularized logistic regression and random forest models nevertheless attained adequate positive predictive value and F1 scores of 0.86.

The weighted ensemble approaches all performed similarly across performance metrics for both ECDO and GAQU, and displayed greater performance variation for VERD (**Table 3. 3c**). The ensemble classifier weighted by F1 score, upon which we chose to focus in advance, was a top-performing model for ECDO and GAQU on all metrics, producing scores ranging between 0.98 and 1.00 (ECDO), and from 0.90 to 1.00 (GAQU). For VERD, the ensemble classifier weighted by F1 was slightly outperformed by the regularized logistic regression and random forest classifiers.

ROC curves (**Fig. 3. 5**) of the training and testing data generated acceptable areas under the curve (AUC) in most cases, aside from the radial support vector machine's performance for GAQU and VERD, which was indistinguishable from that of a random guess. Precision-Recall curves (**Fig. 3. 6.**) exhibited high performance for ECDO (all AUC >= 0.99), high performance for GAQU despite the low amount of training data (aside from the radial support vector machine, all test set AUC >= 0.91), and variable performance for VERD, though the random forest and F1-weighted ensemble methods both met or exceeded AUC of 0.93 on the test set.

## Objective 3: Assess the performance of a trained and tested classifier on new detections

From April 2016 to May 2017, the template screening phase resulted in 9,501 new detections: 4,104 ECDO, 1,402 GAQU, and 3,995 VERD. Applying the trained and tested ensemble classifiers to these data yielded classifier sensitivities of 0.70 (ECDO), 0.67 (GAQU) and 0.81 (VERD) (**Fig. 3. 7**), compared to sensitivities of 1 in the template screening phase. Classifier specificities were 0.98 (ECDO), 0.85 (GAQU), and 0.99 (VERD), compared to specificities of 0 for the template screening phase.

Overall positive predictive values from the classification phase were 0.75 (ECDO), 0.965 (GAQU), and 0.69 (VERD), compared to positive predictive values of 0.06 (ECDO), 0.865 (GAQU), and 0.02 (VERD) for the template screening phase (**Fig. 3.7**). F1 scores improved from 0.12 to 0.725 (ECDO) and from 0.04 to 0.75 (VERD) with the classifier system, but declined from 0.93 to 0.79 in the GAQU model (**Fig. 3. 7**).

The majority of false alarms for ECDO stemmed from wind and anthrophonic sources such as faraway highway traffic noise, though several false cases were prompted by vocalizations from Greater Roadrunner (*Geococcyx californianus*), White-Winged Dove (*Zenaida asiatica*), and Mourning Dove (*Zenaida macroura*). Most GAQU false alarms resulted from electromagnetic inference, with a few due to Common Raven (*Corvus corax*) and Phainopepla (*Phainopepla nitens*). VERD false alarms occurred overwhelmingly as a consequence of electromagnetic interference, though some were caused by crickets and other songbirds.

## 3.5. Discussion

We demonstrate that statistical learning approaches can be used to mitigate false detections acquired within an automated acoustic wildlife monitoring dataset while retaining sufficient true detections for inference about species occurrence status. Compared to a basic template-matching system, the ability to identify false alarms improved, and positive predictive values increased in all cases demonstrated here, though there was a tradeoff in capacity to identify all target signals: we observed a decrease in the F1 score for GAQU, though F1 scores for ECDO and VERD increased markedly. Since GAQU is known to be a highly gregarious and vocally available species (Gee et al. 2013), the observed increase in positive predictive value to 0.96 at the expense of sensitivity is likely a desirable tradeoff. For a rare or acoustically cryptic species, this tradeoff in comparative sensitivity with respect to detected events would not be advantageous.

Three main concepts emerge from this work: First, although other auspicious classification methods implicitly strive to minimize false positives (e.g., Heinicke et al. 2015, Bas et al. 2017, Corrada-Bravo et al. 2017, Ranjard et al. 2017), none that we know of explicitly address false positive mitigation within the context of template-based or threshold-based detection. In addition to making binary predictions about each detection's class, this method also has the advantage of producing probability values for each detection, which may be aggregated to predict the overall probability of species occurrence (Balantic et al. in prep*b*).

Second, an advantage of this method is the opportunity to create ensemble classifiers that overtly capture a research program's monitoring needs with regard to vocalization characteristics of focal species. For example, researchers might opt for a positive predictive value-weighted ensemble classifier for gregarious species, or a sensitivity-weighted ensemble for rare or cryptic species. Research groups could make a variety of decisions about which classification method(s) to employ in production based on research objectives, characteristics of focal species, and classifier performance during the training and testing phase. Systematic decision tools do not presently exist in this arena, and the interpretation of classifier assessment metrics persists as an underappreciated challenge when applying statistical learning approaches to real-world problems.

Lastly, template creation, including selection of the score threshold, is a highly influential component of the detection and classification process. The balance of target signals and false alarms occurring in a dataset is a function of the quality of data from which a template is constructed (Katz et al. 2016, Knight & Bayne 2018), the score threshold selected (Brauer et al. 2016, Katz et al. 2016, Knight et al. 2017), verification standards for manual labeling of target signal and false alarm training data, soundscape features such as non-target noise sources that contribute to detections, individual variation in sounds produced by the target species, and overall vocal availability of the target species, much of which is difficult to know in advance. Low template score thresholds may be necessary for research programs pursuing rare or vocally elusive species, or for circumstances where there is considerable uncertainty around how the template will perform in practice; it follows that large numbers of false alarms are possible, though there is little consistency across detection methodologies for detection threshold selection (Shonfield & Bayne, 2017). In practice, for species with multiple well-described vocalization types, a different template can be deployed for each vocalization type *a priori*, and overall false positive reduction at the detection level can be aggregated across the template portfolio up to the site level for use in occupancy models.

Increasing use of automated methods for detecting target species signals from audio recordings demonstrates the growing importance of accessible automated detection methods. Template-based software methods like spectrogram cross-correlation and binary point matching present an accessible approach with a low barrier to entry for researchers (Hafner & Katz 2018), but factors like inappropriate score detection thresholds, an unwittingly poor template choice, noisy soundscapes, and acoustic features of the target signal may conspire to generate unacceptably high numbers of false alarms. Here, we investigated statistical learning methods that allow researchers to semi-automatically eliminate large numbers of false alarms, and showed that these methods may improve the monitoring quality of automated detection data from template-based detection systems.

## 3.6. Acknowledgments

This project was supported by the U.S. Bureau of Land Management. Cathleen Balantic was supported by NSF IGERT grant 1144388. We thank Mark Massar (BLM) for providing the opportunity to work in the Riverside East Solar Energy Zone, Jonathan Katz for assistance with R programming, and reviewers for their helpful comments. Use of trade names or products does not constitute endorsement by the U.S. Government. The Vermont Cooperative Fish and Wildlife Research Unit is jointly supported by the U.S.

Geological Survey, University of Vermont, Vermont Department of Fish and Wildlife, and Wildlife Management Institute.

## 3.7. References

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J. & Aide, T.M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics, 4, 206–214. doi: https://doi.org/10.1016/j.ecoinf.2009.06.005

Agranat, I. D. (2009). Automatically Identifying Animal Species from their Vocalizations. Wildlife Acoustics, Inc., Concord, MA.

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. PeerJ 1:e103. doi: https://doi.org/10.7717/peerj.103

Avisoft Bioacoustics e.K. (2016). Avisoft-SASLab Pro version 5.2.10 [Computer Software]. URL http://www.avisoft.com/.

Balantic C.M., Katz, J., & T. M. Donovan. (Unpublished results) In Prep*a*. AMMonitor R Package.

Balantic, C.M, & T. M. Donovan. In Prep*b*. (Unpublished results) Dynamic wildlife occupancy models using automated acoustic monitoring data

Bas, Y., Bas, D. & Julien, J.-F. (2017). Tadarida: A Toolbox for Animal Detection on Acoustic Recordings. Journal of Open Research Software. 5(1), p.6. doi: http://doi.org/10.5334/jors.154

Bioacoustics Research Program. (2015). Raven Pro 1.5: Interactive Sound Analysis Software [Computer Software]. URL http://www.birds.cornell.edu/raven.

Bishop, C.M. (2006). Pattern recognition and machine learning. Springer. ISBN: 0387310738 9780387310732.

Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory - COLT '92, pp. 144–152. ACM Press, New York, New York, USA.

Brauer, C., T. Donovan, R. Mickey, J. Katz, & Mitchell, B. (2016). A comparison of acoustic monitoring methods for common anurans of the northeastern United States. Wildlife Society Bulletin 40:140-149. doi: 10.1002/wsb.619

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32. doi: https://doi.org/10.1023/A:1010933404324

Buxton, R. T., & Jones, I.L. (2012). Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration. Journal of Field Ornithology 83:47-60. doi: http://dx.doi.org/10.1111/j.1557-9263.2011.00355.x

Cerqueira, M. C., & Aide, M.T. (2016). Improving distribution data of threatened species by combining acoustic monitoring and occupancy modeling. Methods in Ecology and Evolution. 7(11), 1340-1348. doi: 10.1111/2041-210X.12599

Corrada-Bravo, C.J.C., Berrios, R.A, & Aide, T.M. (2017). Species-specific audio detection: a comparison of three template-based detection algorithms using random forests. PeerJ Computer Science 3:e113. doi: https://doi.org/10.7717/peerj-cs.113

CinixSoft. (2014). CinixSoft Remote Schedule Voice Recorder v4.2.0. [Android App]. URL http://www.cinixsoft.com/

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13, 21–27. doi: 10.1109/TIT.1967.1053964

Davis, J. & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine Learning, Pittsburg, PA.

Digipom (2016). Easy Voice Recorder Pro [Android App]. URL http://www.digipom.com/portfolio-items/easy-voice-recorder/

Duan, S., Zhang, J., Roe, P, Wimmer, J., Dong, X., Truskinger, A. & Towsey, M (2013). Timed Probabilistic Automaton: a bridge between Raven and Song Scope for automatic species recognition. In Munoz-Avila, Hector, & Stracuzzi, D.J. (Eds). Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference, AAAI, Bellevue, Washington, USA, pp. 1519-1524.

Figueroa, H. (2012). XBAT [Computer Software]. Bioacoustics Research Program. URL http://www.xbat.org.

Furnas, B.J. & Callas, R.L. (2014). Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. Journal of Wildlife Management 79(2): 325-337. doi: 10.1002/jwmg.821

Gee, J., Brown, D.E., Hagelin, J.C., Taylor, M. & Galloway, J. (2013). *Gambel's Quail* (Callipepla gambelii), The Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. doi: 10.2173/bna.321

Hafner S. & Katz J. (2018). monitoR: Acoustic template detection in R. R package version 1.0.7, URL: http://www.uvm.edu/rsenr/vtcfwru/R/?Page=monitoR/monitoR.htm.

Heinicke, S., Kalan, A.K., Wanger, O.J., Mundry, R., Lukashevich, H., & Kuhl, H.S. (2015). Assessing the performance of a semi-automated acoustic monitoring system for primates. Methods in Ecology and Evolution, 6(7): 753-763. doi: 10.1111/2041-210X.12384

Katz, J., Hafner, S.D. & Donovan, T. (2016). Assessment of Error Rates in Acoustic Monitoring with the R package monitoR. Bioacoustics, 25, 177–196. doi: https://doi.org/10.1080/09524622.2015.1133320

Knight, E. C., Hannah, K.C., Foley, G., Scott, C., Mark Brigham, R., & Bayne, E. (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology 12(2):14. doi: https://doi.org/10.5751/ACE-01114-120214

Knight, E.C. & Bayne. E.M. 2018. Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. Bioacoustics. doi: 10.1080/09524622.2018.1503971

Kuhn, M. (2016). caret: Classification and Regression Training. R package version 6.0-71. http://CRAN.R-project.org/package=caret

Miller, D.A.W., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology, 92, 1422–1428. doi: 10.1890/10-1396.1

Ovaskainen, O., Moliterno de Camargo, U., & Somervuo, P. (2018). Animal Sound Identifier (ASI): software for automated identification of vocal animals. *Ecology letters*, 21(8): 1244-1254. doi: https://doi-org.ezproxy.uvm.edu/10.1111/ele.13092

Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L., & Sauer, J.R. (2002). Large scale wildlife monitoring studies: statistical methods for design and analysis. Environmetrics, 13(2): 105-119. doi: https://doi.org/10.1002/env.514

Potamitis, I., Ntalampiras, S., Jahn, O. & Riede, K. (2014). Automatic bird sound detection in long real-field recordings: applications and tools. Applied Acoustics 80:1-9. doi: https://doi.org/10.1016/j.apacoust.2014.01.001

Powers, D.M.W. (2007). Evaluation: From Precision, Recall, and F-Factor to ROC, Informedness, Markedness & Correlation. School of Informatics and Engineering, Flinders University. Adelaide, Australia. Technical Report SIE-07-001.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ranjard, L., Reed, B.S., Landers, T.J., Raynar, M.J., Friesen, M.R., Sagar, R.L., & Dunphy, B.J. (2017). MatlabHTK: a simple interface for bioacoustics analyses using hidden Markov models. Methods in Ecology and Evolution 8(5): 615-621. doi: 10.1111/2041-210X.12688

Romagosa, C.M. (2012). Eurasian Collared-Dove (Streptopelia decaocto), The Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology; Retrieved from the Birds of North America: https://birdsna.org/Species-Account/bna/species/eucdov. doi: 10.2173/bna.630

Royle, J.A. & Link, W.A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. Ecology, 87, 835–41. doi: 10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2

Ruiz-Gutierrez, V., Hooten, M.B. & Campbell Grant, E.H. (2016). Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias (N. Yoccoz, Ed.). Methods in Ecology and Evolution, 7, 900–909. doi: 10.1111/2041-210X.12542

Shonfield, J. & Bayne, E.M. (2017). Autonomous recording units in avian ecological research: current use and future applications. Avian Conservation and Ecology, 12(1): 14. DOI: 10.5751/ACE-00974-120114

Stowell, D., Wood, M., Stylianou, Y. & Glotin, H. (2016). Bird detection in audio: a survey and a challenge. IEEE International Workshop on Machine Learning for Signal Processing, Salerno, Italy. doi: 10.1109/MLSP.2016.7738875

Sueur J., Aubin T., & Simonis, C. (2008). Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics, 18: 213-226. doi: http://dx.doi.org/10.1080/09524622.2008.9753600

Towsey, M., Planitz, B., Nantes, A., Wimmer, J. & Roe, P. (2012). A toolbox for animal call recognition. Bioacoustics, 21, 107–125.doi: http://dx.doi.org/10.1080/09524622.2011.648753

Webster, M.D. (1999). Verdin (Auriparus flaviceps), The Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology; Retrieved from the Birds of North America: *https://birdsna.org/Species-Account/bna/species/verdin. doi: 10.2173/bna.470*

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473-1480).

Wildlife Acoustics. (2016). Kaleidescope [Computer Software]. URL http://www. wildlifeacoustics.com.

Zhu, X. & Davidson, I. (Eds.). (2007). Knowledge Discovery and Data Mining. IGI Global.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

## 3.8. Tables

Table 3. 1. Confusion Matrix Examples to distinguish between true and false positives at the vocalization occurrence level vs. the detection level.

1a. Following from Fig. 3. 2, a confusion matrix at the ‘vocalization occurrence level’ summarizes all vocalizations issued by the target species and captured within the recording. Three vocalizations are correctly detected and referred to as ‘true positives’ (TP = 3), while one non-vocalization is flagged as signal from the target species, which is a ‘false positive’ (FP = 1). Meanwhile, two vocalizations are missed by the system and are ‘false negatives’ (FN = 2). Lastly, approximately 38 time bands within the recording are appropriately ignored since they contain no vocalizations from the target and are ‘true negatives’ (TN = 38). We highlight the top row to indicate that only the true and false positive detections from Table 3. 1a are considered within this paper.

| | Actual Class:<br>Vocalization | Actual Class:<br>No Vocalization | |
|---|---|---|---|
| Predicted Class:<br>Vocalization | TP = 3 | FP = 1 | 4 |
| Predicted Class:<br>No Vocalization | FN = 2 | TN = 38 | 40 |
| | 5 | 39 | 44 |

3. 1b. Confusion Matrix for Detections Only.

The four detections highlighted in the top row of Table 3. 3. 1a are now subject to ‘detection level’ classification, in which an algorithm is used to reclassify events in an effort to minimize false alarms. The goal of reclassification is to maximize sensitivity, specificity, positive predictive value and F1 score. Ideal conditions are given in brackets next to the actual condition.

| | Actual Class Label:<br>Target Signal | Actual Class Label:<br>False Alarm | Row Sum: |
|---|---|---|---|
| Predicted Class:<br>Target Signal | TP = 3 **[3]** | FP = 1 **[0]** | 4 **[3]**<br>Pos Pred. Value = ¾ = 0.75<br>**[Pos. Pred. Value = 3/3 = 1]** |
| Predicted Class:<br>False Alarm | FN = 0 **[0]** | TN = 0 **[1]** | 0 **[1]** |
| Column Sum: | 3 **[3]**<br>Sensitivity = 3/3 = 1<br>**[Sensitivity = 3/3 = 1]** | 1 **[1]**<br>Specificity = 0/1=0<br>**[ Specificity = 1/1 = 1]** | 4<br>F1 = 2*(0.75*1)/(0.75 +1) = 0.86<br>**[F1 = 2*(1*1) / (1+1) = 1 ]** |

Table 3. 2. Number of manually verified detections from March 2016 used as classifier training and testing data for three focal species: Eurasian Collared-Dove (ECDO), Gambel’s Quail (GAQU), and Verdin (VERD).

| **Template** | **Total N** | **Total True** | | **Total False** | |
|---|---|---|---|---|---|
| **ECDO** | 323 | 135 | | 188 | |
| | | Training | Testing | Training | Testing |
| | | 95 | 40 | 132 | 56 |
| **GAQU** | 62 | 34 | | 28 | |
| | | Training | Testing | Training | Testing |
| | | 24 | 10 | 20 | 8 |
| **VERD** | 246 | 49 | | 197 | |
| | | Training | Testing | Training | Testing |
| | | 35 | 14 | 138 | 59 |

Table 3. 3. Assessment of Classifier Performance on the Test Data**.** The classifier performance metrics in this table can take on values between 0 (worst) and 1 (best). Rows indicate classifiers, and columns indicate performance metrics.

a. ECDO Models:

| **Classifier** | **Sensitivity** | **Specificity** | **Pos. Pred. Value** | **F1** |
|---|---|---|---|---|
| Regularized Logistic Regression | 1.00 | 0.98 | 0.98 | 0.99 |
| Linear Support Vector Machine | 1.00 | 0.95 | 0.93 | 0.96 |
| Radial Support Vector Machine | 1.00 | 0.96 | 0.95 | 0.98 |
| Random Forests | 1.00 | 0.98 | 0.98 | 0.99 |
| Kernelized k-Nearest Neighbors | 1.00 | 0.96 | 0.95 | 0.98 |
| Ensemble weighted by Sensitivity | 1.00 | 0.98 | 0.98 | 0.99 |
| Ensemble weighted by Specificity | 1.00 | 0.98 | 0.98 | 0.99 |
| Ensemble weighted by Pos. Pred. Value | 1.00 | 0.98 | 0.98 | 0.99 |
| Ensemble weighted by F1 | 1.00 | 0.98 | 0.98 | 0.99 |

b. GAQU Models:

| **Classifier** | **Sensitivity** | **Specificity** | **Pos. Pred. Value** | **F1** |
|---|---|---|---|---|
| Regularized Logistic Regression | 0.90 | 1.00 | 1.00 | 0.95 |
| Linear Support Vector Machine | 0.80 | 1.00 | 1.00 | 0.89 |
| Radial Support Vector Machine | 1.00 | 0.00 | 0.56 | 0.71 |
| Random Forests | 0.90 | 1.00 | 1.00 | 0.95 |
| Kernelized k-Nearest Neighbors | 0.90 | 1.00 | 1.00 | 0.95 |
| Ensemble weighted by Sensitivity | 0.90 | 1.00 | 1.00 | 0.95 |
| Ensemble weighted by Specificity | 0.90 | 1.00 | 1.00 | 0.95 |
| Ensemble weighted by Pos. Pred. Value | 0.90 | 1.00 | 1.00 | 0.95 |
| Ensemble weighted by F1 | 0.90 | 1.00 | 1.00 | 0.95 |

c. VERD Models:

| Classifier | Sensitivity | Specificity | Pos. Pred. Value | F1 |
|---|---|---|---|---|
| Regularized Logistic Regression | 0.86 | 0.97 | 0.86 | 0.86 |
| Linear Support Vector Machine | 0.71 | 0.95 | 0.77 | 0.74 |
| Radial Support Vector Machine | 0.00 | 1.00 | NA | NA |
| Random Forests | 0.86 | 0.97 | 0.86 | 0.86 |
| Kernelized k-Nearest Neighbors | 0.57 | 0.95 | 0.73 | 0.64 |
| Ensemble weighted by Sensitivity | 0.79 | 0.97 | 0.85 | 0.81 |
| Ensemble weighted by Specificity | 0.71 | 0.98 | 0.91 | 0.80 |
| Ensemble weighted by Pos. Pred. Value | 0.79 | 0.97 | 0.85 | 0.81 |
| Ensemble weighted by F1 | 0.79 | 0.97 | 0.85 | 0.81 |

## 3.9. Figures

Figure 3. 1. a. Verdin songbird vocalization within a recording. b. Example template created from Verdin vocalization occurring at ~24 seconds.

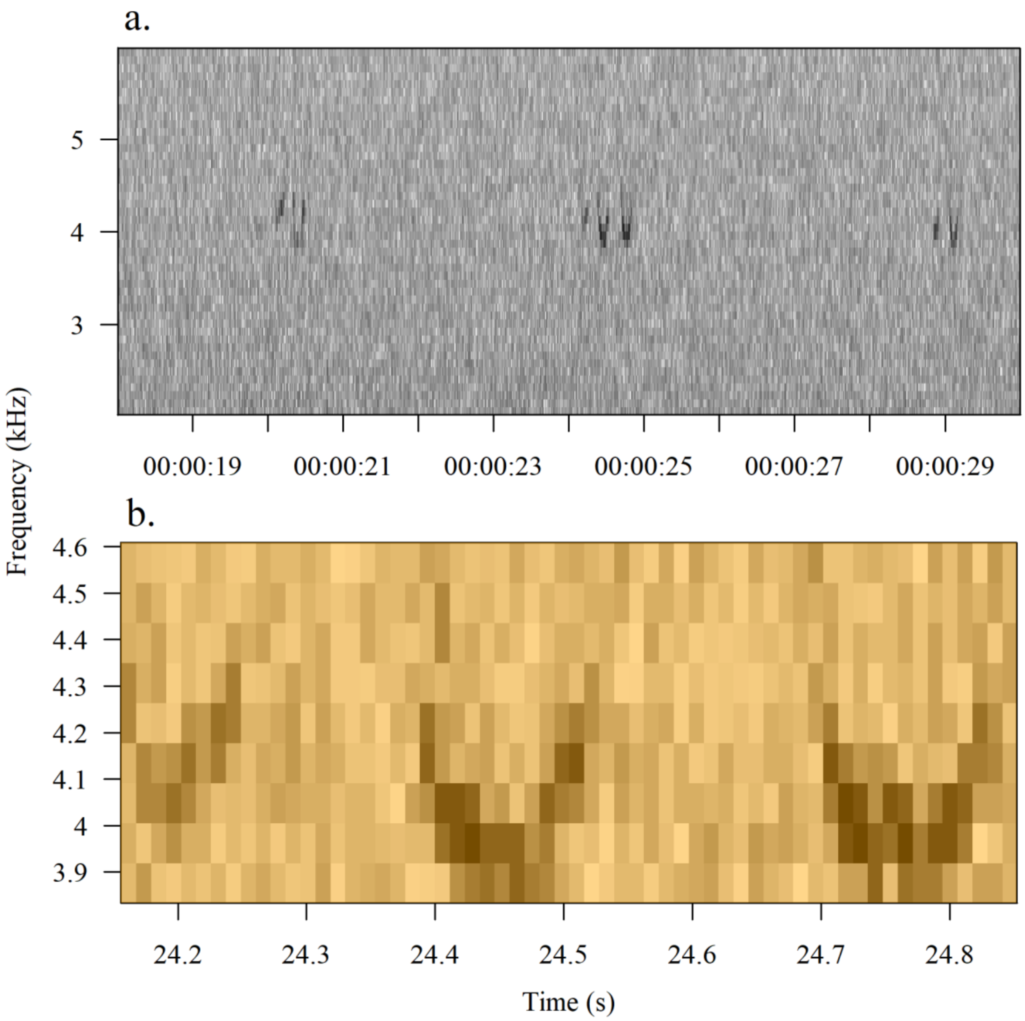


Figure 3. 2. Illustration of event detection via template matching paired with a score threshold. Red boxes in the top panel denote detections, while the red line in the bottom panel indicates a selected threshold (0.3). The first red box is a false alarm produced as a result of electromagnetic interference. The last three red boxes are all target signals wherein the Verdin is actually vocalizing. Note also two occurrence-level false negatives, in which the species is vocalizing but no detection occurred.

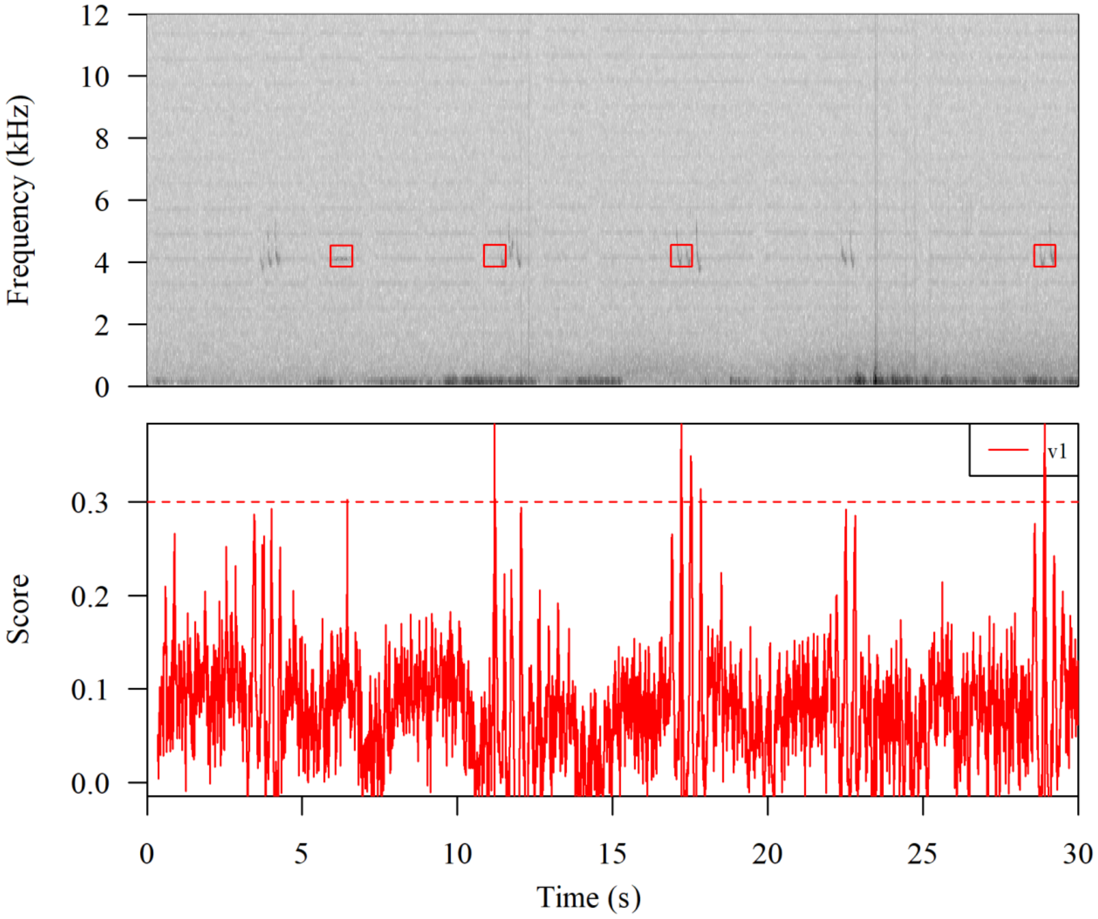


Figure 3. 3. Vocalization examples (a-c) and templates (d-f) for Eurasian Collared-Dove, Gambel’s Quail, and Verdin, respectively.

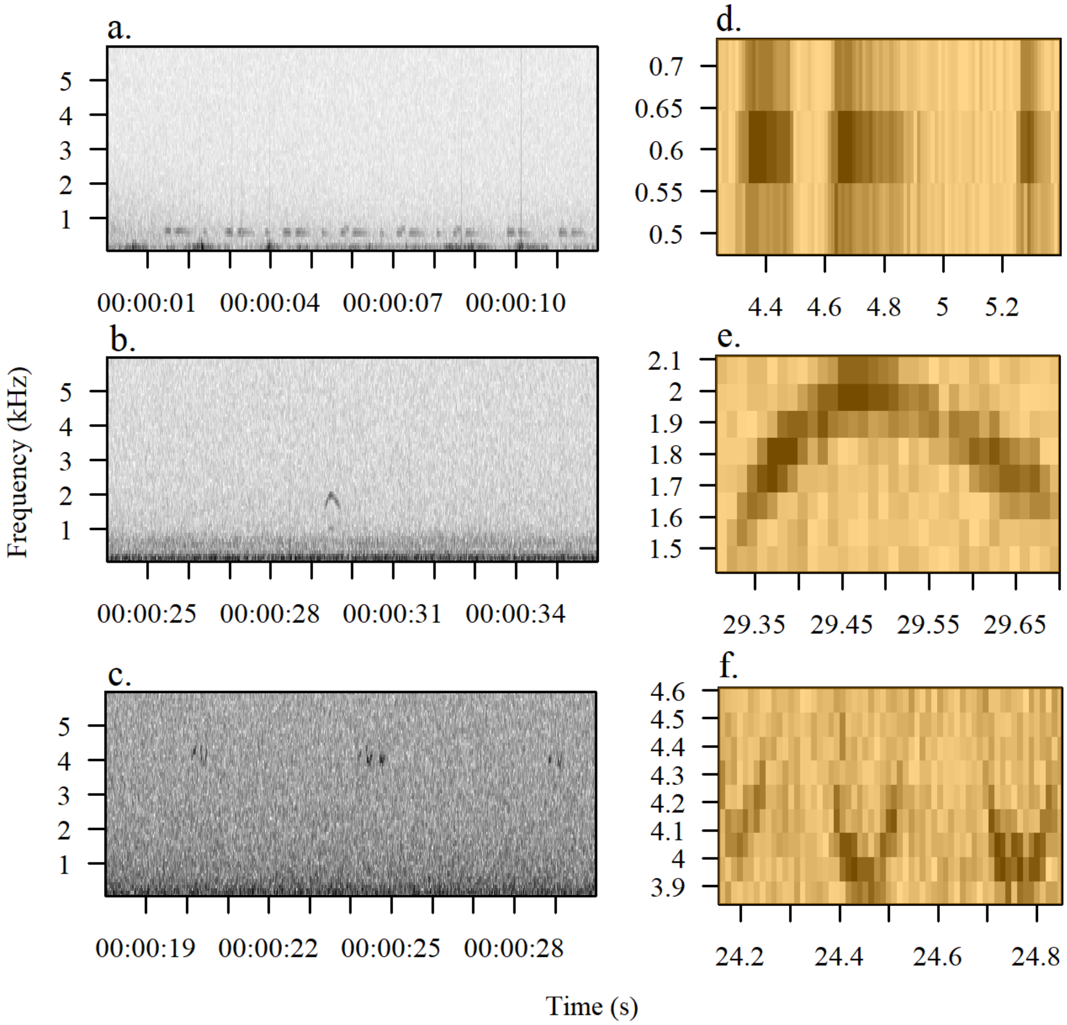


Figure 3. 4. Visual summary of all manually verified detections used as training and testing data. Templates used to collect detections (a-c) are juxtaposed against average spectrograms for all verifications (d-f), all target signal verifications (g-i), and all false alarm verifications (j-L).

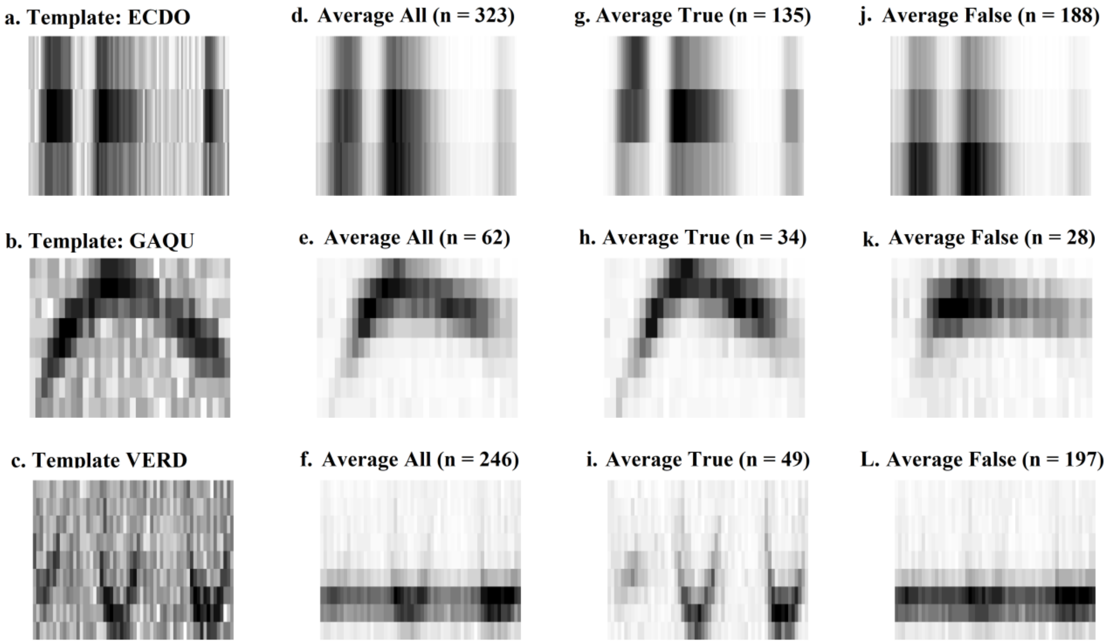
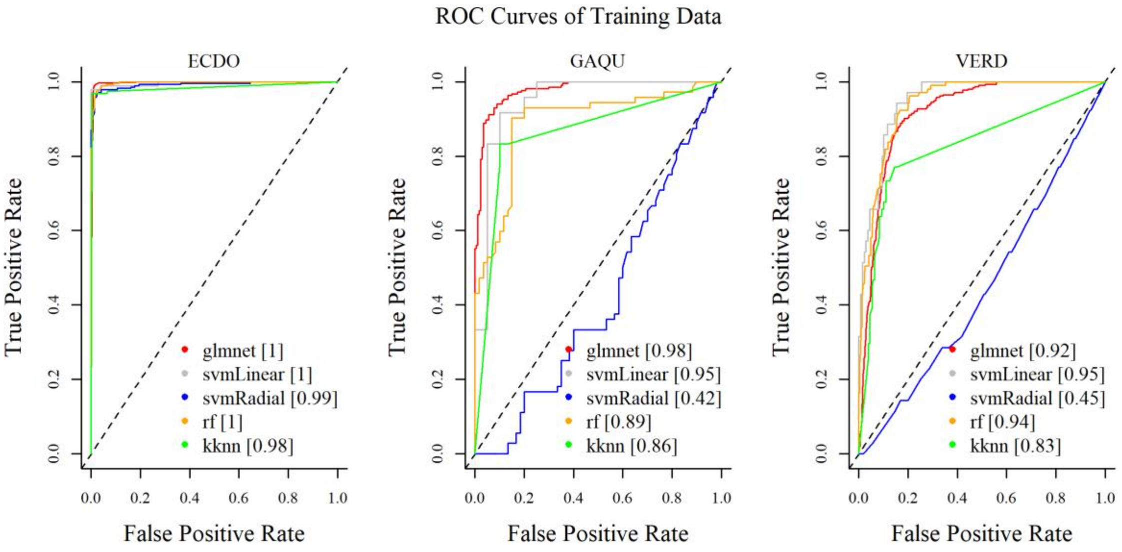


Figure 3. 5. Receiver-Operator Characteristic (ROC) Curves describing classifier performance during the training and testing phases. The upper panel shows ROC curves on the 10-fold cross-validated training data for the five classifiers. The bottom panel shows ROC curves on the test data. The ensemble classifiers only make predictions in the test phase, so the bottom panel also demonstrates the ensemble classifier weighted by F1 score. Area under the curve (AUC) is denoted next to each model's name in square brackets. Curves that reach into the upper left corner, with AUC values close to 1, show the best classification performance.


ROC Curves of Training Data
ECDO
GAQU
VERD
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
glmnet [1]
svmLinear [1]
svmRadial [0.99]
rf [1]
kknn [0.98]
glmnet [0.98]
svmLinear [0.95]
svmRadial [0.42]
rf [0.89]
kknn [0.86]
glmnet [0.92]
svmLinear [0.95]
svmRadial [0.45]
rf [0.94]
kknn [0.83]

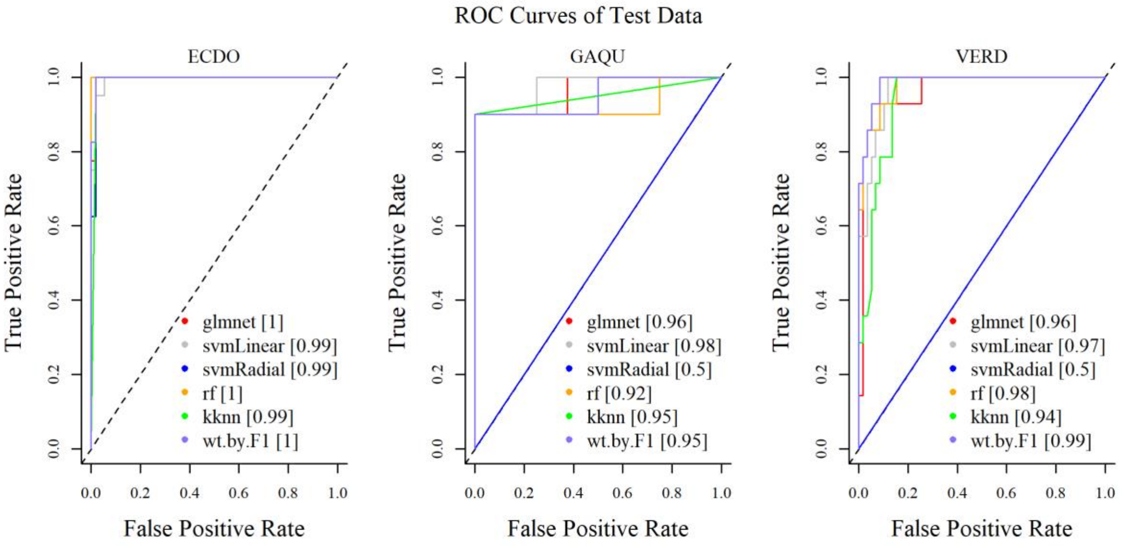
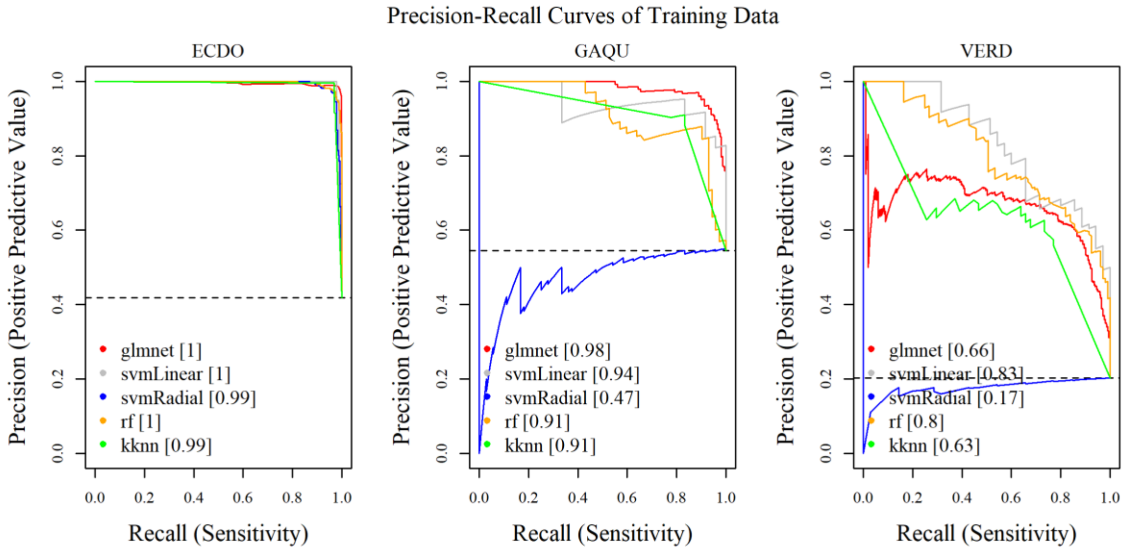

ROC Curves of Test Data
ECDO
GAQU
VERD
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
glmnet [1]
svmLinear [0.99]
svmRadial [0.99]
rf [1]
kknn [0.99]
wt.by.F1 [1]
glmnet [0.96]
svmLinear [0.98]
svmRadial [0.5]
rf [0.92]
kknn [0.95]
wt.by.F1 [0.95]
glmnet [0.96]
svmLinear [0.97]
svmRadial [0.5]
rf [0.98]
kknn [0.94]
wt.by.F1 [0.99]

Figure 3. 6. Precision-Recall Curves describing classifier performance during the training and testing phases. The upper panel shows PR curves on the 10-fold cross-validated training data for the five original classifiers. The bottom panel shows PR curves on the test data. The ensemble classifiers only make predictions in the test phase, so the bottom panel also demonstrates performance of the ensemble classifier weighted by F1 score. Area under the curve (AUC) is denoted next to each model's name in square brackets. Curves that reach into the upper right corner, with AUC values close to 1, show the best classification performance.


Precision-Recall Curves of Training Data
ECDO
GAQU
VERD
Precision (Positive Predictive Value)
Recall (Sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
glmnet [1]
svmLinear [1]
svmRadial [0.99]
rf [1]
kknn [0.99]
glmnet [0.98]
svmLinear [0.94]
svmRadial [0.47]
rf [0.91]
kknn [0.91]
glmnet [0.66]
svmLinear [0.83]
svmRadial [0.17]
rf [0.8]
kknn [0.63]

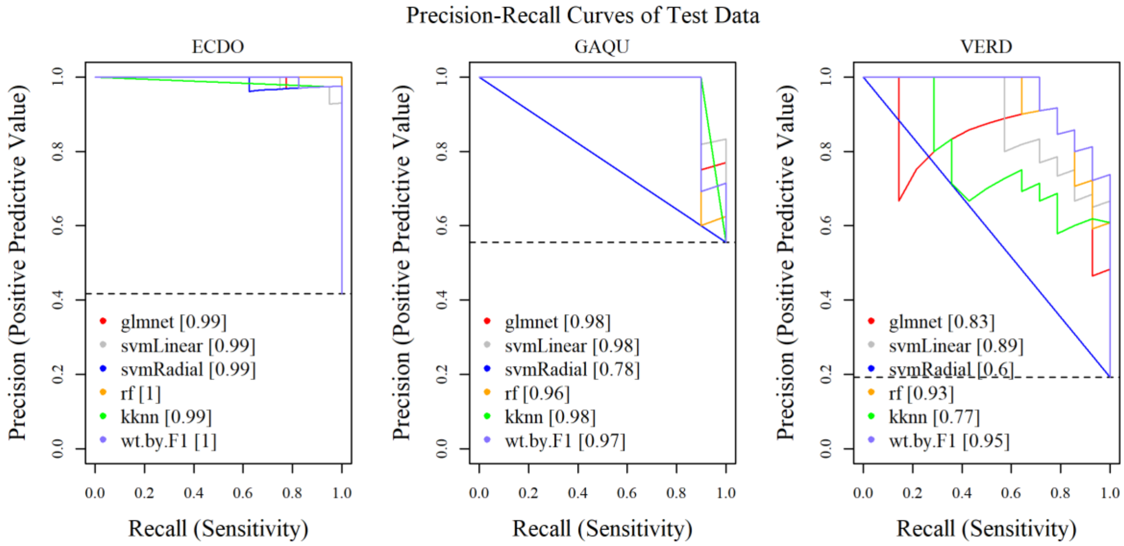
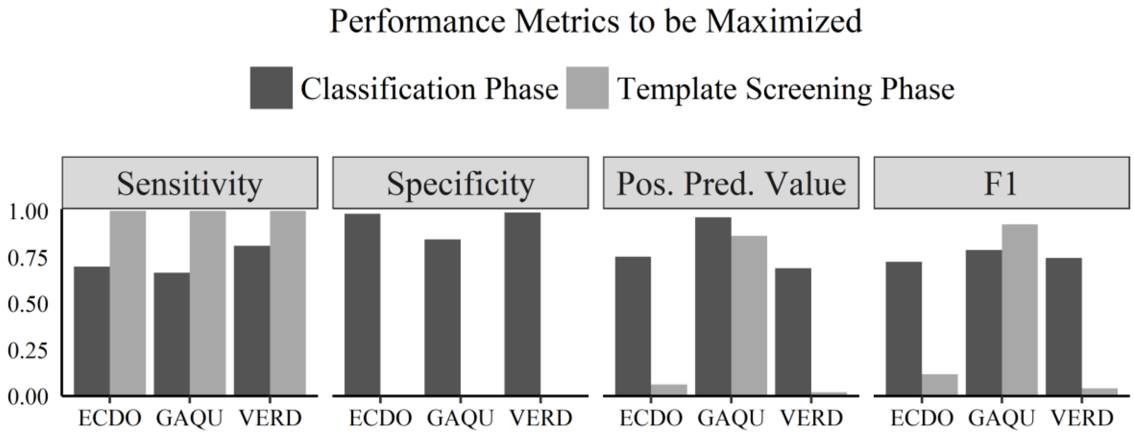

Precision-Recall Curves of Test Data
ECDO
GAQU
VERD
Precision (Positive Predictive Value)
Recall (Sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
glmnet [0.99]
svmLinear [0.99]
svmRadial [0.99]
rf [1]
kknn [0.99]
wt.by.F1 [1]
glmnet [0.98]
svmLinear [0.98]
svmRadial [0.78]
rf [0.96]
kknn [0.98]
wt.by.F1 [0.97]
glmnet [0.83]
svmLinear [0.89]
svmRadial [0.6]
rf [0.93]
kknn [0.77]
wt.by.F1 [0.95]

Figure 3. 7. Comparison of performance metrics for the classification and template screening phases. Scores closest to 1 are desired for all metrics.



# CHAPTER 4: DYNAMIC WILDLIFE OCCUPANCY MODELS USING AUTOMATED ACOUSTIC MONITORING DATA

Cathleen Balantic[1*]

Therese Donovan[2]

[1] Corresponding Author: cathleen.balantic@uvm.edu; Vermont Cooperative Fish and Wildlife Research Unit, 302 Aiken Center, 81 Carrigan Drive, University of Vermont, Burlington, VT 05405, USA

[2] U.S. Geological Survey, Vermont Cooperative Fish and Wildlife Research Unit, Rubenstein School of Environment and Natural Resources, University of Vermont, Burlington, VT 05405, USA

*Prepublication Disclaimer:*

*This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy*

## 4.1. Abstract

Automated acoustic monitoring of wildlife has been used to characterize populations of sound-producing species across large spatial scales. However, false negatives and false positives produced by automated detection systems can compromise the utility of these data for researchers and land managers, particularly for research programs endeavoring to describe colonization and extinction dynamics that inform land use decision-making. To investigate the suitability of automated acoustic monitoring for dynamic occurrence models, we simulated underlying occurrence dynamics, calling patterns, and the automated acoustic detection process for a hypothetical species under a range of scenarios. We investigated an automated species detection aggregation method that considered a suite of options for creating encounter histories. From these encounter histories, we generated parameter estimates and computed bias for occurrence, colonization, and extinction rates using a dynamic occupancy modeling framework that accounts for false positives via small amounts of manual confirmation. We were able to achieve relatively unbiased estimates for all three state parameters under all scenarios, even when the automated detection system was simulated to be poor, given particular encounter history aggregation choices. However, some encounter history aggregation choices resulted in unreliable estimates; we provide caveats for avoiding these scenarios. Given specific choices during the detection aggregation process, automated acoustic monitoring data may provide an effective means for tracking species occurrence, colonization, and extinction patterns through time, with the potential to inform adaptive management at multiple spatial scales.

Key Words

Automated acoustic monitoring; occupancy; dynamics; false positives; encounter history; parameter bias

## 4.2. Introduction

Remote automated acoustic monitoring of sound-producing wildlife provides a means for characterizing status and trends in species occurrence across landscapes (Cerquiera & Aide 2016). In a typical remote acoustic monitoring program, autonomous recording units (ARUs) are installed at study locations to capture recordings of the environment over time, based on a schedule input to the device by the research team. Vast quantities of audio data may be collected in a short amount of time, from which sounds produced by target monitoring species may be detected. Remote acoustic monitoring offers the potential to efficiently gather occurrence data for sound-producing wildlife species across regional spatial scales (Furnas & Callas 2005). Long-term, large-scale acoustic monitoring programs may be positioned to engage in systematic adaptive management research, wherein iterative learning reduces uncertainty over time to improve management decisions amid climate change and rapidly changing land uses (Williams et al. 2009).

Characterization of occurrence, or occupancy, requires only species presence-absence data – or more precisely, detection-nondetection data, since the probability of detecting a truly present species (*p*) is often less than 1, and false negatives transpire when a species is present but not detected (MacKenzie et al. 2002). To characterize false negatives, a site must be surveyed more than once. The fundamental building block of an occupancy model is thus the encounter history, a binary string indicating whether a species was detected or not detected on each survey occasion. Any combination of zeroes and ones is possible; an encounter history of 001, for example, indicates that a site was surveyed three times, and the species was only detected on the third occasion. The original single-season occupancy model has been expanded upon in several crucial ways, including dynamic (multiple season) models (MacKenzie et al. 2003), models that account for both false negatives and false positives (Royle & Link 2006, Miller et al. 2011, Miller et al. 2013, Chambert et al. 2015), and numerous other advances (Bailey et al. 2014). The dynamic occupancy model is particularly suitable for research seeking to understand trends over time. In addition to initial occupancy status (ψ), this model characterizes local extinction (ε) and colonization (γ) patterns between survey seasons, as well as covariates that influence the initial state and state changes.

Although acoustic monitoring is suited to capturing dynamic occupancy data for sound-producing species, the ease of acoustic data acquisition can overwhelm research programs with massive audio data streams. Accordingly, audio recordings may be rapidly processed using computer algorithms for automatically detecting species by their calls. For example, we have used the R package *AMMonitor* to create customized call templates for target species combined with statistical learning classifiers for this purpose (Katz et al. 2016, Hafner & Katz 2018, Balantic et al. in prep*a*). Numerous other software solutions exist for automated detection, such as Wildlife Acoustics Kaleidescope (Wildlife Acoustics 2018), Raven Pro (Bioacoustics Research Program 2015), the Arbimon platform (Aide et al. 2013), MatlabHTK (Ranjard et al. 2016), Tadarida (Bas et al. 2017), and Animal Sound Identifier (ASI) (Ovaskainen et al. 2018), though none provide means for aggregating detections into encounter histories.

Regardless of the software or detection method used, computer algorithms for automated detection may not detect a species when it is present (“false negative”) or may incorrectly detect an absent species (“false positive”). We take care to distinguish between detection mistakes occurring at the “event level” and those occurring at the “survey level.” Event-level mistakes are perpetrated by the automated detection method; these occur when the algorithm flags a detection not actually from the target species (“event-level false positive”) (**Fig. 4. 1-a-i**), or when the algorithm fails to detect an existing signal from the target species (“event-level false negative” (**Fig. 4. 1-a-iii**). Event-level detection mistakes represent a ubiquitous and well-documented challenge in automated acoustic monitoring research problems (Acevedo et al. 2009, Katz et al. 2016a, Brauer et al. 2016, Stowell et al. 2016, Shonfield & Bayne 2017). Survey-level mistakes, on the other hand, originate as a consequence of aggregating event-level mistakes into an encounter history for occupancy analysis. Ambiguity around how to combine event-level detections to create survey-level encounter histories is an area of emergent interest in automated acoustic monitoring, and is made especially vexing by acoustic monitoring’s capacity to generate high numbers of surveys compared with traditional field methods. Robust detection aggregation methodology is paramount for dynamic models, where the consequences of survey-level detection mistakes are amplified: when false positive detection errors are ignored, estimated extinction and colonization rates can become so biased and imprecise as to render results useless (McClintock et al. 2010, Miller et al. 2015, Ruiz-Gutierrez et al. 2016).

To address concerns about survey-level false negatives and false positives, Miller et al. (2013) introduced the “multiple detection states” dynamic occupancy model (hereafter, the Miller model). Survey-level detection states in the Miller model are categorized as “certain” or “uncertain,” making this framework amenable to cases where humans can later verify a subset of automated detections. In this work, within automated acoustic monitoring contexts, we presume that *all* audio recordings are scanned for the target species using an automated acoustic detection system. Event-level detections are then aggregated into survey outcomes, which compose the encounter history (**Fig. 4. 1**). Any survey-level detections that result via automation only (state 1) are denoted by a ‘1’ to indicate an uncertain detection (**Fig. 4. 1-a**). A subset of surveys, however, is allocated for *a posteriori* verification, wherein we assume no false positives exist. Surveys with automated event-level detections corroborated by manual verification (state 2) are given a ‘2’ to represent a certain detection at the survey level (**Fig. 4. 1-b**). A survey with no detections is assigned a ‘0’, indicating uncertain absence (**Fig. 4. 1-c**).

To illustrate, the history 120 000 suggests two demographic seasons, each surveyed three times. In the first season – assumed closed to demographic change across surveys – the species was detected in the first survey via automation only (producing an uncertain detection), detected with certainty in the second survey via automation with manual verification, and undetected in the third survey. In the second season, also assumed closed to demographic change, the species was not detected in any of the three surveys, all of which connote uncertain absences. The parameters estimated by the Miller model are the initial probability of occupancy ($\psi$), the state transition probabilities for colonization ($\gamma$) and extinction ($\varepsilon$), and a family of detection probability parameters. For uncertain detections, which are acquired via the automated acoustic detection system, $p_{10}$ represents the probability of incorrectly detecting the species at an unoccupied site (survey-level false positive), while $p_{11}$ is the probability of correctly detecting a present species at an occupied site (survey-level true positive). Certain detections are denoted by the parameter *b*, the probability of detecting a species via automated detection paired with manual verification, given presence. The probability of each observed encounter history can be computed for each site given the model parameters, and the product of those probabilities across all sites can be used to estimate parameters with maximum likelihood analysis (Miller et al. 2013).

Although the Miller model may have high utility for acoustic monitoring programs endeavoring to describe local extinction and colonization dynamics, two chief challenges remain. First, minimal guidance exists for aggregating large quantities of event-level automated acoustic detections into survey-level encounter histories (though see Chambert et al. 2017; Newson et al. 2017), and we are unaware of any previous efforts to explicitly exploit existing properties of automated detection algorithms for this purpose. Secondly, it is unclear how encounter history aggregation decisions affect the reliability of the Miller model for producing the unbiased, precise state parameter estimates ($\psi$, $\gamma$, and $\varepsilon$) necessary to make informed monitoring and management decisions. Without a comprehensive framework for moving from audio data collection to mistake-sensitive dynamic occupancy analysis, acoustic monitoring programs will be hobbled in their capacity to effectively leverage the opportunities offered by long-term, large-scale monitoring research, yielding compromised inference about population trends and limited model utility for subsequent adaptive management decisions.

Objectives:

The goal of this paper was to explore methodology for using acoustic data in dynamic occupancy models where observed data may include both false positives and false negatives. Our objectives were to: 1) Simulate latent occurrence dynamics and calling production, as well as the automated acoustic detection process for a hypothetical target species across *N* sites in a hypothetical monitoring area, 2) Introduce an event-level detection aggregation method that leverages properties of automated acoustic monitoring to create encounter histories under a suite of detection aggregation time frames, detection thresholds, and confirmation capacities, and 3) Generate parameter estimates and compute the bias for occurrence, colonization, and extinction rates using the Miller Model.

## 4.3. Materials and Methods

### Simulation of occurrence dynamics, species calling production, and detection process

We simulated two 30-day seasons of underlying occupancy dynamics and species sound production for a hypothetical target species across 100 sites. For simplicity, we did not use site or survey-level covariates to model any processes. To create dynamic occupancy scenarios, we used the four simulation cases outlined by Miller et al. 2015 (**Fig. 4. 2a**): high initial occurrence with high turnover (HH) ($\psi = 0.60$, $\gamma = \varepsilon = 0.25$), high initial occurrence with low turnover (HL) ($\psi$ = 0.60, $\gamma = \varepsilon$ = 0.05), low initial occurrence with high turnover (LH) ($\psi$ = 0.20, $\gamma = \varepsilon$ = 0.25), and low initial occurrence with low turnover (LL) ($\psi$ = 0.20, $\gamma = \varepsilon$ = 0.05).

For each dynamic occupancy scenario, we simulated the underlying sound production process wherever the species was present (**Fig. 4. 2b**). Individual calling rates vary widely based on species, breeding stage, and environmental conditions (Catchpole & Slater 2008), with overall abundance driving the total number of target signals available in the soundscape (Royle & Nichols 2003). We condensed these elements into an average species calling rate per hour, $\lambda_c$, and investigated two cases: (1) a low call production scenario that averaged 20 calls per hour (or 0.33 calls per minute) ($\lambda_c$ = 20), and (2) a high call production scenario that averaged 100 calls per hour (or 1.67 calls per minute) ($\lambda_c$ = 100). For each sampled minute of the season, we used a Poisson process to generate the true number of calls produced by the species.

Next, we simulated the automated acoustic detection process (**Fig. 4. 2c**) where we addressed three components: (1) the timing and frequency of audio recordings, (2) the existence of “false alarm” sources within the recording soundscape, and (3) the general aptitude of the automated detection method. First, we chose a recording scheme of five minutes of audio sampling per day, presumed to occur during ideal windows for capturing target species call production. Second, we used a Poisson process to inject false alarms ($\lambda_f$) into each minute of audio recording. The false alarm rate, $\lambda_f$, acts as an analog to the call production rate ($\lambda_c$), in that it connotes underlying sources of false alarms present in the soundscape, which fool an automated detector into generating event-level false positives. We selected a rate of $\lambda_f = 48$ false alarms per hour (0.8 false alarms per minute), based on the false alarm rate we computed from a field study using automated detections across 675 hours of real field recordings (Balantic et al. in prep*a*).

Finally, we simulated the production of event-level detections in each recording within an automated acoustic monitoring framework. Automated detection algorithms, also known as classifiers, may be constructed to produce the probability that an event-level detection is truly a signal from the target species (Balantic et al. in prep*a*). For example, in **Fig. 4. 3**, to each event-level detection, a trained statistical learning classifier has assigned a probability that the event is a signal from the target species. Hereafter, we refer to this attribute as the "target signal probability" of any event-level detection. We simulated two alternative classifiers ("good" and "bad"), each defined by a mixture of two beta distributions. The good classifier was likely to assign high target signal probabilities to true target signals (which are produced by $\lambda_c$) and low target signal probabilities to false alarms (which are produced by $\lambda_f$) ($\alpha = 4$, $\beta = 1$ for target signals; $\alpha = 1$, $\beta = 4$ for false alarms). The bad classifier was represented by a beta distribution with $\alpha = \beta = 3$ for both target signals and false alarms, yielding average target signal probabilities of 0.5 across all detections (**Fig. 4. 2c**). **Table 4. 1** provides an example of event-level detections with target signal probabilities assigned by good and bad classifiers.

In summary, the simulated acoustic environment consisted of four underlying species occurrence dynamics cases, each with two levels of calling production. All eight of these scenarios had the same underlying false alarm rate. Finally, we simulated the automated detection process with two types of classifiers, good and bad, which output the target signal probability associated with each event-level detection. The Objective 1 simulations thus produced 16 scenarios for subsequent evaluation.

## Encounter History Aggregation

To analyze the 16 scenarios produced by Objective 1 under the Miller model occupancy framework, we collapsed event-level detections into encounter histories (**Fig. 4. 2d**). Using a capture-recapture framework (*sensu* Otis et al. 1978), target signal probabilities associated with event-level detections were aggregated to produce the overall probability that at least one target signal had been detected within a particular survey period, which yields the survey-level detection. We take care not to conflate the occupancy term “survey” with an individual audio recording: multiple audio recordings might be amassed to collectively constitute the survey based on a chosen unit of survey closure, which is informed by research goals and life history of the target species. To demonstrate, imagine an occupancy survey closure period defined as all of the audio recordings taken in a single day (in our simulation, five minutes of recordings per day are combined into a single survey). Suppose that on the first day (survey 1) three events are detected by the classifier, with target signal probabilities of 0.15, 0.04, and 0.11 (**Fig. 4.3a**). In this case, we aggregate the probabilities as (1-0.15)*(1-0.04)*(1-0.11), which gives the probability that *all* detected events are false alarms, 0.73. Next, 1-0.73 gives 0.27, the probability that *any* of these detected events is truly a signal from the target. Applying a user-chosen threshold of 0.95, we log a ‘0’ in the encounter history for this survey and presume that the species was not detected unless we are 95% sure that we captured at least one target signal. In the next day's survey (**Fig. 4. 3b**), three events are detected by the automated algorithm, with target signal probabilities of 0.56, 0.88, and 0.71. The probability that all detected events are false alarms is (1-0.56)*(1-0.88)*(1-0.71), resulting in a probability of ~0.015 that all events are false alarms, and probability of ~0.98 that at least one true target signal has been captured by the automated system. Applying the same 0.95 threshold, we log a '1' for this survey to reflect that the target species has been detected in the uncertain state (that is to say, it has been detected automatically and without manual verification). Taken together, the two surveys in this example return an encounter history of 01.

To generate encounter histories for each of the 16 scenarios from Objective 1, we examined eight alternative scenarios defined by three factors: (1) the survey-level detection threshold, (2) the aggregation period, and (3) the percentage of manually confirmed surveys, which comprise state 2 of the Miller model (**Fig. 4. 2d**). First, we investigated two survey-level detection thresholds: 0.8 and 0.95 (i.e., we were 80% or 95% certain that at least one target signal was detected during the survey period). Second, we examined two aggregation options, in which recordings were lumped into a single survey based on a desired unit of closure. We used survey aggregation periods of either one day or three days across each 30-day monitoring period. Thus, over a 30-day season, a single season encounter history for the 1-day aggregation period would yield a string of 30 surveys, represented by a 0, 1, or 2 (with a total of 60 surveys over two seasons). A single season encounter history for the 3-day aggregation period would produce a string of 30-day season / 3-day aggregation period = 10 surveys (20 surveys total over two seasons). Lastly, we examined two scenarios for the total proportion of surveys that were confirmed *a posteriori* to serve as the “certain” state (state 2) of the Miller model: 0.025 or 0.05. In other words, we randomly assigned 2.5% or 5% of surveys to be manually confirmed to produce a certain state. For practical purposes, with 100 sites across two 30-day seasons, at a rate of 5 minutes of recording per site per day, this would equate to manual verification of event-level detections from 12.5 hours (2.5%) or 25 hours (5%) worth of recordings, regardless of the *N*-day survey aggregation period used.

In summary, the 16 acoustic scenarios generated from Objective 1 were each subjected to 8 alternative scenarios for developing encounter histories, resulting in 16 * 8 = 128 total scenarios to be analyzed with the Miller model in Objective 3. To summarize the outcome of the simulation, we calculated survey-level true and false positive rates based on the survey window aggregation length, classifier type, survey-level detection threshold, and species calling rate. The survey-level true positive rate indicated the proportion of surveys in the encounter history where the species was present and detected (with rates closest to 1 most desirable). The survey-level false positive rate signified the proportion of sites where the species was absent but mistakenly detected (with rates closest to 0 most desirable) (note that this *survey*-level false positive rate should not be confused with the *event*-level false alarm rate).

### Generate state parameter estimates and compute bias under different scenarios

The 128 scenarios were simulated 500 times each (64,000 replicates total). To generate parameter estimates for occupancy, colonization, and extinction, we fit intercept models for each of the 64,000 replicates using RPresence V.12.10 (**Fig. 4. 2e**) (Hines 2018). All simulation models were fit using informed initial parameter values to aid convergence to a global minimum in the negative log-likelihood, because preliminary testing showed that results were sensitive to starting values. We compared these state parameter estimates to the simulated dynamics values to compute raw bias, as well as the mean and standard deviation across each scenario's 500 replicates (**Fig. 4. 2f**). Although we focused on the outcomes of the state parameter estimates, we also computed estimate bias for the detection parameters, $p_{11}$, $p_{10}$, and $b$. We recorded model convergence rates for all scenarios.

## 4.4. Results

### Occurrence Dynamics and Sound Simulation

For the four occurrence-turnover states and two sound production rates, we summarized daily available sound production averaged across occupied and unoccupied sites in **Fig. 4. 4**. The total number of species target signals is contingent on occupancy status, given five minutes of recording daily – low occurrence rates naturally produce a lower number of target signals available for automated capture across sites. Meanwhile, the average number of available false alarms is the same regardless of occupancy status and species sound production rate. The good classifier assigned an average target signal probability of 0.80 (+/- 0.002 sd) for target signals, and 0.20 (+/- 0.001 sd) for false alarms. The bad classifier assigned an average target signal probability of 0.50 (+/- 0.003 sd) for target signals, and 0.50 (+/- 0.001 sd) for false alarms.

## Encounter History Aggregation

To create encounter histories from the signals produced in Objective 1, recall that we investigated eight alternative scenarios (factors = aggregation day length: 1 vs. 3; survey-level detection threshold:  0.8 vs 0.95; and human-verified confirmation level: 2.5% vs. 5%). Each survey within an encounter history was assigned either a 0 (uncertain absence), 1 (uncertain detection produced by the automated system) or 2 (certain detection produced by the automated system with manual confirmation) to denote species detection/non-detection status. Survey-level true and false positive rates produced by the automated method differed based on aggregation length, survey-level detection threshold, and classifier (**Fig. 4. 5**). Encounter histories using 1-day aggregation produced 30 surveys in total, because the 30-day monitoring duration was split into survey periods lasting one day. Meanwhile, encounter histories using 3-day aggregation had 10 total surveys, because the 30-day monitoring duration was split into survey periods lasting three days. Overall, the confirmation levels we chose did not yield any appreciable difference in true and false positive rates.  However, rates varied substantially depending on species calling rate, aggregation level, detection threshold, and classifier performance.

For 1-day aggregation (**Fig. 4. 5a**), the 0.95 survey-level detection threshold produced encounter histories with lower underlying true positive rates than those created by the 0.80 threshold, particularly when a good classifier is used against a low call rate

(**Fig. 4. 5a**; upper left panel). It is logical to expect inferior true positive rates for the 0.95 threshold and the good classifier if there are few target signals within a survey period – these conditions foster a higher overall standard that surveys must meet before meriting a score of '1'. As a result of this high standard, for both species calling rates, false positive rates for 1-day aggregation approached zero when using the 0.95 threshold and good classifier; the rate rose to 0.17 when using the 0.80 threshold and good classifier (**Fig. 4.5a**; upper right panel). The differences caused by higher and lower-standard detection systems illustrate the tradeoff inherent in striving for a high site-level true positive rate while keeping false positives to a minimum. Overall, the bad classifier generated much higher false positive rates, ranging from 0.46 (0.95 threshold) to 0.75 (0.80 threshold) (**Fig. 4. 5a**; lower right panel).

The tradeoff between a high true positive rate and a low false positive rate is magnified in the 3-day aggregation scenarios (**Fig. 4. 5b**). Both survey-level detection thresholds and both classifiers generated encounter histories with true positive rates of 1 or nearly so. For false positive rates, however, the good classifier had false positive rates as low as 0.44 using the stricter detection threshold (0.95) and false positive rates as high as 0.85 using the lenient threshold (0.8) (**Fig. 4. 5b**; upper right panel). For the bad classifier, false positive rates were near 1 for all encounter history scenarios (**Fig. 4. 5b**; lower right panel). In summary, although 3-day aggregation generated encounter history scenarios with higher underlying survey-level true positive rates overall, these encounter history scenarios also carried higher underlying false positive rates. Meanwhile, 1-day aggregation produced encounter history scenarios with lower overall true positive rates, but also much lower false positive rates.

## Bias of state parameter estimates under different scenarios

Summarized across all dynamics and calling scenarios, encounter histories generated with the 1-day survey aggregation period generally produced the least biased estimates across the three state parameters, with bias values closest to zero being most desirable (**Fig. 4. 6**). The superiority of the smaller aggregation period held true across both survey-level detection thresholds, both the good and bad classifiers, and both proportions of *a posteriori* survey confirmation. Under 1-day aggregation, neither confirmation level nor survey-level detection threshold made an appreciable difference in the bias estimates (**Fig. 4. 6a,c,e,g**). Under 3-day aggregation, the higher confirmation level (5%) reduced both the bias and variation in bias (compare **Fig. 4. 6b** to **4. 6d**, and **Fig. 4. 6f** to **Fig. 4. 6h**). From the big picture view, when a good classifier was used, bias was comparable across all **Fig. 4. 6** scenario panels except for 3-day aggregation, 2.5% confirmation, and 0.8 survey-level detection threshold (**Fig. 4. 6f**), where even the good classifier's estimates tended to be more erratic and biased. Thus, although 1-day aggregation outperformed 3-day aggregation overall, 3-day aggregation is competitive with increasing survey-level detection thresholds and/or confirmation levels, especially if the classifier is good, demonstrating that longer aggregation windows can retain utility if adequately balanced by higher automated detection standards and higher manual confirmation effort.

Under most scenarios, state parameter estimates tended to have wide ranges in variation. Zooming in on the most conservative encounter history aggregation conditions (1-day aggregation, 0.95 survey-level detection threshold, **Fig. 4. 7**), mean estimates fell within 3% of simulated truth, though with deviation out to 10% in both directions for some parameters. The lower confirmation level (**Fig. 4. 7a**) was generally competitive with the higher confirmation level (**Fig. 4. 7b**). The high occurrence-low turnover (HL) scenario had the least biased and most precise estimates across all three parameters under all scenarios and both confirmation levels, and the low occurrence-low turnover (LL) scenario would have approached this level of precision if not for the tendency to overestimate the extinction parameter ($\varepsilon$). The higher turnover scenarios (HH, LH) generally produced more variation in the bias. The influence of species calling rate was minimal overall. Estimates for the detection parameters $p_{11}$, $p_{10}$, and $b$ were generally much less biased and more precise than the state parameter estimates, with mean biases and standard deviations all falling well within 3% of simulated truth (**Fig. 4. 8**).

The Miller model convergence rate across all scenarios was 59%. Only replicates that converged were included in the **Fig. 4. 6** and **Fig. 4. 7** results. Convergence rates generally mirrored the bias results. The number of aggregation days had the clearest impact on convergence: 1-day aggregation had an 84% convergence rate, while 3-day aggregation only converged 34% of the time. Classifier type also affected convergence, with the good classifier (68% convergence rate) outperforming the bad classifier (51% convergence rate). Survey detection level also affected convergence rates: models that used a 0.95 threshold converged 66% of the time, whereas models with a 0.80 threshold converged at a rate of 53%. The impact of confirmation level was minimal (2.5% confirmation: 57% converged, 5% confirmation: 61% converged). Convergence rates also differed based on the underlying state dynamics, with the high turnover scenarios (HH, LH) converging at greater rates overall (HH: 67%, HL: 57%, LH: 64%, and LL: 50%). The high calling rate (55%) converged substantially less frequently than the low calling rate (64%).

## 4.5. Discussion

The capacity to understand and predict long-term, large-scale species occurrence dynamics is critical against a backdrop of climate change and rapidly shifting land uses (Nichols et al. 2015). Although automated acoustic monitoring provides a means for cost-effective and efficient collection of species occurrence data, minimal guidance exists for translating enormous streams of raw audio data into dynamic occurrence models that provide actual utility for wildlife researchers and land managers. We introduced a novel method for aggregating detected events into encounter histories for use in dynamic models meant to capture changes in occurrence patterns over time. When automated detection algorithms are constructed to yield the probability that a detected event is produced by a target signal, these event-level probabilities may be aggregated within a capture-recapture framework to provide the probability that *any* detected event within a survey period is a sound from the target species (Otis et al. 1978). We believe this work is the first to unify the concepts of automated acoustic data collection with analysis for mistake-sensitive dynamic occupancy modeling, although single-season approaches have been implemented (Cerquiera & Aide 2016, Chambert et al. 2017). Where classifier-assigned target signal probabilities are not available, Chambert et al. (2017) offer a method contingent on the total abundance of event-level detections. Other alternatives include automated target event detection followed by manual removal of false positives (Cerquiera & Aide 2016), or deploying machine learning approaches to identify and remove false positives automatically (Balantic et al. in prep*a*). In contrast to removing false positives manually or automatically, the method we describe here takes full advantage of the information provided by target signal probabilities associated with each detected event.

To explore the utility of the Miller model framework for dynamic occupancy models constructed from automated acoustic monitoring data, we investigated our probability aggregation method in 128 scenarios that spanned a range of occurrence dynamics, species sound production rates, classifier performances, and encounter history aggregation. Our results demonstrate that the Miller model was able to produce state parameter estimates within an average of 3% of simulated truth estimates for occurrence, colonization, and extinction for all latent conditions (dynamics and sound production) and all observation conditions (good vs. bad classifier), given specific encounter history aggregation choices. We also applied our probability aggregation method to the false positive-ignorant dynamic model (Mackenzie et al. 2003) but found that it was generally not competitive with the Miller model. In a 100-repetition test, the model of Mackenzie et al. 2003 tended to overestimate initial occurrence and underestimate extinction rates (**Appendix C**), suggesting that the Miller model is the stronger choice for automated acoustic monitoring.

In our simulation, narrow frames of survey aggregation produced the least biased parameter estimates. These shorter survey aggregation periods, in turn, produce a larger number of surveys. Single-day aggregation outperformed 3-day aggregation because longer aggregation periods are more likely to result in “uncertain” survey-level false positives in the encounter history, particularly if the survey-level detection threshold is not high enough. Longer aggregation periods lead to a greater number of detected events. If a species is absent from a site, even if a good classifier is used, the target signal probabilities assigned to false alarms may not be low enough to overcome the effects of many probabilities ultimately being multiplied together. The (multiplied) product of too many probabilities may be an unacceptably high number of survey-level false positives within the probability aggregation scheme (as in **Fig. 4. 5b**). Although the 3-day aggregation period slightly outperforms 1-day aggregation on the survey-level true positive rate, the accompanying bloated false positive rates are too high to overcome without bias when fitting the dynamic occupancy model. Thus, smaller windows of probability aggregation may perform better in general, though the narrowness of this aggregation window must be balanced against practical considerations for the period of survey closure for a target species.

These general results bode well for an automated acoustic monitoring program. Automated acoustic monitoring – like camera trapping and other remote automated methods – boasts a unique position in the occupancy modeling realm, where many traditional study methods tend to be “survey poor” (Mackenzie & Royle 2005). In contrast, automated acoustic monitoring provides the opportunity to be “survey rich” if audio recordings occur regularly over an extended time, a benefit of the flexibility around gathering audio recordings into distinct survey periods. Since false positives can inflate the number of recommended surveys (Clement 2016), the opportunity to be survey-rich is useful for a monitoring methodology where false positives are prevalent.

Unsurprisingly, a higher quality classifier will better serve an acoustic monitoring program than a poor classifier. In our experiment, the good classifier was typically able to provide minimally biased results even when coping with a long aggregation period (e.g. **Fig. 4. 6d**), or low survey-level detection thresholds provided that the aggregation period was short enough (**Fig. 4. 6e,g**), while the bad classifier was often less robust under these conditions. For the good classifier, as long as the aggregation period was short, the confirmation levels we examined made little difference. While we expect that no research team would intentionally deploy a bad classifier, the performance of an automated detection system during the testing phase can be markedly different from its performance on new audio data, which can introduce emergent challenges such as seasonal changes to the soundscape, turnover of individual animals that contributed training data to the automated detection system, and cultural drift of vocalization behavior over time (Williams et al. 2013). Researchers should take caution in the deployment of automated detection systems that have not been thoroughly field-tested (Russo & Voigt 2016). To moderate the impacts of these potential changes on classifier performance over time, a monitoring team may opt for higher confirmation levels, or might choose to place their automated detection and classification models in a Bayesian framework, updating them regularly at intervals appropriate for the target species and study landscape. Additionally, if target species issue multiple types of call or song signals, multiple detectors and classifiers can be used to scan audio files; our method can easily incorporate such cases.

Compared with short-term ecological monitoring studies, long-term studies have a disproportionately large impact on scientific knowledge and policy (Hughes et al. 2017), and research programs engaging in long-term, large-scale automated acoustic monitoring of wildlife have the potential to contribute to this type of knowledge. However, the utility of long-term, large-scale acoustic monitoring will be undercut without a means for moving from raw acoustic data to population models from which inference may be gained and land management decision-making supported. Generation of occupancy model encounter histories from large data streams is a salient challenge in automated acoustic monitoring. In this work, we conducted all simulations using the acoustic monitoring data management framework described in Balantic et al. (in prep*b*), which contains functions to support machine learning assignment of target signal probability values to automatically-detected events (Balantic et al. in prep*a*,*b*).

## 4.6. Acknowledgments

This project was supported by the U.S. Bureau of Land Management. Cathleen Balantic was supported by NSF IGERT grant 1144388. We thank Mark Massar (BLM) for providing the opportunity to work in the Riverside East Solar Energy Zone, Jonathan Katz for assistance with R programming for the *AMMonitor* package, and reviewers for their helpful comments. Use of trade names or products does not constitute endorsement by the U.S. Government. The Vermont Cooperative Fish and Wildlife Research Unit is jointly supported by the U.S. Geological Survey, University of Vermont, Vermont Department of Fish and Wildlife, and Wildlife Management Institute.

## 4.7. References

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J. & Aide, T.M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics,* 4, 206–214. DOI: 10.1016/j.ecoinf.2009.06.005

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. PeerJ 1:e103. DOI: https://doi.org/10.7717/peerj.103

Bailey, L.L., MacKenzie, D.I. & Nichols, J.D. (2014) Advances and applications of occupancy models (E. Cooch, Ed.). *Methods in Ecology and Evolution*, **5**, 1269–1279. DOI:  10.1111/2041-210X.12100

Balantic, C.M. & Donovan, T.M. In Prep*a.* (Unpublished results) Statistical learning mitigation of false positive detections in automated acoustic wildlife monitoring.

Balantic, C.M., Katz, J., & Donovan, T.M. In Prep*b*. (Unpublished results) AMMonitor R Package.

Bas, Y., Bas, D. & Julien, J.-F., (2017). Tadarida: A Toolbox for Animal Detection on Acoustic Recordings. Journal of Open Research Software. 5(1), p.6. DOI: http://doi.org/10.5334/jors.154

Bioacoustics Research Program. (2015). Raven Pro 1.5: Interactive Sound Analysis Software [Computer Software]. URL http://www.birds.cornell.edu/raven.

Brauer, C., T. Donovan, R. Mickey, J. Katz, & Mitchell, B. (2016). A comparison of acoustic monitoring methods for common anurans of the northeastern United States. *Wildlife Society Bulletin*, **40**,140-149. DOI: 10.1002/wsb.619

Catchpole, C.K. & Slater, P.J.B. (2008). Bird Song: Biological Themes and Variations, 2nd Ed. Cambridge University Press, Cambridge, UK.

Cerquiera, M. & Aide, T.M. (2016). Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling. *Methods in Ecology and Evolution*, **7**, 1340-1348. DOI: 10.1111/2041-210X.12599

Chambert, T., Miller, D.A.W., & Nichols, J.D. (2015). Modeling false positive detections in species occurrence data under different study designs. Ecology, 96(2), 332-339. DOI: https://doi.org/10.1890/14-1507.1

Chambert, T., Waddle, J.H., Miller, D.A.W., Walls, S.C., & Nichols, J.D. (2017). A new framework for analyzing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, **9**, 560-570. DOI: 10.1111/2041-210X.12910

Clement, M. (2016). Designing occupancy models when false positive detections occur. Methods in Ecology and Evolution, 7, 1538-1547. DOI: 10.1111/2041-210X.12617

Furnas, B. J., & Callas, R.L. (2015). Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. Journal of Wildlife Management 79:325-337. DOI: http://dx.doi.org/10.1002/jwmg.821

Hughes, B.B., Beas-Luna, R., Barner, A.K., Brewitt, K., Brumbaugh, D.R., Cerny-Chipman, E.B., Close, S.L., Coblentz, K.E., de Nesnera, K.L., Drobnitch, S.T., et al. (2017). Long-term studies contribute disproportionately to ecology and policy. *Bioscience,* **67**(3), 271-281. DOI: 10.1093/biosci/biw185

Hafner S. & Katz J. (2018). monitoR: Acoustic template detection in R. R package version 1.0.7, URL: http://www.uvm.edu/rsenr/vtcfwru/R/?Page=monitoR/monitoR.htm.

Hines, J. (2018). RPresence for PRESENCE: Software to estimate patch occupancy and related parameters. Version 12.10. https://www.mbr-pwrc.usgs.gov/software/presence.html

Katz, J., Hafner, S.D. & Donovan, T. (2016). Assessment of Error Rates in Acoustic Monitoring with the R package monitoR. *Bioacoustics*, **25**, 177–196. DOI: 10.1080/09524622.2015.1133320

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255. DOI: 10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, **84**, 2200–2207. DOI: 10.1890/02-3090MacKenzie, D. I., & Royle, J. A. (2005). Designing occupancy studies: general advice and allocating survey effort. *Journal of applied Ecology*, *42*(6), 1105-1114. https://doi.org/10.1111/j.1365-2664.2005.01098.x

McClintock, B.T., Bailey, L.L., Pollock, K.H. & Simons, T.R. (2010). Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, **91**, 2446–2454. DOI: https://doi.org/10.1890/09-1287.1

Miller, D.A.W., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology*, **92**, 1422–1428. DOI: 10.1890/10-1396.1

Miller, D.A.W., Nichols, J.D., Gude, J.A., Rich, L.N., Podruzny, K.M., Hines, J.E. & Mitchell, M.S. (2013). Determining Occurrence Dynamics when False Positives Occur: Estimating the Range Dynamics of Wolves from Public Survey Data. *PLoS ONE*, **8**, e65808. DOI: 10.1371/journal.pone.0065808

Miller, D.A.W., Bailey, L.L., Grant, E.H.C., McClintock, B.T., Weir, L.A. & Simons, T.R. (2015). Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known (O. Gimenez, Ed.). *Methods in Ecology and Evolution*, **6**, 557–565. DOI: 10.1111/2041-210X.12342

Newson, S.E., Bas, Y., Murray, A., & Gillings, S. (2017). Potential for coupling the monitoring of bush-crickets with established large-scale acoustic monitoring of bats. *Methods in Ecology and Evolution*, **8**, 1051-1062. DOI: 10.1111/2041-210X.12720

Nichols, J. D., Johnson, F. A., Williams, B. K., & Boomer, G. S. (2015). On formally integrating science and policy: walking the walk. *Journal of Applied Ecology*, *52*(3), 539-543. DOI: 10.1111/1365-2664.12406

Otis, D. L., Burnham, K.P., White, G.C., & Anderson, D.R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs,* **62**, 1-135.

Ovaskainen, O., Moliterno de Camargo, U., & Somervuo, P. (2018). Animal Sound Identifier (ASI): software for automated identification of vocal animals. *Ecology letters*, 21(8): 1244-1254. doi: https://doi-org.ezproxy.uvm.edu/10.1111/ele.13092

Ranjard, L., Reed, B.S., Landers, T.J., Raynar, M.J., Friesen, M.R., Sagar, R.L., Dunphy, B.J. (2017). MatlabHTK: a simple interface for bioacoustics analyses using hidden Markov models. Methods in Ecology and Evolution 8(5): 615-621. doi: 10.1111/2041-210X.12688

Royle, J.A. & Link, W.A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, **87**, 835–41. DOI: 10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2

Royle, J.A. & Nichols, J.D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology* **84**(3), 777-790. DOI: https://doi.org/10.1890/0012-9658(2003)084[0777:EAFRPA]2.0.CO;2

Ruiz-Gutierrez, V., Hooten, M.B. & Campbell Grant, E.H. (2016). Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias (N. Yoccoz, Ed.). *Methods in Ecology and Evolution*, 7, 900–909. DOI: 10.1111/2041-210X.12542

Russo, D., & Voigt, C.C. (2016). The use of automated identification of bat echolocation calls in acoustic monitoring: A cautionary note for a sound analysis. *Ecological Indicators*, **66,** 598-602. DOI: https://doi.org/10.1016/j.ecolind.2016.02.036

Shonfield, J. & Bayne, E.M. (2017). Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology*, **12**(1): 14. DOI: 10.5751/ACE-00974-120114

Stowell, D., Wood, M., Stylianou, Y. & Glotin, H. (2016). Bird detection in audio: a survey and a challenge. IEEE International Workshop on Machine Learning for Signal Processing, Salerno, Italy.

Wildlife Acoustics. (2018). Kaleidescope [Computer Software]. URL http://www. wildlifeacoustics.com.

Williams, B.K., Szaro, R.C., & Shapiro, C.D. (2009). Adaptive Management: The U.S. Department of the Interior Technical Guide, 2nd Edition. Adaptive Management Working Group, U.S. Department of the Interior, Washington DC.

Williams, H., Levin, I.I., Norris, D.R., Newman, A.E.M., & Wheelwright, N.T. (2013). Three decades of cultural evolution in Savannah sparrow songs. *Animal Behaviour*, **85** (1): 213 DOI: 10.1016/j.anbehav.2012.10.028

## 4.8. Tables

Table 4. 1. Illustration of the detection portion of the simulation. Event-level detections can be target signals or false alarms, generated according to $\lambda_c$ and $\lambda_f$ , respectively. For each event-level detection, a classifier assigns a probability that the detection is actually a target signal. The good classifier typically assigns higher target signal probabilities to actual target signals, and lower probabilities to false alarms. The bad classifier makes no such distinction. Both classifiers randomly sample probabilities from their respective beta distributions visualized in Fig. 4. 2c.

| | | Target Signal Probability | |
|---|---|---|---|
| Minute | Sound Type | Good Classifier | Bad Classifier |
| 1 | False Alarm | 0.31 | 0.72 |
| 1 | False Alarm | 0.27 | 0.67 |
| 2 | Target Signal | 0.80 | 0.52 |
| 2 | False Alarm | 0.04 | 0.29 |
| 2 | Target Signal | 0.75 | 0.46 |
| 3 | Target Signal | 0.60 | 0.71 |
| 4 | Target Signal | 0.87 | 0.61 |
| 4 | False Alarm | 0.07 | 0.36 |
| 5 | Target Signal | 0.88 | 0.35 |
| 5 | False Alarm | 0.27 | 0.09 |
| | **Mean False Alarm** | **0.19** | **0.43** |
| | **Mean Target Signal** | **0.78** | **0.53** |

## 4.9. Figures

Figure 4. 1. Construction of an encounter history for the Miller model is illustrated with three “surveys” represented by audio recording spectrograms. Event-level detections by a hypothetical automated method for a target species (red boxes) can be event-level false positives (a-i) or event-level true positives (a-ii). Event-level false negatives occur where the automated method misses a target signal (a-iii). Event-level detections from all recordings within a survey period are collected to produce a single value that describes survey-level detection status according to the Miller model (0, 1, or 2). In ‘a’, which used the automated method only, aggregation produces a ‘1’ in the encounter history to indicate an uncertain detection. Survey ‘b’ underwent *a posteriori* manual verification; all event-level detections within this survey are checked by hand. We assign this survey a ‘2’ to denote a certain detection at the survey level. Survey ‘c’ yielded zero event-level detections and is assigned a ‘0’ at the survey level. Together, surveys a, b, and c produce an encounter history of 120 for the season.

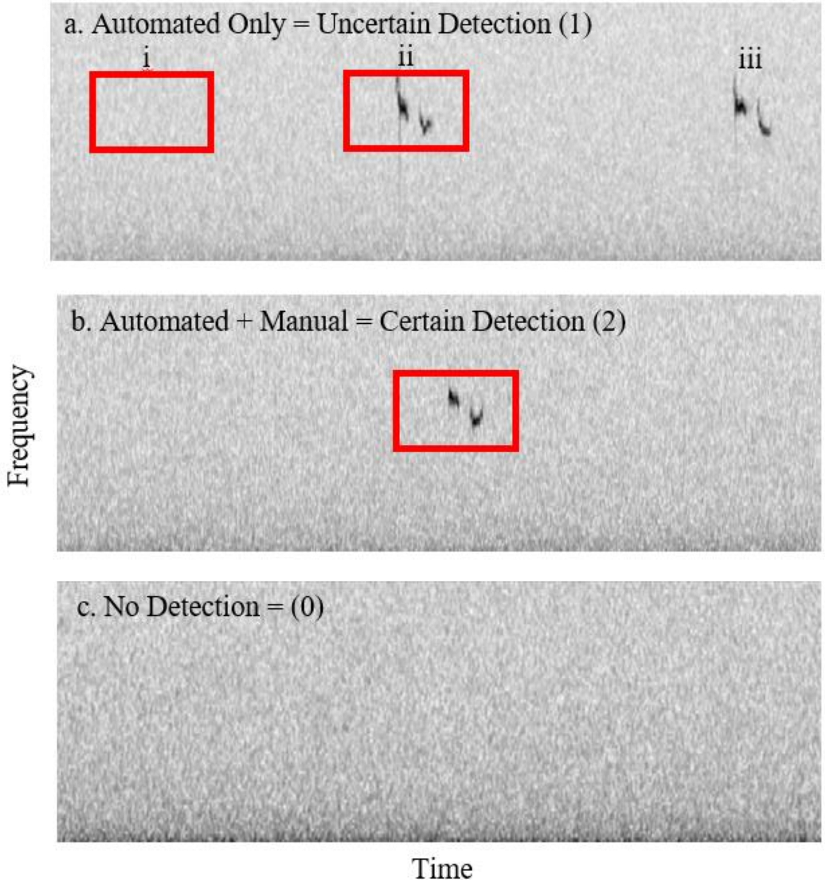

a. Automated Only = Uncertain Detection (1)
i
ii
iii
b. Automated + Manual = Certain Detection (2)
c. No Detection = (0)
Frequency
Time

Figure 4. 2. Simulation and parameter estimation workflow for simulating dynamics (a), simulating species calling production (b), simulating the automated detection process (c), building encounter histories (d), generating parameter estimates (e), and computing bias (f).

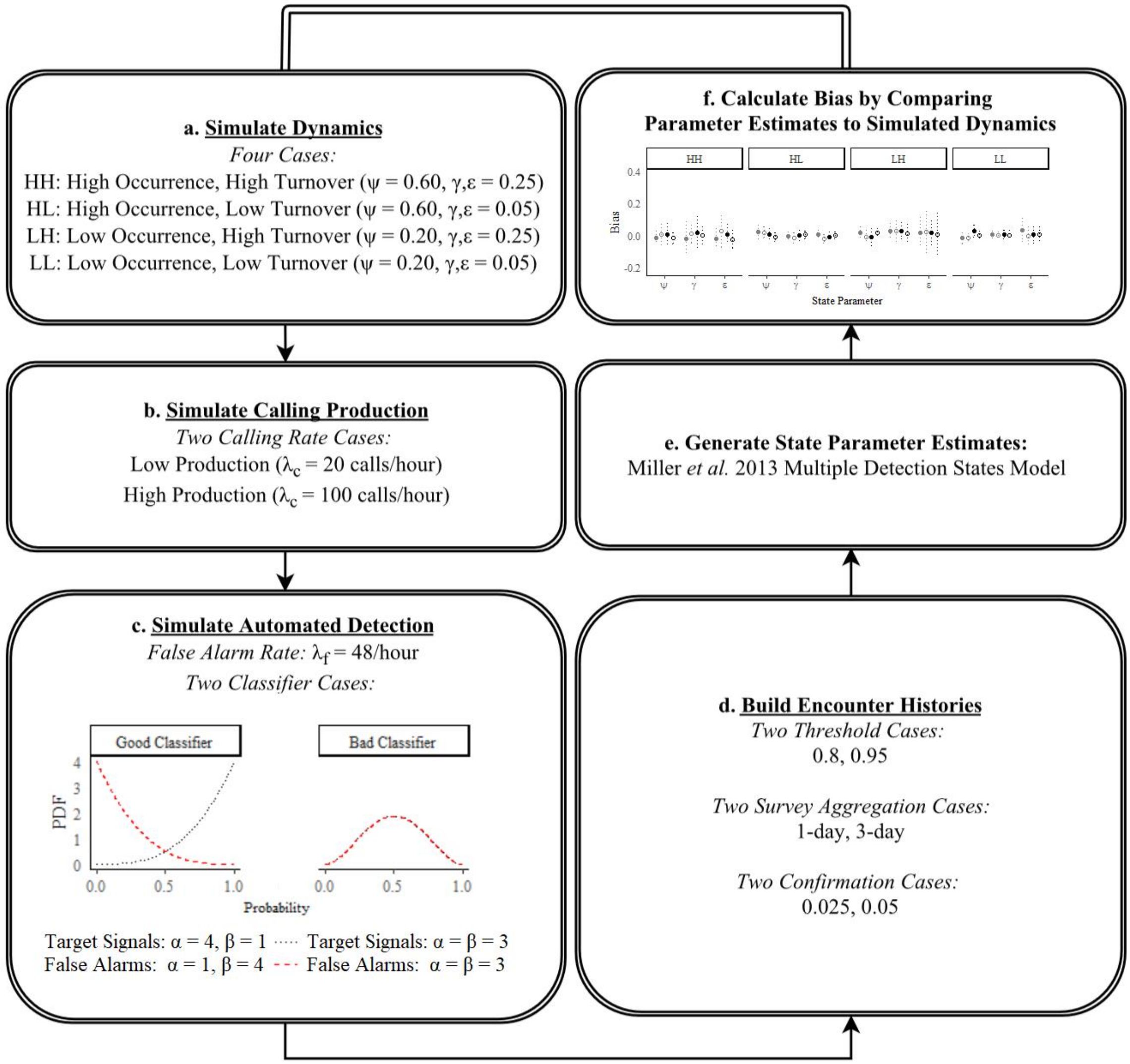

a. Simulate Dynamics
Four Cases:
HH: High Occurrence, High Turnover (ψ = 0.60, γ,ε = 0.25)
HL: High Occurrence, Low Turnover (ψ = 0.60, γ,ε = 0.05)
LH: Low Occurrence, High Turnover (ψ = 0.20, γ,ε = 0.25)
LL: Low Occurrence, Low Turnover (ψ = 0.20, γ,ε = 0.05)
b. Simulate Calling Production
Two Calling Rate Cases:
Low Production (λc = 20 calls/hour)
High Production (λc = 100 calls/hour)
c. Simulate Automated Detection
False Alarm Rate: λf = 48/hour
Two Classifier Cases:
Good Classifier
Bad Classifier
PDF
0
1
2
3
4
0.0
0.5
1.0
Probability
Target Signals: α = 4, β = 1 ····· Target Signals: α = β = 3
False Alarms: α = 1, β = 4 - - - False Alarms: α = β = 3
d. Build Encounter Histories
Two Threshold Cases:
0.8, 0.95
Two Survey Aggregation Cases:
1-day, 3-day
Two Confirmation Cases:
0.025, 0.05
e. Generate State Parameter Estimates:
Miller et al. 2013 Multiple Detection States Model
f. Calculate Bias by Comparing
Parameter Estimates to Simulated Dynamics
HH
HL
LH
LL
Bias
0.4
0.2
0.0
-0.2
ψ
γ
ε
State Parameter

Figure 4. 3. Event-level detections (red boxes) and their associated target signal probabilities can be aggregated into survey-level detection outcomes for an encounter history by multiplying together 1 minus the event-level probabilities within each survey, and then 1 minus this outcome to yield the probability of at least one target signal within the survey (italicized text). If the result exceeds a user-defined threshold, such as 0.95, a 1 is assigned at the survey level. Otherwise, the survey is assigned a 0.

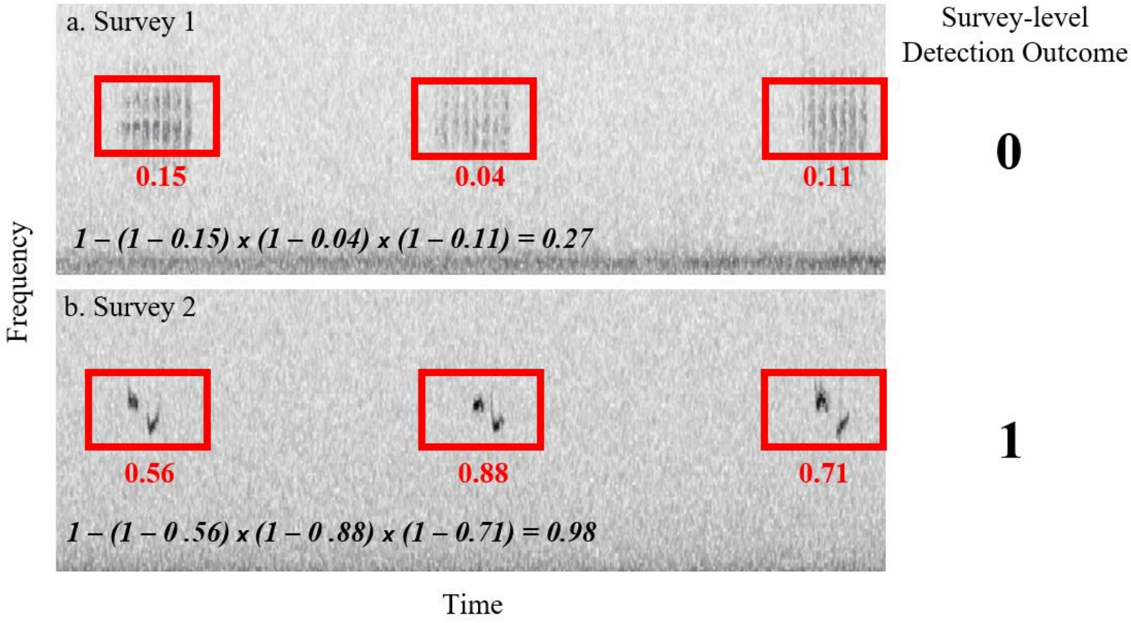

a. Survey 1
Survey-level
Detection Outcome
0.15
0.04
0.11
0
*1 – (1 – 0.15) x (1 – 0.04) x (1 – 0.11) = 0.27*
Frequency
b. Survey 2
0.56
0.88
0.71
1
*1 – (1 – 0 .56) x (1 – 0 .88) x (1 – 0.71) = 0.98*
Time

Figure 4. 4. The average daily number of target signals captured by the automated system (dark gray), vs. the average daily number of false alarms captured by the automated system (light gray), across all sites, under both low (20 calls per hour) and high (100 calls per hour) species calling production scenarios, and under all four underlying dynamics scenarios (Fig. 4. 2a): high occurrence-high turnover (HH), high occurrence-low turnover (HL), low occurrence-high turnover (LH), and low occurrence-low turnover (LL).

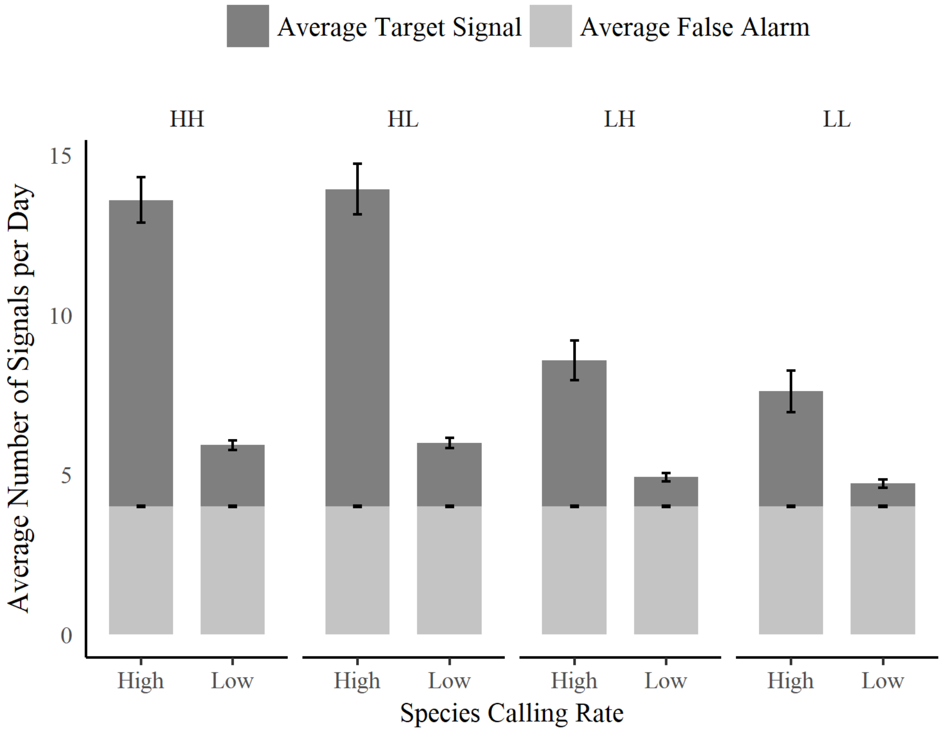


Figure 4. 5. Boxplots of survey-level true positive and false positive rates for 1-day aggregation (a) and 3-day aggregation (b) by classifier type, survey-level detection threshold, and species calling rate (x-axis).

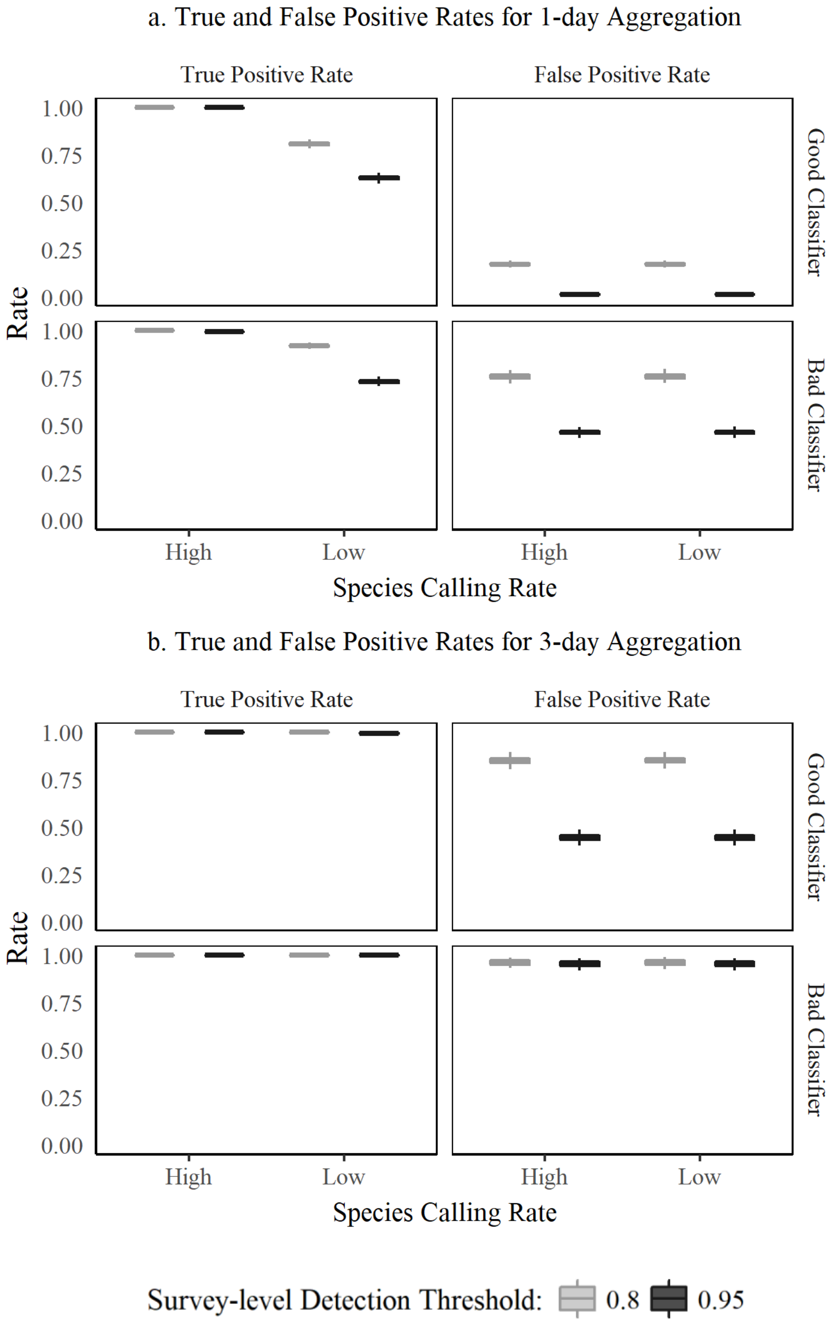
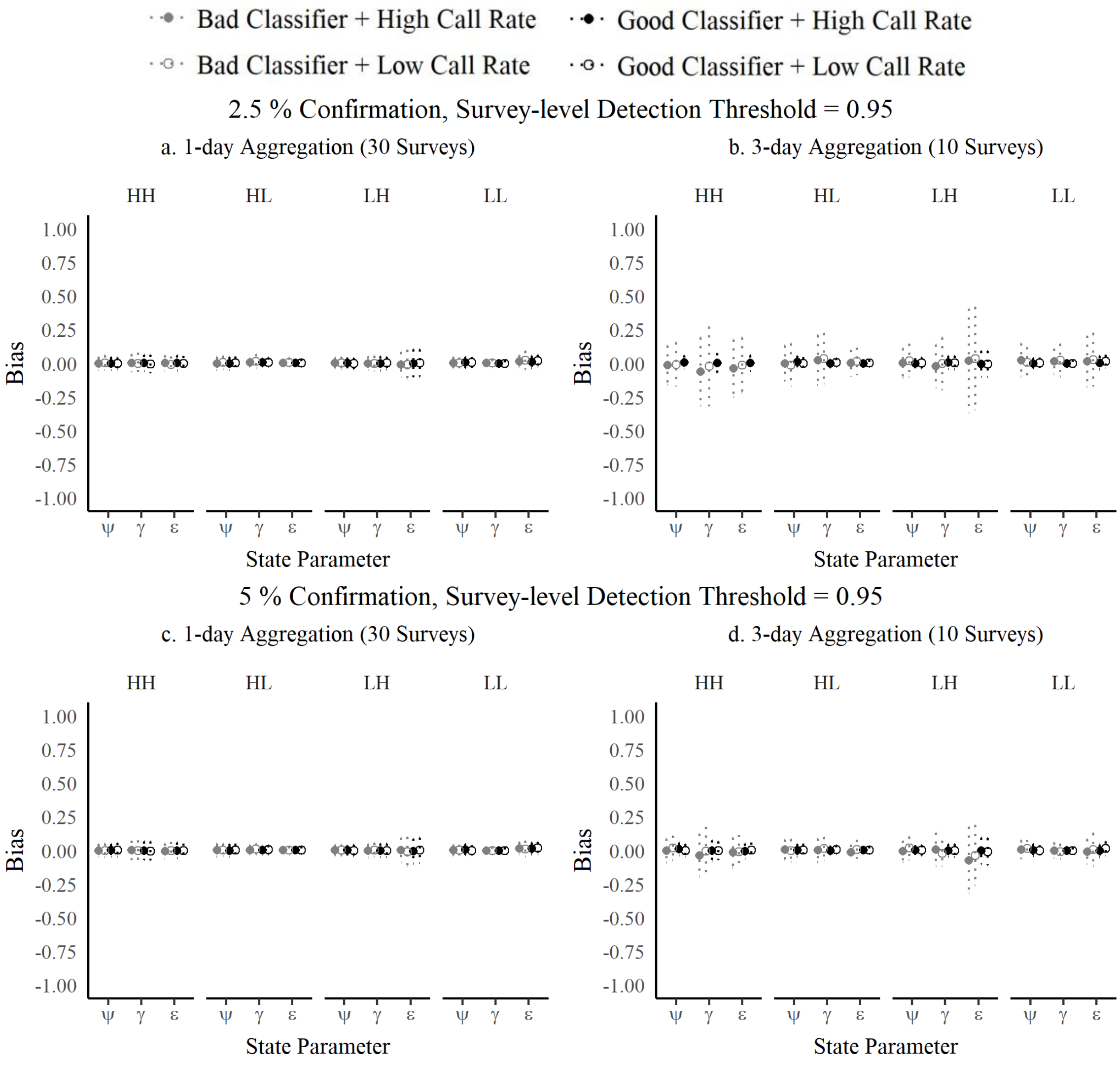
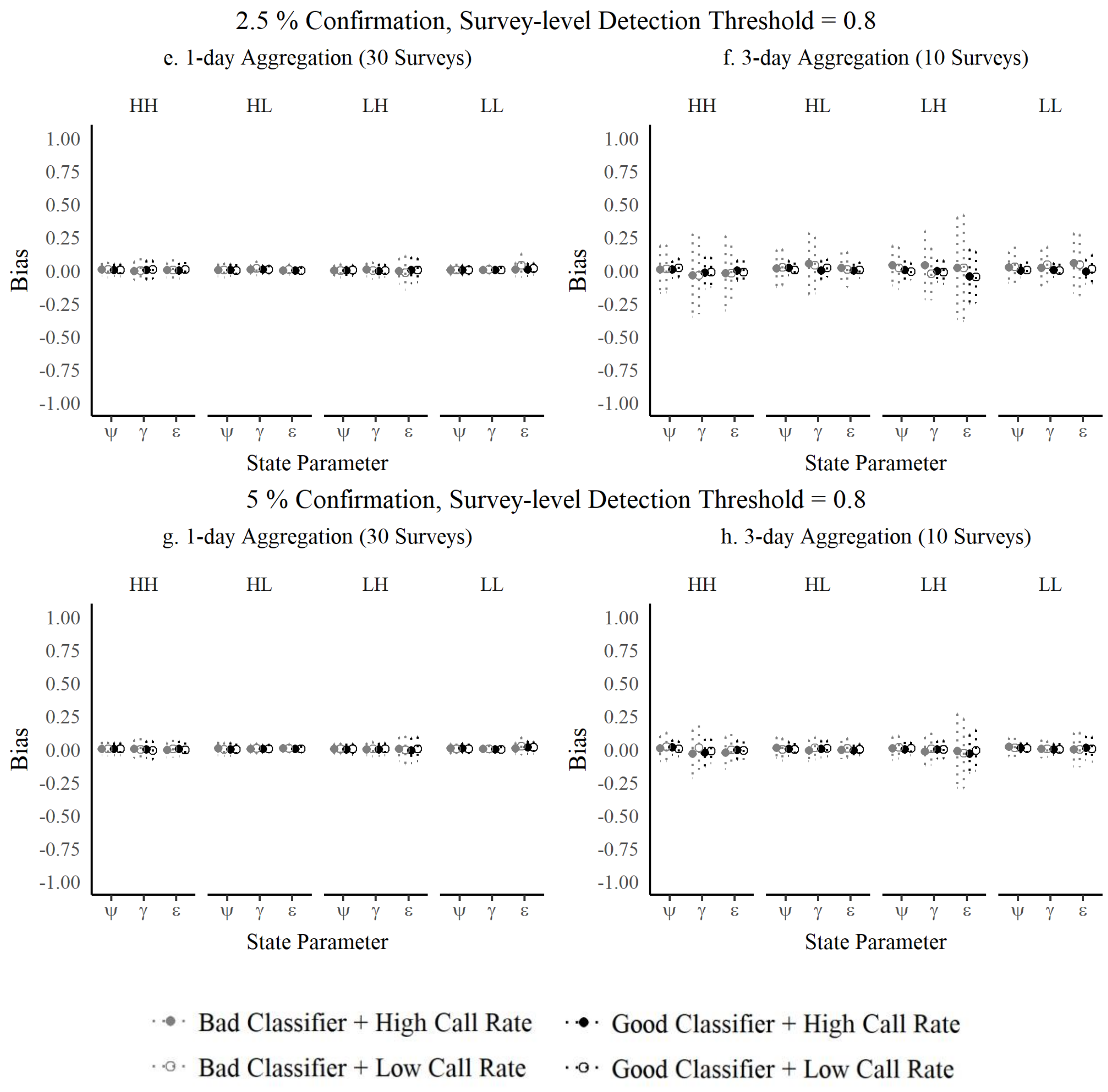


Figure 4. 6. Summary of state parameter estimate bias across occurrence dynamics, species call rates, classifier performance, aggregation frames, survey-level detection thresholds, and confirmation percentages. Circles indicate the mean bias, with dotted vertical bars showing standard deviations. Open circles denote scenarios with a low call rate. Closed circles denote a high call rate. Gray circles denote the bad classifier, and black circles denote the good classifier.





2.5 % Confirmation, Survey-level Detection Threshold = 0.8

e. 1-day Aggregation (30 Surveys)

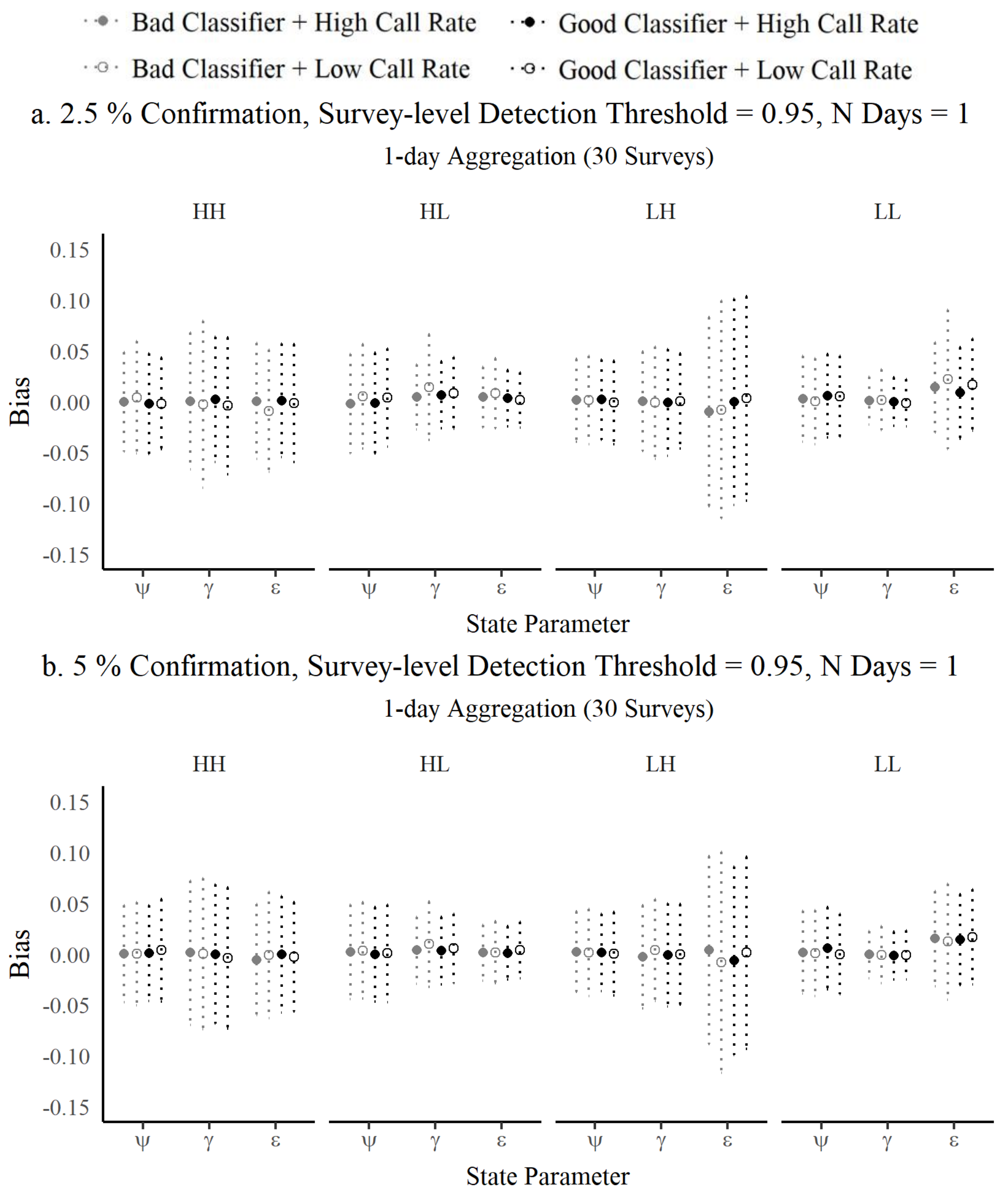
f. 3-day Aggregation (10 Surveys)

5 % Confirmation, Survey-level Detection Threshold = 0.8

g. 1-day Aggregation (30 Surveys)

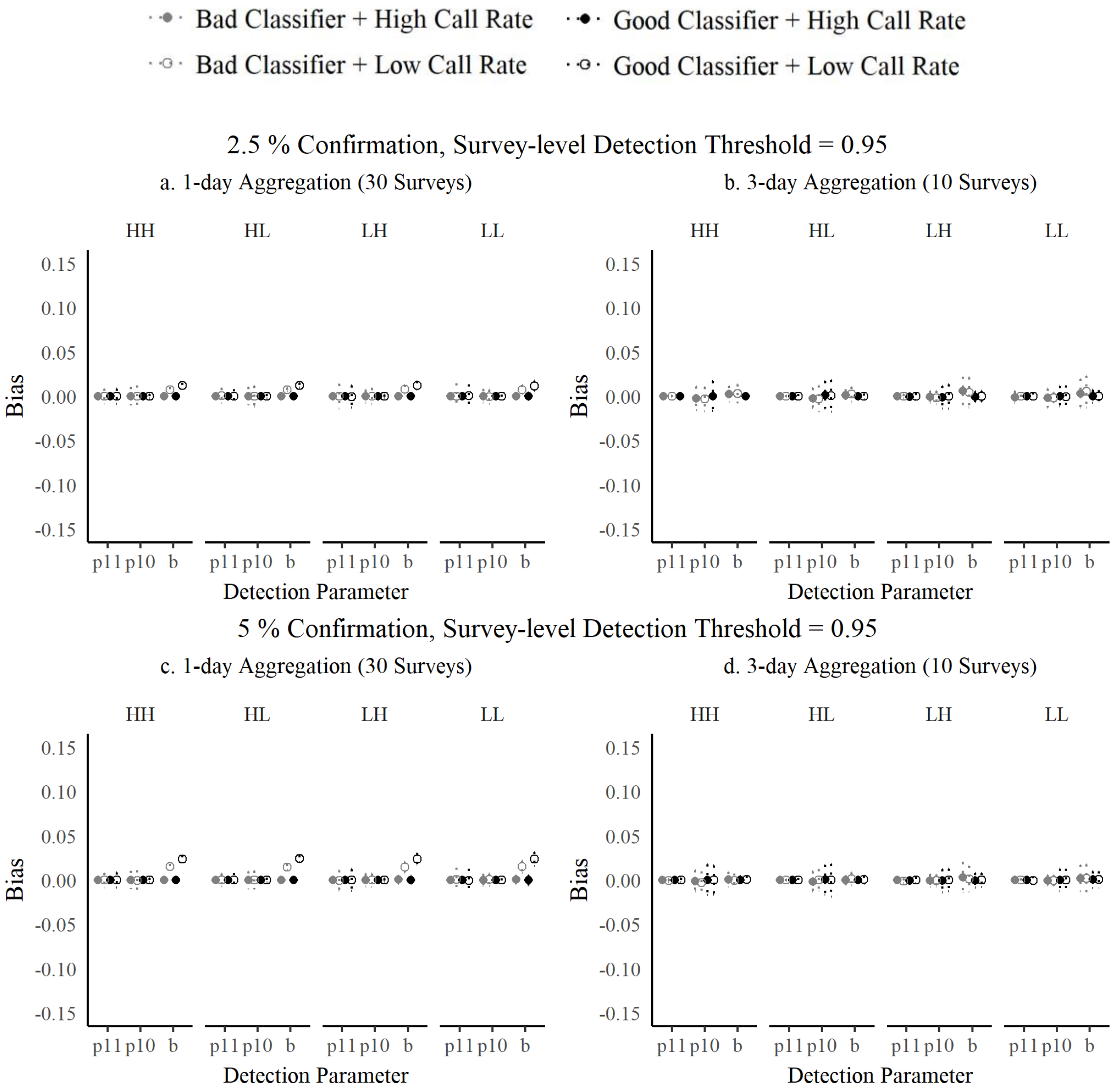
h. 3-day Aggregation (10 Surveys)

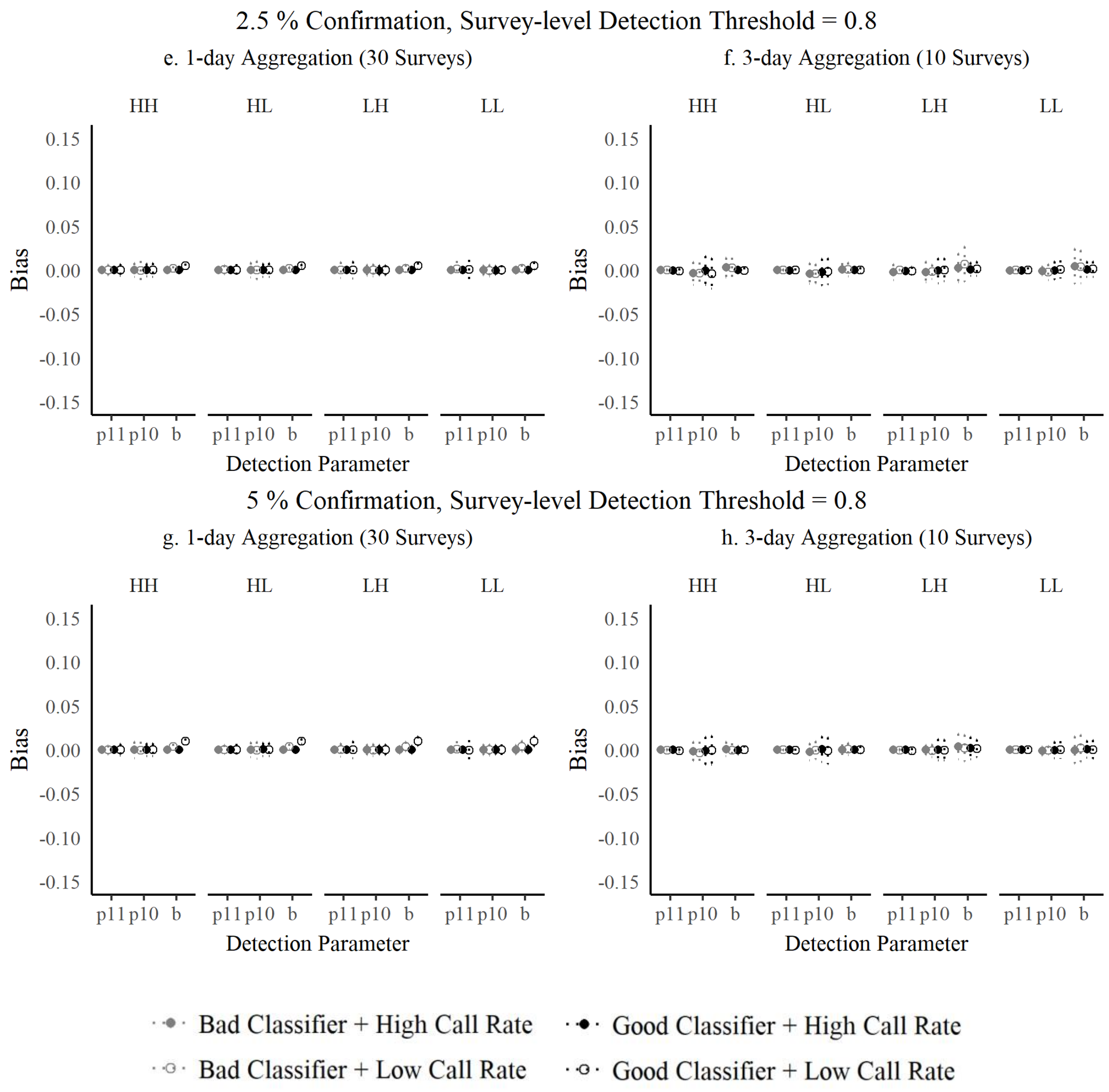
Figure 4. 7. Summary of state parameter estimate bias across different dynamics, call production, classifier performance, and confirmation percentages, for the most conservative survey aggregation circumstances (aggregate days = 1, survey-level detection threshold = 0.95). Circles indicate the mean bias, with dotted vertical bars showing standard deviations. Note that the y-axis has narrowed to range from -0.15 to 0.15.


a. 2.5 % Confirmation, Survey-level Detection Threshold = 0.95, N Days = 1

b. 5 % Confirmation, Survey-level Detection Threshold = 0.95, N Days = 1

Figure 4. 8. Summary of detection parameter estimate bias across different dynamics, call production, classifier performance, and confirmation percentages, for the most conservative survey aggregation circumstances (aggregate days = 1, survey-level detection threshold = 0.95). Circles indicate the mean bias, with dotted vertical bars showing standard deviations. Note that the y-axis has narrowed to range from -0.15 to 0.15.




2.5 % Confirmation, Survey-level Detection Threshold = 0.8
e. 1-day Aggregation (30 Surveys)
f. 3-day Aggregation (10 Surveys)
5 % Confirmation, Survey-level Detection Threshold = 0.8
g. 1-day Aggregation (30 Surveys)
h. 3-day Aggregation (10 Surveys)
HH
HL
LH
LL
Bias
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
p11
p10
b
Detection Parameter
Bad Classifier + High Call Rate
Good Classifier + High Call Rate
Bad Classifier + Low Call Rate
Good Classifier + Low Call Rate

## COMPREHENSIVE BIBLIOGRAPHY

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J. & Aide, T.M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics, 4, 206–214. doi: https://doi.org/10.1016/j.ecoinf.2009.06.005

Adams M.D., Law B.S., & Gibson M.S. (2010). Reliable automation of bat call identification for Eastern New South Wales, Australia, using classification trees and anascheme software. Acta Chiropterologica, 12(1):231–245.

Agranat, I. D. (2009). Automatically Identifying Animal Species from their Vocalizations. Wildlife Acoustics, Inc., Concord, MA.

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. PeerJ 1:e103. doi: https://doi.org/10.7717/peerj.103

Alippi, C., Anastasi, G., Galperti, C., Mancini, F., & Roveri, M. (2007). Adaptive sampling for energy conservation in wireless sensor networks for snow monitoring applications. In Mobile Adhoc and Sensor Systems, 2007. MASS 2007. IEEE International Conference on (pp. 1-6). IEEE.

Anastasi, G., Conti, M., Di Francesco, M., & Passarella, A. (2009). Energy conservation in wireless sensor networks: A survey. Ad hoc networks, 7(3), 537-568.

Anderson, S. E., Dave, A. S. & Margoliash, D. (1996). Template-based automatic recognition of birdsong syllables from continuous recordings. J Acoustical Soc America, 100 (2) Part 1, pp. 1209–1219.

Araya-Salas, M. & Smith-Vidaurre, G. (2017). warbleR: an r package to streamline analysis of animal acoustic signals. Methods Ecol Evol. 8, 184-191.

Astaras, C., Linder, J.M., Wrege, P., Diotoh Orume, R., Macdonald, D.W. (2017). Passive acoustic monitoring as a law enforcement tool for Afrotropical rainforests. Frontiers in Ecology and the Environment, 15(5).

Avisoft Bioacoustics e.K. (2016). Avisoft-SASLab Pro version 5.2.10 [Computer Software]. URL http://www.avisoft.com/.

Bailey, L.L., MacKenzie, D.I. & Nichols, J.D. (2014) Advances and applications of occupancy models (E. Cooch, Ed.). Methods in Ecology and Evolution, 5, 1269–1279. DOI: 10.1111/2041-210X.12100

Bas, Y., Bas, D. & Julien, J.-F. (2017). Tadarida: A Toolbox for Animal Detection on Acoustic Recordings. Journal of Open Research Software. 5(1), p.6. doi: http://doi.org/10.5334/jors.154

Bender, D.J., Bayne, E.M., & Brigham, R.M. (1996). Lunar condition influences coyote (*Canis latrans*) howling. American Midland Naturalist, 136(2), 413-417. doi: 10.2307/2426745

Bioacoustics Research Program. (2018). Raven Pro 1.5: Interactive Sound Analysis Software [Computer Software]. URL http://www.birds.cornell.edu/raven.

Bishop, C. H., Etherton, B. J., & Majumdar, S. J. (2001). Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. Monthly weather review, 129(3), 420-436.

Bishop, C.M. (2006). Pattern recognition and machine learning. Springer. ISBN: 0387310738 9780387310732.

Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory - COLT '92, pp. 144–152. ACM Press, New York, New York, USA.

Brauer, C., T. Donovan, R. Mickey, J. Katz, & Mitchell, B. (2016). A comparison of acoustic monitoring methods for common anurans of the northeastern United States. Wildlife Society Bulletin 40:140-149. doi: 10.1002/wsb.619

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32. doi: https://doi.org/10.1023/A:1010933404324

Bucher, C. G. (1988). Adaptive sampling—an iterative fast Monte Carlo procedure. Structural safety, 5(2), 119-126.

Buxton, R. T., & Jones, I.L. (2012). Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration. Journal of Field Ornithology 83:47-60. http://dx.doi.org/10.1111/j.1557-9263.2011.00355.x

Bye, S. L., Robel, R. J., & Kemp, K. E. (2001). Effects of human presence on vocalizations of grassland birds in Kansas. The Prairie Naturalist, 33(4):249–256.

Campbell, M., & Francis, C.M. (2011). Using stereo-microphones to evaluate observer variation in North American Breeding Bird Survey point counts. Auk 128:303-312. http://dx.doi.org/10.1525/auk.2011.10005

Catchpole, C.K. & Slater, P.J.B. (2008). Bird Song: Biological Themes and Variations, 2nd Ed. Cambridge University Press, Cambridge, UK.

Cerqueira, M. C., & Aide, M.T. (2016). Improving distribution data of threatened species by combining acoustic monitoring and occupancy modeling. Methods in Ecology and Evolution. 7(11), 1340-1348. doi: 10.1111/2041-210X.12599

Chambert, T., Miller, D.A.W., & Nichols, J.D. (2015). Modeling false positive detections in species occurrence data under different study designs. Ecology, 96(2), 332-339. DOI: https://doi.org/10.1890/14-1507.1

Chambert, T., Waddle, J.H., Miller, D.A.W., Walls, S.C., & Nichols, J.D. (2017). A new framework for analyzing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, **9**, 560-570. DOI: 10.1111/2041-210X.12910

Charney, N.D., Kubel, J.E., Eiseman, & Eiseman, C.S. (2015). Temporally adaptive sampling: a case study in rare species survey design with marbled salamanders (*Ambystoma opacum*). PloS One, 10(3), e0120714. doi: https://doi.org/10.1371/journal.pone.0120714

CinixSoft. (2014). CinixSoft Remote Schedule Voice Recorder (v4.2.0). [Android App]. URL http://www.cinixsoft.com/

Clement, M. (2016). Designing occupancy models when false positive detections occur. Methods in Ecology and Evolution, 7, 1538-1547. DOI: 10.1111/2041-210X.12617

Cornell Lab of Ornithology. (2017). BirdCast: Bird migration forecasts in real-time. URL http://www.birdcast.info

Corrada-Bravo, C.J.C., Berrios, R.A, & Aide, T.M. (2017). Species-specific audio detection: a comparison of three template-based detection algorithms using random forests. PeerJ Computer Science 3:e113. doi: https://doi.org/10.7717/peerj-cs.113

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13, 21–27. doi: 10.1109/TIT.1967.1053964

Dark Sky. (2017). Dark Sky API [Application Programming Interface]. URL https://darksky.netDawson, D. K., & Efford, M.G. (2009). Bird population density estimated from acoustic signals. Journal of Applied Ecology 46:1201-1209. doi: http://dx.doi.org/10.1111/j.1365-2664.2009.01731.x

Davis, J. & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine Learning, Pittsburg, PA.

Dawson, D. K., & M. G. Efford. (2009). Bird population density estimated from acoustic signals. Journal of Applied Ecology 46:1201-1209. http://dx.doi.org/10.1111/j.1365-2664.2009.01731.x

Digipom (2016). Easy Voice Recorder Pro [Android App]. URL http://www.digipom.com/portfolio-items/easy-voice-recorder/

Duan, S., Zhang, J., Roe, P., Wimmer, J., Dong, X., Truskinger, A., & Towsey, M. (2013) Timed Probabilistic Automaton : a bridge between Raven and Song Scope for automatic species recognition. In Muñoz-Avila, Hector & Stracuzzi, David J. (Eds.). Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference, AAAI, Bellevue, Washington, USA, pp. 1519-1524.

Dugger, K. M., Forsman, E. D., Franklin, A. B., Davis, R. J., White, G. C., Schwarz, C. J., et al., & Doherty Jr, P. F. (2015). The effects of habitat, climate, and Barred Owls on long-term demography of Northern Spotted Owls. The Condor, 118(1), 57-116.

Dyo, V., Ellwood, S. A., MacDonald, D. W., Markham, A., Trigoni, N., Wohlers, R., Mascolo, C., Pasztor, B., Scellato, S., & Yosef, K. (2012). WILDSENSING: Design and deployment of a sustainable sensor network for wildlife monitoring. ACM Transactions on Sensor Networks 8 (4), Article 29. doi: 10.1145/2240116.2240118

Fagerlund S. (2007). Bird species recognition using support vector machines. EURASIP Journal on Applied Signal Processing 2007(1):1–8.

Farnsworth, A. (2005). Flight calls and their value for future ornithological studies and conservation research. The Auk, 122(3), 733-746.

Ferguson, P. F., Conroy, M. J., & Hepinstall-Cymerman, J. (2015). Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys. Methods in Ecology and Evolution, 6(12), 1395-1406.

Figueroa, H. (2012). XBAT [Computer Software]. Bioacoustics Research Program. URL http://www.xbat.org.

Fiske, I., & Chandler, R. (2011). unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. Journal of Statistical Software, 43(10), 1-23. URL http://www.jstatsoft.org/v43/i10/.

Frey, S., Fisher, J.T., Cole Burton, A., & Volpe, J.P. (2017). Investigating animal activity patterns and temporal niche partitioning using camera-trap data: challenges and opportunities. Remote Sensing in Ecology and Conservation, 3(3).

Frick, W. F., Stepanian, P. M., Kelly, J. F., Howard, K. W., Kuster, C. M., Kunz, T. H., & Chilson, P. B. (2012). Climate and weather impact timing of emergence of bats. PLoS One, 7(8), e42737.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York, NY, USA: Springer series in statistics.

Furnas, B. J., & Callas, R.L. (2015). Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. Journal of Wildlife Management 79:325-337. DOI: http://dx.doi.org/10.1002/jwmg.821

Gage, S.H., Napoletano, B., & Cooper, M., (2001). Assessment of ecosystem biodiversity by acoustic diversity indices. J. Acoust. Soc. Am. 109 (5), 2430.

Gage, S. H., W. Joo, E. P. Kasten, J. Fox, & S. Biswas. (2015). Acoustic observations in agricultural landscapes. Pages 360-377 in S. K. Hamilton, J. E. Doll, & G. P. Robertson, editors. The Ecology of Agricultural Landscapes: Long-Term Research on the Path to Sustainability. Oxford University Press, New York, New York, USA. ISBN: 978-0199773350.

Gage, S.H. & Farina, A. (2017). Ecoacoustics Challenges. Pages 313-320 in A. Farina & S.H. Gage, editors. Ecoacoustics: The Ecological Role of Sounds. John Wiley & Sons. ISBN: 978-1-119-23069-4.

Gaynor, K.M., Hojnowski, C.E., Carter, N.H., & Brashares, J.S. (2018). The influence of human disturbance on wildlife nocturnality. Science, 360 (6394) 1232-1235. DOI: 10.1126/science.aar7121

Gee, J., Brown, D.E., Hagelin, J.C., Taylor, M. & Galloway, J. (2013). *Gambel's Quail* (Callipepla gambelii), The Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. doi: 10.2173/bna.321

Goëau, H., Glotin, H., Vellinga, W.-P., Planqué, R. & Joly, A. (2016). LifeCLEF bird identification task 2016: The arrival of deep learning. In Working Notes of CLEF 2016-Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016, 440–449.

Gutzwiller, K. J., & Marcum, H. A. (1997). Bird reactions to observer clothing color: implications for distance-sampling techniques. Journal of Wildlife Management 61:935-947. http://dx.doi.org/10.2307/3802203

Hafner S. & Katz J. (2018). monitoR: Acoustic template detection in R. R package version 1.0.7, URL: http://www.uvm.edu/rsenr/vtcfwru/R/?Page=monitoR/monitoR.htm.

Hayes, A. R., & Huntly, N. J. (2005). Effects of wind on the behavior and call transmission of pikas (Ochotona princeps). Journal of mammalogy, 86(5), 974-981.

Heinicke, S., Kalan, A.K., Wanger, O.J., Mundry, R., Lukashevich, H., & Kuhl, H.S. (2015). Assessing the performance of a semi-automated acoustic monitoring system for primates. Methods in Ecology and Evolution, 6(7): 753-763. doi: 10.1111/2041-210X.12384

Hill, A. P., Prince, P., Piña Covarrubias, E., Doncaster, C. P., Snaddon, J. L., & Rogers, A. (2018). AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. Methods in Ecology and Evolution, 9(5), 1199-1211.

Hines, J. (2018). RPresence for PRESENCE: Software to estimate patch occupancy and related parameters. Version 12.10. https://www.mbr-pwrc.usgs.gov/software/presence.html

Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, & Van Wilgenburg, S.L. (2002). Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. Wildlife Society Bulletin 30:709-720.

Hughes, B.B., Beas-Luna, R., Barner, A.K., Brewitt, K., Brumbaugh, D.R., Cerny-Chipman, E.B., Close, S.L., Coblentz, K.E., de Nesnera, K.L., Drobnitch, S.T., et al. (2017). Long-term studies contribute disproportionately to ecology and policy. Bioscience, 67(3), 271-281. DOI: 10.1093/biosci/biw185

Hutto, R. L., & Stutzman, R.J. (2009). Humans versus autonomous recording units: a comparison of point-count results. Journal of Field Ornithology 80:387–398.

Jain, A., & Chang J.Y. (2004). Adaptive sampling for sensor networks, in: Proc. 1st international workshop on Data management for sensor networks (DMSN 2004), Toronto, Canada, August 30th, 2004, pp. 10–16.

Johnson, F. A., Boomer, G. S., Williams, B. K., Nichols, J. D., & Case, D. J. (2015). Multilevel learning in the adaptive management of waterfowl harvests: 20 years and counting. Wildlife Society Bulletin, 39(1), 9-19.

Katz, J., Hafner, S.D. & Donovan, T. (2016). Assessment of Error Rates in Acoustic Monitoring with the R package monitoR. Bioacoustics, 25, 177–196. doi: https://doi.org/10.1080/09524622.2015.1133320

Knight, E. C., Hannah, K.C., Foley, G., Scott, C., Mark Brigham, R., & Bayne, E. (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology 12(2):14. https://doi.org/10.5751/ACE-01114-120214

Knight, E.C. & Bayne. E.M. 2018. Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. Bioacoustics. doi: 10.1080/09524622.2018.1503971

Kroodsma, D.E., & Miller, E.H. (eds). (1996). Ecology and evolution of acoustic communication in birds. Comstock Pub, University of Michigan. ISBN: 0801430496, 9780801430497

Kuhn, M. (2016). caret: Classification and Regression Training. R package version 6.0-71.  http://CRAN.R-project.org/package=caret

Leecaster, M. K., & Weisberg, S. B. (2001). Effect of sampling frequency on shoreline microbiology assessments. Marine pollution bulletin, 42(11), 1150-1154.

Lévy, C., Linarès, G., & Nocera, P. (2003). Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems. In Workshop on DSP in Mobile and Vehicular Systems, Nagoya, Japan.

Luther, D., & Baptista, L. (2010). Urban noise and the cultural evolution of bird songs. Proceedings of the Royal Society B: Biological Sciences 277: 469–473.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A. & Langtimm, C.A. (2002). Estimating site occupancy rates when detection probabilities are less than one. Ecology, 83, 2248–2255. doi: 10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology, 84, 2200–2207. DOI: 10.1890/02-3090

MacKenzie, D. I., & Royle, J. A. (2005). Designing occupancy studies: general advice and allocating survey effort. Journal of applied Ecology, 42(6), 1105-1114.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006). Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Academic Press. **ISBN:** 9780120887668.Manley, P.N., Zielinski, W.J., Schlesinger, M.D., & Mori, S.R. (2004). Evaluation of a multiple-species approach to monitoring species at the ecoregional scale. Ecological Applications, 14(1) 296-310. doi: https://doi.org/10.1890/02-5249

Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R., & Anderson, J. (2002). Wireless sensor networks for habitat monitoring. In Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications (pp. 88-97). Acm.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D. & Tyack, P. L. (2012). Estimating animal population density using passive acoustics. Biological Reviews.

Mayhew, W.W. (1965). Adaptations of the amphibian, *Scaphiopus couchi,* to desert conditions. American Midland Naturalist, 74(1), 95-109. doi: 10.2307/2423123

McClintock, B.T., Bailey, L.L., Pollock, K.H. & Simons, T.R. (2010). Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, **91**, 2446–2454. DOI: https://doi.org/10.1890/09-1287.1

McKown, M.W. (2012). A wireless acoustic sensor network for monitoring wildlife in remote locations. The Journal of the Acoustical Society of America, 132, 2036. DOI: https://doi.org/10.1121/1.4755484

Mennill, D. J., & Vehrencamp, S.L. (2008). Context-dependent functions of avian duets revealed by microphone-array recordings and multispeaker playback. Current Biology 18:1314-1319. http://dx.doi.org/10.1016/j.cub.2008.07.073

Miller, D.A.W., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology, 92, 1422–1428. doi: 10.1890/10-1396.1

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology, 92(7), 1422-1428.

Miller, D. A., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., & Mitchell, M. S. (2013). Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data. PLoS one, 8(6), e65808.

Miller, D.A.W., Bailey, L.L., Grant, E.H.C., McClintock, B.T., Weir, L.A. & Simons, T.R. (2015). Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known (O. Gimenez, Ed.). *Methods in Ecology and Evolution*, **6**, 557–565. DOI: 10.1111/2041-210X.12342

Moore, A., McCarthy, M. Parris, K., & Moore., J.L. (2014). The Optimal Number of Surveys when Detectability Varies. PLoS ONE, 9, e115345. doi: https://doi.org/10.1371/journal.pone.0115345

Moore, A.L. & McCarthy, M.A. (2016). Optimizing ecological survey effort over space and time. Methods in Ecology and Evolution, 7, 891–899. doi: 10.1111/2041-210X.12564

Newson, S.E., Bas, Y., Murray, A., & Gillings, S. (2017). Potential for coupling the monitoring of bush-crickets with established large-scale acoustic monitoring of bats. *Methods in Ecology and Evolution*, **8**, 1051-1062. DOI: 10.1111/2041-210X.12720

Nichols, J. D., Runge, M. C., Johnson, F. A., & Williams, B. K. (2007). Adaptive harvest management of North American waterfowl populations: a brief history and future prospects. Journal of Ornithology, 148(2), 343-349.

Nichols, J. D., Karanth, K. U., & O'Connell, A. F. (2011). Science, conservation, and camera traps. In Camera Traps in Animal Ecology (pp. 45-56). Springer, Tokyo.

Nichols, J.D., Yackulic, C.B., Reid, J., Hines, J.E., Davis, R, & Forsman, E. (2015). Dynamic occupancy modeling for conservation. Presented at Ecological Society of America Annual Meeting, Baltimore, MD, August 2015.

Noad, M. J., Cato, D. H., Bryden, M. M., Jenner, M. N., & Jenner, K. C. S. (2000). Cultural revolution in whale songs. Nature, 408(6812), 537.

Otis, D. L., Burnham, K.P., White, G.C., & Anderson, D.R. (1978). Statistical inference from capture data on closed animal populations. Wildlife Monographs, 62, 1-135.

Ovaskainen, O., Moliterno de Camargo, U., & Somervuo, P. (2018). Animal Sound Identifier (ASI): software for automated identification of vocal animals. *Ecology letters,* 21(8): 1244-1254. doi: https://doi-org.ezproxy.uvm.edu/10.1111/ele.13092

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., et al. (2011). Soundscape ecology: the science of sound in the landscape. BioScience, 61(3), 203-216.

Pollock, K. H., Nichols, J. D., Simons, T. R., Farnsworth, G. L., Bailey, L. L., & Sauer, J. R. (2002). Large scale wildlife monitoring studies: statistical methods for design and analysis. Environmetrics, 13(2), 105-119.

Porter, J., Arzberger, P., Braun, H. W., Bryant, P., Gage, S., Hansen, T., et al. & Michener, W. (2005). Wireless sensor networks for ecology. AIBS Bulletin, 55(7), 561-572.

Potamitis, I., Ntalampiras, S., Jahn, O. & Riede, K. (2014). Automatic bird sound detection in long real-field recordings: applications and tools. Applied Acoustics 80:1-9. doi: https://doi.org/10.1016/j.apacoust.2014.01.001

Powers, D.M.W. (2007). Evaluation: From Precision, Recall, and F-Factor to ROC, Informedness, Markedness & Correlation. School of Informatics and Engineering, Flinders University. Adelaide, Australia. Technical Report SIE-07-001.

Priyadarshani N., Marsland S., Castro I., & Punchihewa A. (2016). Birdsong Denoising Using Wavelets. PLoS ONE 11(1): e0146790. doi:10.1371/ journal.pone.0146790

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raghunathan, V., Ganeriwal, S., & Srivastava, M. (2006). Emerging techniques for long lived wireless sensor networks. IEEE Communications Magazine, 44(4), 108-114.

Ralph, C. J., Sauer, J. R., & Droege, S. (1995). Monitoring bird populations by point counts. Gen. Tech. Rep. PSW-GTR-149. Albany, CA: US Department of Agriculture, Forest Service, Pacific Southwest Research Station. 187 p, 149.

Ranjard, L., Reed, B.S., Landers, T.J., Raynar, M.J., Friesen, M.R., Sagar, R.L., & Dunphy, B.J. (2017). MatlabHTK: a simple interface for bioacoustics analyses using hidden Markov models. Methods in Ecology and Evolution 8(5): 615-621. doi: 10.1111/2041-210X.12688

Robbins, C. S., D. Bystrak, & P. H. Geissler. (1986). The breeding bird survey: its first fifteen years, 1965–1979. Resource Publication No. 156, U.S. Fish and Wildlife Service, Washington, DC, USA.

Romagosa, C.M. (2012). Eurasian Collared-Dove (Streptopelia decaocto), The Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology; Retrieved from the Birds of North America: https://birdsna.org/Species-Account/bna/species/eucdov. doi: 10.2173/bna.630

Root-Gutteridge, H., Bencsik, M., Chebli, M., Gentle, L. K., Terrell-Nield, C., Bourit, A., & Yarnell, R. W. (2014). Identifying individual wild Eastern grey wolves (Canis lupus lycaon) using fundamental frequency and amplitude of howls. Bioacoustics, 23(1), 55-66.

Rosenstock, S. S., D. R. Anderson, K. M. Giesen, T. Leukering, & M. F. Carter. (2002). Landbird counting techniques: current practices and an alternative. Auk 119:46-53. http://dx.doi.org/10.1642/0004-8038(2002)119[0046:LCTCPA]2.0.CO;2

Royle, J.A. & Nichols, J.D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology* **84**(3), 777-790. DOI: https://doi.org/10.1890/0012-9658(2003)084[0777:EAFRPA]2.0.CO;2

Royle, J. A. (2006). Site occupancy models with heterogeneous detection probabilities. Biometrics, 62(1), 97-102.

Ruiz-Gutierrez, V., Hooten, M.B. & Campbell Grant, E.H. (2016). Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias (N. Yoccoz, Ed.). Methods in Ecology and Evolution, 7, 900–909. doi: 10.1111/2041-210X.12542

Russo, D., & Voigt, C.C. (2016). The use of automated identification of bat echolocation calls in acoustic monitoring: A cautionary note for a sound analysis. *Ecological Indicators*, **66,** 598-602. DOI: https://doi.org/10.1016/j.ecolind.2016.02.036

Salamon, J., J. Pablo Bello, A. Farnsworth, M. Robbins, H. Klinck, S. Keen, & Kelling, S. (2016). Towards the automatic classification of avian flight calls for bioacoustic monitoring. PLoS ONE. doi: 10.1371

Shonfield, J. & Bayne, E.M. (2017). Autonomous recording units in avian ecological research: current use and future applications. Avian Conservation and Ecology, 12(1): 14. DOI: 10.5751/ACE-00974-120114

Sidie-Slettedahl, A. M., K. C. Jensen, R. R. Johnson, T. W. Arnold, J. E. Austin, & Stafford, J.D. (2015). Evaluation of autonomous recording units for detecting 3 species of secretive marsh birds. Wildlife Society Bulletin 39:626-634. http://dx.doi.org/10.1002/wsb.569

Sieve Analytics. (2018). Products. Url: https://www.sieve-analytics.com/products

Smith, D. R., Conroy, M. J., & Brakhage, D. H. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. Biometrics, 777-788.

Stowell, D., Wood, M., Stylianou, Y. & Glotin, H. (2016). Bird detection in audio: a survey and a challenge. IEEE International Workshop on Machine Learning for Signal Processing, Salerno, Italy. doi: 10.1109/MLSP.2016.7738875

Stowell, D., Stylianou, Y., Wood, M., Pamuła, H., & Glotin, H. (2018). Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. arXiv preprint arXiv:1807.05812.

Sueur J., Aubin T., Simonis C. (2008). Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics, 18: 213-226

Sueur, J., & Farina, A. (2015). Ecoacoustics: the ecological investigation and interpretation of environmental sound. Biosemiotics, 8(3), 493-502.

Sugiyama, M. & Kawanabe, M. (2012). Machine learning in non-stationary environments: Introduction to covariate shift adaptation. MIT press.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., & Kelling, S. (2009). eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142: 2282-2292. doi: https://doi.org/10.1016/j.biocon.2009.05.006

Suthers, R.A., Fitch, W.W., Fay, R.R., & Popper, A.N. (eds) (2016). Vertebrate sound production and acoustic communication. Springer International Publishing, Switzerland. DOI: https://doi.org/10.1007/978-3-319-27721-9 ISBN: 978-3-319-27721-9

Swiston, K. A., & Mennill, D.J. (2009). Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-Billed, and putative Ivory-Billed Woodpeckers. Journal of Field Ornithology 80:42-50. http://dx.doi.org/10.1111/j.1557-9263.2009.00204.x

Tasker (2015). Tasker: Total Automation for Android (v4.8). [Android App]. URL http://tasker.dinglisch.net/

Thompson, S.K. & Seber, G.A.F. (1994). Detectability in Conventional and Adaptive Sampling. Biometrics, 50, 712. doi: 10.2307/2532785

Thompson, W.L., White, G.C., & Gowan, C. (1998). Monitoring Vertebrate Populations. Academic Press, San Diego. ISBN: 0126889600

Thompson, W. L. (2004). Sampling rare or elusive species: concepts and techniques for estimating population parameters. Island Press, Washington, D.C., USA. ISBN: 9781559634519

Towsey, M., Planitz, B., Nantes, A., Wimmer, J. & Roe, P. (2012). A toolbox for animal call recognition. Bioacoustics, 21, 107–125.doi: http://dx.doi.org/10.1080/09524622.2011.648753

Turk, P. & Borkowski, J.J. (2005). A review of Adaptive Cluster Sampling: 1990-2003. Environmental and Ecological Statistics, 12(1), 55-94. doi: https://doi.org/10.1007/s10651-005-6818-0

U.S. Bureau of Land Management. (2016). Riverside East Solar Energy Zone Long Term Monitoring Strategy: Final Report. Prepared by Environmental Science Division, Argonne National Laboratory, for the U.S. Department of the Interior Bureau of Land Management.

U.S. Geological Survey. (2001). North American Breeding Bird Survey Methodology: Methods and Requirements. Patuxent Wildlife Research Center, U.S. Department of the Interior, U.S. Geological Survey. URL https://www.pwrc.usgs.gov/bbs/participate/training/11.html

Villanueva-Rivera, L.J, & Pijanowski, B. (2018). Soundecology: Soundscape Ecology. R Package. URL: https://cran.r-project.org/web/packages/soundecology/index.html

Webster, M.D. (1999). Verdin (Auriparus flaviceps), The Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology; Retrieved from the Birds of North America*:* https://birdsna.org/Species-Account/bna/species/verdin. doi: 10.2173/bna.470

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473-1480).

White, G. C., Anderson, D. R., Burnham, K. P. & Otis, D. L. (1982). Capture-recapture and removal methods for sampling closed populations. LA-8787-NERP, Los Alamos National Laboratory, Los Alamos, NM. 235pp.

Whytock, R. C., & Christie, J. (2017). Solo: an open source, customizable and inexpensive audio recorder for bioacoustic research. Methods in Ecology and Evolution, 8(3), 308-312.

Wildlife Acoustics. (2016). Song Meter SM4 [Acoustic Recording Hardware]. URL https://www.wildlifeacoustics.com/products/song-meter-sm4

Wildlife Acoustics. (2018). Kaleidescope [Computer Software]. URL http://www. wildlifeacoustics.com.

Williams, B.K., Szaro, R.C., & Shapiro, C.D. (2009). Adaptive Management: The U.S. Department of the Interior Technical Guide, 2nd Edition. Adaptive Management Working Group, U.S. Department of the Interior, Washington DC.

Williams, H., Levin, I.I., Norris, D.R., Newman, A.E.M., & Wheelwright, N.T. (2013). Three decades of cultural evolution in Savannah sparrow songs. *Animal Behaviour*, **85** (1): 213 DOI: 10.1016/j.anbehav.2012.10.028

Woods, C. P., Csada, R.D., & Brigham, R.M. (2005). Common Poorwill (*Phalaenoptilus nuttallii*), The Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. doi: 10.2173/bna.32

Wrege, P.H., Rowland, E.D., Bout, N., & Doukaga, M. (2012). Opening a larger window onto forest elephant ecology. African Journal of Ecology, 50 (2).

Wrege, P.H., Rowland, E.D., Keen, S., & Shiu, Y. (2017). Acoustic monitoring for conservation in tropical forests: examples from forest elephants. Methods in Ecology and Evolution, 8 (10).

Xu, Y., Choi, J., Dass, S., & Maiti, T. (2011). Bayesian prediction and adaptive sampling algorithms for mobile sensor networks. In American Control Conference (ACC), 2011 (pp. 4195-4200). IEEE.

Zamora-Gutierrez, V., Lopez-Gonzalez, C., MacSwiney Gonzalez, M.C., Fenton, B., Jones, G., Kalko, E. K. V., Puechmaille, S.J., Stathopoulos, V., Jones, K.E. (2016). Acoustic identification of Mexican bats based on taxonomic and ecological constraints on call design. Methods in Ecology and Evolution, 7(9).

Zhou, J & De Roure, D. (2007). FloodNet: coupling adaptive sampling with energy aware routing in a flood warning system. Journal of Computer Science and Technology, 22 (1), pp. 121-130.

Zhu, X. & Davidson, I. (Eds.). (2007). Knowledge Discovery and Data Mining. IGI Global.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

# Appendix A

A. 1. Comprehensive $p*_{max}$ Achievement Results. The value in red denotes the only case in which the fixed schedule achieved p*max prior to the optimized schedule.

**Date $p*_{max}$ Achieved (Full Year)**

| Species | Optimized | Fixed | Difference (Days) | Effort |
|---|---|---|---|---|
| BTGN | 12/28/2016 | - | - | 2 |
| ECDO | 5/24/2016 | - | - | 2 |
| GAQU | 5/11/2016 | - | - | 2 |
| PHAI | 4/28/2016 | - | - | 2 |
| VERDI | 5/16/2016 | - | - | 2 |
| BTGN | 4/8/2016 | 5/23/2016 | -45.2 | 5 |
| COPO | 7/15/2016 | - | - | 5 |
| ECDO | 3/2/2016 | 4/22/2016 | -51.4 | 5 |
| GAQU | 2/26/2016 | 4/10/2016 | -43.8 | 5 |
| PHAI | 2/19/2016 | 3/30/2016 | -39.6 | 5 |
| VERD | 2/25/2016 | 4/8/2016 | -42.5 | 5 |
| BTGN | 2/19/2016 | 3/20/2016 | -30.0 | 10 |
| COPO | 3/7/2016 | 4/9/2016 | -32.8 | 10 |
| ECDO | 1/31/2016 | 3/7/2016 | -35.4 | 10 |
| GAQU | 1/29/2016 | 3/1/2016 | -32.3 | 10 |
| PHAI | 1/27/2016 | 2/25/2016 | -28.9 | 10 |
| VERD | 1/28/2016 | 2/28/2016 | -30.9 | 10 |
| BTGN | 1/24/2016 | 2/25/2016 | -32.4 | 20 |
| COPO | 2/2/2016 | 2/13/2016 | -10.8 | 20 |
| COYOTE | 10/11/2016 | 10/3/2016 | **8.9** | 20 |
| ECDO | 1/17/2016 | 2/13/2016 | -27.3 | 20 |
| GAQU | 1/15/2016 | 2/8/2016 | -23.1 | 20 |
| LENI | 3/20/2016 | 9/16/2016 | -180.0 | 20 |
| PHAI | 1/15/2016 | 2/2/2016 | -18.1 | 20 |
| VERD | 1/15/2016 | 2/6/2016 | -22.8 | 20 |
| BTGN | 1/16/2016 | 2/8/2016 | -23.1 | 30 |
| COPO | 1/23/2016 | 1/31/2016 | -7.5 | 30 |
| COYOTE | 6/15/2016 | 8/10/2016 | -55.0 | 30 |
| ECDO | 1/12/2016 | 2/3/2016 | -21.7 | 30 |
| GAQU | 1/11/2016 | 1/29/2016 | -18.0 | 30 |
| LENI | 2/23/2016 | 4/12/2016 | -49.7 | 30 |
| PHAI | 1/10/2016 | 1/25/2016 | -15.2 | 30 |

| | | | | |
|---|---|---|---|---|
| TOAD | 11/16/2016 | - | - | 30 |
| VERD | 1/10/2016 | 1/27/2016 | -16.4 | 30 |
| BTGN | 1/15/2016 | 1/24/2016 | -9.5 | 40 |
| COPO | 1/19/2016 | 1/24/2016 | -5.1 | 40 |
| COYOTE | 4/12/2016 | 5/16/2016 | -33.2 | 40 |
| ECDO | 1/10/2016 | 1/21/2016 | -11.4 | 40 |
| GAQU | 1/9/2016 | 1/19/2016 | -10.2 | 40 |
| LENI | 2/11/2016 | 3/8/2016 | -25.4 | 40 |
| PHAI | 1/8/2016 | 1/18/2016 | -9.2 | 40 |
| TOAD | 9/27/2016 | - | - | 40 |
| VERD | 1/8/2016 | 1/18/2016 | -9.4 | 40 |

**Summary Statistics for Difference in Days (Full Year)**

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|
| -180.0 | -34.3 | -25.4 | -29.7 | -13.3 | 8.9 |

**Date *p*max* Achieved (March Only)**

| Species | Optimized | Fixed | Difference (Days) | Effort |
|---|---|---|---|---|
| No species achieved p*max below an effort of 10 samples | | | | |
| ECDO | 3/30/2016 | - | - | 10 |
| GAQU | 3/27/2016 | - | - | 10 |
| PHAI | 3/23/2016 | - | - | 10 |
| VERD | 3/28/2016 | - | - | 10 |
| BTGN | 3/25/2016 | - | - | 20 |
| ECDO | 3/18/2016 | - | - | 20 |
| GAQU | 3/14/2016 | 3/31/2016 | -16.7 | 20 |
| PHAI | 3/12/2016 | 3/25/2016 | -13.2 | 20 |
| VERD | 3/16/2016 | - | - | 20 |
| BTGN | 3/14/2016 | 3/30/2016 | -16.2 | 30 |
| COPO | 3/25/2016 | - | - | 30 |
| ECDO | 3/10/2016 | 3/26/2016 | -15.6 | 30 |
| GAQU | 3/10/2016 | 3/22/2016 | -12.3 | 30 |
| PHAI | 3/9/2016 | 3/18/2016 | -9.2 | 30 |
| VERD | 3/10/2016 | 3/23/2016 | -13.9 | 30 |
| BTGN | 3/11/2016 | 3/20/2016 | -9.2 | 40 |
| COPO | 3/19/2016 | 3/27/2016 | -7.8 | 40 |
| ECDO | 3/8/2016 | 3/17/2016 | -8.4 | 40 |

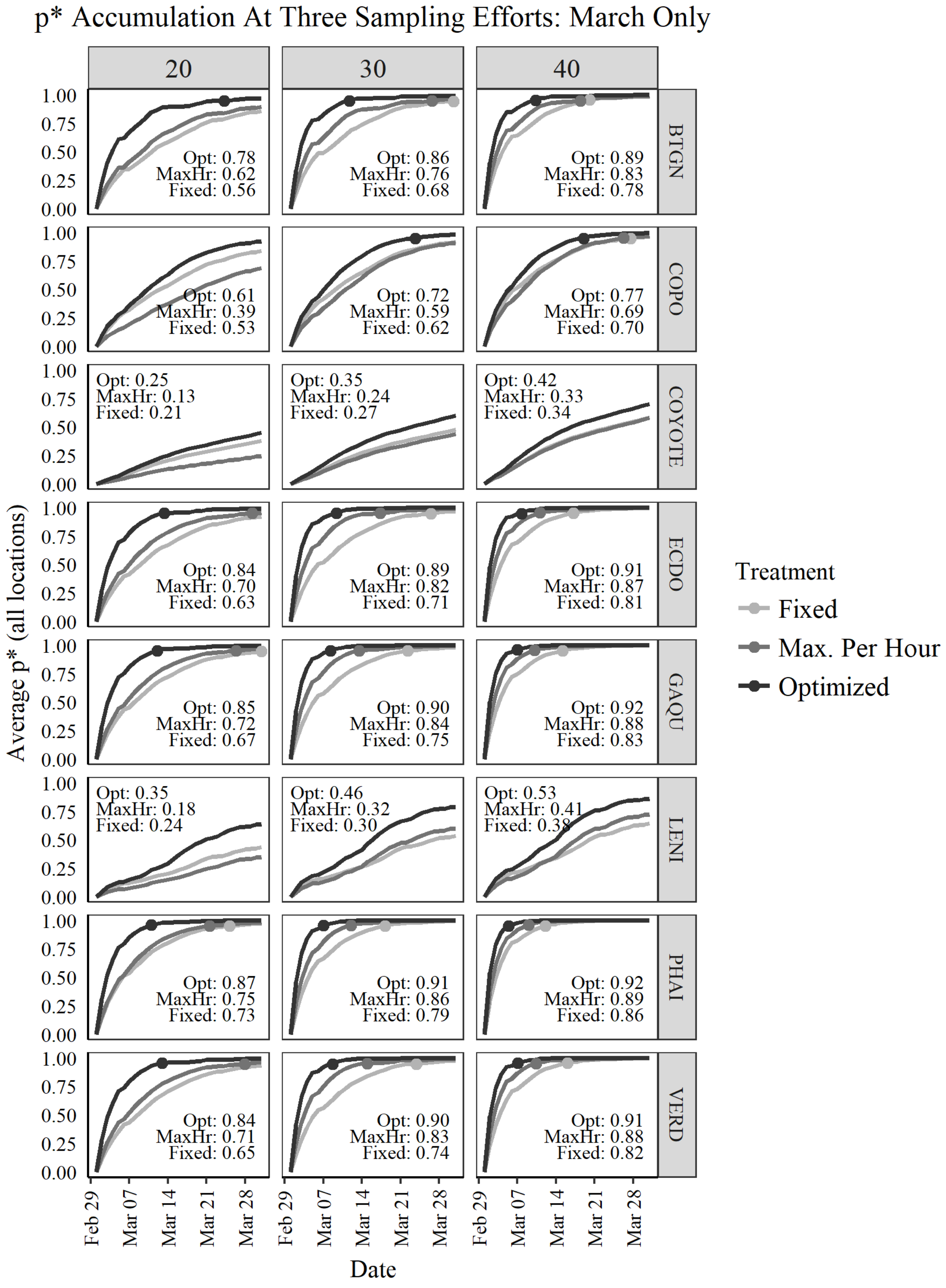
| | | | | |
|---|---|---|---|---|
| GAQU | 3/8/2016 | 3/15/2016 | -7.1 | 40 |
| PHAI | 3/7/2016 | 3/12/2016 | -5.2 | 40 |
| VERD | 3/8/2016 | 3/16/2016 | -8.0 | 40 |

**Summary Statistics for Difference in Days (March Only)**

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|
| -16.7 | -13.9 | -9.2 | -11.0 | -8.0 | -5.2 |

## A. 2. Comparison of Simple Optim vs. Max Per Hour vs. Fixed Schedule

In the *scheduleOptim()* function's 'max per hour' option, end users may specify the maximum allowable number of samples that may be distributed within any given hour. This option exists for end users who do not want all their sampling power allotted into a single hour. In an exploratory experiment, we used a study duration of March 2016 (31 days), and selected a maximum number of 10 samples allowable per hour. We looked at sampling efforts of $S$ = 20, 30 and 40 one-minute samples per day (because at $S$ = 2, 5, and 10, the simple optimization and 'max per hour' options will perform identically). We kept the fixed schedule in for comparison. The figure below demonstrates that the 'simple' optimized schedule method, which preferentially allocates sampling power into the highest scoring hour, outperforms the 'max per hour' optimization method in the simulation. In some cases, for nocturnal species (COPO, COYOTE, and LENI), the fixed schedule outperforms the 'max per hour' method on the $p$* AUC measurement.


p* Accumulation At Three Sampling Efforts: March Only
20
30
40
BTGN
COPO
COYOTE
ECDO
GAQU
LENI
PHAI
VERD
1.00
0.75
0.50
0.25
0.00
Average p* (all locations)
Date
Feb 29
Mar 07
Mar 14
Mar 21
Mar 28
Opt: 0.78 MaxHr: 0.62 Fixed: 0.56
Opt: 0.86 MaxHr: 0.76 Fixed: 0.68
Opt: 0.89 MaxHr: 0.83 Fixed: 0.78
Opt: 0.61 MaxHr: 0.39 Fixed: 0.53
Opt: 0.72 MaxHr: 0.59 Fixed: 0.62
Opt: 0.77 MaxHr: 0.69 Fixed: 0.70
Opt: 0.25 MaxHr: 0.13 Fixed: 0.21
Opt: 0.35 MaxHr: 0.24 Fixed: 0.27
Opt: 0.42 MaxHr: 0.33 Fixed: 0.34
Opt: 0.84 MaxHr: 0.70 Fixed: 0.63
Opt: 0.89 MaxHr: 0.82 Fixed: 0.71
Opt: 0.91 MaxHr: 0.87 Fixed: 0.81
Opt: 0.85 MaxHr: 0.72 Fixed: 0.67
Opt: 0.90 MaxHr: 0.84 Fixed: 0.75
Opt: 0.92 MaxHr: 0.88 Fixed: 0.83
Opt: 0.35 MaxHr: 0.18 Fixed: 0.24
Opt: 0.46 MaxHr: 0.32 Fixed: 0.30
Opt: 0.53 MaxHr: 0.41 Fixed: 0.38
Opt: 0.87 MaxHr: 0.75 Fixed: 0.73
Opt: 0.91 MaxHr: 0.86 Fixed: 0.79
Opt: 0.92 MaxHr: 0.89 Fixed: 0.86
Opt: 0.84 MaxHr: 0.71 Fixed: 0.65
Opt: 0.90 MaxHr: 0.83 Fixed: 0.74
Opt: 0.91 MaxHr: 0.88 Fixed: 0.82
Treatment
Fixed
Max. Per Hour
Optimized

# Appendix B

B. 1. Summary of acoustic features used as inputs to classification models that predict whether a detection is a true or false positive.

| **Feature** | **Description** |
|---|---|
| Raw amplitude values | Acquired by way of Fourier Transform. Every single raw amplitude value (in dB) in the matrix of a detected event. Each amplitude value is a measure of signal intensity at that point, and is rendered in colored shading on the spectrogram. |
| Correlation Score | Correlation score produced by moving window analysis during template matching. |
| Zero-Crossing Rate for each time bin | zcr = 0.5 * mean(abs(sgn(x(t+1)) - sgn(x(t))))<br>with: N the length of the signal x, and where: sgn(x(t)) = 1 if x(t) >= 0 and sgn(x(t)) = -1 if x(t) < 0. |
| Time Contours for each time bin | Amplitude probability mass function for each time bin |
| Frequency Contours for each frequency bin | Amplitude probability mass function for each frequency bin |
| Time.P1 | Time initial percentile based on cumulative distribution function generated from time probability mass function |

| | |
|---|---|
| Time.M | Time median based on cumulative distribution function generated from time probability mass function |
| Time.P2 | Time terminal percentile based on cumulative distribution function generated from time probability mass function |
| Time.IPR | Time interpercentile range based on cumulative distribution function generated from time probability mass function |
| Freq.P1 | Frequency initial percentile based on cumulative distribution function generated from frequency probability mass function |
| Freq.M | Frequency median based on cumulative distribution function generated from frequency probability mass function |
| Freq.P2 | Frequency terminal percentile based on cumulative distribution function generated from frequency probability mass function |
| Freq.IPR | Frequency interpercentile range based on cumulative distribution function generated from frequency probability mass function |
| Spectral Mean | Sum of the product of the spectrogram intensity (in dB) and the frequency, divided by the total sum of spectrogram intensity. |
| Spectral Standard Deviation | Standard deviation of the mean frequency |
| Spectral Median | The value of the halfway point in ordered frequency values in the data set |
| Spectral Mode | Dominant frequency of the amplitude matrix |

| | |
|---|---|
| Q1: First quartile (0.25 quantile) | The first quartile; a measure of statistical dispersion. Value that divides the lowest 25% of data from the highest 75%. |
| Q3: Third quartile (0.75 quantile) | The third quartile; a measure of statistical dispersion. Value that divides the highest 25% of data from the lowest 75%. |
| Interquartile range (IQR) | $IQR = Q3 - Q$. A statistical dispersion (variability) measure based on dividing the detected event into quartiles. |
| Spectral Centroid | $C = sum(x*y)$<br>with $x$ = frequencies, $y$ = relative amplitude of the $i$ frequency, and $N$ = number of frequencies. |
| Spectral Skewness | A measure of signal asymmetry.<br>$S = sum((x-mean(x))^3)/(N-1)/sd^3$<br>Spectrum asymmetry increases with \|S\|. |
| Spectral Kurtosis | A measure of signal peakedness.<br>$K = sum((x-mean(x))^4)/(N-1)/sd^4$ |
| Spectral Flatness | $F = N*(prod(y_i)^{(1/N)} / sum(y_i))$<br>With $y$ = relative amplitude of the $i$ frequency, and $N$ = number of frequencies. Ratio between geometric mean and arithmetic mean. Flatness of noisy signals are closer to 1; flatness of pure tone signal is closer to 0. |
| Spectral Entropy (Shannon's) | $S = -sum(ylogy)/log(N)$.<br>Noisy signals have S closer to one, while pure tone signals have S closer to 0. |

B. 2. Example of weighted average ensemble probability computation

1. For a single detection, take the vector of true positive class probabilities for all five classifiers, [P]:

   [P] = [p1, p2, p3, p4, p5]

   [P] = [0.02, 0.29, 0.20, 0.29, 0.09]

2. Gather each classifier's score on the metric of interest (e.g., Sensitivity) in vector [S]

   [S] = [0.86, 0.77, 0.00, 0.86, 0.73]

3. Compute a vector representing how proportionally close each score is to the highest score:

   [D] = [S] / max[S]

   [D] = [0.86, 0.77, 0.00, 0.86, 0.73] / 0.86 = [1.00, 0.895, 0.00, 1.00, 0.849]

4. Compute a vector of weights normalized to add to 1, [N]:

   [N] = [D] / sum([D])

   [N] = [1.00, 0.895, 0.00, 1.00, 0.849] / 3.74 = [0.27, 0.24, 0.00, 0.27, 0.23]

5. Compute dot-product of the vector of probabilities [P] times the vector of normalized weights [N] to get a single weighted-average value for the class probability, $p_W$.

$p_w = [P] \bullet [N] = [0.02, 0.29, 0.20, 0.29, 0.09] \bullet [0.27, 0.24, 0.00, 0.27, 0.23] = (0.02*0.27) + (0.29*0.24) + (0.20*0.00) + (0.29*0.27) + (0.09*0.23) =$ **0.17**

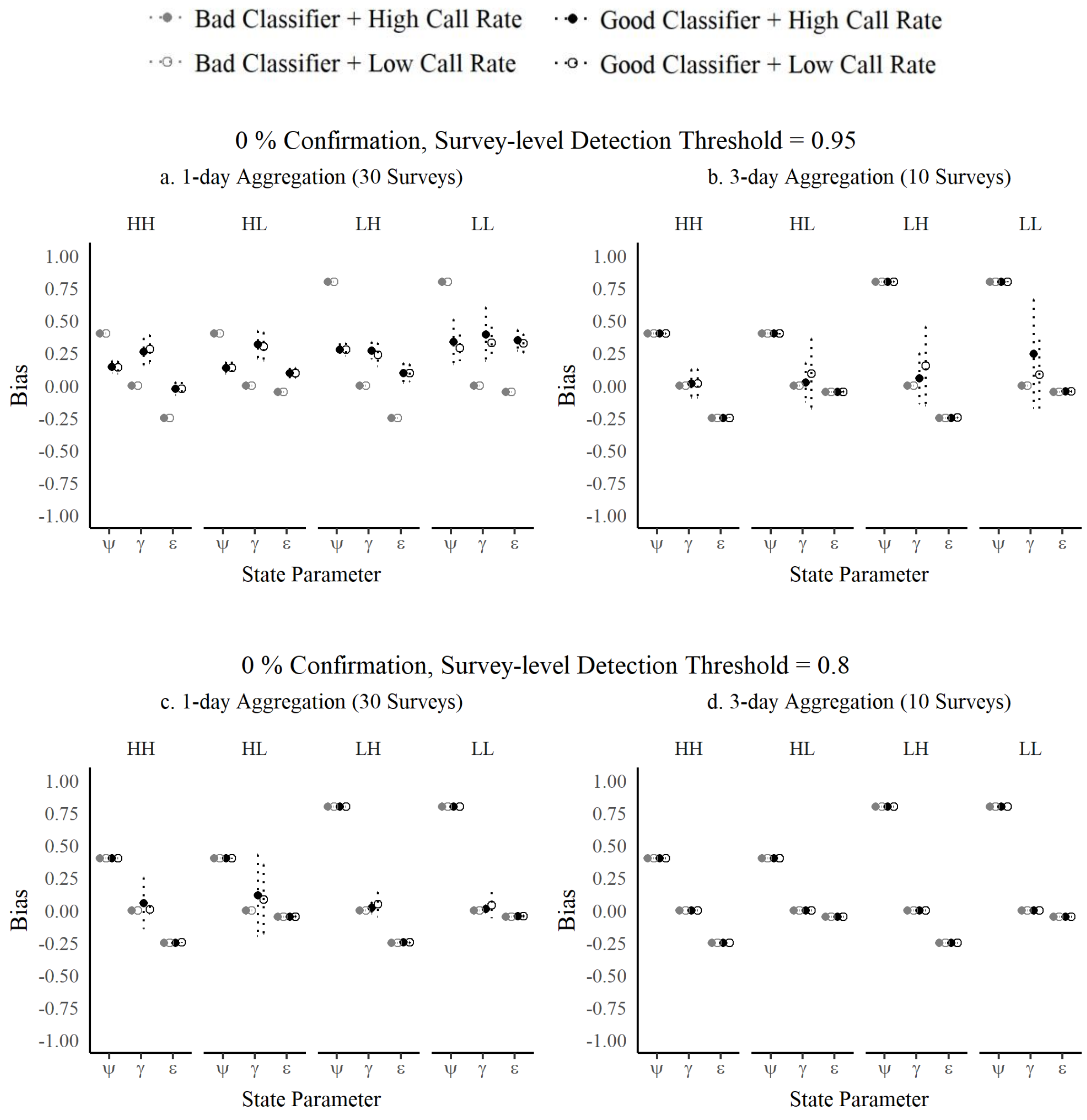
6. For $p_w < 0.5$, class = false alarm. For $p_w > 0.5$, class = target signal. If ties, coinflip for class.

   **$p_w = 0.17$ = false alarm class**

## Appendix C

Performance of the classic dynamic occupancy model that ignores false positives (Mackenzie *et al.* 2003) in a 100-replicate experiment. Summary of parameter estimate bias across occurrence dynamics, species call rates, classifier performance, aggregation frames, survey-level detection thresholds, and confirmation percentages. Circles indicate the mean bias, with dotted vertical bars showing standard deviations. Open circles denote scenarios with a low call rate. Closed circles denote a high call rate. Gray circles denote the bad classifier, and black circles denote the good classifier.

Performance of the classic (false positive-ignorant) dynamic model in a 100-replicate experiment: state parameter results.



0 % Confirmation, Survey-level Detection Threshold = 0.95

a. 1-day Aggregation (30 Surveys)

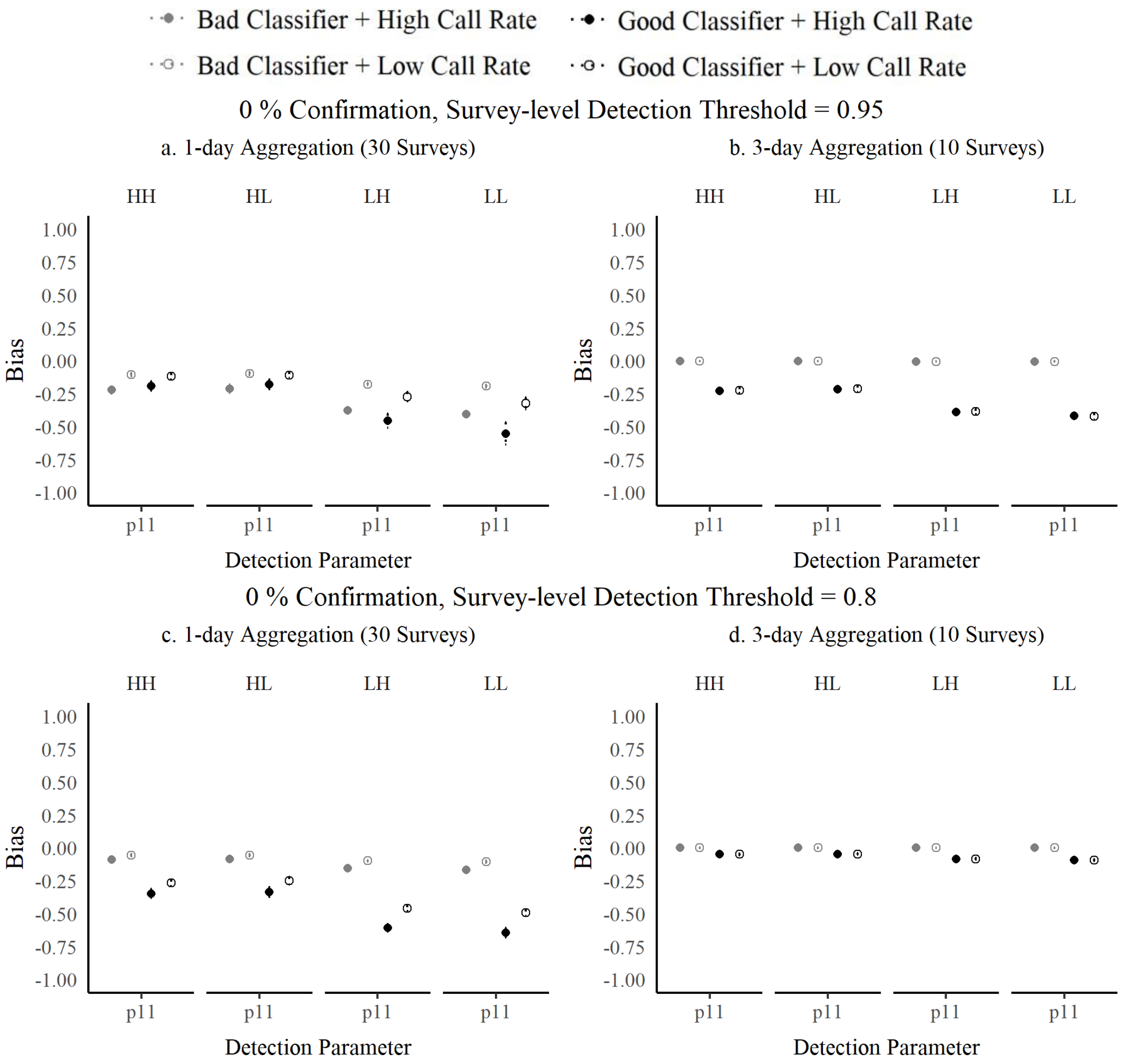
b. 3-day Aggregation (10 Surveys)

0 % Confirmation, Survey-level Detection Threshold = 0.8

c. 1-day Aggregation (30 Surveys)

d. 3-day Aggregation (10 Surveys)

Performance of the classic (false positive-ignorant) dynamic model in a 100-replicate experiment: detection parameter results.


Bad Classifier + High Call Rate
Good Classifier + High Call Rate
Bad Classifier + Low Call Rate
Good Classifier + Low Call Rate
0 % Confirmation, Survey-level Detection Threshold = 0.95
a. 1-day Aggregation (30 Surveys)
b. 3-day Aggregation (10 Surveys)
0 % Confirmation, Survey-level Detection Threshold = 0.8
c. 1-day Aggregation (30 Surveys)
d. 3-day Aggregation (10 Surveys)
HH
HL
LH
LL
Bias
p11
Detection Parameter