BRINGING FINER RESOLUTION TO WILDLIFE MONITORING: ACCOUNTING FOR MISPERCEPTION AND UNCOVERING SEASONAL VARIATION IN ECOLOGICAL PROCESSES

By

John D. J. Clare

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Forestry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2020

Date of final oral examination: 12/8/2020

The dissertation is approved by the following members of the Final Oral Committee:

Philip Andrew Townsend, Professor, Forest and Wildlife Ecology

Benjamin Zuckerberg, Associate Professor, Forest and Wildlife Ecology

Timothy Van Deelen, Professor, Forest and Wildlife Ecology

Jonathan Pauli, Associate Professor, Forest and Wildlife Ecology

Volker Radeloff, Professor, Forest and Wildlife Ecology

Jennifer Stenglein, Wisconsin Department of Natural Resources

Abstract

BRINGING FINER RESOLUTION TO WILDLIFE MONITORING: ACCOUNTING FOR MISPERCEPTION AND UNCOVERING SEASONAL VARIATION IN ECOLOGICAL PROCESSES

John D. J. Clare

Under the supervision of Professor Philip Andrew Townsend and Associate Professor Benjamin Zuckerberg

At the University of Wisconsin-Madison

We have entered the age of ecological macro and micromonitoring, as continued developments in tagging and remote sensing technologies provide ecologists with vast volumes of data at unprecedented combinations of scale and resolution. Yet although data collection is increasingly unbound, leveraging it to improve conservation and management implementation still poses challenges. Data volume often forces researchers to consider trade-offs between veracity (i.e., the accuracy of the data) and processing speed (i.e., how quickly data can be used). The extent and resolution of new sampling techniques can uncover new ecological patterns, but the novelty of such patterns can make it difficult to conceptualize useful models to describe them. Here, I try to take steps towards surmounting some of these issues.

Chapters 1 and 2 focus on issues of data veracity. Although measurement and classification error are ubiquitous in ecological data, these problems have become more visible as researchers increasingly depend upon algorithms or recruited volunteer scientists to perform data collection and classification tasks. Chapter 1 describes a general framework to guide researchers undertaking data quality assessments and implementing remediation actions that is rooted in ecological inference rather than error incidence. Chapter 2 focuses on expanding the statistical tool-kit that ecologists can use to account for misclassified detection/non-detection data, and demonstrates that previously developed approaches focusing on occupancy estimation are easily extensible to essentially any parametric model class reliant on species occurrence data.

Chapters 3 and 4 focus on longer-standing ecological questions, but bringing new seasonal resolution to bare. Chapter 3 focuses on quantifying deer behavioral responses to predation risk, and seeks to disentangle competing hypotheses for how such responses are structured. Deer respond to proximal measures of potential wolf predation risk in ways that might be expected to have cascading vegetation effects in some environmental contexts, but not others. In particular, deer responses were strongly mediated by seasonal environmental variables, suggesting a potential 'phenology of fear'. Chapter 4 seeks to delineate wildlife communities and uncover the primary environmental factors that structure their occurrences. Snow appears to be a particularly powerful driver of species distributions, and wildlife responses to changes in snow-depth and vegetation greenness across the year drive distinct seasonal variation in patterns of species richness: such temporal variation (or partitioning of the "seasonal" niche) may play a key role in maintaining community diversity.

Acknowledgments

My co-advisors Phil Townsend and Ben Zuckerberg set this degree and dissertation in motion and played key roles in its development. Feedback from them and my other committee members—alphabetically, Jon Pauli, Volker Radeloff, Jen Stenglein, and Tim Van Deelen—was constructive, rigorous (but humane!), and greatly improved the focus and content of the work here. I have genuinely enjoyed interacting with and learning from each of you. Also, being a person whom is admittedly…uh, somewhere between cranky and not at all enjoyable to be around, great thanks for your patience.

My dissertation is but one small piece of a much larger WNDR project, Snapshot Wisconsin. Innumerable people have played critical roles in its implementation and the end production of the data used herein. Without the efforts and vision of Jen, Phil, Ben, Tim, Christine Anhalt-Depies, Christina Locke, Aditya Singh, Nanfeng Liu, Young Lee, Vivek Malleshappa, Susan Frett, Sarah Cameron, Emily Buege Donavan, Taylor Peltier, Neil Gilbert, Karl Martin, probably many other individuals at WDNR, and thousands of volunteers within Wisconsin and across the world, neither the project nor any of the work here would have been possible. Chrissy Grebe, Mirelle Goetz, Meredith McLoon, Sawyer Boldt, Lena Carlson, and Jake Berger did fantastic work collecting and processing data.

I also wish to thank the graduate students/personnel within the Townsend and Zuckerberg labs and the Department of Forest and Wildlife Ecology for making this whole experience enjoyable. There are too many to fully name, but let me raise particular cheers to: Jen Cruz, Mike Hardy, Beckett Hills, Alyse Krueger, Chris Latimer, Marie Martin, Colleen Miller, Larry Warner, Michael Wheeler, Evan Wilson, and Connor Wood (and their respective partners and pet companions); to those from thousands of miles and what feels like decades ago that I did not end up getting the opportunity to thank (Brian Allen, Samantha Brown, Erik Blomberg, Mo Correll, Lisa Izzo, Dan Linden, Ellie Mangelinckx, George Maynard, Brian Rolek, Kate Ruskin, Lindsay Seward, Jonathan Watson); to the many others whom put in time on the Nat or NB courts (shouts to Phil Manlick, Brendan Hobart, Elena Razenkova, Isabel Rojas), or at other establishments of varying repute; to those who (for reasons that remain unclear) thought I

might have something to contribute to their own varied research efforts (Neil Gilbert, Gavin Jones, Lucas Olson); and to the many others whom were happy to share their collective wisdom (Tedward Erker, Monika Shea, Diana Guzman-Colon, Omar Ohrens, Adam Chlus, Johanna Buchner, Ben Spaier, Sarika Mittra). Many people here and above belong in multiple classes. Be well.

Thanks to my genetic relatives, who exhibited great patience throughout this process and have had the decency (wisdom? indifference?) to avoid asking questions that would be tedious to answer, while otherwise being incredibly supportive.

Saving the largest thanks for last, I could not have completed this without the incredible support of my immediate family and two favorite beings, Jenny and Wren McCabe. From coast to coast, living with you each has been the brightest spot in each day. As the lyric goes, I can't help it if I'm lucky.

Table of Contents

Introduction

The past decade has born witness to massive changes in the nature of ecological data collection. With increases in sensing capacity driven by enhanced technology, the widespread recruitment of volunteer scientists, and increased collaboration and dedication to data-sharing, previously unparalleled capacity to describe biological patterns across massive spatial extents (e.g., nations, continents) and incredibly fine spatiotemporal resolutions (meters, days, hours, minutes) has become fairly routine. Capacity to capture biological nuance has also increased: for example, airborne and spaceborne sampling efforts can monitor canopy leaf chemistry or arctic foodwebs, while groups of dedicated volunteers can track the irregular migration paths of single organisms.

The excitement regarding the degree to which increased volume, extent, and resolution of biological data provided by automated recording units, volunteer scientists, and air or space-borne platforms can improve biodiversity conservation and management efforts partially derives from existing monitoring limitations. One of the most pronounced challenges for (and criticisms of) conservation and management is generating information that can reliably guide decisions (Aceves-Bueno et al. 2015, Artelle et al. 2018): nearly all traditional monitoring platforms suffer gaps across space, time, or taxa and are constrained to employ specific and potentially non-optimal spatial, temporal, and biological resolutions (Jetz et al. 2019). Such constraints propagate uncertainty across several steps of the decision-making process ranging from recognizing the need to make decisions (e.g., realizing a population is in decline) through specifying and optimizing potential actions (e.g., selecting an effective action, Fuller et al. 2020, Wright et al. 2020). Using and networking new platforms provides opportunity to fill these information gaps and enhance the exent and resolution of biological monitoring, and potentially, conservation and management decision-making (Turner 2014, Townsend et al. 2020).

Unsurprisingly, agencies are increasingly using large and open remotely sensed biodiversity datasets (Sullivan et al. 2009, Ahumada et al. 2019) or developing their own platforms. One example of the latter is Snapshot Wisconsin (Locke et al. 2019, Townsend et al. 2020), a monitoring program run by the Wisconsin Department of Natural Resources in which volunteers are recruited to deploy and maintain

trail cameras across the state of Wisconsin and classify images on a combination of crowdsourced and agency-developed web-applications. Snapshot Wisconsin has proven to be a massive success with respect to volunteer recruitment and data collection, generating millions of classified images for use in support of management decision-making, and is a powerful demonstration that agencies can tractably develop and manage such programs.

Despite optimism, observation network approaches to biological monitoring pose their own technical and conceptual challenges. Although informatics—capacity to classify, store, collate, and otherwise make data analysis-ready—can pose severe barriers and receive the majority of attention from practitioners, other issues are equally germane (e.g., Lindenmayer et al. 2018, LaSorte et al. 2018, Bayraktarov et al. 2019). Data volume and uncertainty with respect to sampling parameters and data veracity mandate the development of statistical models that are efficient but also robust to cryptic but ubiquitous types of sampling, measurement, and process error. Because larger datasets can often be to fit models of considerable complexity, there is further need to ensure that model-based inference remains accessible and interpretable. Finally, improved ability to sample ecological patterns and process often reveals considerable conceptual limitations: what is a useful model for a phenomena that has not previously been studied?

The focus of this dissertation is to take steps towards addressing some of these challenges using Snapshot Wisconsin as a template for other 'big data' biological monitoring programs. Two chapters focus on technical challenges, and two focus on using the project's larger sampling frame to make insights into previously ecological patterns.

The first chapter focuses on a mixture of informatics and statistical challenges. Traditional assessments of ecological study design have focused upon optimizing sampling to, for example, maximize statistical power and minimize expenditure (e.g., Clare et al. 2015). These considerations are often less germane for sensor or volunteer-based monitoring, where increasing sampling effort may induce little added expense (or indeed, sampling effort fall beyond researcher control). More pertinent questions focus on optimizing combinations of data veracity and the speed with which data is assimilated.

Concerns about elevated rates of measurement or classification error associated with volunteer-based data collection and algorithm-driven data interpretation have motivated many researchers to quantify the rate or incidence of such errors in empirical data. Yet commonly, these efforts fail to adequately define the inferential costs of the estimated error incidence, which makes it essentially impossible to define how accurate data need to be and whether existing data are sufficient or not. We lay out a repeatable framework for jointly assessing and remediating data errors based upon targets grounded in ecological inference rather than relatively uninterpretable error rates. We found that baseline error incidence associated with the species classification of Snapshot Wisconsin images renders estimation using occupancy models unreliable, but that the application of either design or model-based approaches to mitigate false-positive error permits acceptable inference. In doing so, we further found that attributes of different species may better explain errors than attributes of the classification task or platform: in particular, volunteers appeared predisposed to falsely identify rare species. Simulation results and assessements of previous efforts to quantify the indidence of species misclassification suggest that many (perhaps most) datasets likely require error remediation to estimate species distributions without bias.

Over the course of analyzing and writing this first chapter, it became clear that model-based solutions to misclassification/measurement error (i.e., explicitly modeling classification uncertainty within the statistical model of interest) were the most effective and least effort-intensive way to reduce biases. Design-based solutions (i.e. approaches focused on improving baseline classification performance or identifying and censuring problematic data prior to analysis) often require iterative steps including with assessing sensitivity to error, implementing the treatment, and assessing its efficacy. As ecological modeling efforts become increasingly bespoke and ambitious, this iterative process can become burdensome. Yet by and large, ecological misclassification models were largely constrained to occupancy estimation.

In chapter 2, we describe ways in which several types of misclassification models developed for binary data within occupancy models (Miller et al. 2011, Chambert et al. 2015) to account for false-positive and false-negative errors are easily extensible across several model classes that employ this sort

of data. Using simulation, we demonstrate that parameters other than occupancy probabilities are biased by relatively small incidences of false positive error, that our proposed extensions greatly reduce bias even with relatively small amounts of ancillary (correctly classified) data, and that when false positive error has relatively low incidence, model misspecification of the false positive process can have limited consequence. The implication is that many studies (even those using potentially complex and bespoke integrated models) can account for false positive error in detection/non-detection or presence/background data relatively easily, although model-based solutions more finely-tuned to the specific sampling procedures and misclassification process are likely to be optimal.

Chapter 3 delineates a break from the technical focus of the previous chapters. Here, we evaluated white-tailed deer behavioral responses to metrics of potential predation risk. Motivated by conflicting expectations based upon previous empirical evidence and existing theory, we assessed support for 4 non-exclusive hypotheses for the system: that it is bottom-up driven (Ford and Goheen 2015), that deer exhibit weak and environmentally homogenous responses to active predators in the system (Schmitz 2004), that deer exhibit stronger and more environmentally heterogeneous responses to risk because predators in the system tend to practice hunting modes more closely aligned with stalking as a result of Wisconsin's largely forested landscapes (Flagel et al. 2016), or whether predators in the system, while exhibiting active hunting modes, tend to target specific linear features or other landscape attributes in ways that would produce more concentrated risk cues and stronger deer responses.

We found that deer responses to metrics of potential predation risk (primarily from wolves) were environmentally heterogeneous, implying limited support for the first two hypotheses. Wolves reduced the intensity of deer use more strongly in areas with greater surrounding forest cover, but wolves themselves not appear to visit such areas any more frequently. Thus, while wolf hunting strategies do not appear to center on seeking cover that might facilitate stalking, there was some evidence that deer percieved wolves as more dangerous in these areas. Instead, wolf occurrence patterns suggest that their hunting strategies focus upon using linear features that enable faster movement and presumably increase the likelihood of prey encounter (Dickie et al. 2016, Dickie et al. 2020), and deer allocated less time to

foraging in these locations. Moreover, deer avoidance of wolves was mediated by changes in snow depth and vegetation greenness, and graphical exploration of the effects suggest a seasonal cycle in in deer responses.

These findings have several implications. First, they suggest that, as suggested by previous studies (Callan et al. 2013, Flagel et al. 2016), wolves could be enacting a non-consumptive tri-trophic cascade across the Great Lakes region. More broadly, they suggest that definitions of predator hunting mode and habitat domain may deserve more careful definition, and that predator hunting mode, per se, may not be a particularly powerful predictor of prey responses to predation risk without considering environmental context. Finally, seasonal variation in deer responses suggest an emergent line of research. Given the degree to which seasonality impacts the energetic landscape, and organism foraging decisions, life-histories, and physical states, 'phenologies of fear'might be a widely realized phenomena.

The final chapter focuses on understanding seasonal shifts in communities and elucidating species responses to a broad set of factors (snow conditions, plant phenology, land cover, and nighttime lights/urbanization) undergoing global changes. We use spatiotemporal multi-species occupancy models to estimate environmental associations, predict species distributions across the year, and generate insights into general community patterns. We find that only a few species (primarily organisms that practice torpor) exhibit pronounced broad-scale variability in their distributions over the year. Despite this, species richness was spatiotemporally variable (tending to peak in spring and again in fall), largely because species occupancies were driven by variation both static in snow depth and vegetation greenness. Indeed, across the community, snow depth had the most substantial (and generally negative) effect on species distributions. On average, species negatively impacted by snow were less negatively or positively associated with vegetation greenness, suggesting Wisconsin's mammalian (and gallanecuous bird) community might be primarily defined along an axis describing seasonal adaptations.

Given this, it is unsurprising that patterns in species distributions largely suggest a mix of primarily northerly or primarily southerly located organisms, and that annually integrated patterns in species richness suggest most species in parts of central Wisconsin. However, annually integrated species

richness was greater in areas with more pronounced variation between growing season and winter richness, and tended to less in areas where, on average, there was less seasonal variation in richness and tended to be greater richness at any given time.

In sum, Wisconsin's wildlife communities may be more strongly structured by seasonal variation than by other environmental factors. This adds to a growing body of evidence highlighting climate change as the primary driver of biotic change, and suggests that conservation and management bodies within Wisconsin should prioritize actions aimed at assessing biotic vulnerability to altered climatic conditions and developing strategies to mitigate or adapt to pending effects. A key driver of climatically-driven community shifts may relate to the breakdown of factors that seasonally partition species niches: such partitioning may be criticial for maintaining richer communities.

The dissertation below presented as a series of articles for publication in different scientific journals; redundant information in introductory sections and any differences in formatting are intentional and conform to journal-specific standards.

References

Aceves-Bueno, E., A. S. Adeleye, D. Bradley, W. T. Brandt...et al. 2015. Citizen science as an approach for overcoming insufficient monitoring and inadequate stakeholder buy-in in adaptive management:criteria and evidence. Ecosystems 18:493-506.

Ahumada, J. A., E. Fegraus, T. Birch, N. Flores...e al. 2020. Wildlife Insights: a platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. Environmental Conservation 47:1-6.

Artelle, K. A., J. D. Reynolds, A. Treves, J.C. Walsh, P. C. Paquet, and C. T. Darimont. 2018. Hallmarks of science missing from North American wildlife management. Science Advances 4:eaao0167.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E.L., Nguyen, H.A., McRae, L., Possingham, H.P. and Lindenmayer, D.B., 2019. Do big unstructured biodiversity data mean more knowledge? Frontiers in Ecology and Evolution 6: art 239.

Bonney, R., C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk. 2009. Citizen Science: a developing tool for expanding science knowledge and scientific literacy. BioScience 59:977-984.

Chambert, T., D.A.W. Miller, and J. D. Nichols. 2015. Modeling false positive detections in species occurrence data under different study designs. Ecology 96:332-339.

Callan, R., N. P. Nibbelink, T. P. Rooney, J. E. Wiedenhoeft, and A. P. Wydeven. 2013.

Recolonizing wolves trigger a trophic cascade in Wisconsin (USA). Journal of Ecology 101:837-845.

Dickie, M., R. Serrouya, R. S. McNay, and S. Boutin. 2017. Faster and farther: wolf movement on linear features and implications for hunting behavior. Journal of Applied Ecology 54:253-263.

Dickie, M., S. R. McNay, G. D. Sutherland, M. Cody, and T. Avgar. 2020. Corridors or risk? Movement along, and use of, linear features varies predictably among large mammal predator and prey species. Journal of Animal Ecology 89:623-634.

Flagel, D. G., G. E. Belovsky, and D. E. Beyer, Jr. 2016. Natural and experimental tests of trophic cascades: gray wolves and white-tailed deer in a Great Lakes forest. Oecologia: 180:1183-1194.

Ford, A. T., and J. R. Goheen. 2015. Trophic cascades by large carnivores: a case for strong inference and mechanism. Trends in Ecology and Evolution 30:725-735.

Jetz, W., M. A. McGeoch, R. Guralnick, S. Ferrier,..et al. 2019. Global biodiversity variables for mapping and monitoring species populations. Nature Ecology and Evolution 3:539-551.

La Sorte, F. A. , C. A. Lepcczyk, J.L Burnett, A. H. Hurlbert, M. W. Tingley, and B. Zuckerberg. 2018. Opportunities and challenges for big data ornithology. Condor 120:414-426.

Lindenmayer, D. B., G. E. Likens, and J. F. Franklin. 2018. Earth observation networks (EONs): finding the right balance. Trends in Ecology and Evolution 33:1-3.

Locke, C.M., C.M. Anhalt-Depies, S. Frett, J. L. Stenglein, S. Cameron, V. Malleshappa, T. Peltier, B. Zuckerberg, and P.A. Townsend. Managing a large citizen science project to monitor wildlife. Wildlife Society Bulletin 43: 4:10.

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

Palmer, M. S., J. Fieberg, A. Swanson, M. Kosmala, and C. Packer. 2017. A ‘dynamic’ landscape of fear: prey responses to spatiotemporal variation in predation risk across the lunar cycle. Ecology Letters 20:1364-1373.

Schmitz, O. J., V. Krivan, and O. Ovadia. 2004. Trophic cascades: the primacy of trait-mediated indirect interactions. Ecology Letters 7:153-163.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282-2292.

Steenweg, R., M. Hebblewhite, R. Kays, J. Ahumada, ...et al. 2017. Scaling up camera traps: monitoring the planet’s biodiversity with networks of remote sensors. Frontiers in Ecology and the Environment 15:26-34.

Townsend, P. A., J. Clare, N. Liu, J. L. Stenglein, C. Anhalt-Depies, T. R. Van Deelen, N. A. Gilbert, A. Singh, K. J. Martin, and B. Zuckerberg. Integrating remote sensing within

jurisdictional observation networks to improve the resolution of ecological management. doi: 10.1101/2020.06.08.140848.

Turner, W. 2014. Sensing biodiversity. Science 346: 301-302.

Chapter 1 -Making inference with messy (citizen science) data: when are data accurate enough and how can they be improved?

John D. J. Clare[1], Philip A. Townsend[1], Christine Anhalt-Depies[1], Christina Locke[2], Jennifer L. Stenglein[2], Susan Frett[2], Karl J. Martin[3], Aditya Singh[1], Timothy R. Van Deelen[1], and Benjamin Zuckerberg[1].

[1]*Department of Forest and Wildlife Ecology, University of Wisconsin – Madison, Madison, Wisconsin*

[2]*Office of Applied Sciences, Wisconsin Department of Natural Resources, Madison, Wisconsin*

[3]*Division of Cooperative Extension, University of Wisconsin Extension, Madison, Wisconsin*

Citation:

Clare, J. D. J, P. A. Townsend, C. Anhalt-Depies, C. Locke, J. L. Stenglein, S. Frett, K. J. Martin, A. Singh, T. R. Van Deelen, and B. Zuckerberg. 2019. Making inference with messy (citizen science) data: when are data accurate enough and how can they be improved? Ecological Applications 29:e01849.

Abstract

Measurement or observation error is common in ecological data: as citizen scientists and automated algorithms play larger roles as processors of growing volumes of data to address problems at large scales, concerns about data quality and strategies for improving it have received greater focus. However, practical guidance pertaining to fundamental data quality questions for data users or managers—how accurate do data need to be and what is the best or most efficient way to improve it?—remains limited. We present a generalizable framework for evaluating data quality and identifying remediation practices, and demonstrate the framework using trail camera images classified using crowdsourcing to determine acceptable rates of misclassification and identify optimal remediation strategies for analysis using occupancy models. We used expert validation to estimate baseline classification accuracy and simulation to determine the sensitivity of two occupancy estimators (standard and false-positive extensions) to different empirical misclassification rates. We used regression techniques to identify important predictors of misclassification and prioritize remediation strategies. More than 93% of images were accurately classified, but simulation results suggested that most species were not identified accurately enough to permit distribution estimation at our predefined threshold for accuracy (< 5% absolute bias). A model developed to screen incorrect classifications predicted misclassified images with > 97% accuracy—enough to meet our accuracy threshold. Occupancy models that accounted for false positive error provided even more accurate inference even at high rates of misclassification (30%). As simulation suggested occupancy models were less sensitive to additional false-negative error, screening models or fitting occupancy models accounting for false positive error emerged as efficient data remediation solutions. Combining simulation-based sensitivity analysis with empirical estimation of baseline error and its variability allows users and managers of potentially error-prone data to identify and fix problematic data more efficiently. It may be particularly helpful for "big data" efforts dependent upon citizen scientists or automated classification algorithms with many downstream users, but given the ubiquity of

observation or measurement error, even conventional studies may benefit from focusing more attention upon data quality.

Introduction

Applied ecologists increasingly study phenomena and tackle problems occurring at massive spatial scales (e.g., LaSorte et al. 2018). This shift has partially been driven by increased rates of data collection provided by citizen scientists or automated recording devices (Sullivan et al. 2009, Steenweg et al. 2017), increased capacity to store and share data (e.g., Bonney et al. 2009), and new tools to process growing volumes of data more quickly (Swanson et al. 2016, Nourrouzzadeh et al. 2018). Reassigning data processing or classification previously performed by trained experts to groups of volunteers or machine-learning algorithms can be time and cost-effective, but potentially introduces additional measurement or observation error that can result in biased or more uncertain inference (Dickinson et al. 2010, Gardiner et al. 2012, Kosmala et al. 2016, McShea et al. 2016, Abra et al. 2018). Ensuring sufficient data quality is an intrinsic component of most automated data processing efforts and established citizen science programs (Bonter et al. 2012, Kosmala et al. 2016), and carries important consequences for broad-scale biodiversity and ecological monitoring (Gardiner et al. 2012).

To improve the quality of data processed by either citizen scientist or machine-learning algorithms, practitioners can choose from a few general approaches. Practitioners can reduce the complexity of the classification task or, more specific to citizen scientists, alter the classification interface (Kosmala et al. 2016). They can attempt to improve baseline performance by altering training protcols, like providing an algorithm or volunteer a larger pool of data to learn from, or increasing the number of parameters that an algorithm uses for classification (Nourrouzzadeh et al. 2018, Tabak et al. 2018). They can attempt to manipulate data accuracy after collection or classification, by, for example, determining indicators of unreliable data so that it can be censured from further analysis (Alldredge et al. 2007, Bonter et al. 2012, Swanson et al. 2016). Implementing these actions and evaluating their success generally requires having reference data (produced by expert verification or under controlled experimental settings)

to gauge accuracy (Crall et al. 2011, Miller et al. 2015). Finally, researchers can use more sophisticated analyses that explicitly account for additional sampling error types or sources of error variability. These can be parameterized by assuming error types or variability exist without explicit knowledge of their structure (Royle and Link 2006, Bird et al. 2014), or parameters may be informed by the results of an experimental evaluation exercise or post hoc evaluation (Chambert et al. 2015, Ruiz-Gutiérrez et al. 2016).

While data quality assurance is an important component of any ecological investigation, most empirical evaluations exhibit two major limitations. First, most studies that evaluate data quality use estimates of measurement error or changes in measurement error within the raw data to quantify baseline quality or the improvements induced by an intervention (but see Gardiner et al. 2012, Butt et al. 2013). However, the impetus for improving data quality is not to produce better or more accurate data for its own sake, but to improve ecological inferences made after analyzing the data. Metrics reported by many applications (e.g., misclassification rates, measurement variance) describe how accurate a given dataset is or has become, but do not necessarily effectively describe how *useful* it or has become for addressing focal questions. We contend that data quality is better conceptualized as a mixture of data accuracy and planned analyses, and should thus be defined as a threshold for accuracy that allows one to achieve a specific analytical objective.

A second limitation, more specific to citizen science, is a focus on the efficacy of a single method for improving data quality (e.g., data screening, Bonter et al. 2012, Swanson et al. 2016; considering alternative analysis structures, Isaac et al. 2014) rather than considering multiple approaches (e.g., improving volunteer proficiency vs. using a more complex statistical model). Evaluating data quality carries costs associated with expert verification or experimental calibration, and evaluating subsequent remediation actions require further resources. Identifying potential action or actions with a strong likelihood of success is critical for efficiently achieving and maintaining data quality. Many citizen science projects employ multiple methods for ensuring data quality (Wiggins and Crowston 2015), which

both suggests that many projects currently have data that could be used to rank or prioritize potential remediation actions, and that many projects may be implementing inefficient remediation actions.

These specific limitations can be summarized as questions that ecologists increasingly feel pressure to address when leveraging big data: how accurate does a dataset need to be and what is the best way to remediate existing error? As project managers attempting to implement a citizen science project designed to support natural resource management decisions, we found that although there was a great deal of literature that described specific components of data quality or remediation (Bird et al. 2014, Lewandowski and Specht 2015, Ruiz-Gutiérrez et al. 2016, Swanson et al. 2016) or highlighted the general importance of these concepts (Kosmala et al. 2016), guidance for pragmatic implementation was limited. As researchers wishing to use the data, we wanted to ensure that the questions we (or future downstream users) wished to ask could be reliably answered.

We present a generalizable framework for evaluating data quality and data remediation practices and apply it to improve the design and implementation of a broad-scale survey and monitoring effort using camera trap data collected and classified by citizen scientists. Our goals were to determine baseline data accuracy, determine data quality by evaluating how current levels of misclassification influenced species distribution inferences made using occupancy models (MacKenzie et al. 2002), and evaluate the potential efficacy of alternative strategies that might be employed to improve inferences. Our framework integrates the needs of project managers, data curators, analysts, and ecologists into a complete platform for assessing data quality.

Methods

Framework

A complete data and remediation evaluation process will generally follow a six-step sequential framework (Figure 1). To evaluate data, investigators must 1) define desired data quality explicitly in terms of study objectives grounded in specific analyses or estimates, 2) estimate existing levels of

accuracy or error within the dataset, and 3) estimate a requisite level of accuracy or error within the raw data that allows study objectives to be achieved. Remediation evaluation includes 4) identifying possible actions, 5) exploring important sources of variation in error within a dataset to target a specific action or set of actions to evaluate, and 6) implementing and evaluating candidate actions to determine whether any meet the defined data quality objective. This process is likely to be iterative and adaptive over the duration of the study (Kosmala et al. 2016). Below, we describe each step and our implementation in more detail.

Defining data quality

We view the definition of data quality as the most fundamental component of any evaluation process. It requires investigators to codify research objectives and planned analyses: what is to be estimated, how is it going to be estimated, and how well does it need to be estimated? These decisions are analogous to standard study design decisions (e.g., choosing type I vs. type II errors) and will largely depend upon project goals and how specific components of any estimation process are prioritized or weighted.

The Wisconsin Department of Natural Resources (WDNR) implemented Snapshot Wisconsin to support wildlife management decision-making by documenting rare or endangered species and providing information about the spatial and temporal population variability in species of managerial interest. There are several distinct analyses that are likely to be used to accomplish project objectives, we treat occupancy estimation as our planned analysis for evaluating data quality, as it is of direct interest for rare or incidental species, can provide insights into spatial variation in population size for low-density and solitary species (Linden et al. 2017), and may provide information about population changes over time if certain assumptions hold (Ellis et al. 2014).

We defined adequate data (or adequate data improvement) as that which permitted us to estimate occurrence probability with less than 5% absolute expected bias and less than 10% root-mean-square error given that an occupancy model was correctly parameterized. We defined a second, less stringent,

definition of adequacy as being able to correctly estimate the directional effect of important predictors (here, predictors with log-odds effect of 1 or 2 per sd unit change) with > 95% power at $\alpha = 0.05$. These definitions characterize project capacity to produce outputs that might serve as sufficient baselines for subsequent monitoring, or alternatively, capacity to identify regions where species were relatively more or less common and important distribution correlates (Guillera-Arroita et al. 2015).

Estimating existing data accuracy

Once data quality has been defined, determining whether existing data are sufficient requires both estimating existing rates of error using controlled experimental settings or post-hoc verification (Crall et al. 2011, Miller et al. 2015), and estimating requisite rates of error that translate to data of acceptable quality.

We reviewed 19,212 images each classified by multiple volunteers on a crowdsourcing platform to determine the "true" species in each image (Appendix S2 contains more detail, also see Data S2). We used the results of this review to estimate species-specific probabilities that a species was classified as present when not (false-positive error probability) or was missed when present (false-negative error probability). We used a Bayesian approach to estimate these parameters, assuming correct (or incorrect) classifications $y_i \sim$ Bernoulli (θ), and defined a prior distribution for θ as Beta (1, 1). This conjugate parameterization permitted us to analytically derive the posterior distribution of error parameters $\hat{\theta}$ as Beta $(1 + \sum_{i=1}^{n} I(y_i = 1), 1 + \sum_{i=1}^{n} I(y_i = 0))$.

Estimating requisite data accuracy

In some cases, deriving requisite data accuracy is as straightforward as evaluating moments or summaries of the data. For example, if data quality is defined as being able to achieve < 10% absolute bias in the prevalence of some binary phenomena using a logistic regression, then data are sufficient if the difference between false positive and false negative classification error is < 10%. Because data produced by citizen scientists and automated detectors or algorithms is often aggregated in varied ways and analyzed using

more complex techniques that make it difficult to understand the relationship between sample error and estimator error, simulation may be required to translate data accuracy into data quality.

We used simulation (see Data S1) to evaluate the sensitivity of two occupancy estimators to image misclassification and to determine target error thresholds that would permit sufficiently accurate estimates. Simulations were fixed as having 25 temporal replicates (each equivalent to a 24-hour period, a sampling interval commonly used in analyses of trail camera images) and 500 spatial replicates, levels of survey effort that approximate the minimum sampling effort we might use for an occupancy analysis. Site-specific values for occupancy and detection parameters were implemented as logit-linear values: logit $(\psi_{i,sim}) = \beta_0 + \beta_1 X_{i,sim,1} + \beta_2 X_{i,sim,2}$, where β_0 = either -1 or 1, $\beta_1 = -2$, and $\beta_2 = 1$; logit $(p_{i,sim}) = \alpha_0 + \alpha_1 X_{i,sim,1} + \alpha_2 X_{i,sim,2}$, where α_0 = either -2 or -3, $\alpha_1 = -1$, $\alpha_2 = 1$, and i indexes specific sites. All covariate values were simulated as Normal (0, 1). Thus, at an average site where X_1 and $X_2 = 0$, expected occupancy probability (ψ) was roughly 26% or 73%, expected per-sample detection probability given presence (p) was roughly 5% or 12%, and expected cumulative detection probabilities over the 25 d sampling duration (P^*) were roughly 76% or 94%. Average parameter values were selected to represent differences between relatively rare and common species based on derivations from previous camera-based occupancy studies in the state (Clare et al. 2015, Clare et al. 2016).

Observations were initially generated as $y_{i,sim} \sim$ Bernoulli $(z_{i,sim,} \times p_{i,sim})$, where $z_{i,sim}$ is the occupancy state for a site/simulation combination and was generated as $z_{i,sim} \sim$ Bernoulli$(\psi_{i,sim})$. We then induced additional false-negative and false-positive classification error within each simulated dataset: 3%, 10%, or 30% of the true detections were thinned, and additional false-positive detections were distributed across all sampling intervals such that 3%, 10%, or 30% of all detections were false positives (Appendix S3). Assuming each sampling interval contains at most one true and one false positive detection, this translates empirical estimates of error percentages at the observational level to model inputs (see Appendices SI3 and SI4 for more discussion of this issue). We simulated 300 data sets for each combination of parameter values, and fit occupancy models to each dataset using Markov chain Monte

Carlo simulation (3 chains each consisting of 2000 adaptation steps and 3000 samples) using JAGS v3.4 (Plummer 2003) through the R library 'jagsUI' (Kellner 2015). This analysis and all others were performed using R v3.2 (R Core Team 2015). We assumed convergence if $\hat{r} < 1.1$ (Gelman and Rubin 1992) and traceplots indicated adequate mixing. We evaluated sensitivity to misclassification error using the mean error and relative bias of finite-sample occupancy estimates (the proportion of occupied sampled sites, PAO, Royle and Kery 2007), the relative bias of $\hat{\beta}$, and empirical power to detect the correct directional effect of beta parameters.

For a subset of simulated scenarios (Appendix S3), we evaluated false positive occupancy models as statistical data remediation action following the observation-confirmation protocol described by Chambert et al. (2015). Chambert et al.'s (2015) model assume that at a subset of sites, all temporally replicated observations are confirmed after the fact as either containing no detections, only true positive detection(s), only false positive detection(s), or both true and false positive detections. The validation process allows estimation of parameters s_0 and s_1, which reflect the probabilities of recording > 0 false positive or true detections at a site during a specific sampling interval j. At sites lacking verification, the observation process is treated as $y_{i,j}$~Bernoulli ($z_i \times p_{11} + [1-z_i] \times p_{10}$), where p_{11} and p_{10} are true and false probabilities of detection derived from the parameters s_0 and s_1. We modified the original model description to reflect a more efficient and realistic validation process for our project by only subjecting sampling intervals containing positive detections to simulated validation and simulating the validation process as randomly occurring across sampling intervals rather than at all intervals at specific sites. Because the parameters s_0 and s_1 are unknown prior to model-fitting, and in most settings, investigators are more likely to have a sense of misclassification rates or probabilities within the raw data, we induced false-positive and additional false-negative error as before (equivalent false-positive and negative rates of 3%, 10%, or 30%). We fixed the proportion of simulated samples that were validated as either 10%, 30%, or 50% of detections, and evaluated estimator sensitivity to error as described above. We defined prior distributions as Uniform (0 ,1) for probability parameters or intercepts on the logit^{-1} scale, and Normal (0,

2.5) for coefficients. Model sensitivity point estimates and uncertainty intervals were derived using the mean and 95% highest density intervals of the posterior distribution.

Identifying and narrowing candidate remediation actions

Many (non-statistical) remediation actions can be used to achieve a desired level of data quality. Fully evaluating the efficacy of manipulating a project interface or altering training protocols can be time consuming. One way to narrow the list of potential remediation actions is to compare how effectively variables associated with different actions explain error. Because there may be many potential variables deserving consideration, initially focusing on factors that encompass several more detailed predictors can expedite the remediation process.

We considered four general remediation strategies. First, there were differences in sampling protocols as the program evolved over time; images were uploaded and classified as sequential non-overlapping batches ("seasons" hereafter). There were season-specific differences in image quality (lower in one season due to camera firmware settings), camera models (Reconyx HC600 and HC500 models vs. Bushnell Trophy Cam Pro models), camera placement strategies (seasons used variably focused upon sampling aquatic mammal monitoring or ungulates), how images were presented to online citizen classifiers (single photographs vs. sequences of 3-affiliated triggers), and minor changes to the user interface. If classification error varied strongly by season, it would suggest error was sensitive to changes in data collection protocols and how data were presented for classification. This would further imply that modifications to the interface or overarching project protocols deserved prioritization as means to reduce dataset error, and that specific terms associated with protocol differences could be used to screen data or model misclassification.

Second, we hypothesized that intrinsic differences in the placement of specific cameras might be a cause of data classification error. This would suggest that changes to the specific guidelines for camera

placement, or including random error terms for distinct camera locations or locational covariates (e.g., camera-specific height) when trying to predict misclassification could be useful strategies.

Finally, we hypothesized that error structure might result from inherent interspecific differences in false-negative error (difficulty correctly identifying certain species) or false-positive error (volunteers more likely to default to certain species given uncertainty). If error was better explained by the true species in the image, it would indicate that additional training aimed at helping volunteers distinguish species might be most useful, as the true species in an image is typically unknown without further evaluation and thus impractical to use a term to predict error. If classification error was best predicted by the crowd-reported consensus, it would indicate that a strategy focusing on predicting misclassification error including terms for the reported consensus species (as well as terms associated with other general factors considered) might be optimal. Alternatively, it might suggest that interface modifications that allowed volunteers to report metrics of classification uncertainty might be useful.

We fit generalized linear models with a binary response (crowdsourced consensus classification correct or not) and a single factorial predictor (season, camera site, true species identification, or the consensus species). We used Akaike's Information Criterion (AIC) to rank the prioritization of each general remediation strategy deserving more detailed follow-up analysis (Burnham and Anderson 2002). Data here were 17,139 images that we considered identifiable (i.e., the "true" species was not unknown) that had sufficient metadata to allow more targeted follow up analysis.

Implementing and evaluating remediation action

After narrowing the list of remediation strategies, next steps often include identifying specific variables to manipulate, implementing an action or correction, and then evaluating whether the action improves data quality. For example, had "season" been identified as the most important variable for explaining misclassifications in our data, we would have evaluated variability in error as a function of specific interface components, altered components in the classification interface strongly associated with error,

and reviewed subsequent crowdsourced classifications to determine the whether those changes had been effective. In some cases, the steps above can be considered simultaneously. In this case, the single best explanatory factor for misclassification was the reported species identity, and simulations suggested that the influence of additional false negative error was negligible (see Results). Thus, developing a screening model to predict misclassified images for subsequent review or censure served jointly as a more detailed exploration of error and as a remediation action that we could directly evaluate based upon model performance.

We split the data into training (10,270 images, 60%) and testing partitions (6,869 images, 40%), and considered several specific predictors that we hypothesized were directly contributing to image misclassification (Table 1). These included predictors reflecting the proportion of volunteers whom selected the consensus classification (the strength of consensus), variation in camera placement settings, date effects to capture seasonal variation in the appearance of species, time effects to capture diel variation in lighting and camera flash mode, and image settings or qualities. Finally, we considered a predictor that would capture variation in error as a function of volunteers viewing images at random: sudden changes in the reported chronicity of species at a specific camera location. We fit candidate generalized linear (mixed) models that either shared intercepts and slopes across species (sensu Swanson et al. 2016), allowed intercepts to randomly vary across species, or allowed intercepts and slopes to randomly vary across species using Hamiltonian Markov Chain Monte Carlo via R library "rstanarm" (Gabry and Goodrich 2016). We used default priors (intercept and coefficient priors for scaled data were respectively N(0, 10) and N(0, 2.5)), and simulation settings consisted of 4 chains with 1000 burn-in and 1000 posterior samples, or if necessary for convergence, 4000 burn-in and 4000 posterior samples each.

We compared models and assessed screening performance using out-of-sample measures of the Receiver Operating Characteristic area under the curve (AUC), partial area under the curve up to a false positive threshold of 0.1 (pAUC, McClish 1989), the maximum value of Matthews correlation coefficient (MCC) at any cut-point (Matthews 1975), and the positive predictive value (PPV) at a classification cut-

point of 0.5. These metrics (implemented for the full test partition and different subsets of interest) provide direct information about the accuracy of the data that might enter an occupancy model after a potential screening process and was implemented information about how many true positive detections might be discarded during a screening process. Point estimates and uncertainty intervals were derived from the mean and 95% highest density intervals of the posterior predictive distribution. We used test subsets to explore trade-offs between false-positive and false negative error relative to our simulation results (i.e., how many true detections would be lost during a screening process to achieve an acceptable level of false positive error?).

Results

Estimating existing data accuracy

Across the full dataset, the accuracy of crowdsourced species classifications was 93.4%, but false positive and false negative error varied considerably across species (Figure 2). More commonly encountered species were generally subject to less false-positive and false-negative error (Figure 3). Exceptions include lagomorphs, as brown phenophase snowshoe hare (*Lepus americanus*) were commonly misclassified as cottontails (*Sylvilagus floridanus*, Table S1 and Figure S3 in Appendix S1), and "unknown" species-without consensus (often clearly identifiable to experts).

Estimating requisite data accuracy

Simulation results suggested that all false positive rates considered led to overestimation of species distribution using the base occupancy model and shrank estimates of occurrence associations (Figure 4). These were more pronounced when species were more easily detected and narrowly distributed. Still, our criteria for data adequacy (expected absolute bias < 0.05) was met when false positive rates were 3%, and most models fit to simulated data estimated the directional covariate effect correctly (empirical power was as low as 96%, but most commonly 100%). In contrast, additional false negative error had little influence on estimator performance (Appendix S1, Table S3). Importantly, if a 3% false positive proportion was

used to define adequate data quality, only 4 species appeared to be classified with sufficient baseline accuracy (Figure 2). Occupancy models accounting for false-positive error provided unbiased inference across the error rates considered (Figure 5). Estimator performance improved as more sampling intervals were validated (Figure 6), but the rate of improvement decreased as more samples were validated. That is, the largest gains in performance were associated with shifting from a standard occupancy model to one accommodating false-positive error.

Identifying and narrowing candidate remediation actions

Interspecific factors (the true species or reported species in the image) explained far more misclassification variability than differences in season or camera location (ΔAIC > 1000, Table S4 in Appendix S1), indicating species identity was more strongly associated with classification error than elements of the classification interface or camera placement. The reported putative species within the image explained error more effectively than the true species (AIC ω_i = 1), implying more interspecific variability in false-positive error than false-negative error and that implementing data screening to flag potential false positive classifications was a potentially useful remediation strategy.

Implementing and evaluating remediation action

The best performing misclassification screening model performed very strongly on out of sample data (AUC = 0.97, 95% CRI = 0.96−0.97; pAUC = 0.80, 95% CRI = 0.77−0.83; PPV = 0.97, 95% CRI = 0.97−0.98; MCC = 0.68, 95% CRI = 0.66−0.69, Table S5 in Appendix S1), suggesting that across all species, censuring images predicted to be misclassified provided adequate data without substantial removal of correct classifications. It included random intercepts and coefficients (using reported consensus species as the grouping effect) associated with a quadratic effect of day of year, the proportion of users voting for the consensus, and the effect of sudden changes in the chronology of species at specific camera station (definitions in Table 1). The probability of the crowdsourced consensus being correct increased as more volunteers agreed on the consensus species and was less likely if the species

reported at a specific camera changed over rapid intervals (e.g., bear present, deer present, bear present within one minute; Figure 7).

Screening performance varied substantively across organismal groups (Appendix S1, Table S6). We were better able to discriminate between true and false classifications of common species that were intrinsically classified with greater accuracy. For example, to achieve a false positive rate of less than 3% within test-partitioned black bear (*Ursus americanus*) pictures required censuring less than 2% of the data; to achieve the same false positive threshold for canids required censuring 52.3% of the recorded observations and discarding more than 40% of the true positive classifications in the process (i.e., enacting an additional false-negative error beyond what was simulated). Post-hoc simulations corresponding to this scenario (70% of true detections removed and 3% false-positive detections induced) suggested the base occupancy estimator still performed adequately under simulated sampling conditions after severe data censuring (mean error = 0.02, RMSE < 0.04).

Discussion

Ecologists have always faced sampling limitations and imperfections. Emprical comparisons of sampling methods (Clare et al. 2017), power analysis and related simulation approaches (Ellis et al. 2014), and other techniques are commonly used to determine how to allocate sampling effort or resources most efficiently. Determining how much data are needed and how more data can be collected have historically been preeminent study design foci, and they remain important considerations. Although our titular questions are analogous, they have seen have seen less attention by practitioners as a whole (Miller et al. 2015), which is problematic because measurement or observation error is found within nearly every study in which it is directly evaluated (e.g., McClintock et al. 2010, Butt et al. 2013). Our specific results are most germane for the growing number of independent efforts that use automated detection devices, citizen scientists, or both (e.g., there are more than 20 trail camera projects hosted by Zooniverse). However, ensuring data quality is more broadly important for broad-scale or even global efforts that are

scarcely feasible without the participation of citizen scientists or the use of automated detection or classification techniques (Chandler et al. 2017, Steenweg et al. 2017, Kissling et al. 2018).

So, how accurate do data need to be? We have contended throughout that this depends upon specific research or monitoring objectives and as such, is likely to be distinct to specific studies. However, our implementation provides some insights with regards to one of the most ubiquitous data processing tasks (species identification), one of the most common goals in ecology (estimating species distributions), and one of the most widely used models for estimating species distribution. The first data quality concern associated with estimating species distributions is that although professionals, volunteers, crowdsourced aggregates, and machine learning algorithms commonly identify species accurately overall (> 95%, e.g., McClintock et al. 2010, Swanson et al. 2016, Nourouzaddeh et al. 2018), overall species identification accuracy is often weighted by a few very common and easily identified species and the accuracy of individual species is highly variable. The range of misclassification we considered here (3 – 30%) is not unique to our study; similar rates of misidentification are documented across a range of methodologies for classifying trail camera images (McShea et al. 2016, Swanson et al. 2016, Nourouzaddeh et al. 2018, Tabak et al. 2018) or recorded calls (Simons et al. 2007, McClintock et al. 2010, Farmer et al. 2012, Mac Aohda et al. 2018, Priyadarshani et al. 2018). This suggests that despite the overall accuracy of many datasets processed by humans with limited training (volunteer or not) or automated algorithms, there is a non-trivial risk of substantially overestimating the distributions of many species using many commonly used data types. Furthermore, motivation to further expedite data processing has motivated development of compound approaches in which citizen scientist classifications are used to train algorithms (Willi et al. 2018), which is likely to further compound existing errors. In short, the aggregated accuracy measures often reported are not necessarily accurate gauges of data accuracy itself.

The subsequent problem is that associations between data accuracy and estimator accuracy can be extremely variable, and as such, even if data accuracy is correctly described, it can be a poor index for

data quality. There are several underlying reasons for this. First, estimator sensitivity to error depends upon how error is being measured or parameterized. There can be substantively less bias when false positive error constitutes 3% of all detections rather than, say, happening at 3% across sites and sampling intervals. This is likely one reason that our simulations suggest the occupancy estimator is less sensitive to false positive error than previous empirical or simulation studies (Miller et al. 2013, Miller et al. 2015, Ruiz-Gutiérrez et al. 2016); false-positive parameters are often distinct from than how species identification accuracy is typically reported, and our estimates of s_0 were generally far smaller than the fraction of detections that were simulated as false positives. Secondly, although we generally ignore it here, model sensitivity also depends upon how observations are aggregated for analysis (see Appendix S3). Third, the relationship between an estimator's relative bias or error and data error varies as a function of the attributes of the sampled species; less widespread species were more sensitive to false positive error in our simulations. Finally, different estimators exhibit entirely distinct sensitivities to different amounts or types of error. For many models, the association between data error and estimator error may be nonlinear and disproportionate. For example, 5% more detections may translate to 25% more animals estimated to exist (Clare et al. 2018). For other models, the overall amount of detections rather than their locations may be more important. Although classification error appears to have reshuffled species observations across locations, the overall prevalence of species within our dataset was largely preserved (Table S1 and Figure S1 in Appendix S1). Had we considered a random encounter model (Rowcliffe et al. 2008) as our planned analysis, we may have come to different data quality conclusions.

So, if data will be less reliable and models not as robust as desired, what can be done? In the worst-case scenario, data accuracy or reliability cannot be quantified and no auxiliary information that might inform the estimation of error has been collected. Here, practitioners can default to cautious interpretation and conservative analyses (Bird et al. 2014, Isaac et al. 2014). Our results suggest that even with severe observation error less occurring at random, patterns in estimated occurrence can still be

monotonically correlated with the true state, and such information may still be useful for spatially delineating areas of managerial concern (Guillera-Arroita et al. 2015). Alternatively, following previous recommendations (Miller et al. 2015), practitioners can fit estimators in which all observations are treated as uncertain and false positive errors are a latent component of the model (Royle and Link 2006). These considerations also deserve attention from users of professionally collected or classified data, which is typically comparably accurate and less thoroughly vouched (Lewandowski and Specht 2015, Kosmala et al. 2016).

Simply having some measurement of data uncertainty, such as the confidence of an identification algorithm or agreement between multiple human classifiers, allows investigators to use more (and more effective) remediation actions. Uncertainty measures can be used to delineate between more and less reliable data prior to a species distribution analysis (e.g., a data censure), within an analysis as a distinct data type (e.g., the multiple detection state model described by Miller et al. 2011), or as a covariate for error for latent error (analogous to metrics of observer proficiency used by Johnston et al. 2018). The efficacy of these remediation actions depends upon how strongly confidence correlates with accuracy. Within our study, agreement among citizen scientists was associated with but not equivalent to the expected accuracy of the classification (see also Swanson et al. 2016). The confidence of a trained algorithm when applied to distinct data can be similarly unreliable (Tabak et al. 2018).

In general, investigating data accuracy more directly and deeply provides researchers more opportunities for effective remediation. Investigators using experiments or post-hoc data verification to quantify error and variability in error will have more information about how general project components that can be manipulated (volunteers, protocols, interfaces) differentially contribute to error, and will be able to make more informed and effective decisions about how to manipulate these. In our case study, the classification “season” explained less variability in classification error than the other general project components, suggesting that potential manipulations associated with differences in the platform across seasons (e.g., minor changes to filter options, or as a more expensive example, switching to different

camera models) held limited potential. Evidence for interspecific variation as a major driver of classification accuracy informed the use of specifies-specific random effects terms within screening models that greatly outperformed models without random effect terms. In turn, predictors identified as useful while exploring variation in error can also be directly incorporated within false-positive occupancy models (Chambert et al. 2015, Ruiz-Gutiérrez et al. 2016), and can make these models even more effective.

Perhaps our most strident motivation entering this study was the contention that data quality and remediation should be evaluated within a single process. It is difficult to fix data without knowing how it is going to be used. The viability of data censuring (rather than adopting the more intensive task of directly reviewing all questionable data) was directly contingent upon evidence suggesting relative estimator insensitivity to additional false negative error. Simultaneously assessing data quality and remediation (and evaluating multiple remediation actions) also caries synergistic benefits. Models that are effective for screening misclassifications are also likely to be useful parameterizations for false positive error within an occupancy model. Similarly, exploring data-censoring models and sources of error provided insights into the potential of different interface manipulations. Quantifying inter-specific variability in error and user agreement as useful indicators of accuracy directly informed protocol changes such as highlighting commonly confused species (Figure S3 in Appendix S1) within the classification interface and focusing communications with volunteers towards providing feedback on species identified as easily confused or difficult to classify. The effects of these actions have not been evaluated but enacting them required trivial effort. While the best strategy for our stated objective appears to be using occupancy models incorporating false-positive error such extensions have not been described for many other potential analyses, and data censuring or other actions may circumstantially be more effective. Although we have focused on remediation as a matter of ensuring data quality for a specific problem, effective remediation efforts may require multiple actions to provide investigators the flexibility to

achieve different objectives (Kosmala et al. 2016), and implementing specific actions effectively can make subsequent actions easier to implement.

Similarly, we we believe that combining data and remediation evaluations can have synergistic benefits for data users and managers. Certainly, researchers whom explicitly attempt to quantify data needs will have a stronger understanding of what questions can be answered, and projects that quantify data accuracy have a better sense of which data are worth collecting, but perhaps the greatest benefits may come from sharing such information across platforms. Projects that present quantitative information about data reliability make it easier for researchers to select suitable data or choose suitable questions, and researchers that share specific data needs make it easier for projects to set concrete targets, and in turn, may make it easier for researchers to acquire sufficient data.

Evaluating data quality and varied remediation actions is not without cost. Analyzing simulations, verifying data, and conducting calibration experiments all require time and expense, and some projects may have few samples that can be verified. Quantifying data accuracy is likely the most costly component, and we acknowledge that the classification of trail camera images can be evaluated relatively expediently, whether via post-hoc verification of images or by calibrating volunteer performance on known samples (sensu Ruiz-Gutiérrez et al. 2016). Tabak et al. (2018) report experts were able to classify 200 images per hour; anecdotally, careful verification of images seems to be somewhat slower (30 to 50 sequences of three images per hour). Still, verifying thousands of classifications, even if individual samples can be quickly processed, is not a trivial undertaking, and we expect that many efforts have been dissuaded from performing data evaluations by the perceived amount of requisite effort. The optimal size of a data evaluation sample is difficult to generally quantify, because it depends on the desired inferential objectives and properties of error within the data. If data collection is complete, the ideal size of the validation or calibration sample may be that which provides the investigator sufficient confidence that data are adequate (e.g., 95% CI associated with error estimates in the baseline data or associated with a screening model's predictions indicate that baseline or censured data are sufficient to use). Projects with

ongoing data collection are likely to benefit from evaluating data iteratively (Kosmala et al. 2016), and the requisite sample should take into account the ability to detect changes in baseline performance as procedures change over time.

But although specific guidelines for designing data evaluation efforts are difficult to provide, we wish to emphasize that a verification or calibration sample does not need to be enormous to effectively characterize error, and that any effort allocated towards evaluating data quality represents improvement over allocating no effort. In fact, there are almost certainly diminishing returns associated with increasing the size of a data evaluation sample. The difference in precision between estimates of the overall probability of a white-tailed deer, snowshoe hare, or sandhill crane image being a false positive (respectively, 95% CRI = 0.015 – 0.020, 0.001 – 0.024, and 0.001 – 0.077) was disproportionate to the difference in effort (11,650, 272, and 45 images evaluated, respectively). The primary difference between validating 50 simulated sampling intervals vs. 750 simulated sampling intervals when fitting an occupancy model incorporating false positives was a small gain in estimate precision. That is, a 15-fold increase in effort allocated towards validating sampling intervals or a 40-fold increase in effort allocated to validating deer images vs. snowshoe hare images made little difference. Gains associated with using more complex models to screen or describe error similarly diminished. For example, incorporating random intercepts for species led to substantive gains in out-of-sample predictive performance for screening models, but gains associated with further considering random slope terms were far smaller (Appendix S1, Table S5). We discuss further ways in which our own data evaluation effort may have been implemented more efficiently in Appendix S2.

Citizen scientists, automated detectors, classification algorithms, and a commitment to data sharing have the collective capacity to revolutionize the scope and scale of ecological inquiry. Applied ecologists now have means to efficiently produce or concatenate data permitting sound inference at both fine resolution and across extents not only meaningful to management decision making, but more broadly, cross-jurisdictional extents that reflect the massive scales that many important ecological drivers and

biodiversity threats operate at (Princé and Zuckerberg 2015, Steenweg et al. 2017). Whether the contributions made by many existing or developing studies or monitoring programs leveraging these techniques achieve the ambitions of these programs will partially depend upon how willingly and widely principles of data quality described herein are adopted.

Acknowledgments

We acknowledge funding and other support from the Wisconsin Citizen-based Monitoring Network Partnership Program, NASA Ecological Forecasting #NNX14AC36G, and NESSF #NNX16A061H, the University of Wisconsin Cooperative Extension, and a grant from the Federal Aid in Wildlife Restoration act awarded to WDNR. This publication uses data generated via the Zooniverse.org platform, funded by in part by a grant from the Alfred P. Sloan Foundation and a Global Impact Award from Google. We thank the Department of Forest and Wildlife Ecology for their support. We thank A. Johnston, A. Wiggins, and V. Radeloff for comments that greatly improved the manuscript.

References

Abra, F. D., M. P. Huijser, C. S. Pereira, and K. Ferraz. 2018. How reliable are your data? Verifying species identification of road-klled mammals recorded by road maintenance personnel in Sao Paulo State, Brazil. Biological Conservation 225:42-52.

Alldredge, M. W., T. R. Simons, and K. H. Pollock. 2007. A field evaluation of distance measurement error in auditory avian point count surveys. Journal of Wildlife Management 71:2759-2766.

Bird, T. H., A. E. Bates, J. S. Lefcheck, N.A. Hill, R. J. Thomson, G. J. Edgar, …S. Frusher. 2014. Statistical solutions for error and bias in global citizen science datasets. Biological Conservation 173:144-154.

Bonter, D. N., C. B. Cooper, M. Gardiner, L. Allee, P. Brown, J. Losey, H. Roy, and R. Smyth. 2012. Data validation in citizen science: a case study from Project FeederWatch. Frontiers in Ecology and the Environment 10:305-307.

Burnham, K.P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York.

Butt, N., E. Slate, J. Thompson, Y. Malhi, and T. Tiutta. 2013. Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. Ecological Applications 23: 936-943.

Chambert, T., D.A.W. Miller, and J. D. Nichols. 2015. Modeling false positive detections in species occurrence data under different study designs. Ecology 96:332-339.

Chandler, M., L. See, K. Copas, A. M. Z. Bonde…et al. 2017. Contribution of citizen science towards international biodiversity monitoring. Biological Conservation 213:280-284.

Clare, J. D. J., E.M. Anderson, and D. M. MacFarland. 2015. Estimating bobcat abundance at a landscape scale and evaluating occupancy as a density index. Journal of Wildlife Management 79:469-480.

Clare, J. D. J., D.W. Linden, E.M. Anderson, and D. M. MacFarland. 2016. Do the antipredator strategies of shared prey mediate intraguild predation and mesopredator suppression? Ecology and Evolution 6:3884-3897.

Clare, J., S. T. McKinney, J.E. DePue, and C.S. Loftin. 2017. Pairing field methods to improve inference in wildlife surveys while accommodating detection covariance. Ecological Applications 27:2031-2047.

Clare, J., P. Townsend, and B. Zuckerberg. 2018. Generalized sample verification models to estimate ecological state variables with detection-nondetection data while accounting for imperfect detection and false positive errors. bioRxiv: https://doi.org/10.1101/422527

Clements, M. J., T, J. Rodhouse, P.C. Ormsbee, J. M. Szewczak, and J. D. Nichols. 2014. Accounting for false-positive acoustic detections of bats using occupancy models. Journal of Applied Ecology 51:1460-1467.

Crall, A. W., G. J. Newman, T. J. Stohlgren, K. A. Holfelder, J. Graham, and D. M. Waller. 2011. Assessing citizen science data quality: An invasive species case study. Conservation Letters 4:433-442.

Dickinson, J., B. Zuckerberg, and D. Bonter. 2010. Citizen science as an ecological research tool: Challenges and benefits. Annual Review of Ecology, Evolution, and Systematics 41:149-172.

Ellis, M.M., J. S. Ivan, and M. K. Schwartz. 2014. Spatially explicit power analyses for occupancy-based monitoring of wolverine in the US Rocky Mountains. Conservation Biology 28:52-62.

Farmer, R. G., M. L. Leonard, and A. G. Horn. 2012. Observer effects and avian call count survey quality: rare-species biases and overconfidence. Auk 129:76-86.

Gabry, J., and B. Goodrichh. 2016. rstanarm: Bayesian applied regression modeling via Stan. https://cran.r-project.org/web/packages/rstanarm. Accessed March 23, 2016.

Gardiner, M. M., L. L. Allee, P. M. Brown, J. E. Losey, H. E. Roy, and R. R. Smyth. 2012. Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. Frontiers in Ecology and the Environment 10:471-476.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7:457-472.

Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M.A. McCarthy, R. Tingley, and B. A. Wintle. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24:276-292.

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

Johnston, A., D. Fink, W. M. Hochachka, and S. Kelling, 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods in Ecology and Evolution 9:88-97.

Kellner, K. (2015). jagsUI: a wrapper around rjags to streamline JAGS analyses. <github.com/kenkellner/**jagsUI.**> Accessed 15 July 2015.

Kissling, W. D., J. A. Ahumada, A. Bowser, M. Fernandez…et al. 2018. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biological Reviews 93:600-625.

Kosmala, M., A. Wiggins, A. Swanson, and B. Simmons. 2016. Assessing data quality in citizen science. Frontiers in Ecology and the Environment 14:551-560.

La Sorte, F. A. , C. A. Lepcczyk, J.L Burnett, A. H. Hurlbert, M. W. Tingley, and B. Zuckerberg. 2018. Opportunities and challenges for big data ornithology. Condor 120:414-426.

Lewandowski, E., and H. Specht. 2015. Influence of volunteer and project characteristics on data quality of biological surveys. Conservation Biology 29:713-723.

Linden, D. W., A. K. Fuller, J. A. Royle, and M. P. Hare. 2017. Examining the occupancy-density relationship for a low-density carnivore. Journal of Applied Ecology 54:2043-2052.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248-2255.

Mac Ohda, O., R. Gibb, K. E. Barlow, E. Browning…et al. 2018. Bat detective—deep learning tools for bat acoustic signal detection. PLoS Computational Biology 14: e1005995.

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Bichemica et Biophysica Acta 405:442-451.

McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010. Experimental investigation of observation error in anuran call surveys. Journal of Wildlife Management 74:1882-1893.

McClish, D. K. 1989. Analyzing a portion of the ROC curve. Medical Decision Making 9:190-195.

McShea, W. J., T. Forrester, R. Costello, Z. He, and R. Kays. 2016. Volunteer-run cameras as distributed sensors for macrosystem mammal research. Landscape Ecology 31:55-66.

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

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

Miller, D. A. W., L. L. Bailey, E. H. C. Grant, B. T. McClintock, L. A. Weir, and T. R. Simons. 2015. Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known. Methods in Ecology and Evolution 6:557-565.

Nourouzzadeh, M.S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune. 2018. PNAS 115:E5716-E5725.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using GIBBS sampling. Proceedings of the 3rd international workshop on distributed statistical computing.

Princé, K., and B. Zuckerberg. 2015. Climate change in our backyards: the reshuffling of North America's winter bird communities. Global Change Biology 21:572-585.

Priyadarshani, N., S. Marsland, and I. Castro. 2018. Automated birdsong recognition in complex acoustic environments: a review. Journal of Avian Biology 49:jav-01447.

Rowcliffe, J. M., J. Field, S. T. Turvey, and C. Carbone. 2008. Estimating animal density using camera traps without the need for individual recognition. Journal of Applied Ecology 45:1228-1236.

Royle, J. A., and M. Kery. 2007. A Bayesian state-space formulation of dynamic occupancy models. Ecology 88:1813-1823.

Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835-841.

Ruiz-Gutiérrez, V., M. B. Hooten, and E. H. Campbell Grant. 2016. Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias. Methods in Ecology and Evolution 7:900-909.

Simons, T. R., M.W. Alldredge, K. H. Pollock, and J. M. Wettroth. 2007. Experimental analysis of the auditory detection process on avian point counts. Auk 124:986-999.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282-2292.

Steenweg, R., M. Hebblewhite, R. Kays, J. Ahumada, J.T. Fisher, C. Burton…et al. 2017. Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors. Frontiers in Ecology and the Environment 15:26-34.

Swanson, A., M. Kosmala, C. Lintott, and C. Packer. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology 30:520-531.

Tabak, M. A., et al. Machine learning to classify animal species in camera trap images: applications in ecology. 2018. BioRxiv 346809: https://doi.org/10.1101/346809

Wiggins, A., and K. Crowston. 2015. Surveying the citizen science landscape. First Monday 20: http://dx.doi.org/10.5210/fm.v20i1.5520

Willi, M. R. T. Pitman, A. W. Cardoso, C. Locke, A. Swanson, A. Boyer, M. Veldthuis, and L.

Fortson. *In press.* Identifying animal species in camera trap images using deep learning and citizen science. Methods in Ecology and Evolution doi:10.1111/2041-210X.13099

Table 1. Candidate covariates considered within generalized linear mixed-modeling of crowdsourced species classification error of trail camera images.

Predictor	Description
User Proportion[A]	Proportion of users voting for consensus species
jday[A]	Julian Day
dectime[A]	Decimal hour photo was taken (military time)
TWSC[A]	Time-weighted species change[B]
Height	Camera height above ground level (ft)
Distance	Camera distance to target trail (ft)
Evenness	Pielou Evenness Index of individual classifications
Sequence type	Dummy variable for presentation as a sequence (vs. individual image)
Resolution	Dummy variable indicating a low resolution image

[A]Covariate was used within candidate model for predicting classification error. Other covariates in table were considered, but not ultimately included within the modeling effort due to limited support in exploratory analyses or collinearity with other predictors.

[B]Time-weighted species change (TWSC) is derived based upon the chronology of crowd-reported species at a specific camera location. Let $i_{x,b}$ serve as an indicator variable representing whether the reported species in sequential image x and image x-1 are different (1) or the same (0), with $i_{x,a}$ serving analogously for image x and $x + 1$, and let $t_{x,b}$ and $t_{x,a}$ respectively represent the decimal time (in hours) separating image x and image x-1 and for image x and $x + 1$. TWSC is calculated as $i_{x,b} \times \frac{1}{t_{x,b}} + i_{x,a} \times \frac{1}{t_{x,a}}$. A larger value of TWSC indicates a sudden change in the species recorded at a specific camera location (the maximum value occurs when images A, B, C are each separated by the minimum trigger interval of 15 s and record species A, B, A or B, A, B).

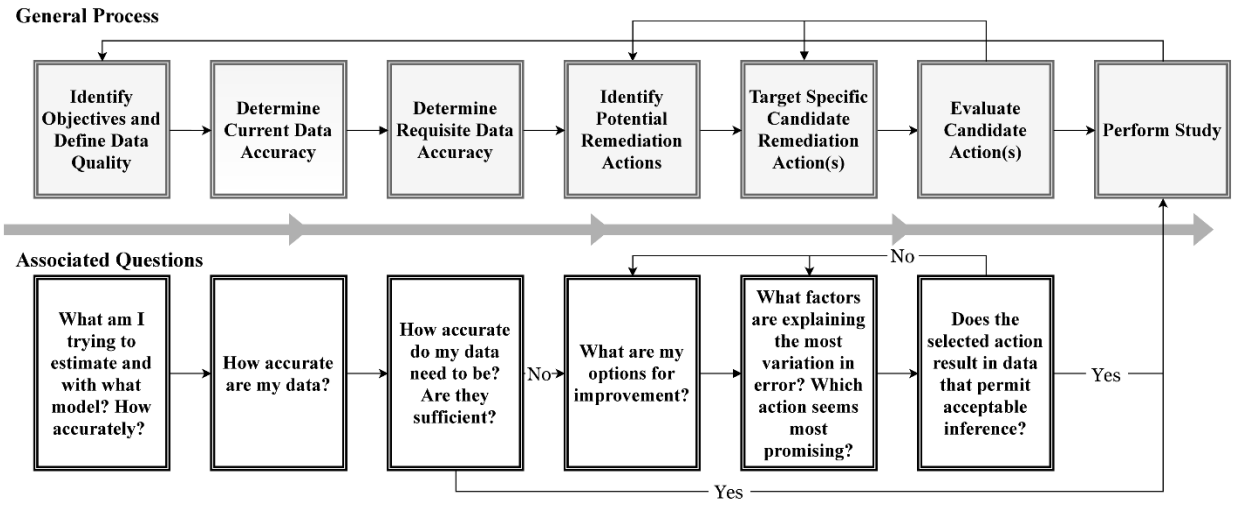

Figure 1. Conceptual diagram of the sequential process described for evaluating data quality and data remediation actions described within the main text.

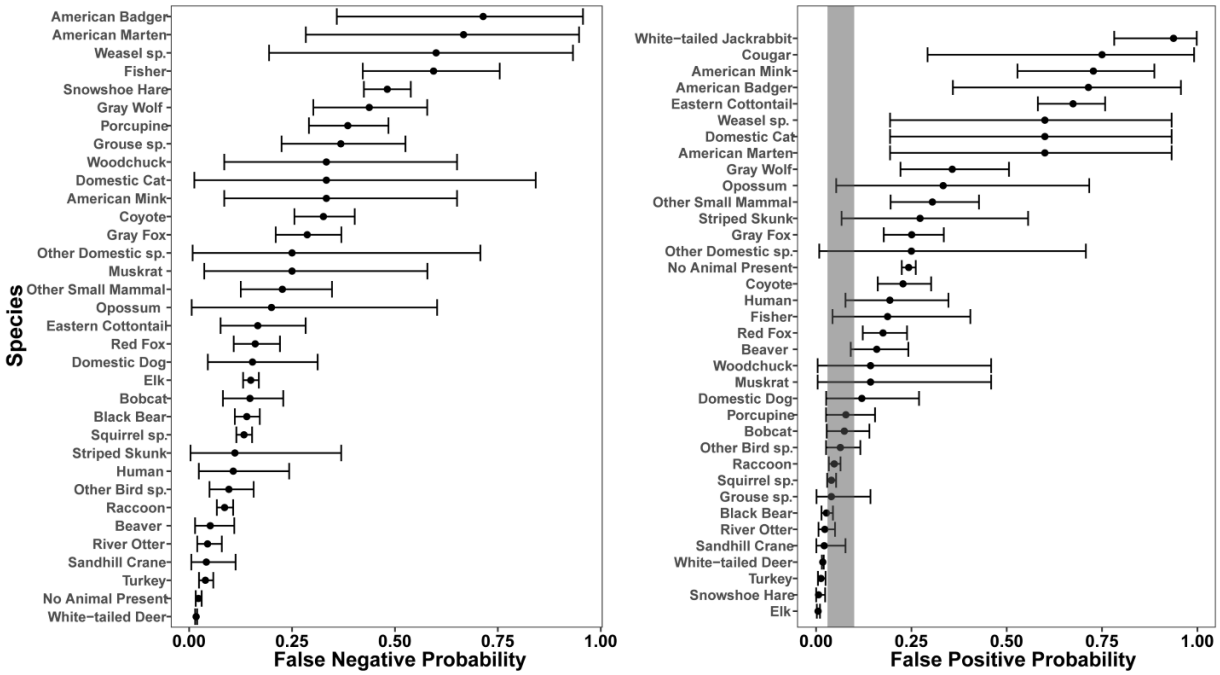

Figure 2. False negative (left) and false positive (right) probabilities estimated with expert validation of crowdsourced trail camera image classification. Whiskers represent 95% credible intervals. The gray shaded area on the right panel contains a threshold for false positive error that simulation suggested was requite for < %5 bias using the standard occupancy estimator, and highlights that using baseline classification results without addressing false positive error was likely to lead to substantial bias for many species.

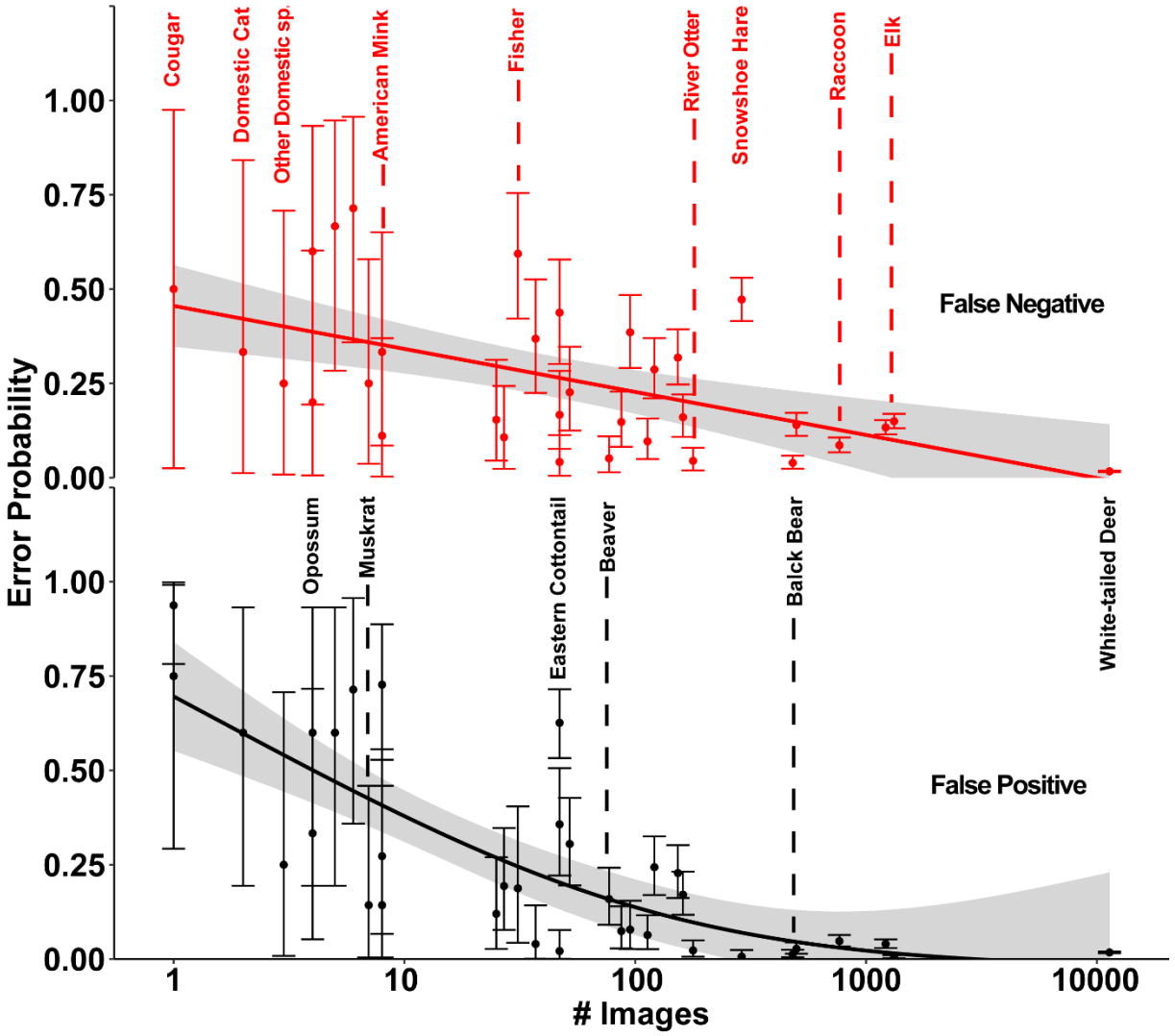

Figure 3. Negative association between species prevalence in the dataset (total number of images; log-transformed) and false positive (black) and false negative (red) classification error probabilities, which suggests that the distribution of rare or cryptic species was more likely to be estimated with substantial bias. Error bars represent 95% credible intervals. Solid lines and shaded area denote fitted additive model with smoothing selected using cross-validation, and 95% CI strictly for visualization purposes.

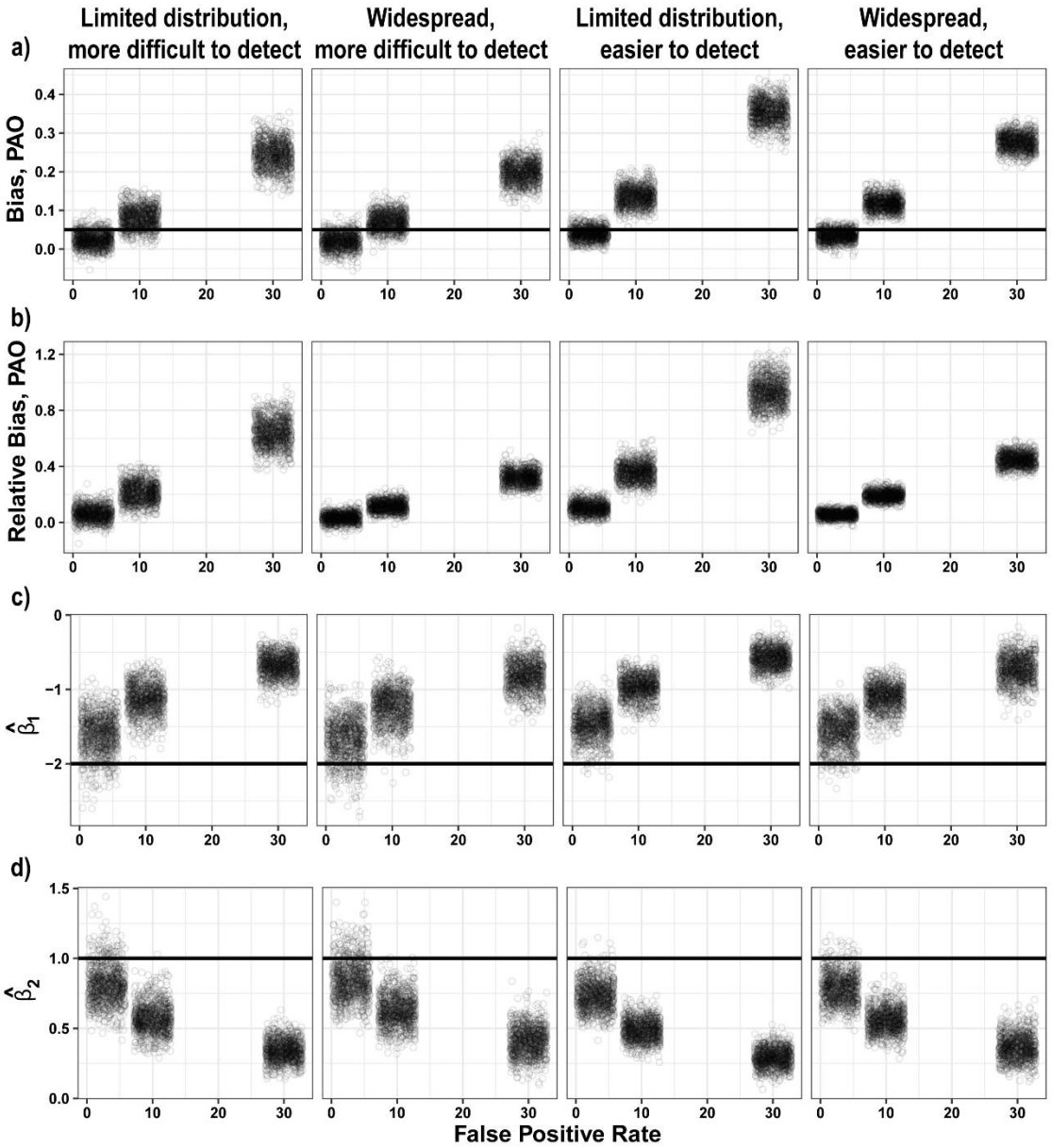

Figure 4. False positive classification error is the strongest determinant of mean error or bias (A) and relative bias (B) in finite sample estimates (PAO indicates proportion of area occupied). Solid line in A represents the predefined threshold described in the text. Occupancy coefficient estimates are displayed in panels C and D, and false positive error shrinks coefficients towards zero (true values indicated with solid lines). These effects are strongest when actual occurrence is lower ($\text{logit}^{-1}[\psi_{\text{intercept}}] = 0.26$ vs. 0.74) and detection probability is higher ($\text{logit}^{-1}[p_{\text{intercept}}] = 0.12$ vs. 0.05). X-coordinates are jittered for visualization.

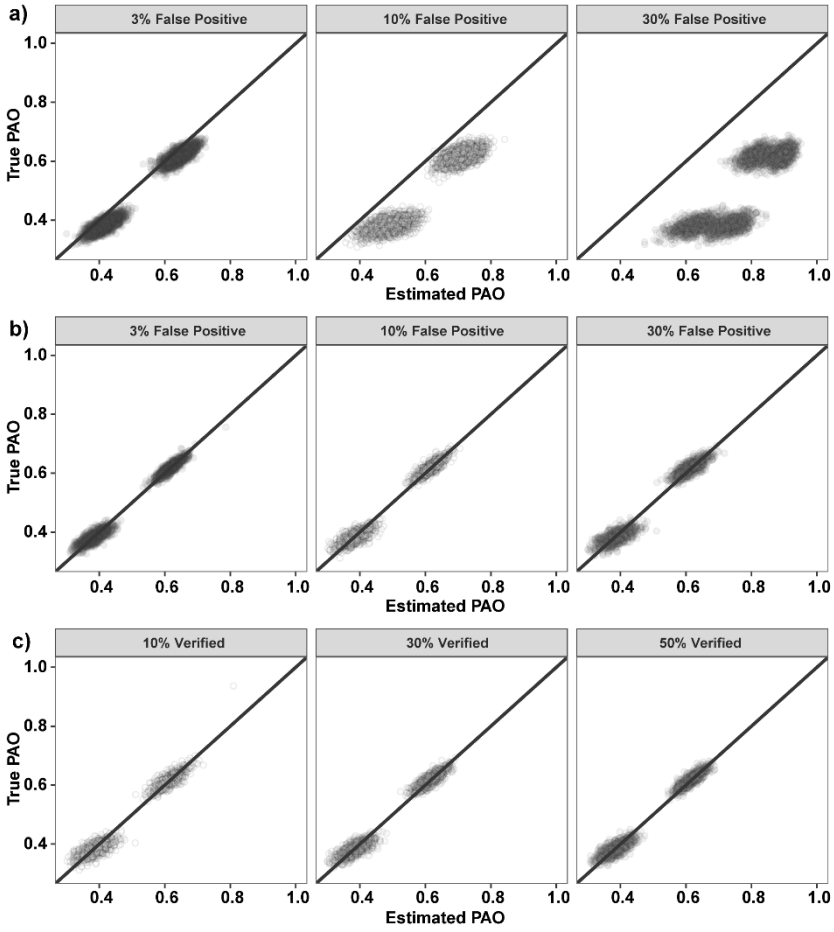

Figure 5. Standard occupancy models assuming only false-negative error are strongly biased as the proportion of false positive observations within the sample increases (A). Models that also incorporate false-positive error estimated using sample validation are generally accurate even when error rates are 30% (B). False positive models were unbiased when 10%, 30%, or 50% of the samples were verified (C) across all levels of baseline (3%, 10%, and 30% false-positive error all plotted here).

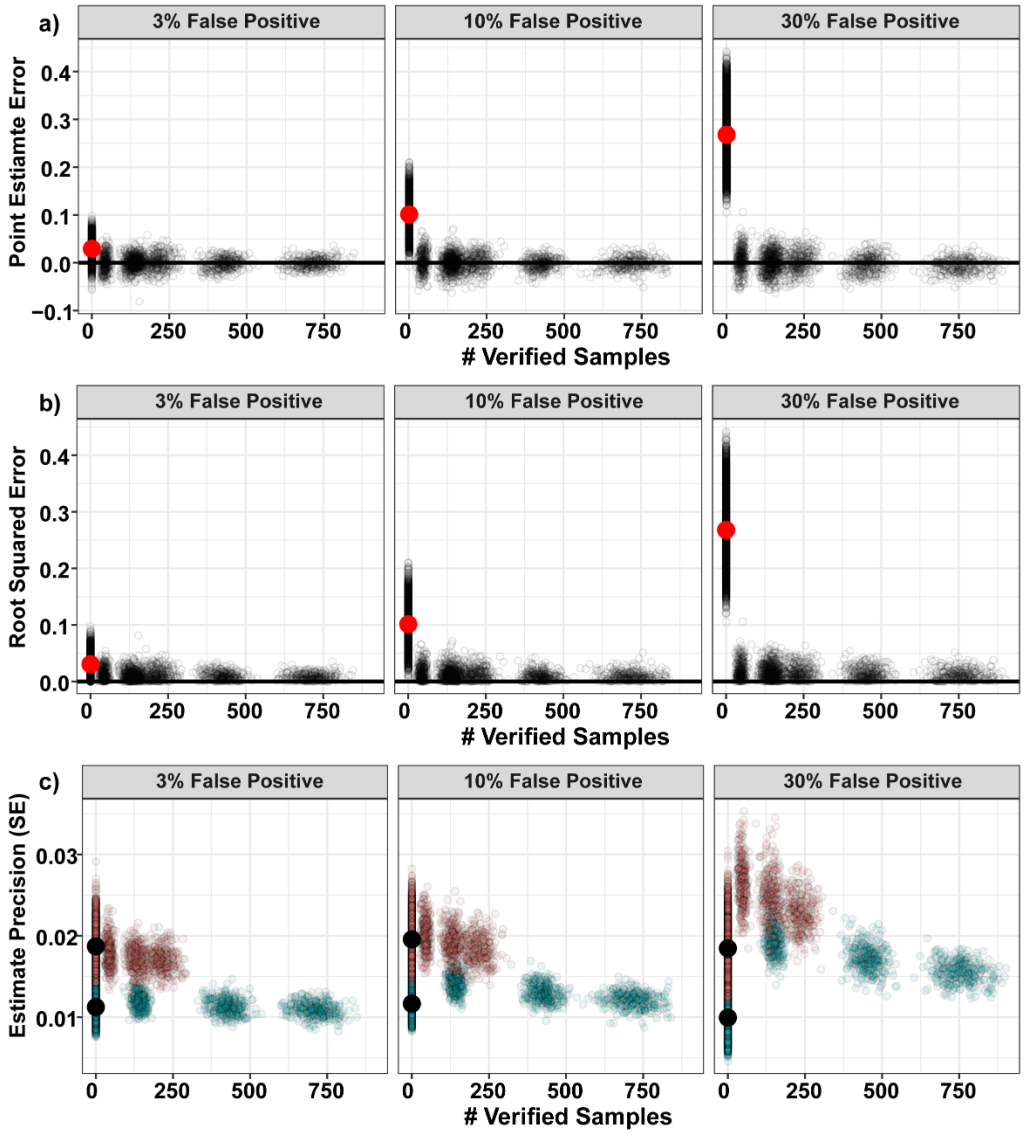

Figure 6. Point estimate bias (A) and root-squared error (B) of finite-sample occupancy estimates decreased as more simulated samples were verified, but also that the expected decrease in either loss function decelerated. Using a model incorporating false-positives inflates estimate uncertainty (C; standard error is approximated by standard deviation of the posterior distribution). Points corresponding to 0 verified samples reflect estimates from the standard occupancy model, while results corresponding to > 0 verified samples reflect estimates from an occupancy model incorporating false positive error. Red points in panels A and B reflect mean values when no samples were verified. In panel C, black points reflect mean values when no samples were verified, and red and blue dots correspond to simulation settings where the probability of detection was 0.047 (red) and 0.12 (blue).

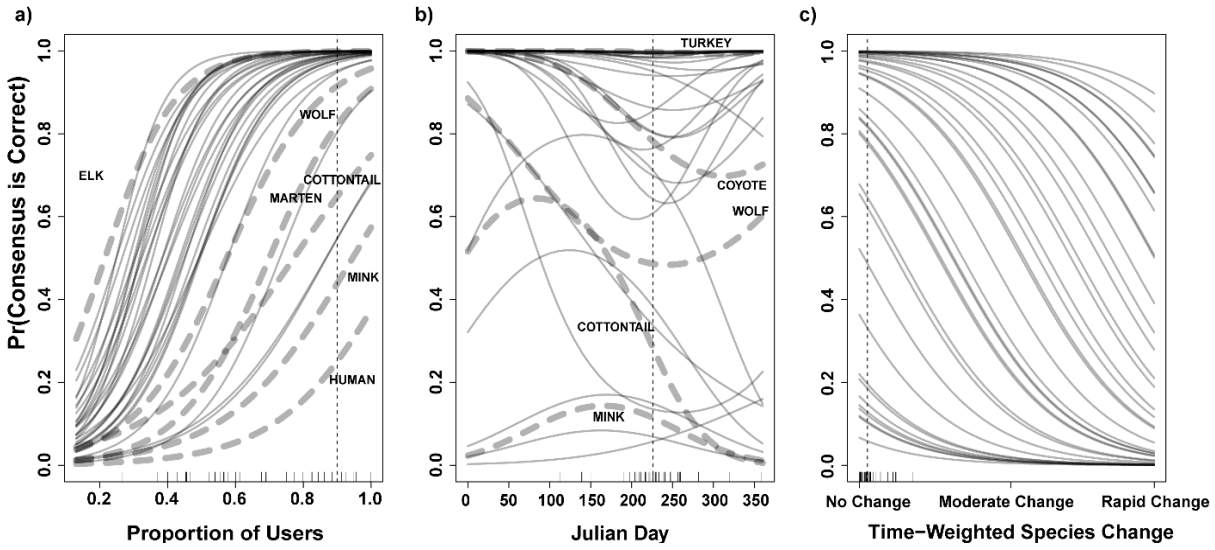

Figure 7. Marginal modeled effects suggest that the consensus crowdsourced classification was more likely to be correct as the proportion of users voting for the consensus species increased (left panel), interspecifically variable depending upon the Julian day on which the image was taken (center), and less likely to be accurate if the classifications immediately previous and/or subsequent at a given camera station reported different species in quick succession (right). Each line represents response of a different animal species: each effect is depicted with other terms held at species-specific means. Rug plots along the bottom depict the distribution of species-specific mean values; vertical line depicts the mean value across the entire dataset. In the left panel, the divergent response associated with a human classification is likely a function of different retirement rules associated with human images.

Appendix S1: Supporting Tables & Figures.

Table S1. Baseline estimates of species-specific false positive error probability (1-Pr[Correct Classification|Species reported]) and false negative error probability (1- Pr[Correct Classification|Species actually in image]) based upon crowdsourced consensus classification of trail camera images.

Species	Selection Choice	False Positive Probability (95% CRI)	False Negative Probability (95% CRI)	True #	Reported #
Taxidea taxa	American Badger	0.714 (0.359-0.957)	0.714 (0.359-0.957)	5	5
Martes americana	American Marten	0.600 (0.194-0.932)	0.667 (0.284-0.947)	4	3
Neovison vison	American Mink	0.727 (0.528-0.887)	0.333 (0.085-0.651)	7	20
Castor canadensis	Beaver	0.159 (0.091-0.242)	0.051 (0.014-0.110)	76	86
Ursus americanus	Black Bear	0.027 (0.014-0.044)	0.140 (0.111-0.172)	498	440
Lynx rufus	Bobcat	0.074 (0.028-0.140)	0.148 (0.082-0.229)	86	79
Puma concolor	Cougar	0.750 (0.292-0.992)	0.500 (0.025-0.975)	0	2
Canis latrans	Coyote	0.228 (0.162-0.302)	0.326 (0.256-0.402)	154	134
Felis catus	Domestic Cat	0.600 (0.194-0.932)	0.333 (0.013-0.842)	1	3
Canis familiaris	Domestic Dog	0.120 (0.027-0.270)	0.154 (0.045-0.312)	24	23
Sylvilagus floridanus	Eastern Cottontail	0.674 (0.582-0.758)	0.167 (0.076-0.283)	41	105
Cervus elaphus	Elk	0.005 (0.002-0.010)	0.150 (0.131-0.169)	1321	1129
Pekaniai pennanti	Fisher	0.188 (0.043-0.405)	0.594 (0.422-0.755)	30	14
Urocyon cinereoargenteus	Gray Fox	0.251 (0.177-0.335)	0.287 (0.210-0.370)	119	113
Canis lupus	Gray Wolf	0.357 (0.221-0.506)	0.438 (0.302-0.578)	46	40
Multiple	Grouse	0.040 (0.001-0.142)	0.368 (0.225-0.525)	36	23
Homo sapiens	Human	0.194 (0.077-0.347)	0.107 (0.024-0.243)	26	29
Ondatra zibethicus	Muskrat	0.143 (0.004-0.459)	0.250 (0.037-0.579)	6	5
NA	No Animal Present	0.243 (0.225-0.261)	0.023 (0.016-0.030)	1634	2110
Didelphis virginiana	Opossum	0.333 (0.053-0.716)	0.200 (0.006-0.602)	3	4
Multiple	Other Bird	0.064 (0.026-0.116)	0.096 (0.050-0.157)	112	108
Multiple	Other Domestic	0.250 (0.008-0.708)	0.250 (0.008-0.708)	2	2
Multiple	Other Small Mammal	0.305 (0.195-0.427)	0.226 (0.125-0.347)	51	57
Erethizon dorsatum	Porcupine	0.078 (0.026-0.155)	0.385 (0.291-0.484)	94	62
Procyon lotor	Raccoon	0.048 (0.033-0.064)	0.086 (0.067-0.107)	765	734
Vulpes vulpes	Red Fox	0.175 (0.123-0.239)	0.160 (0.108-0.221)	159	162
Lutra canadensis	River Otter	0.023 (0.006-0.050)	0.045 (0.020-0.079)	177	173
Grus canadensis	Sandhill Crane	0.021 (0.001-0.077)	0.042 (0.005-0.113)	46	45
Lepus americanus	Snowshoe Hare	0.006 (0.000-0.024)	0.481 (0.424-0.538)	293	152
Sciurid spp.	Squirrel or Chipmunk	0.040 (0.029-0.052)	0.133 (0.115-0.153)	1214	1096
Mephitis mephitis	Striped Skunk	0.273 (0.067-0.556)	0.111 (0.003-0.369)	7	9
Meleagris gallopavo	Turkey	0.013 (0.005-0.025)	0.039 (0.024-0.058)	480	467
Mustela spp.	Weasel	0.600 (0.194-0.932)	0.600 (0.194-0.932)	3	3
Odocoileus virginianus	White-tailed Deer	0.018 (0.015-0.020)	0.017 (0.015-0.019)	11638	11650
Lepus townsendii	White-tailed Jackrabbit	0.938 (0.782-0.998)	0.500 (0.025-0.975)	0	14
Marmota monax	Woodchuck	0.143 (0.004-0.459)	0.333 (0.085-0.651)	7	5

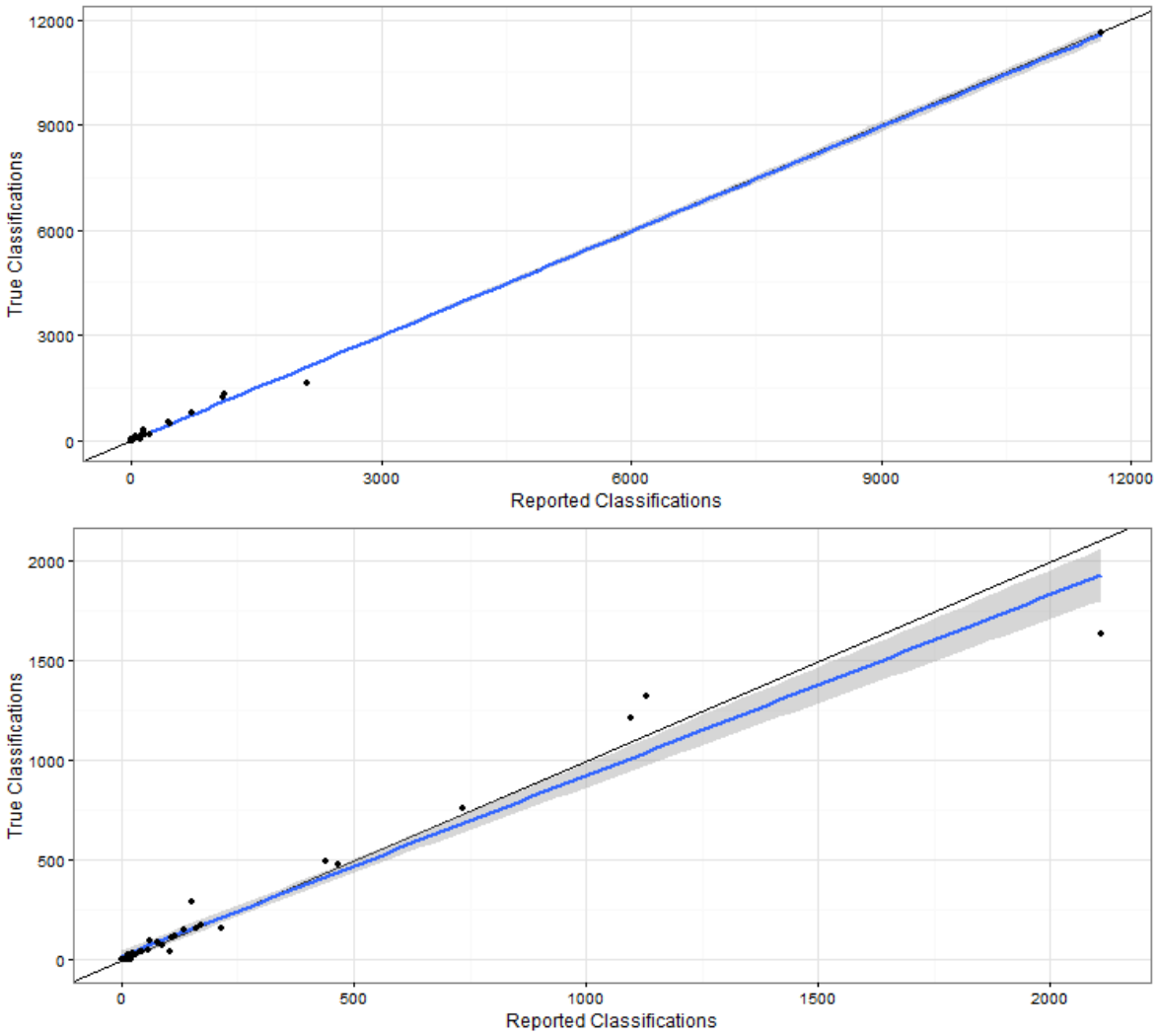

Figure S1. The association between the species-specific true prevalence in the full data set and reported prevalence based upon crowdsourced consensus classifications is linear and very close to 1:1 (top); excluding white-tailed deer, the association suggests that rare species were slightly overstated and widespread species slightly understated by crowdsourced consensus. Blue lines and shaded areas represent generalized additive model predictions and confidence intervals with smoothing based upon cross-validation.

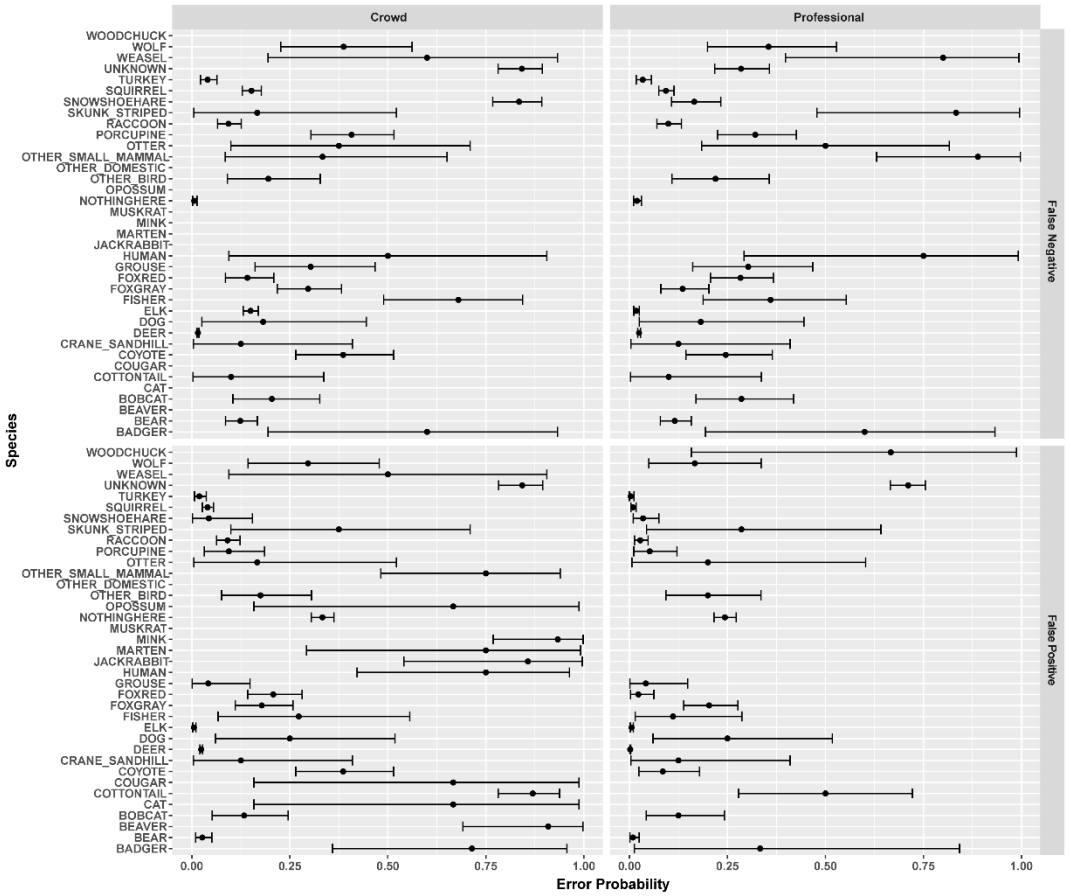

Figure S2. Crowdsourced vs. professional classification errors within a subset of the total dataset for which they could be compared. The crowd introduced more unobserved species than professionals and generally exhibited slightly more error than professionals, although even professionals did not classify many species accurately enough to meet our predefined threshold.

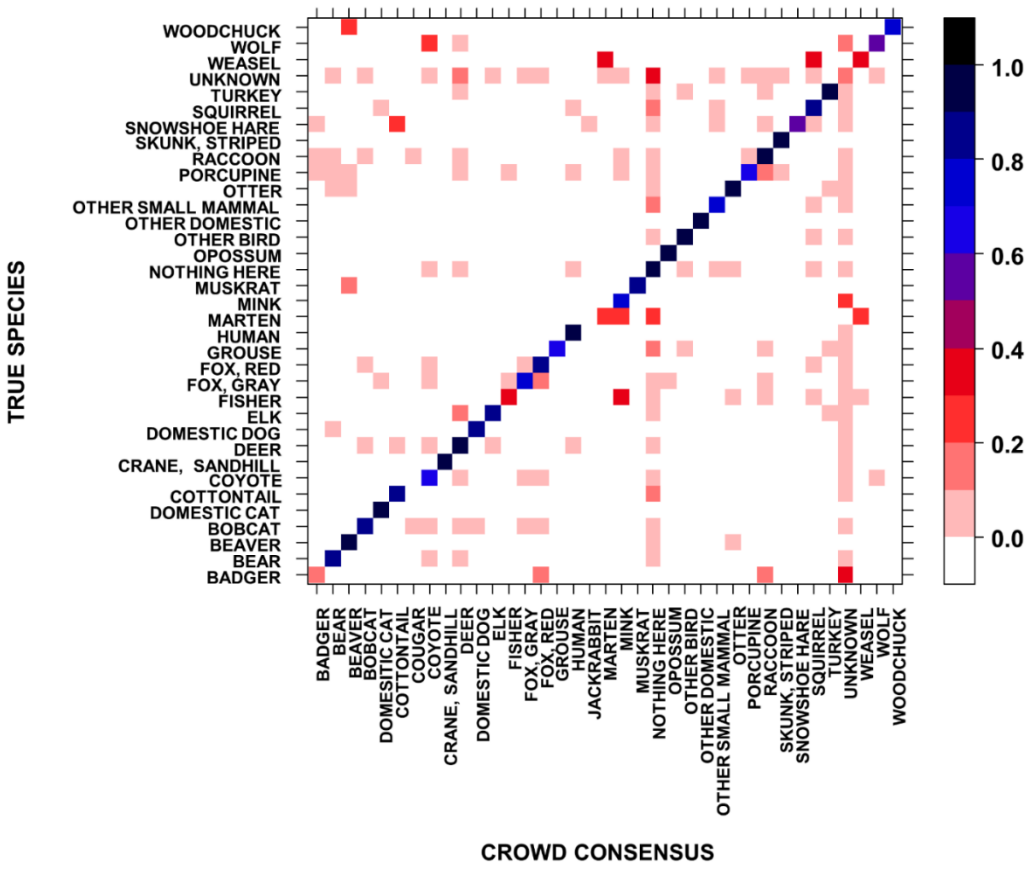

Figure S3. The proportion of images by species that were designated a particular species as assigned by a crowdsourced consensus. The prominence of the diagonal (not exact because certain species were not encountered within the dataset but were crowd-assigned) indicates that crowdsourced classification was generally correct, but interspecifically variable; the off-diagonal values highlight interspecific variability in classification confusion.

Table S2. Datasets (Season) containing the images used to evaluate crowdsourced classification accuracy. Wisconsin Wildlife Watch S2 was not used for additional data screening analysis due to lack of relevant metadata associated with confidentiality conditions for the data collection process.

Season	# Images	Camera Make/Image Resolution	Image Presentation	Within Sample Accuracy
Wisconsin Wildlife Watch S1	3847	Reconyx/High	Single Image	95.37%
Wisconsin Wildlife Watch S2	2182	Reconyx/High	Single Image	93.12%
Snapshot Wisconsin S1	12341	Bushnell/Low	Sequence of 3	92.36%
Snapshot Wisconsin S2	313	Bushnell/High	Sequence of 3	99.68%
Snapshot Wisconsin S3	638	Bushnell/High	Sequence of 3	98.27%

Table S3. Occupancy estimator performance across full range of simulation settings. If % detections verified is not indicated with a dash, the fitted model incorporated false positive observation error as well.

					Mean Error	RMSE	Relative Bias			Power	
$\psi_{average}$	$p_{average}$	% FP	% Additional FN	% Verified	PAO	PAO	PAO	β_1	β_2	β_1	β_2
0.26	0.12	3	3	-	0.05	0.05	0.12	0.27	-0.26	1.00	1.00
0.73	0.12	3	3	-	0.04	0.04	0.06	0.21	-0.21	1.00	1.00
0.26	0.047	3	3	-	0.03	0.03	0.07	0.19	-0.18	1.00	1.00
0.73	0.047	3	3	-	0.02	0.03	0.03	0.16	-0.14	1.00	1.00
0.26	0.12	10	10	-	0.15	0.15	0.39	0.53	-0.53	1.00	1.00
0.73	0.12	10	10	-	0.12	0.12	0.19	0.45	-0.44	1.00	1.00
0.26	0.047	10	10	-	0.08	0.08	0.21	0.44	-0.43	1.00	1.00
0.73	0.047	10	10	-	0.07	0.07	0.11	0.37	-0.37	1.00	1.00
0.26	0.12	30	30	-	0.36	0.37	0.96	0.72	-0.72	0.99	0.99
0.73	0.12	30	30	-	0.28	0.28	0.45	0.65	-0.65	0.96	0.97
0.26	0.047	30	30	-	0.25	0.25	0.66	0.67	-0.67	1.00	1.00
0.73	0.047	30	30	-	0.20	0.20	0.33	0.60	-0.60	0.98	0.97
0.26	0.12	10	30	-	0.13	0.13	0.35	0.52	-0.51	1.00	1.00
0.73	0.12	10	30	-	0.12	0.12	0.19	0.45	-0.45	1.00	1.00
0.26	0.047	10	30	-	0.08	0.09	0.22	0.45	-0.45	1.00	1.00
0.73	0.047	10	30	-	0.07	0.08	0.12	0.40	-0.38	1.00	1.00
0.26	0.12	30	10	-	0.35	0.35	0.92	0.71	-0.71	1.00	1.00
0.73	0.12	30	10	-	0.27	0.27	0.44	0.64	-0.63	1.00	1.00
0.26	0.047	30	10	-	0.24	0.24	0.63	0.66	-0.65	1.00	1.00
0.73	0.047	30	10	-	0.20	0.20	0.32	0.59	-0.59	0.99	0.99
0.26	0.12	10	3	-	0.13	0.13	0.34	0.49	-0.50	1.00	1.00
0.73	0.12	10	3	-	0.12	0.12	0.19	0.45	-0.44	1.00	1.00
0.26	0.047	10	3	-	0.08	0.08	0.21	0.41	-0.41	1.00	1.00
0.73	0.047	10	3	-	0.07	0.08	0.12	0.37	-0.36	1.00	1.00
0.26	0.12	30	3	-	0.35	0.35	0.91	0.71	-0.70	1.00	1.00
0.73	0.12	30	3	-	0.28	0.28	0.45	0.63	-0.63	0.99	0.99
0.26	0.047	30	3	-	0.24	0.24	0.63	0.65	-0.66	1.00	1.00
0.73	0.047	30	3	-	0.19	0.20	0.31	0.59	-0.59	1.00	1.00
0.26	0.12	3	30	-	0.04	0.04	0.09	0.27	-0.28	1.00	1.00
0.73	0.12	3	30	-	0.03	0.04	0.05	0.22	-0.22	1.00	1.00
0.26	0.047	3	30	-	0.02	0.03	0.06	0.20	-0.21	1.00	1.00
0.73	0.047	3	30	-	0.02	0.03	0.03	0.17	-0.17	1.00	1.00
0.26	0.12	3	10	-	0.04	0.04	0.10	0.25	-0.26	1.00	1.00
0.73	0.12	3	10	-	0.03	0.04	0.06	0.22	-0.21	1.00	1.00
0.26	0.047	3	10	-	0.02	0.03	0.06	0.19	-0.19	1.00	1.00
0.73	0.047	3	10	-	0.02	0.03	0.03	0.15	-0.15	1.00	1.00
0.73	0.047	3	70	-	0.02	0.04	0.04	0.23	-0.21	1.00	1.00

Table S3 (Continued)

					Mean Error	RMSE	Relative Bias			Power	
$\psi_{average}$	$p_{average}$	% FP	% Additional FN	% Verified	PAO	PAO	PAO	β_1	β_2	β_1	β_2
0.26	0.047	3	3	10	0	0.02	0	-0.03	0.03	1	1
0.73	0.12	3	3	10	0	0.01	0	0	0.02	1	1
0.26	0.047	3	3	30	0	0.02	0	-0.03	0.04	1	1
0.73	0.12	3	3	30	0	0.01	0	-0.02	0.02	1	1
0.26	0.047	3	3	50	0	0.02	0	-0.02	0.04	1	1
0.73	0.12	3	3	50	0	0.01	0	-0.02	0.01	1	1
0.26	0.047	10	10	10	0	0.02	0.01	-0.02	0.03	1	1
0.73	0.12	10	10	10	0	0.02	-0.01	-0.01	0.02	1	1
0.26	0.047	10	10	30	0	0.02	0	-0.03	0.04	1	1
0.73	0.12	10	10	30	0	0.01	0	-0.01	0.02	1	1
0.26	0.047	10	10	50	0	0.02	0	-0.03	0.05	1	1
0.73	0.12	10	10	50	0	0.01	0	-0.01	0.02	1	1
0.26	0.047	30	30	10	0.01	0.03	0.02	-0.04	0.04	1	1
0.73	0.12	30	30	10	0	0.02	0	-0.02	0.04	1	1
0.26	0.047	30	30	30	0	0.02	0.01	-0.02	0.03	1	1
0.73	0.12	30	30	30	0	0.02	-0.01	0	0.02	1	1
0.26	0.047	30	30	50	0	0.02	0	-0.04	0.05	1	1
0.73	0.12	30	30	50	-0.01	0.02	-0.01	0	0.01	1	1

Table S4. Candidate factorial predictors (Predictor) evaluated as potential random grouping factors for subsequent modeling of classification error within trail camera images, Akaike Information Criterion (AIC), and model support (w_i).

Predictor	AIC	ΔAIC	w_i
Crowd Reported Species	4591	0	1
True Species	5842	1250	0
Camera Location	7262	2670	0
Season	7558	2966	0

Table S5. Candidate models used to identify accurately classified trail camera images, and predictive performance upon withheld data (AUC = area under the Receiver Operating Characteristic curve, pAUC = corrected area under the ROC curve with false positive rates ranging from 0 to 0.1, MCC=Matthews correlation coefficient, PPV = positive predictive value). Bold values associated with the selected model model. [A]All model terms with a squared superscript were entered as a quadratic, e.g., $jday^2$ denotes that $jday+jday^2$ were used.

Model	AUC	pAUC	MCC	PPV
1|species	0.863	0.259	0.491	0.949
1|species +User Proportion+TWSC+$jday^2$+ $dectime^{2A}$	0.965	0.786	0.659	0.972
1|species +User Proportion+$TWSC^2$+$jday^2$+ $dectime^2$	0.964	0.784	0.656	0.973
1|species +User Proportion	0.961	0.776	0.646	0.971
1|species +User Proportion+$jday^2$	0.963	0.783	0.649	0.971
1|species +User Proportion+TWSC+$jday^2$	0.966	0.795	0.656	0.973
1|species +User Proportion+TWSC	0.965	0.79	0.653	0.973
1|species +User Proportion+$TWSC^2$	0.964	0.787	0.651	0.973
1|species +User Proportion+ $jday^2$+ $dectime^2$	0.962	0.77	0.653	0.972
1|species+User Proportion + $TWSC^2$+$jday^2$	0.965	0.794	0.653	0.972
(1+ User Proportion+TWSC)|species	0.964	0.786	0.653	0.973
(1+ User Proportion+TWSC)|species+ $TWSC^2$	0.963	0.778	0.652	0.972
(1+ User Proportion+$TWSC^2$)|species	0.962	0.771	0.651	0.973
(1+ User Proportion)|species	0.961	0.771	0.647	0.972
(1+ User Proportion+ $jday^2$)|species + TWSC	**0.968**	**0.804**	**0.675**	**0.973**
(1+ User Proportion+ $jday^2$)|species	0.965	0.787	0.668	0.972
(1+ User Proportion)|species + TWSC	0.964	0.786	0.654	0.973
(1+ User Proportion)|species + $TWSC^2$	0.964	0.781	0.653	0.973
(1+ User Proportion+ $jday^2$)|species + $TWSC^2$	0.967	0.801	0.672	0.973
1+User Proportion+TWSC+$jday^2$+ $dectime^2$	0.908	0.424	0.554	0.959
1 +User Proportion+$TWSC^2$+$jday^2$+ $dectime^2$	0.908	0.419	0.55	0.959
1+User Proportion	0.892	0.311	0.53	0.954
1+User Proportion+$jday^2$	0.881	0.201	0.535	0.956
1+User Proportion+TWSC+$jday^2$	0.912	0.457	0.553	0.959
1+User Proportion+TWSC	0.912	0.463	0.554	0.959
1+User Proportion+$TWSC^2$	0.912	0.463	0.547	0.959
1+User Proportion+ $jday^2$+ $dectime^2$	0.892	0.317	0.532	0.956
1+User Proportion+ $jday^2$+ $TWSC^2$	0.906	0.404	0.548	0.959

Table S6. Performance of candidate models across subsets of the test partition. Bold values are associated with the selected model; note that the selected model is not always the best across all metrics.

Model	PPV Deer	PPV Bear	PPV Canid	PPV Rare	MCC Deer	MCC Bear	MCC Canid[A]	MCC Rare[B]
1|species +User Proportion+TWSC +jday^2+ dectime^2	0.985	0.971	0.861	0.887	0.454	0.745	0.522	0.682
1|species +User Proportion+TWSC^2 +jday^2+ dectime^2	0.985	0.981	0.865	0.889	0.449	0.862	0.522	0.692
1|species +User Proportion	0.984	0.981	0.861	0.881	0.430	0.740	0.447	0.742
1|species +User Proportion+jday^2	0.984	0.981	0.855	0.881	0.427	0.862	0.444	0.743
1|species +User Proportion+TWSC +jday^2	0.986	0.971)	0.870	0.883	0.471	0.808	0.502	0.674
1|species +User Proportion+TWSC	0.985	0.971	0.873	0.881	0.481	0.808	0.522	0.664
1|species +User Proportion+TWSC^2	0.985	0.971	0.875	0.883	0.481	0.862	0.525	0.677
1|species +User Proportion+ jday^2+ dectime^2	0.984	0.980	0.855	0.889	0.427	0.740	0.464	0.745
1|species+User Proportion + TWSC^2+jday^2	0.985	0.971	0.872	0.885	0.475	0.862	0.508	0.688
(1+ User Proportion+TWSC) |species	0.986	0.980	0.867	0.879	0.479	0.740	0.489	0.692
(1+ User Proportion+TWSC) |species+ TWSC^2	0.986	0.981	0.868	0.882	0.480	0.862	0.492	0.703
(1+ User Proportion+TWSC^2)|species	0.986	0.981	0.865	0.881	0.479	0.740	0.478	0.710
(1+ User Proportion)|species	0.984	0.981	0.860	0.879	0.430	0.740	0.444	0.726
(1+ User Proportion+ jday^2)|species + TWSC	**0.986**	**0.980)**	**0.877**	**0.873**	**0.491**	**0.740**	**0.525**	**0.653**
(1+ User Proportion+ jday^2)|species	0.985	0.981	0.868	0.877	0.449	0.862	0.479	0.719
(1+ User Proportion)|species + TWSC	0.986	0.980	0.873	0.881	0.479	0.740	0.521	0.662
(1+ User Proportion)|species + TWSC^2	0.986	0.981	0.873	0.881	0.481	0.862	0.522	0.671

Table S6 (Continued)

Model	PPV Deer	PPV Bear	PPV Canid	PPV Rare	MCC Deer	MCC Bear	MCC Canid[A]	MCC Rare[B]
(1+ User Proportion+ $jday^2$)|species + $TWSC^2$	0.986	0.981	0.877	0.875	0.492	0.862	0.522	0.664
1+User Proportion+TWSC +$jday^2$+ $dectime^2$	0.986	1	0.873	0.767	0.462	0.808	0.459	0.463
1 +User Proportion+$TWSC^2$ +$jday^2$+ $dectime^2$	0.987	0.99	0.874	0.763	0.465	0.745	0.459	0.449
1+User Proportion	0.984	0.99	0.869	0.763	0.43	0.74	0.418	0.444
1+User Proportion+$jday^2$	0.985	0.99	0.868	0.776	0.44	0.74	0.465	0.444
1+User Proportion+TWSC +$jday^2$	0.986	1	0.878	0.763	0.47	0.808	0.51	0.43
1+User Proportion+TWSC	0.986	1	0.878	0.763	0.47	0.808	0.498	0.433
1+User Proportion+$TWSC^2$	0.987	0.99	0.88	0.769	0.479	0.74	0.507	0.433
1+User Proportion+ $jday^2$+ $dectime^2$	0.984	0.99	0.868	0.789	0.437	0.74	0.42	0.449
1+User Proportion+ $jday^2$+ $TWSC^2$	0.987	0.99	0.879	0.767	0.477	0.74	0.53	0.43

[A]Includes coyote, gray fox, gray wolf, red fox.
[B]Includes bobcat, beaver, domestic cat, domestic dog, fisher, white-tailed jackrabbit, American marten, American mink, Virginia opossum, river otter, porcupine, striped skunk, and weasel species.

Appendix S2 – Details associated with data evaluation

Description of the verification process

The images collected by Snapshot Wisconsin trail cameras are processed (or classified) on a crowdsourcing platform hosted by the Zooniverse (www.snapshotwisconsin.org, and precursor www.wisconsinwildlifewatch.org). We used post-hoc evaluation of processed data to estimate the baseline accuracy with which volunteers classified species within trail camera images. The classification interface consisted of a single image or a series of three images (hereafter jointly referred to as images or sequences) that a volunteer could view and classify as containing one or more of 42 potential species, with subsequent classification options related to species-specific counts, behaviors, and or types. Our evaluation focused upon species identification because it the only task germane to estimating occurrence. Volunteers received guidance from reference photographs, descriptions, and a series of filters that could be implemented to show only species with selected sizes, body shapes, or coloration (Swanson et al. 2016). Images were randomly viewed and classified by multiple volunteers until specific retirement criteria were met (1 volunteer reported a human present, first 3 volunteers or 5 total volunteers reported the image as having no animals present, 7 volunteers selected the same species within the photo, or once 15 volunteers had contributed classifications). We defined the crowd consensus classification as the species that received the most votes. We considered ties to be equivalent to a consensus of unknown species, although volunteers lacked an explicit option to classify an image as unknown because previous research suggested this option was overused (Swanson et al. 2016).

The 19,212 images considered here were classified by volunteers on the Zooniverse platform following these rules/interface. WDNR professional staff (n = 13) independently classified 12,232 images through an internal agency classification interface that featured the same classification options as the Zooniverse platform except that 1) professionals were allowed to tag an image as containing an unknown or unidentifiable species, and 2) images were viewed chronologically at specific camera locations rather than at random. Each image was classified by a single WDNR professional. Previous studies have defined experts and professionals interchangeably (Lewandowski and Specht 2015): we distinguish experts

(individuals with extensive experience classifying trail camera images from within the region) from professionals (professional employees of a natural resource agency with a background in ecology but variable task-specific experience or proficiency).We did not treat professional classifications as truth because not all agency professionals had previous experience classifying trail camera images, and classification occurred over prolonged continuous periods that may have induced observer fatigue (Swanson et al. 2016).

Experts (n = 1, or 2 if the first was uncertain; authors JC & CA-D) verified 1,051 images where the crowdsourced classification differed from the professional classification, 381 images where crowdsourced and professional classifications agreed upon a rarely detected or indistinct species (anything but white-tailed deer, turkey, elk, raccoon, sciurid, or black bear), and the expert's classification was treated as truth (7 of these consensus images were incorrect). We define JC and CA-D as experts given extensive previous experience classifying trail camera images (*c* 500,000) primarily collected within Wisconsin. Experts further reviewed a sample of 300 images where crowdsourced and professional classifications converged upon more commonly detected or distinct species: none of these were incorrect, and we operationally assume that joint classifications of these species are correct. This assumption is likely not strictly true, but we do not believe we are substantively overestimating classification accuracy given that consensus between professionals and experts for "more difficult" species was 98.2 % accurate (374/381 images correct), and assuming the 300 images sampled are reasonably representative, there is less than 5% probability that the underlying accuracy of the image populations is < 99 % if error follows a beta distribution. When unsure, CA-D and JC defaulted to an unknown classification, and we assume these images are correctly classified. We assume that images experts defined as unknown were truly unidentifiable (although technically this is a false-negative error, error could not be assigned to any given species).

JC further jointly classified 6,980 images with a crowdsourced classification on the Zooniverse platform. Because these classifications contributed to image retirement, we removed the expert vote

before determining the crowd consensus. We assume these classifications exhibit no false positive error because JC only classified images when confident.

Thus, we define the true species in an image as the expert classification for 8,712 out of 19,212 (45.3%), and otherwise as professional classifications that also concurred with crowdsourced classifications upon a fairly common or distinguishable species (white-tailed deer, turkey, elk, raccoon, sciurid, or black bear) that a more limited expert validation suggested were reliably classified (10,500 images, 54.7%).

Additional comments related to the case study's verification process.

Table S1 (Appendix S1) provides estimates of crowdsourced classification false positive and false negative error by species. Although the overall accuracy of crowdsourced classifications across all data considered was 93.2%, there was stronger agreement between crowdsourced classifications and gold-standard expert classifications using the same interface (96.1% crowd accuracy). Within the pool of images (n = 12341) evaluated both by professionals and crowdsourcing, professional species-level classifications were 94.1% accurate, and crowdsourced classifications were 93.4% accurate. Because classifications were performed under different circumstances/interfaces, we emphasize that these values can only be compared cautiously. We suspect that viewing images serially carries a substantive advantage in terms of classification accuracy but warn that it makes dealing with misclassification more complex (see Appendix S3). Figure S2 (Appendix S1) provides a side by side comparison of crowdsourced and professional classification error within the dataset that was not gold standard. Professional and crowdsourced species-specific classification accuracy generally correlated (Figures S2 and S3 in Appendix S1). Professionals used the "unknown" option available to them and were slightly less likely to falsely introduce species than the crowd, although this did lead to (in many of these cases, the image was identifiable). Importantly, professional classification was also variable across species. We reiterate previous warnings that relying upon paid employees (commonly students or technicians) with some

natural resource training to classify trail camera images or collect other ecological data does not invariably produce reliable data.

Although rare species were subject to greater false positive and false negative crowdsourced classification error, there was strong correlation between the true prevalence of species within the dataset and the reported prevalence of species within the dataset (i.e., correlation between the number of true images for different species and the number of consensus classifications by species; $r > 0.99$, Figure S1 in Appendix S1), and there was little association between the true prevalence of a species within the dataset and the degree to which crowdsourced classifications overstated or understated overall prevalence (i.e., the correlation coefficient for the number of true images and the ratio of true images and consensus classifications < 0.02).

But although crowdsourced classifications did not appear to systemically over or under report the prevalence of certain species, error rates were greater for less-prevalent species, and generally speaking, interspecific confusion appeared to increase as species prevalence decreased (Figure S3, Appendix S1). In some cases, there appeared to be strong confusion between a limited number of species (for example, snowshoe hare were commonly misclassified as cottontail, woodchucks were exclusively misclassified as beavers), but other species were subject to less specialized misclassification (porcupine, for example, were misclassified as several species). Although professional classification accuracy as a whole was comparable to crowdsourced accuracy, one notable difference between professional and crowdsourced classifications was that professionals tended to default towards classifying an image as unknown or having nothing in it: as a result, there was much less meaningful false-positive error associated with professional classifications. Although Swanson et al. (2016) recommend that trail camera crowdsourcing interfaces avoid providing volunteers with an “unknown” or “unidentifiable” classification option (i.e., forcing choice *sensu* Raddick et al. 2013), we suggest that the value of such an option depends upon the costs associated with false-positive vs. false negative errors.

In the main text we note that the reported species within the image was the best factorial predictor of false-positive error considered. In the main text, we note that this suggests that volunteers may default to certain species when unsure of its identity. Anecdotally, we believe that observers may generally default to extremely common species when images contained few obvious cues (e.g., bears or elk very close to cameras seemed to be classified as deer more frequently than expected), or towards more charismatic species (e.g., images of deer or bobcats reported as cougars). One potential action (not implemented) that might help with this is providing volunteers the option to report a measure of confidence when classifying (Anton et al. 2018), which might help managers distinguish between informed and speculative decisions.

However, misclassification did vary across all factors considered (the true species in the image, the specific camera location, the season; for example, accuracy varied by 7% across seasons, although this was the least supported candidate factor, Table S4 in Appendix S1). Evaluating data quality is probably most effective as an iterative process (noted by Kosmala et al. 2016 and within the main text). In the initial stages of data evaluation, it may be more desirable and easier to target a single approach where returns appear to be greatest, but as more data is verified, more refined strategies can be developed and identified.

The same issue arises when implementing an action to improve data quality. For analysis presented in the main text, we removed predictors that captured design issues that were no longer germane or exhibited limited predictive performance within initial analyses using gradient-boosted models (Friedman 2001). We ultimately used parametric models to evaluate data screening as a remediation action because random effects structures associated with parametric models more naturally accommodate variance across unbalanced factors (species) with vastly differing sample sizes than recursive-partitioning methods, the models are more transferable (Wegner and Olden 2012), and they provided better performance than machine-learning methods. Although we chose to employ a single model with random effects for data screening given its performance across the dataset for the time being,

its performance varied across species (see Tables S5 and S6 in Appendix S1 for model results). Provided enough samples have been verified, it seems likely that data screening algorithms may be most effective when different species are screened using distinct and independent models.

Investigating reliability of crowdsourced count classifications

Although we do not explore this in the main text, many studies rely upon distinct data classification tasks (e.g., species and count, species and state, etc.). Within the full dataset, we found that consensus between crowdsourced and either expert or professional counts was 95.9%; if consensus was greater if multi-modal crowdsourced counts were all assumed to be the lower candidate number (96.5%) than the larger candidate (95.9%). Here, explicit task and interface differences make defining the "truth" difficult. Some agency staff viewing images (again, 3-trigger sequences) defined the count as it might be perceived on the crowdsourcing platform: the number of individuals within the specific sequence. Other staff seemed to define counts more contextually, and when viewing a serial set of images, appeared to define the count as the total number of unique individuals seen within the series. It is not clear which interpretation is necessarily correct.

Miscounting appeared to positively covary with misclassifying (i.e., the probability that both species and animal were classified correctly was 88.3 %; the expectation under independence is 86.7%). This is partially structurally unavoidable because a classifying an image's species as "nothing present" naturally undercounts the number of animals in the image. Discounting these 0's, crowd-sourced counts were still slightly lower, which is probably preferable to a mixture of zero-inflation and overcounting from the perspective of trying to model error in count data.

Regardless, the combination of count error and classification error presents challenges for practitioners explicitly attempting to use counts within a modeling exercise. We were unable to develop any screening/censure model that was more accurate than the baseline data, so censuring putative "miscounts" would have induced more error into the dataset than initially existed. Strength of

crowdsourced count consensus and variance in the reported count appeared to be important (positively-associated) predictors of count reliability, but the most important predictor was the "true" count (smaller numbers of animals were counted more reliably), which is not a useful indicator for screening data. Furthermore, for our own project (and many others on the Zooniverse platform) the number of animals (or animal states) is a forced-choice task analogous to species classification; that is, the true and reported count follow a categorical/multinomial distribution, which is different than how counts are assumed to be distributed within many ecological models.

Improving the efficiency of data evaluation efforts.

Our data evaluation process was far from perfect, and we make brief suggestions to improve future efforts here. Stratifying verification or calibration across different categorical responses or predictors (e.g., species, observers) may be important if these factors are strongly imbalanced. A single data task (e.g., species identification) may contribute to multiple downstream predictors or responses (e.g., several distribution models). Investigators should keep the analytical usage of the data in mind when designing data evaluation efforts, and seek to maximize replication at the appropriate hierarchical unit: all else being equal, characterizing variance within a few responses or predictors may be less important than characterizing variance between these factors. Furthermore, if the assumed response distribution has unequal variance, it may be beneficial to exert more effort towards evaluating error in categorical parts of the parameter space with greater intrinsic uncertainty because estimates (on the real scale) will naturally be more diffuse: i.e., for a binomial task, focusing more effort towards species or individuals (etc.) with $p = 0.5$. These issues coalesced within our study and may have a more general linkage for species identification tasks: a random sample of observations will provide an imbalanced sample of species, and the species most commonly sampled (or used for training an algorithm) will be identified most accurately (Swanson et al. 2016, Nourouzzadeh et al. 2018). Imbalance within our data and our non-stratified verification sampling strategy provided us the most power to estimate error parameters and explain variation in error for the organisms that were already most accurately identified and least required a large

sample to precisely estimate error parameters. We have focused ongoing verification efforts towards rarer species with greater intrinsic rates of error. Savvy investigators may be able to formalize adaptive sampling principles to guide data evaluation.

Supporting References

Anton, V., S. Artley, A. Geldenhuis, and H. U. Wittmer. 2018. Monitoring the mammalian fauna of urban areas using remote cameras and citizen science. Journal of Urban Ecology 4:juy002.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 1189-1232.

Kosmala, M., A. Wiggins, A. Swanson, and B. Simmons. 2016. Assessing data quality in citizen science. Frontiers in Ecology and the Environment 14:551-560.

Nourouzzadeh, M.S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune. 2018. PNAS 115:E5716-E5725.

Raddick, M. J., G. Bracey, P. L. Gay, C. J. Lintott, C. Cardamone, P. Murray, K. Schawinski, A. S. Szalay, and J. Vandenberg. 2013. Galaxy Zoo: Motivations of Citizen Scientists. arXiv: 1303.6886.

Swanson, A., M. Kosmala, C. Lintott, and C. Packer. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology 30:520-531.

Wenger, S. J., and J. D. Olden. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3:260-267.

Appendix S3, Section 1: Additional Simulation Study Details.

We briefly outline our simulation implementation here (code is found within). As noted in the main text, we first generated a "true" detection history based upon input values for p and ψ by simulating true presence and absence across sites and simulations as $z_{i,sim} \sim \text{Bernoulli}(\psi_{\text{i,sim}})$, and $y_{i,j,sim} \sim \text{Bernoulli}(p_{\text{i,sim}} \times z_{i,sim})$. We derived the empirical proportion of samples in which there was a detection as $\bar{y}_{sim}$ and generated a "ghost" false positive detection history as $x_{i,j,sim} \sim \text{Bernoulli}(\bar{y}_{sim} \times fp)$, where fp is the user-controlled proportion of false positive detections within a dataset. Thus, the actual sample-based false positive detection probability as defined by Chambert et al. (2015) is much lower than fp (instead, this something more similar to an observation level false positive probability). We implement false-positives in this fashion because either site-based or sample-based false positive detection probabilities described by Royle and Link (2006), Miller et al. (2011), Chambert et al (2015), etc. are estimands that are difficult to estimate without fitting a model (because these reflect the probability of detecting an organism at a site where it is not present, or the probability of detecting an organism falsely during any sampling interval, and species presence is not known, while the definition of sampling alters the sample-specific false positive parameter); instead, some overall assessment of observation accuracy is more likely to be known. However, the actual proportion of false-positive detections within a dataset cannot be known without validating all samples (Gardiner et al. 2012), and so we incorporated Binomial variance across simulations to reflect sampling uncertainty in the actual proportion of false positive detections. We implemented additional false-negative error associated with misclassification by randomly manipulating all true positive observations within a dataset to absences as Bernoulli(fn), where fn is a user-defined input (i.e., $y_{i,j,sim} | y_{i,j,sim} = 1$ is thinned as Bernoulli($1 - fn$)). The "actual" detection history used to evaluate the standard occupancy estimator described by MacKenzie et al. (2002) was then derived as the maximum of the thinned $y_{i,j,sim}$ and $x_{i,j,sim.}$

We employed a similar generating process to evaluate the observation confirmation false positive occupancy estimator described by Chambert et al. (2015). The hierarchical model can be described as:

$$z_i \sim \text{Bernoulli}(\psi)$$

$$y_{ij} \sim \text{Bernoulli}(z_i p_{11} + (1 - z_i) p_{10})$$

$$z_s \sim \text{Bernoulli}(\psi)$$

$$v_{st} \sim \text{Categorical}(\mathbf{\Omega})$$

Here, indices *i* and *j* refer to sites and sampling intervals where data has not been verified, and *s* and *t* index sites and sampling intervals at stations where data has been completely verified. The vector $\mathbf{\Omega}$ describes the probability that a sampling intervals has been verified as containing no observation, a true detection only, a false detection only, and both true and false detections within an interval and can be derived as $[\{z_s(1 - s_1) + (1 - z_s)s_0\}\ \{z_s s_1 + z_s(1 - s_0)\}\ \{(1 - z_s)s_0 + z_s s_0(1 - s_1)\}\ \{z_s s_0 s_1\}]$. Parameters p_{11} and p_{10} are derived as $p_{11} = s_1 + s_0 - (s_1 s_0)$, and $p_{10} = s_0$. (Note that we use different ordering for *v* within the markdown document that contains code for replication).

We altered the model such that *y* and *v* have the same indexing. We envision that most verification within our study, and perhaps many others, will happen across a variety of sites haphazardly rather than a complete verification at a subset of sites: verifying observed absences could become very inefficient if many other species are observed, or vegetation is repeatedly triggering a camera, etc. However, $\boldsymbol{y}$ and $\boldsymbol{v}$ must be mutually exclusive, because the observations within a sampling interval are either verified or not (i.e., an observed value of $\boldsymbol{v}$ must correspond to missing data within $\boldsymbol{y}$ and vice versa). We generated $\boldsymbol{v}$ as encompassing a fixed proportion of positive detection samples within each simulation (e.g.., taking 10% of all positive values of $\boldsymbol{y}$ for simulated verification and replacing these values with missing data). This means that within the simulations, $\boldsymbol{v}$ only took on categorical values corresponding to the validated sample containing only a false positive detection, only a true positive detection, or both a true positive detection and a false positive detection. Importantly, $\mathbf{\Omega}$ must be defined to include a probability that $\boldsymbol{v}$ might contain neither true nor false detections--although this was never a simulated outcome--because it was requisite for valid estimation of detection parameters associated with

$\boldsymbol{y}$. These slight alterations do not appear to alter the efficacy of the estimator (Table S4 in Appendix S1 contains summary statistics pertaining to estimator performance for all scenarios considered).

Appendix S3, Section 2: "Observation" level error and false positive models

It is possible that a given sampling interval at a specific site may contain multiple observations depending upon how intervals are defined, how an organism is being recorded/detected, how common a species is, and other factors. We ignored this heterogeneity within our simulation study—in essence, we treated the verified data as including a single image that was correctly classified, a single image that was incorrectly classified, or one of each—for simplicity. As shown in Figure S1, if true detections are associated with a large number of constituent true positive outcomes observed within a sampling occasion, and false positive detections continue to happen at random, the bias associated with a given proportion of false positive observations increases. In contrast, if false positive observations "cluster" within sampling occasions more than true-positive observations, bias associated with misclassification is reduced. Although the data are not formally presented here, the organisms within our study that tend to produce clustered observations within sampling intervals were identified accurately enough to withstand any effects associated with a misbalanced distribution of true and false positive observations within sampling intervals. However, we emphasize that our simulation results pertaining to the sensitivity of the base occupancy estimator under varied proportions of false-positive observations should be interpreted cautiously. The observation confirmation model described by Chambert et al. (2015) and used here remains useful when sampling occasions contain multiple observations, and presented results remain valid provided that all *observations* within *sampling occasions* are used for verification. However, as sampling intervals typically contain an increasing number of observations, both the estimator itself and the verification process may become more inefficient, as it will require more effort to verify all observations in a given sampling interval, and because the observed categorical value of v is likely to depend upon the number of observations.

Thus, in some cases, it may be preferable for convenience or estimation purposes to model false positives directly at the observation level. Perhaps the most straightforward way to do so is to model the encounter process as the observation level, and an advantage of this approach is that coefficient estimates (and uncertainty intervals) produced by an observation screening exercise are directly translatable and could be used to produce informed priors within an occupancy (or other) model. Chambert et al. (2018) describe an observation level model for acoustic detectors in which true and false positive observations arise from distinct Poisson processes within sampling intervals. In contrast, remote cameras typically produce highly overdispersed observation totals within intervals because observations results from Markovian movement and/or Markovian residence due to bait responses. The most effective way to formulate an encounter model for camera observations may be as a Markov modulated process (either in discrete time, Hines et al. 2010, or continuous time, Guillera-Arroita et al. 2011). Extending these models to account for false positive observations could be a useful research avenue, particularly if investigators are interested in evaluating variability in space-use at high temporal resolution (Dorazio and Karanth 2017).

It is also possible to deal with error at the observation level error while maintaining an estimator rooted in repeated observations summarized within sampling intervals. Chambert et al. (2015) briefly discuss possible extensions to deal with varying numbers of observations within intervals, and we formalize a description here. As with the standard occupancy model, let z_i denote the binary occupancy state at site *i* and assume it is distributed as Bernoulli (ψ); let *p* denote the probability of detecting an organism conditional upon it being present at site *i* during sampling interval *j*. Let *v* denote a sequence of *o* verified observations, where $v = 1$ indicates a correctly reported observation, and $v = 0$ indicates a false observation. A false positive observation occurs with probability $(1 - r_1)$, such that $v_{site[o]} \sim \text{Bernoulli}(z_i \times p)$. Let s_0 and s_1 represent the respective probabilities that all observations within a sampling interval are false positive or the unconditional probability of recording > 0 true positive observations within a sampling interval at a site. Respectively, these can be derived as $s_{0,i,j} = I(nobs_{i,j}) \times (1 - r_1)^{nobs_{i,j}}$, where

$nobs_{i,j}$ is the number of recorded observations within interval j at site i, and $I(nobs_{i,j})$ denotes an indicator function that takes a value of 1 if $nobs_{i,j} > 0$ and a value of 0 otherwise; and $s_{1,i,j} = z_i \times p$; unverified presences or absences $y_{i,j} \sim$ Bernoulli($s_{1,i,j} + s_{0,i,j}$). That is, a verified observation $v_{site[o]}$ can be a true positive if and only if the species is present at the site, detected during the interval, and not falsely identified, and a species can be recorded at a non-verified sampling occasion at a specific site either if the species is present, detected, and correctly identified within > 0 detections, or if all observations are false-positive. Sample code in the BUGS language is provided below. Although the existence of multiple observations within a sampling interval is treated as a nuisance, because error is modeled at the observation level, the coefficients associated with a data screening exercise could be translated into model terms here, too: e.g., logit($r_{0,o}$) = $\boldsymbol{\beta X_o}$, with derivation of the sample level false positive parameter requiring a vector or matrix of predictors associated with each of $nobs_{i,j}$.

One critical underlying assumption in effectively all described models accounting for false-positive detections explicitly is that these errors are independent. As noted in SI2, serial image classification may be more accurate because it provides context for distinct observations. But as a trade-off, it may complicate models for dealing with misclassification: i.e., it may requisite to treat *observation error* as Markovian process if images or related observations are classified serially rather than independently.

Supporting References

Chambert, T., D.A.W. Miller, and J. D. Nichols. 2015. Modeling false positive detections in species occurrence data under different study designs. Ecology 96:332-339.

Chambert, T., J. H. Waddle, D. A. W. Miller, S.C. Walls, and J.D. Nichols. 2018. A new framework for analyzing automated acoustic species detection data: occupancy estimation and optimization.of recordings post-processing. Methods in Ecology and Evolution DOI: 10.1111/2041-210X.12910.

Dorazio, R.M., and K.U. Karanth. 2017. A hierarchical model for estimating the spatial distribution and abundance of animals detected by continuous-time recorders. PLoS One 12:e0176966.

Gardiner, M. M., L. L. Allee, P. M. Brown, J. E. Losey, H. E. Roy, and R. R. Smyth. 2012. Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. Frontiers in Ecology and the Environment 10:471-476.

Guillera-Arroita, G., B.J.T. Morgan, M.S. Ridout, and M. Linkie. 2011. Species occupancy modeling for detection data collected along a transect. Journal of Agricultural, Biological, and Environmental Statistics 16:301-317.

Hines, J.E., J.D. Nichols, J.D. Royle, D.I. MacKenzie, A.M. Gopalaswamy, N.S Kumar, and K.U. Karanth. 2010. Tigers on trails: occupancy modeling for cluster sampling. Ecological Applications 20:1456-1466.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248-2255.

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

Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835-841.

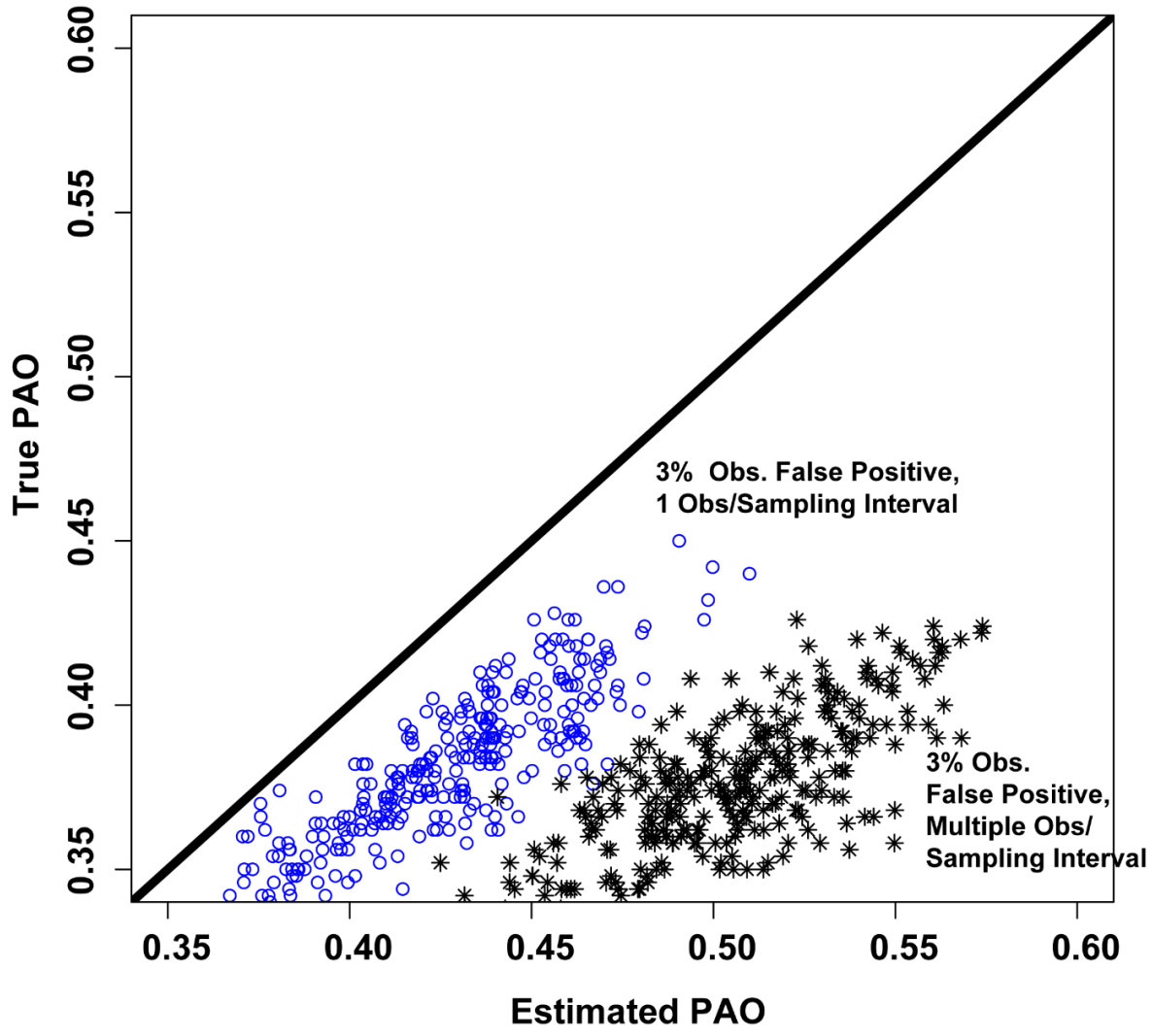

Figure S1. The sensitivity of the base occupancy model to different proportional levels of false positive observations within a dataset depends upon how observations are allocated within sampling intervals. Blue dots show the sensitivity of the estimator when 3% of detections are false positive, but there is at most 1 true and 1 false positive observation per interval (the settings used throughout the paper); black stars show that the model is more sensitive to the same proportion of error when, for example, true positive observations are distributed within sampling intervals as z[i]*p[i]*(Poisson(2)+1), while false positive observations occur at most one time within a sampling interval.

###Psuedocode in BUGS language for an observation confirmation false positive ###occupancy model described above in section 2.

```
model{
    psi0~dbeta(1,1)
    B0<-logit(psi0)
    B1~dnorm(0, .1)
    p0~dbeta(1, 1)
    A0<-logit(p0)
    A1~dnorm(0, .1)
    r1~dbeta(1, 1)
    for (i in 1:nsites){
      logit(psi[i])<-B0+B1*Psi.cov[i]
      logit(p[i])<-A0+A1*p.cov[i]
      z[i]~dbern(psi[i])
      for (j in 1:nvisits){
        s0[i, j]<-ifelse(nobs[i, j] > 0, (1-r1)^nobs[i, j], 0)
            ###prob all observations in the sampling interval for site i are
            ###false positive
        s1[i, j]<-z[i]*p[i] ###standard prob of detection
        y[i, j]~dbern(s1[i, j]+s0[i, j])
###Note, if y[i, j] contains verified sample that has been confirmed as a
###true positive, than the datum should be removed from the likelihood to
###avoid double usage (i.e., set to NA): that the species has been detected
###and confirmed is contained within v. If y[i, j] contains a verified sample
###that has been confirmed as a false positive but there are other samples
###within y[i,j], the confirmed sample should be subtracted from nobs[i,j],
###but the cell should still contain values (it is still possible that one
###observation was not false positive.
      } #j loop
    }#i loops
        for (o in 1:nobs){
        v[o]~dbern(s1[site[o]]*r1)
###This is probability of conditionally detecting and not falsely
###identifying an organism within a single sample

###Note, if the observations are from sites or sampling intervals that are
###distinct from y and not of direct inferential inference, this could be
###formulated as v[o]~dbern(r1) such that the occupancy state or p
###of this site are not considered. Such might be the case if false positives
###are evaluated using calibration with test "sites" for which investigators
###do not wish to infer occurrence.
        }#o loop
}#model end
```

Chapter 2 - Generalized model-based solutions to false positive error in species detection/non-detection data.

John D. J. Clare, Philip A. Townsend, and Benjamin Zuckerberg

Department of Forest and Wildlife Ecology, University of Wisconsin – Madison, Madison, Wisconsin

Citatation:

Clare, J. D. J, P. A. Townsend, and B. Zuckerberg. 2020. Generalized model-based solutions to false positive error in species detection/non-detection data. Ecology. DOI:10.1002/ecy.3241

Abstract

Detection/non-detection data are widely collected by ecologists interested in estimating species distributions, abundances, and phenology, and are often subject to imperfect detection. Recent model development has focused on accounting for both false positive and false negative errors given evidence that misclassification is common across many sampling protocols. To date, however, model-based solutions to false positive error have largely addressed occupancy estimation. We describe a generalized model structure that allows investigators to account for false positive error in detection/non-detection data across a broad range of ecological parameters and model classes, and demonstrate that previously developed model-based solutions are special cases of the generalized model. Simulation results demonstrate that estimators for abundance and migratory arrival time ignoring false positive error exhibit severe (20-70%) relative bias even when only 5-10% of detections are false positives. Bias increased when false positive detections were more likely to occur at sites or within occasions in which true positive detections were unlikely to occur. Models accounting for false positive error following the site confirmation or observation confirmation protocols generally reduced bias substantially, even when few detections were confirmed as true or false positives or when the process model for false positive error was misspecified. Results from an empirical example focusing on gray fox (*Urocyon cinereoargenteus*) in Wisconsin, USA reinforce concerns that biases induced by false positive error can also distort spatial predictions often used to guide decision-making. Model sensitivity to false positive error extends well beyond occupancy estimation, but encouragingly, model-based solutions developed for occupancy estimators are generalizable and effective across a range of models widely used in ecological research.

Introduction

Detection/non-detection data are widely collected by ecologists interested in monitoring populations or elucidating habitat associations (MacKenzie et al. 2002). It is now widely recognized that species detection/non-detection or occurrence data and many types of ecological survey data suffer from imperfect detection: the actual occurrence of species or individuals is rarely perfectly observed. Concerns about false negative error, the failure to detect an individual or species when present, have been recognized for decades and have motivated the development of a broad range of models that explicitly account for it (e.g., MacKenzie et al. 2002, Royle and Nichols 2003).

False positive error, another type of imperfect detection within species occurrence data, can occur when a non-target species or another phenomena (e.g. extrinsic sound) is misclassified as the focal species of interest (McClintock et al. 2010). Molecular assays for infectious agents face similar issues, as sample contamination or non-specific amplification can result in false positive test results (Brost et al. 2018). Across a variety of species and sampling protocols, the incidence of false positive error has been estimated as varying from nearly negligible to constituting 20% of observations or more (e.g., McClintock et al. 2010, Swanson et al. 2016). Simulation results have shown that even relatively few false positive detections can severely bias estimates of species occupancy when models account for false negatives but ignore false positives (Miller et al. 2011, Ruiz-Gutiérrez et al. 2016).

The prevalence of false positive error has spurred investigators to adopt a variety of strategies aimed at ameliorating potential biases. For species detection/non-detection data, strategies include a complete data review after collection, data collection or processing methods aimed at reducing the incidence of false positives, and model-based approaches. Implementing complete data reviews requires all data to be reviewable and can be burdensome or infeasible for large datasets (Gardiner et al. 2012, Ruiz-Gutiérrez et al. 2016). Specific data collection or processing protocols aimed at reducing the incidence of false positives include performing partial data reviews to develop indicators or algorithms for identifying false positives, simplifying classification tasks, or providing additional guidance or training to

human or computer-based classifiers (Miller et al. 2012a, Swanson et al. 2016, Kosmala et al. 2016). Although these approaches can greatly improve data quality, they exhibit certain inefficiencies. It can be time-consuming to quantify how accurately data have been classified, what level of classification accuracy is sufficient for a specific research objective, and whether manipulating training or classification protocols result in sufficient classification accuracy.

Model-based solutions are a more efficient way to ameliorate false positive error, and several variants for detection/non-detection data have been described specifically for occupancy models (Royle and Link 2006, Miller et al. 2011, Chambert et al. 2015, Ferguson et al. 2015, Ruiz-Gutiérrez et al. 2016, Brost et al. 2018). The original occupancy model described by MacKenzie et al. (2002) conceptualizes observed absences as a mixture of true and false negatives. Occupancy models that account for false positive error conceptualize detections as a mixture of true and false positives. Under the 'full' estimator described by Royle and Link (2006) all observed occurrences are of unknown reliability, and disentangling the false negative and false positive mixtures requires constrained priors. Subsequent developments leverage auxiliary data collected under different protocols to improve discrimination between true and false positives. Investigators following the 'site confirmation' protocol are able to unambiguously classify some detections as true positives (e.g., via *a posteriori* confirmation or by paired sampling with using a method free from false positive error), while other detections are ambiguously true positives or false positives (Miller et al. 2011, Ferguson et al. 2015). The 'observation confirmation' protocol is an extension upon the site confirmation protocol in which some detections can be unambiguously classified as true positives or false positives (e.g., via *a posteriori* laboratory tests, Chambert et al. 2015). The 'calibration' protocol involves assessing classification performance within settings in which the ecological state variable is experimentally controlled (e.g., playback experiments or negative laboratory controls, Ruiz-Gutiérrez et al. 2016, Brost et al. 2018). These model-based solutions alleviate bias in occupancy estimators while sparing time associated with performing complete data reviews, making changes to classifier training, or developing and calibrating error indicators.

The development and application of model-based solutions to false positive error has largely focused on occupancy estimation. However, species detection/non-detection data are increasingly used to estimate state variables other than occupancy, including abundance, phenology, and associated dynamics (Royle and Nichols 2003, Roth et al. 2014, Chandler and Clark 2014, Ramsey et al. 2015, Rossman et al. 2016). Presumably, these estimators are similarly biased by false positive error, and these biases could severely hamper a broad set of ecological decisions ranging from assessing the recovery of protected populations to delineating seasonal protections for migratory species.

Here, we show that previously-developed model-based solutions to false positive error can be described as specific cases of a generalized model to account for false positive error in detection/non-detection data. We use simulation to quantify the bias caused by false positive detections across estimators of different state variables like abundance or phenological phenomena such as migratory arrival or emergence from hibernation, and to demonstrate that model-based solutions commonly improve inference across a range of estimation problems. Although generalization is not restricted to any protocol, we primarily focus on study designs following the observation confirmation protocol (Chambert et al. 2015), which is the most applicable when researchers use sampling techniques that produce data that can be reviewed and verified *a posteriori*.

Methods and Results

Generalizing Model-based Solutions for False Positives

Let $\boldsymbol{y}$ denote a matrix of binary observations corresponding to the detection or non-detection of a species at $i = 1, 2,\ldots R$ locations over $j = 1, 2,\ldots T$ discrete sampling occasions. If a species or more specific species-state of interest such as a juvenile is observed, $y_{i,j} = 1$, and $y_{i,j} = 0$ otherwise. We proceed assuming that $\boldsymbol{y}$ is *repeated* detection/non-detection data. However, the concepts apply across data with different dimensions or for data collected slightly differently. For example, $\boldsymbol{y}$ could denote a vector of

presence-background data, where $y_i = 1$ indicates that the species was detected or occurred at location i, and where $y_i = 0$ indicates that location i is part of the randomly selected background sample.

Parametric models for detection/non-detection data typically assume $y_{i,j} \sim$ Bernoulli (θ), where θ is the unconditional probability of detection at a specific place and time. This probability is equivalent to the union of the respective unconditional probabilities of true positive detection (θ_{tp}) and false positive detection (θ_{fp}). If $\theta_{fp} = 0$, $\theta \equiv \theta_{tp}$ and the union between true and false positive detections does not need to be explicitly specified. For example, MacKenzie et al. (2002) define the unconditional probability of detection $\theta_i = z_i \times p$, where z_i is the latent binary occupancy state of site i distributed as Bernoulli (ψ), ψ is a probability of occupancy that might vary in relation to site level covariates, and p is the probability of detecting an organism at a site given that it is present and might vary in relation to either site and occasion level covariates. An equivalent but less compact description is that $\theta_{fp} = 0$, $\theta_{tp,i} = z_i \times p$, and $y_{i,j} \sim$ Bernoulli ($\theta_{tp} \cup \theta_{fp}$). If $\theta_{fp} > 0$, decomposing θ into true and false positive probabilities is critical for unbiased estimation of θ_{tp}, which is typically the focus of ecological inquiry. A model-based solution to false positive error within species detection/non-detection data is simply any model that explicitly assumes

$$y_{i,j} \sim \text{Bernoulli}\ (\theta_{tp} \cup \theta_{fp}) \tag{1}$$

and estimates these probabilities or their constituent parameters. Specific solutions may differ with respect with respect to how θ_{tp}, $\theta_{fp,}$ and their union are defined, and other estimation details.

The union between θ_{tp} and θ_{fp} may take a few forms. The first is what we refer to as an inclusive union or form. In certain sampling situations, false positives might be possible at both the site level (detections at unoccupied sites) and the observation level (false detections at occupied sites). Further, it might be possible for a target species or entity to be truly and falsely detected at the same location during the same sampling occasion. In particular, if a spatial or temporal interval is used to define an occasion (e.g., 1 km of transect, a 24 hr interval), the target species may be detected correctly or falsely multiple times. When collapsing this count of detections into binary data denoting > 0 detections or not, any $y_{i,j} = 1$

may include true and/or false positives. Assuming true and false positives occur independently or are conditionally independent given a set of covariates, $\theta_{tp} \cup \theta_{fp} = \theta_{tp} + \theta_{fp} - (\theta_{tp} \times \theta_{fp})$. In turn, the distribution of the collapsed detection/non-detection data can be described as:

$$y_{i,j} \sim \text{Bernoulli}\ (\theta_{tp} + \theta_{fp} - [\theta_{tp} \times \theta_{fp}]) \quad (2)$$

The union above is a factorization of 3 distinct probabilities (Chambert et al. 2015, Brost et al. 2018): the probability of > 0 true positives only ($\theta_{tp} \times [1 - \theta_{fp}]$), the probability of > 0 false positives only ($\theta_{fp} \times [1 - \theta_{tp}]$), and the probability of > 0 true positives and > 0 false positives ($\theta_{tp} \times \theta_{fp}$). For example, at a single camera trap during a day-long sampling occasion, the target species might be recorded and correctly identified, a non-target species might be recorded and misidentified as the target, or both. An occupancy model accounting for inclusive false positive error is a particular case of (1) in which $\theta_{tp,i} \cup \theta_{fp} = \theta_{tp,i} + \theta_{fp} - (\theta_{tp,i} \times \theta_{fp})$, $\theta_{tp,i} = z_i \times p$, and where θ_{fp} is a constant or some combination of parameters to be estimated.

In other situations where site-level and observation-level false positives are possible, true and false positives might be mutually exclusive events, in which a detection can only be a true positive or a false positive. Here, the product $\theta_{tp} \times \theta_{fp} = 0$, so $\theta_{tp} \cup \theta_{fp} = \theta_{tp} + \theta_{fp}$ and:

$$y_{i,j} \sim \text{Bernoulli}\ (\theta_{tp} + \theta_{fp}) \quad (3)$$

We refer to this as an "observationally exclusive" union or form. For example, a camera trap might be programmed to take a single time-lapse image per day-long sampling occasion, in which case the target species might be detected and correctly classified, or some other species might be detected and misclassified as the target species, but not both (assuming 1 organism within the image). An occupancy model accounting for observationally exclusive false positive error is a particular case of (1) in which $\theta_{tp,i} \cup \theta_{fp} = \theta_{tp,i} + \theta_{fp}$, $\theta_{tp,i} = z_i \times p$ and θ_{fp} is a constant or combination of parameters to be estimated. Note that one can use (3) rather than (2) even if false positive error is inclusive, provided one interprets a false positive as any $y_{i,j} = 1$ where zero observations are true positive. In this case, θ_{fp} represents the

unconditional probability that *all* observations at some location and occasion are false positives. Interpretation and estimation of θ_{tp} is unchanged.

A specific form that is most commonly used within model-based solutions for false positive error makes assumptions we refer to as "conditionally exclusive". As with the observationally exclusive union, $y_{i,j} \sim$ Bernoulli ($\theta_{tp} + \theta_{fp}$). What differentiates this form is that θ_{tp} and θ_{fp} are not only mutually exclusive within any cell $y_{i,j}$, but that there are spatial or temporal conditions under which only true positive observations are possible, and if these conditions are not met, all observations are false positive. This formulation is commonly employed within occupancy models accounting for false positive error following the full, site-confirmation, or calibration protocols. Probabilities p_{tp} and p_{fp} (often referred to as p_{11} and p_{10}) are conditional upon whether site i is occupied ($z_i = 1$) or not ($z_i = 0$), respectively (e.g., Miller et al. 2011). These are particular cases of (1) in which $\theta_{tp,i} \cup \theta_{fp,i} = \theta_{tp,i} + \theta_{fp,i}$, $\theta_{tp,i} = z_i \times p_{tp}$, and $\theta_{fp,i} = (1 - z_i) \times p_{fp}$. Although the assumption that only site-level false positives can occur is probably violated in many sampling situations, occupancy estimation is not strongly biased by such violations (although coverage suffers; Ferguson et al. 2015, Brost et al. 2018).

Estimating θ_{tp} and θ_{fp} following the observation confirmation protocol

The observation confirmation protocol requires investigators to review and confirm all detections within some subset of occasions at some subset of sites *a posteriori*. Confirmed data $\boldsymbol{v}$ are classified such that $v_{i,j} = 1$ if all detections at a site and occasion were confirmed to be true positives, $v_{i,j} = 2$ if all were confirmed as false positives, and $v_{i,j} = 3$ if confirmation reveals both true and false positive detections. If the species is not detected at all and there is nothing to verify, $v_{i,j} = 0$.

Chambert et al. (2015) assume $v_{i,j} \sim$ Categorical ($\boldsymbol{\Omega_i}$) and condition vector $\boldsymbol{\Omega_i}$ upon whether the species occurs at site i or not. Let s_{tp} and s_{fp} denote the respective probabilities that a sampling interval contains > 0 true positive detections at an occupied site or > 0 false positives. If $z_i = 1$ (species occurs), it may not be detected, it can be truly detected only, falsely detected only, or detected both ways, and $\boldsymbol{\Omega_i} =$

$[\{(1 - s_{tp}) \times (1 - s_{fp})\}\ \{(s_{tp} \times (1 - s_{fp})\}\ \{(1 - s_{tp}) \times s_{fp}\}\ \{s_{tp} \times s_{fp}\}]$; brackets {} are used to delineate the distinct scalars in $\mathbf{\Omega_i}$. If $z_i = 0$, the only possible outcomes are no detection or a false positive detection. Removing conditioning upon a specific occupancy state and substituting $\theta_{tp,i}$ for $z_i \times s_{tp}$ and θ_{fp} for s_{fp} yields that $\mathbf{\Omega_i} = [\{(1- \theta_{tp,i}) \times (1- \theta_{fp})\}\ \{\theta_{tp,i} \times (1- \theta_{fp})\}\ \{(1- \theta_{tp,i}) \times \theta_{fp}\ \}\ \{\theta_{tp,i} \times \theta_{fp}\}]$. The first scalar is equal to the probability of non-detection, and the latter three are the probabilities of specific detection outcomes factorized in (2).

In turn, Chambert et al. (2015) assume unconfirmed detection/non-detection data $y_{i,j} \sim$ Bernoulli $(z_i \times [s_{tp}+ s_{fp} - (s_{tp} \times s_{fp})] + [1 - z_i] \times s_{fp})$. If present ($z_i = 1$), a species can be detected ($y_{i,j} = 1| z_i = 1$) either truly, falsely, or both: $s_{tp}+ s_{fp} - s_{tp} \times s_{fp}$ denotes the factorization of these conditional probabilites. If the species is not present ($z_i = 0$), it can only be falsely detected ($y_{i,j} = 1| z_i = 0$). Substituting $\theta_{tp,i,}$ for $z_i \times s_{tp}$ and θ_{fp} for s_{fp} yields that $y_{i,j} \sim$ Bernoulli $(\theta_{tp,i} + \theta_{fp} - [\theta_{tp,i} \times \theta_{fp}])$. This is exactly (2).

That is, an occupancy model following the observation confirmation protocol is a special case of (2) where $\theta_{tp,i} = z_i \times s_{tp}$ and $\theta_{fp} = s_{fp}$. Note that s_{tp} is equivalent to what MacKenzie et al. (2002) call p. A hierarchical description of the complete data likelihood is then:

$$z_i \sim \text{Bernoulli } (\psi)$$

$$\theta_{tp,i} = z_i \times p$$

$$\mathbf{\Omega_i} = [\{(1 - \theta_{tp,i}) \times (1 - \theta_{fp})\}\ \{\theta_{tp,i} \times (1- \theta_{fp})\}\ \{(1- \theta_{tp,i}) \times \theta_{fp}\}\ \{\ \theta_{tp,i} \times \theta_{fp}\}]$$

$$v_{i,j} \sim \text{Categorical } (\mathbf{\Omega_i})$$

$$y_{i,j} \sim \text{Bernoulli } (1 - \Omega_{1,i}\ )$$

A point which we will return to is that $\theta_{tp,i}$ is the same as the unconditional probability of detection—Pr $(y_{i,j} = 1)$—presented by MacKenzie et al. (2002).

First we acknowledge some superficial differences between the presentation here and by Chambert et al. (2015; Table 1). They denote s_{tp} and s_{fp} as s_1 and s_0, respectively. They present p_{10} as the

probability $y_{i,j} = 1 | z_i = 0$, but is redundant ($p_{10} = s_0$ [or s_{fp}] = θ_{fp}) and only one term is needed. Similarly, s_{tp} (or s_1) is redundant with the original model's p (MacKenzie et al. 2002). Chambert et al. (2015) derive $p_{11} = s_{tp} + s_{fp} - (s_{tp} \times s_{fp})$ to describe $\Pr(y_{i,j} = 1 | z_i = 1)$, but this derivation is not strictly necessary. Importantly, notation s_0, p_{10}, and p_{11} also lacks a consistent interpretation within the existing literature (Table 1). The site confirmation and calibration protocols typically condition s_0 and p_{10} on absence, and p_{11} strictly represents a true positive probability conditional upon presence. The observation confirmation protocol defines s_0 as an unconditional probability of falsely detecting a species and defines p_{11} as the probability of detection (either true or false positive) at occupied sites (Chambert et al. 2015).

We assume inclusive false positive error in the extensions described below, but an observationally exclusive formulation could be implemented by assuming $\theta_{tp,i} \times \theta_{fp} = 0$, and reducing $\boldsymbol{\Omega_i}$ to describe three detection states corresponding to no detection, a true positive only, or a false positive only (Appendix S1, Figure S1): $\boldsymbol{\Omega_i} = [\{1 - (\theta_{tp,i} + \theta_{fp})\}\ \{\theta_{tp,i}\}\ \{\theta_{fp}\}]$.

Extension to Non-occupancy Models

If assuming an inclusive or observationally exclusive union between true and false positives and following the observation confirmation protocol, the only difference between the occupancy model described above and a different model accounting for false positive error relates to how $\boldsymbol{\theta_{tp}}$ is defined (Figure 1; Appendix S1, Figure S1). Extension to other model classes requires deriving $\boldsymbol{\theta_{tp}}$ as the unconditional probability of detection presented in the original model description, and appropriately specifying $\boldsymbol{\theta_{tp}}$'s constituent processes. We present two examples below.

Royle-Nichols Model

Royle and Nichols (2003, RN model hereafter) describe the unconditional probability of detection (i.e., $\theta_{tp,i}$) as $1 - (1 - r)^{N_i}$, where r is the probability of detecting an individual during a sampling interval, and N_i, distributed as Poisson (λ), denotes the abundance of a species at site i. The hierarchical likelihood

for a version assuming inclusive false positive error following the observation-confirmation protocol is then:

$$N_i \sim \text{Poisson}\ (\lambda)$$

$$\theta_{tp,i} = 1 - (1 - r)^{N_i}$$

$$\boldsymbol{\Omega_i} = [\{(1 - \theta_{tp,i}) \times (1 - \theta_{fp})\}\ \{\theta_{tp,i} \times (1 - \theta_{fp})\}\ \{(1 - \theta_{tp,i}) \times \theta_{fp}\}\ \{\theta_{tp,i} \times \theta_{fp}\}]$$

$$v_{i,j} \sim \text{Categorical}\ (\boldsymbol{\Omega}_i)$$

$$y_{i,j} \sim \text{Bernoulli}\ (1 - \Omega_{1,i})$$

The only differences between this model and the occupancy model presented in the previous section are within the first two lines: a process model for N_i replaces a process model for z_i, and $\theta_{tp,i}$ is derived as a function of r and N_i rather than z_i and p.

Phenological 'Arrival' Model

Incorporating false positives within an occupancy model designed to estimate the timing of some ephemeral phenomena such as migration arrival or emergence from torpor (Roth et al. 2014; PA model hereafter) follows Chambert et al.'s (2015) description except that organisms can only be truly detected during sampling occasions at occupied sites after arrival. Thus, $\boldsymbol{\theta_{tp}}$ must be described using indexing for i locations and j time periods (i.e., $\theta_{tp,i,j}$). Let arrival time at site i be denoted as x_i and assume that $x_i \sim$ Poisson (φ). To simplify presentation, we define x_i in terms of sampling intervals j rather than specific dates. The hierarchical likelihood is:

$$z_i \sim \text{Bernoulli}(\psi)$$

$$x_i \sim \text{Poisson}\ (\varphi)$$

$$\theta_{tp,i,j} = z_i \times p \times I(j \geq x_i)$$

$$\boldsymbol{\Omega_{i,j}} = [\{(1 - \theta_{tp,i,j}) \times (1 - \theta_{fp})\}\ \{\theta_{tp,i,j} \times (1 - \theta_{fp})\}\ \{(1 - \theta_{tp,i,j}) \times \theta_{fp}\}\ \{\theta_{tp,i,j} \times \theta_{fp}\}]$$

$$v_{i,j} \sim \text{Categorical}(\boldsymbol{\Omega_{i,j}})$$

$$y_{i,j} \sim \text{Bernoulli}\ (1 - \Omega_{1,i,j})$$

Here, $I(j \geq x_i)$ is an indicator function denoting whether occasion j is equal or greater than the specific time of species arrival at site i. Again, the specification of a process model for x_i and the redefinition of $\boldsymbol{\theta_{tp}}$ are the only differences between the model above and the model presented by Chambert et al. (2015).

Only a small sample of the possible extensions are described above. Appendix S1 contains other examples and describes how to account for false positives across model types when using different approaches to estimate false positives.

Exploring Model Sensitivity to False Positive Incidence and the Number of Verified Samples

We undertook a simulation study to evaluate the baseline sensitivity of the RN and PA models to different amounts of false positive error and the performance of extensions using the observation confirmation protocol to account for false positives. We describe the simulation settings and present results below. Throughout, we fixed the simulated sampling effort as 200 sites with 20 sampling occasions each.

RN Model

We first considered six different simulation scenarios representing combinations of across two abundance levels and three different incidences of false positive error. We generated 300 replicate datasets per scenario with site-specific abundances $N_{i,sim} \sim$ Poisson ($\lambda_{i,sim}$) and log ($\lambda_{i,sim}$) = $\beta_0 + \beta_1 X_{1,i,sim}$, where $X_{1,i,sim} \sim$ N (0, 1), β_0 = 0 or -1.5 (3 scenarios each), and β_1 = 1; *sim* indexes a particular simulation replicate. Thus, at a site with an average simulated covariate ($X_{1,i}$ = 0), expected abundance was respectively roughly 0.23 animals or 1 animal. These values were chosen because the RN model tends to perform best when site-specific abundance is low (Kéry and Royle 2016, p. 302) and it is perhaps most commonly applied to low-density species. We first generated 'true' detection data as Bernoulli ($p_{i,sim}$), where $p_{i,sim} = 1 - (1 -$

$r_{i,sim})^{N_i,sim}$, logit ($r_{i,sim}$) = $\alpha_0 + \alpha_1 X_{2,i,sim}$, $X_{2,i,sim} \sim$ N (0, 1), $\alpha_0 = -1.73$, and $\alpha_1 = 1$. Thus, an individual at an average site was expected to be detected with a probability of about 0.15 per sampling occasion.

Within these scenarios (Appendix S2, Table S1), we generated false-positive detections as occurring at random across all site intervals within a simulation (i.e., inclusive false positives). The probability of a false-positive detection within a cell was derived such that out of all detections, approximately 1%, 5%, or 10% were false positives with random Binomial sampling variance (absolute values of $\theta_{fp,sim}$ ranged from < 0.001 to roughly 0.025). We defined θ_{fp} proportionally here rather than explicitly exploring specific values for the parameter itself because few studies provide empirical estimates of unconditional false positive probabilities, many report percentages of observations that are true or false positives (e.g., Simons et al. 2007, Norouzzadeh et al. 2018), and proportional definitions are also commonly used to define thresholds for "accurate" data (e.g., 95% accuracy; Swanson et al. 2016).

Within each scenario, we sampled cells at random to create the verified data $v_{i,j,sim.}$ We considered 10 levels for the number of site × occasion cells in which all detections were verified = {10, 20, 30, 40, 50, 60, 70, 80, 90 100}, with each scenario including 30 simulated datasets for each verification level. Each of the 1800 generated datasets (300 replicate datasets for each of the six simulation scenarios) was used to fit both a standard RN model and an extension accounting for false positive error. For each estimator, we evaluated performance using mean error (absolute bias), root squared error, standard deviation of the posterior distribution, coefficient of variation, and frequentist coverage (% of 95% CIs that included the true value) for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and the finite sample population size ($\widehat{N}^{\text{tot}}$, derived as $\sum_{i=1}^{R} \widehat{N_i}$; we used relative [%] bias rather than mean error for this parameter). Because absence cannot be confirmed, allocating non-detections between $\boldsymbol{v}$ and $\boldsymbol{y}$ is an arbitrary decision: Pr($v_{i,j} = 0$) = Pr($y_{i,j} = 0$), and we left all non-detections within $\boldsymbol{y}$ (Clare et al. 2019). We fit all models here and below using JAGS v 4.0 (Plummer 2003) to perform Markov-Chain Monte Carlo simulation through R v 3.4 (R Core Team 2017), although neither Bayesian estimation nor the complete data likelihood are prerequisite (code using maximum-likelihood estimation is available from the lead author).

As expected, the RN model became increasingly biased as false positives constituted a greater proportion of detections (Figure 2; Appendix S2). Random misclassification across all time periods and locations constituting 1%, 5% or 10% of all detections led to respective relative biases of roughly 10%, 40%, and 70% across both simulated expected abundance levels. Models accounting for false positive error exhibited less bias and root mean squared error regardless of the size of the verified sample. Estimator performance asymptotically improved as more samples were verified, with minimal improvement once detections within between 30 - 50 site × occasion cells were confirmed (Figure 2, Figure S1, Appendix S2).

PA Model

Our exploration of the PA model was similar (Appendix S2, Table S5 outlines simulation scenarios). We first considered three scenarios with the following parameterization: logit ($\psi_{i,sim}$) = $\beta_0 + \beta_1 X_{1,i,sim}$, $X_{1,i,sim} \sim$ N (0, 1), $\beta_0 = 0$, and $\beta_1 = 0.5$; logit ($p_{i,sim}$) = $\alpha_0 + \alpha_1 X_{2,i,sim}$, $X_{2,i,sim} \sim$ N (0, 1), $\alpha_0 = -2$, $\alpha_1 = 0.5$, and average arrival time φ = occasion 6. True observations $y_{i,j,sim}$ were generated as Bernoulli ($z_{i,sim} \times p_{i,sim} \times I(j \geq x_{i,sim})$), where $z_{i,sim} \sim$ Bernoulli ($\psi_{i,sim}$), and site and simulation specific arrival time $x_{i,sim} \sim$ Poisson (φ). That is, the occupancy probability at an average site was 0.50 and the probability of true detection conditional on arrival and occupancy at an average site was roughly 0.12. As before, we simulated 300 replicates per scenario, false positive detections constituted 1%, 5%, or 10% of all detections, and the size of $v_{i,j,sim}$ ranged from 10-100. We fit both the standard PA model and the false-positive extension to each simulation replicate. We evaluated estimator properties with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\hat{\varphi}$ and a finite sample estimate of the proportion of occupied sites ($\widehat{PAO}$, derived for each simulation as $\sum_{i=1}^{R} \hat{z}_i$).

Results for the PA model largely mirrored those for the RN model. Across the range of false positive detection proportions, bias in the estimated proportion of area occupied and arrival time when false positive errors were ignored ranged from 3% - 20%, and 3% - 40%, respectively (Figure 3, top panels). The extended model greatly reduced bias with any amount of verification effort, with little further reduction once 30 samples were confirmed (Figure 3; Appendix S2, Figure S2). Estimates of the

proportion of area occupied were more uncertain when accounting for false positive error, but more precise for arrival time when accounting for false positive error (Figure 3, bottom right), suggesting that false positive error was inducing overdispersion in arrival estimates relative to Poisson expectations.

Evaluating the suitability of site confirmation models

We briefly explored the consequences of lacking the ability to confirm false positives, and incorrectly assuming all false positives were conditionally exclusive (i.e., assuming false positives only occurred at unoccupied sites or sites prior to arrival, and ignoring false positives occurring at occupied sites or occurring concurrently with true positives). Following Chambert et al. (2015), we reconfigured the simulated data described above so that it mimicked the data collected under the site confirmation protocol by considering only confirmed true positives. We fit the RN and PA models using site confirmation approaches that incorrectly assumed false positives only occurred at unoccupied sites or occupied sites prior to arrival. These models were also generally unbiased given 30 – 50 confirmed true positives (Appendix S2, Figures S3 and S4, Tables S2 and S6). However, the site confirmation variant of the RN model occasionally (92 out of 1800 simulations) had difficulty converging, particularly when few true positives were confirmed and simulated abundance was small. The site confirmation RN model also exhibited less than nominal frequentist coverage of finite-sample population size, apparently driven by increased root mean squared error (Appendix S2, Table S3 and S4).

Model Sensitivity to Variability in False Positive Error

False positive detections probably rarely happen randomly across space or time because the misclassified entities (other species or extrinsic phenomena) are not typically randomly distributed. Misclassification in detection/non-detection data is essentially spatiotemporal error in observed species occurrence, and the effects of false positive error upon models using detection/non-detection data are analogous to the effects of spatial error upon models using presence/background or used/available data (Johnson and Gillingham 2008). If false positives always occur within the same sites and occasions as

true positives, inference regarding the true positive process is not affected. In general, one would expect false positive error to result in larger bias if there were greater spatial or temporal (and by extension, environmental) separation between true and false positive detections.

RN Model

We considered two subsequent scenarios (Table S1, Appendix S2) where $\beta_0 = -1.5$ and $\beta_1 = 1$ with $\boldsymbol{\alpha}$ defined as before, and $\text{logit}(\theta_{fp,i}) = -6 + \{-1, 1\}X_{1,i,sim}$. As previously, $\log(\lambda_{i,sim}) = \beta_0 + \beta_1 X_{1,i,sim}$. Thus, false positive observations were either more or less likely in locations with greater expected abundance and probability of true positive observations. The verification protocol was simulated as previously. Empirically, simulated false positive observations constituted about 6% of all observations. We fit three models to each scenario: one that assumed $\theta_{fp} = 0$, one that (incorrectly) assumed θ_{fp} was a constant in order to evaluate the consequences of ignoring variation in false positive error, and one that (correctly) modeled $\theta_{fp,i}$ as varying in relation to $X_{1,i,sim}$.

Abundance estimates using the standard RN model were more biased when false positive detections were more likely to occur at locations where the focal species was less abundant and where true positive detections were less likely (Appendix S2, Figure S5). Abundance estimates produced by extensions accounting for false positive error were nearly unbiased (< 5% relative bias) regardless of the generating model for false positive error, whether it was correctly specified within the fitted model, or the number of confirmed samples (Appendix S2, Figure S5). Again, the performance of models accounting for false positive error asymptotically improved when more samples were confirmed. Although models with a misspecified model for false positive error improved more slowly, all models were essentially unbiased once 100 samples were confirmed.

PA Model

We varied when and where false positive errors occurred relative to the baseline settings within six further scenarios (Table S5, Appendix S2). In the first four scenarios, we defined $\theta_{fp,i,j,sim} = \text{logit}^{-1}(-6 + \{-$

1, 1}$X_{1,i,sim}$) for $j > 4$ and $\theta_{fp,i,j,sim} = 0$ for $j \leq 4$, and $\alpha_0 = \{-2, -1\}$. In the second two, $\theta_{fp,i,j,sim} = \text{logit}^{-1}(-6 + \{-1, 1\}X_{1,i,sim})$ for $j > 2$ and $\theta_{fp,i,j,sim} = 0$ for $j \leq 2$, and $\alpha_0 = -2$. Other values followed the previous description. That is, we used a shared covariate to make false positives either more or less likely to occur at occupied sites, and altered the timing of false positives so that rather than happening at any time, they initiated either 1 or 3 occasions before the average time of arrival. Empirically, these different formulations for false positive error resulted in false positives accounting for between 3% ($\alpha_0 = -2$ and $\theta_{fp,i,j,sim} = 0$ for $j \leq 4$) and 7% ($\alpha_0 = -1$ and $\theta_{fp,i,j,sim} = 0$ for $j \leq 2$) of all detections. As with the RN model, we fit models that assumed $\theta_{fp} = 0$, models that (incorrectly) assumed θ_{fp} was a constant, and models that (correctly) assumed θ_{fp} varied in relation to $X_{1,i,sim}$.

The performance of the PA models accounting for false positive error largely mirrored results seen with the RN model: there was little difference in model performance regardless of whether the probability of false positive error was mis-specified as a constant or allowed to vary spatially, and any differences shrank as the number of verified samples increased (Appendix S2, Figures S6 and S7). For models ignoring false positive error, estimates of the proportion of area occupied became more biased as simulated false positive error started earlier and when false positives were more likely in places where true positives were less likely (Appendix S2, Figure S6). Moreover, occupancy estimation became more biased when the conditional probability of truly detecting a species was lower, indicating some potential sensitivity to small-sample bias.

When false positives were ignored, estimates of arrival time were insensitive to spatial patterns in error, but estimator bias increased as the simulated initiation of false positives occurred earlier relative to the average true arrival time (Appendix S2, Figure S7). In fact, when false positives were simulated as starting only one sampling occasion before true positives, the estimator ignoring false positives exhibited less bias and smaller RMSE with respect to arrival time than the generating estimator. We believe this to be a specific case of offsetting biases induced by false negative and false positive error (see discussion in Appendix S2).

Application: Predicting Gray Fox Relative Abundance across Wisconsin

Models for detection/non-detection data are often used to predict state variables spatially in order to prioritize management or conservation actions (e.g., Guélat and Kéry 2018). As a case study, we focus upon the relative abundance of gray fox (*Urocyon cinerargenteus*) in Wisconsin, USA, where its distribution is poorly understood. We used data from a monitoring program where trail camera images are imperfectly classified via a crowdsourcing platform (Clare et al. 2019) to investigate spatial patterns in fox relative abundance using the RN model. We modeled variation in fox expected abundance using a model accounting for false positive error and one ignoring it. We used indicator variable selection (Kuo and Malick 1998) to identify important predictors and regularize the log-linear coefficients within each model, and made statewide predictions for each by applying the model-averaged posterior predictive distribution across a 2 x 2 km lattice (more detail in Appendix S3).

Out of the images we reviewed, 67% were correctly classified; after further aggregation within 179 distinct 24-hr sampling occasions (i.e., all detections within 179 occasions were reviewed), 60% consisted of only true positives, and 40% consisted of only false positives (either coyote, *Canis latrans* or red fox, *Vulpes vulpes*). Indicator variable selection provided less support for the inclusion of abundance covariates within the standard model than the model accounting for false positive error, and the latter model suggested that false positive error varied spatially in relation to the prevalence of surrounding cropland (Appendix S3, Tables S1 and S2). Consequently, predictions from the standard model exhibit different spatial patterning (Figure 4; although the statistical correlation between pairwise pixel estimates was fairly strong; $r = 0.80$). Furthermore, the point estimate for expected state-wide population size derived via summation across the cells used for prediction was > 300% larger when false positives were ignored (Appendix S3), although estimates overlapped substantially due to imprecision induced by spatial smoothing and the sparsity of observations. Although the estimated probability of a false positive detection per sampling interval was very small (at an average site, 0.0015, 95% CRI 0.0012 – 0.0018,

Appendix S3), this estimate suggests there were > 100 false positive detections given 91,276 total sampling occasions within the dataset.

Discussion

Our results reiterate that when unaccounted for, false positive detections can compromise a broad range of ecological applications and inferences. As expected, simulation results demonstrate that estimates of abundance and phenology can be biased by even moderate amounts of false positives. Although not tested via simulation, our case study suggests that biased parameter estimation can further lead to skewed spatial predictions. Estimation of abundance, which lacks a natural limit, appears particularly sensitive to false positive error. When false positives were randomly generated, inclusive, and constituted 10% of all detections, the RN model (and a related unmarked spatial capture-recapture model, Appendix S1) exhibited 70% relative bias. For comparison, we have observed that occupancy models achieve this level of bias only when false positive detections generated exactly in the same manner constitute 30% of all detections (Clare et al. 2019), and equal incidence of observationally or conditionally exclusive false positive detections would be expected to induce yet greater bias (e.g., Miller et al. 2011). Given that detection/non-detection data often form the backbone of efforts to assess and monitor species populations or phenologies across large scales (e.g., Jetz et al. 2019, Robinson et al. 2019, Sun et al. 2019), techniques to ameliorate these errors are an important need. Luckily, our results suggest that existing model-based solutions designed for occupancy estimation are effective across a broad range of model classes that rely on detection/non-detection data.

We focused on model-based approaches following the observation confirmation protocols. Researchers with the capability to confirm some subset of observations as true and false positives are broadly equipped to deal with false positive error across different model classes by re-specifying $\boldsymbol{\theta_{tp}}$ to reflect the unconditional probability of detection within the model class of interest, specifying a process model for θ_{fp}, and collecting the necessary auxiliary data. We also demonstrated the extensibility of site confirmation approaches, and believe the calibration protocol is similarly flexible (Appendix S1). Two

primary factors underlying the efficacy of these solutions relate to the number of confirmed or calibrated detections and how correctly the models for true and false positives are specified. We discuss these factors below, but preface by acknowledging that the extensibility of model-based approaches makes it challenging to broadly quantify the requisite confirmation effort or model structure for all possible applications. We encourage further simulation across a broader set of model classes (e.g., involving dynamics, data integration, or disease infection intensity; Miller et al. 2012b, Chandler and Clark 2014, Rossman et al. 2016) and protocols to help clarify these considerations.

The amount of auxiliary data required for unbiased estimation of the ecological state variables of interest largely likely depends upon several factors. Our simulation results suggest that when false positive error occurs at random and constitutes 10% or less of all detections, unbiased estimation following the site or observation confirmation protocols may only require confirmed detections within 30 - 50 site by occasion intervals. Brost et al. (2018) demonstrate that the calibration protocol can be similarly reliable given 50 trials with known negatives. As the incidence of false positive error is often less than 10% (e.g., McClintock et al. 2010), many applications may not require an exhaustive confirmation sample. More detections may need to be confirmed or calibrated to achieve unbiasedness if false positives are more common or the observed data is sparse (Ruiz-Gutiérrez et al. 2016). Because sampling efficiency and the incidence of false positive error is often difficult to gauge *a priori*, investigators may be better suited by confirming as many samples as feasible during initial project phases in order to buffer against uncertainty (Clement 2016). As such, perhaps the most important factor to consider when choosing a protocol to account for false positive error is which approach is likely to generate the largest amount of auxiliary data with the least effort or expense (Chambert et al. 2015, Ruiz-Gutiérrez et al. 2016). Note that a single 'confirmation' under the observation confirmation protocol requires that all observations at a site and specific occasion have been confirmed as true, or that all have been confirmed as false positives, or that > 0 true and false positive observations have been confirmed. A confirmation following the site confirmation protocol is simply any site by occasion in which > 0 true

positives have been confirmed: this data may be substantially easier to generate in certain sampling situations.

A connected reason to assimilate more auxiliary data is to better model variation in where and when false positives occur. Modeling variation in $\boldsymbol{\theta_{fp}}$ using covariates or spatiotemporal dependence terms can account for potential estimator biases associated with missing heterogeneity (Miller et al. 2015), and can be critical for reliably predicting spatial patterns or trends in species distributions (*sensu* Guélat and Kéry 2018, case study here). Furthermore, as our simulations demonstrate, understanding the covariance between true and false positive detections can provide insights into the amount of estimator bias likely to be induced by false positives. Although the observation confirmation protocol appeared robust to mis-specifying the false positive process when confirmation followed a random sample and the true positive process was correctly modeled, it seems unlikely that all applications will be as robust. Appropriately modeling variation in true and false positive processes may be particularly critical if the data at hand provide little capacity to differentiate true and false positives. For example, fully latent estimators are particularly sensitive to model structure (Miller et al. 2015). We note these performed well when applied to RN and PA models when provided informed priors for false positive parameters and all generating processes were properly parameterized (Appendix S2). However, while we agree with Miller et al. (2015) that such approaches deserve more consideration if no other options are available, we strongly recommend data confirmation or calibration if possible, as they provide some buffer against the risk of choosing poor prior distributions or specifying poor models for false positive error.

Assumptions regarding where and when false positives can occur may also deserve further consideration. Because the false positive process can be directly quantified, the observation confirmation and calibration protocols allow investigators to assume true and false positive observations are either inclusive or conditionally exclusive. The site confirmation protocol appears to require further constraints to address observation level false positives (Appendix S1). Assuming that false positives can only occur in situations where true positives are impossible constrains the range of possible observation outcomes,

adding precision and making certain approaches identifiable. However, violations of the assumption may carry costs. Here, the PA model was unaffected, but uncertainty intervals for finite-sample population size using the site confirmation RN model were permissive. Brost et al. (2018) found that occupancy estimation remained nearly unbiased but exhibited poor coverage because estimates of true positive detection given occurrence were positively biased. There are likely to be situations where such assumption violations bias ecological parameters as well as observational parameters. For example, if using the RN model to infer abundance at a set of locations that all happened to be occupied, assuming only site-level false positives might not be effective.

The models described here are designed to account for varied types of false positive error across a range of sampling techniques (Chambert et al. 2015). A cost of this generality is that they are not tuned for specific sampling problems. For example, they do not distinguish between different types of misclassification (e.g., pairwise misclassifications of different species, which would be useful for multi-species models and inference), and do not leverage other specific information such the count of observations within an occasion (e.g., Conn et al. 2013, Chambert et al. 2018). Where appropriate, these fine-tuned solutions may be preferable, and can often also be extended to other estimation problems by refining the state process.

Our motivation for pursuing generalizable model-based solutions was grounded in concerns regarding the efficiency of implementing alternative solutions within our own work. Simulation results here reinforce our concerns. Bias associated with false positive error depends on the model employed, the incidence of error, and where and when false positives occur. As such, it may only be safe to ignore false positive detections when making ecological inference if they have extremely low incidence or generally happen at the same time and place as true positive detections. These conditions are difficult to ascertain without collecting information about classification performance that itself could be used to develop a model-based solution (Ruiz-Gutiérrez et al. 2016, Clare et al. 2019). The ability to leverage the efficiency

of model-based solutions across a broader range of model classes should make it substantially easier for investigators to account for the false positive errors that pervade ecological data.

Acknowledgments

Support for this research was provided by NASA ESSF NNX16AO61H and Ecological Forecasting grant NNX14AC36G, and a grant from the Federal Aid in Wildlife Restoration act awarded to Wisconsin Department of Natural Resources. We use data partially generated via the Zooniverse.org platform funded by a grant from the Alfred P. Sloan Foundation and a Global Impact Award from Google. Comments from V. Ruiz-Gutiérrez and 4 anonymous reviewers improved the manuscript.

References

Brost, B. M., B. A. Mosher, and K. A. Davenport. 2018. A model-based solution for observational errors in laboratory studies. Molecular Ecology Resources 18:580-589.

Chambert, T., D.A.W. Miller, and J. D. Nichols. 2015. Modeling false positive detections in species occurrence data under different study designs. Ecology 96:332-339.

Chambert, T., J. H. Waddle, D.A.W. Miller, S.C. Walls, and J. D. Nichols. 2018. A new framework for analyzing acoustic species detection data: occupancy estimation and optimization of recordings post-processing. Methods in Ecology and Evolution 9:560-570.

Chandler, R. B., and J. D. Clark. 2014. Spatially-explicit integrated population models. Methods in Ecology and Evolution 5:1351-1360.

Clare, J.D.J., P. A. Townsend, C. Anhalt-Depies, C. Locke…et al. 2019. Making inference with messy (citizen science) data: when are data accurate enough and how can they be improved. Ecological Applications 29:e01849.

Clement, M. J. 2016. Designing occupancy studies when false-positive detections occur. Methods in Ecology and Evolution 7:1538-1547.

Conn, P. B., B. T. McClintock, M. F. Cameron, D. S. Johnson, E. E. Moreland, and P. L. Boveng. 2013. Accommodating species identification errors in transect surveys. Ecology 94:2607-2618.

Gardiner, M. M., L. L. Allee, P. M. Brown, J. E. Losey, H. E. Roy, and R. R. Smyth. 2012. Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. Frontiers in Ecology and the Environment 10:471-476.

Guélat, J. and Kéry, M. 2018. Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods in Ecology and Evolution 9:1614-1625.

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

Jetz, W., M. A. McGeogh, R. Guralnick, S. Ferrier…et al. 2019. Essential biodiversity variables for mapping and monitoring species distributions. Nature Ecology and Evolution 3:539-551.

Johnson, C. J., and M. P. Gillingham. 2008. Sensitivity of species distribution models to error, bias, and model design: an application to resource selection functions for woodland caribou. Ecological Modeling 213:143-155.

Kéry, M., and J. A. Royle. 2016. Applied hierarchical modeling in ecology, Volume 1. Academic Press, London.

Kosmala, M., A. Wiggins, A. Swanson, and B. Simmons. 2016. Assessing data quality in citizen science. Frontiers in Ecology and the Environment 14:551-560.

Kuo, L., and B. Mallick. 1998. Variable selection for regression models. Sankhyā 60: 65-81.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248-2255.

McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010. Experimental investigation of observation error in anuran call surveys. Journal of Wildlife Management 74:1882-1893.

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

Miller, D. A. W., L. A. Weir, B. T. McClintock, E. H. Campbell Grant, L. L. Bailey, and T. R. Simons. 2012a. Experimental investigation of false positive errors in auditory species occurrence surveys. Ecological Applications 22:1665-1674.

Miller, D. A. W., B. L. Talley, K. R. Lips, and E. H. Campbell Grant. 2012b. Estimating patterns and drivers of infection prevalence and intensity when detection is imperfect and sampling error occurs. Methods in Ecology and Evolution 3:850-859.

Miller, D.A.W., L. L. Bailey, E. H. C. Grant, B. T. McClintock, L. A. Weir, and T. R. Simons. 2015. Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known. Methods in Ecology and Evolution 6:557-565.

Plummer, M. 2003. JAGS: a program for analysis of Bayesian graphical models using GIBBS sampling. Proceedings of the 3rd international workshop on distributed statistical computing.

R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Ramsey, D. S. L., P. A. Caley, and A. Robley. 2015. Estimating population density from presence-absence data using a spatially explicit model. Journal of Wildlife Management 79:491-499.

Robinson, O. J., V. Ruiz-Gutiérrez, D. Fink, R. J. Meese, M. Holyoak, and E. G. Cooch. 2018. Using citizen science data in integrated population models to inform conservation. Biological Conservation 227:361-368.

Rossman, S., C. B. Yackulic, S. P. Saunders, J. Reid, R. Davis, and E. F. Zipkin. 2016. Dynamic N-occupancy models: estimating demographic rates and local abundance from detection/non-detection data. Ecology 97:3300-3307.

Roth, T., N. Strebel, and V. Amrhein. 2014. Estimating unbiased phenological trends by adapting site-occupancy models. Ecology 95:2144-2154.

Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835-841.

Royle, J. A., and J. D. Nichols. 2003. Estimating abundance from repeated presence-absence data or point counts. Ecology 84:777-790.

Ruiz-Gutiérrez, V., M. B. Hooten, and E. H. Campbell Grant. 2016. Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias. Methods in Ecology and Evolution 7:900-909.

Sun, C. C., J. A. Royle, and A. K. Fuller. 2019. Incorporating citizen science data in spatially explicit integrated population models. Ecology 100:e02777.

Swanson, A., M. Kosmala, C. Lintott, and C. Packer. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology 30:520-531.

Table 1. Variables and symbology used within the original descriptions of the observation-confirmation occupancy model and our reformulation here, and where applicable, differences between the observation confirmation protocol and other protocols.

Symbol	Description
z_i	Latent binary occupancy state of site *i*
y_{ij}	Binary detection/non-detection at site *i*, occasion *j*; detections not verified and potentially include false positives.
v_{ij}	Categorical data derived from post-hoc verification of all observations at site *i*, occasion *j*: true positive detection(s) only, false positive detection(s) only, true and false positive detection(s), no detection.
p_{10}	A parameter describing the probability that $y_{ij} = 1 \mid z_i = 0$ (i.e., detection at unoccupied sites). Because detections at unoccupied sites can only be false positive, $p_{10} = s_0$.
p_{11}	A parameter describing the probability that $y_{ij} = 1 \mid z_i = 1$ (i.e., detection at occupied sites). When following the observation conformation protocol, this includes true and false positives ($p_{11} = s_{tp} + s_{fp} - [s_{tp} \times s_{fp}]$). When following the site confirmation or calibration protocols, p_{11} is equal to s_{tp} (or the parameter *p* used in a standard occupancy model), because all detections at occupied sites are assumed to be true positive detections.
s_{tp} (s_1)	Following observation confirmation and calibration protocols, a parameter that describes the probability of a true positive detection (i.e., > 0 confirmed true positive observations within v_{ij}) given that a site is occupied. Equivalent to the parameter *p* used in a standard occupancy model.
s_{fp} (s_0)	Following observation confirmation protocol, a parameter describing the unconditional probability of a false positive detection (i.e., > 0 confirmed false positive observations within v_{ij}). Under calibration protocol, may be conditional upon site being unoccupied (Chambert et al, 2015). Brost et al. (2018) describe the parameter unconditionally and as equivalent to θ_{fp}, but call it φ.
θ_{tp}	Following description here, a derived parameter describing the unconditional probability of a true positive detection within both y_{ij} and v_{ij}. $\theta_{tp,i} = z_i \times s_1$ (or $z_i \times p$).
θ_{fp}	Following description here, a parameter describing the unconditional probability of a false positive detection within both y_{ij} and v_{ij}.

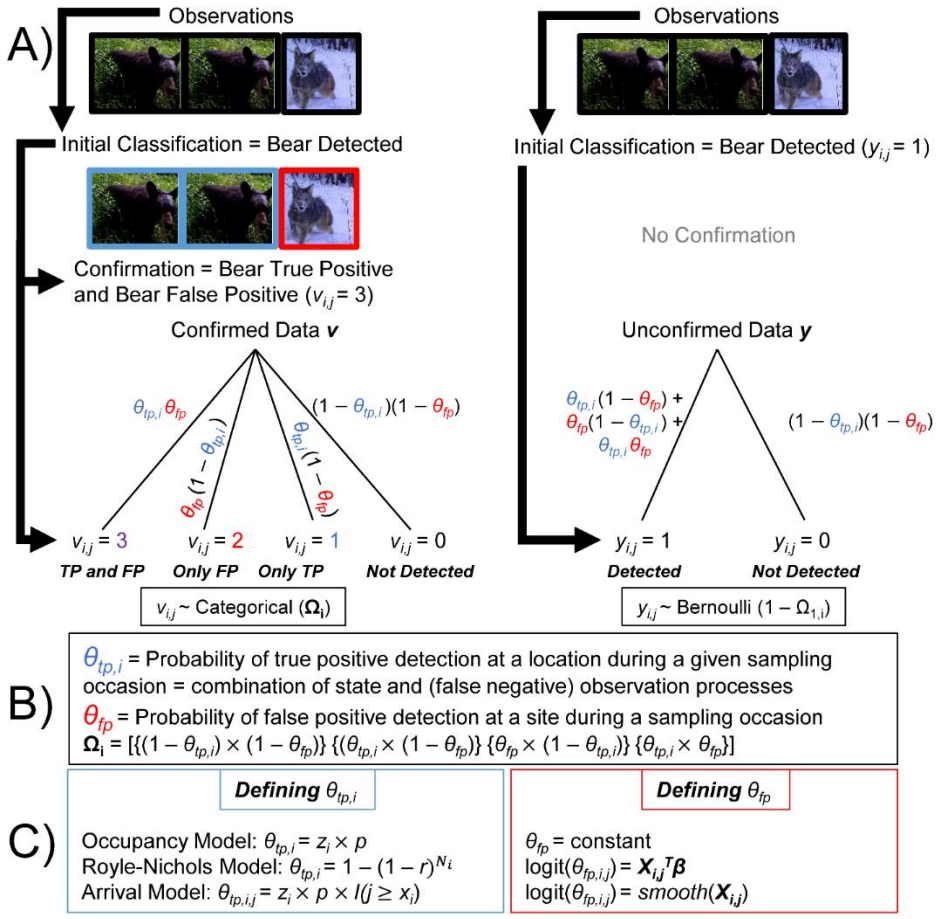

Figure 1. Schematic outlining how the observation confirmation protocol can be implemented to deal with false positive error across several model classes. The protocol presupposes that multiple observations at any site and occasion—which may include both true and false positives following eq. (2) —are collapsed into binary detection/non-detection (A). At left, detections at a subset of sites and sampling occasions are confirmed *a posteriori:* the example here depicts that at a specific site and sampling occasion, both true (2 bear images) and false (1 coyote image) positives occur, and the confirmed observations are collapsed into categorical data: here, v_{ij} = 3. Unconfirmed data y_{ij} (example at right) is either classified as 0 (no detection) or 1(detected). The probability that y_{ij} = 1 is equivalent to the sum of the probabilities that the observations constituting a detection at site *i* and occasion *j* consist of entirely true positives (v_{ij} =1), entirely false positives (v_{ij} = 2), or a mix of both (v_{ij} = 3). The probabilities underpinning ***y*** and ***v*** reflect mixtures of the unconditional probabilities of true (θ_{tp}) and false positive (θ_{fp}) detection (B). The unconditional probability of a true positive detection is the same as the unconditional probability of detection—i.e., Pr(y_{ij} =1)—defined within the base model of interest (C). In turn, because true and false positives are assumed to be independent, the unconditional probability of a false positive detection can be modeled as a constant or as functionally varying in relation to site, occasion, or site-by-occasion covariates irrespective of the model for θ_{tp}. See Appendix S1, Figure S1 for analogous figure following eq. (3).

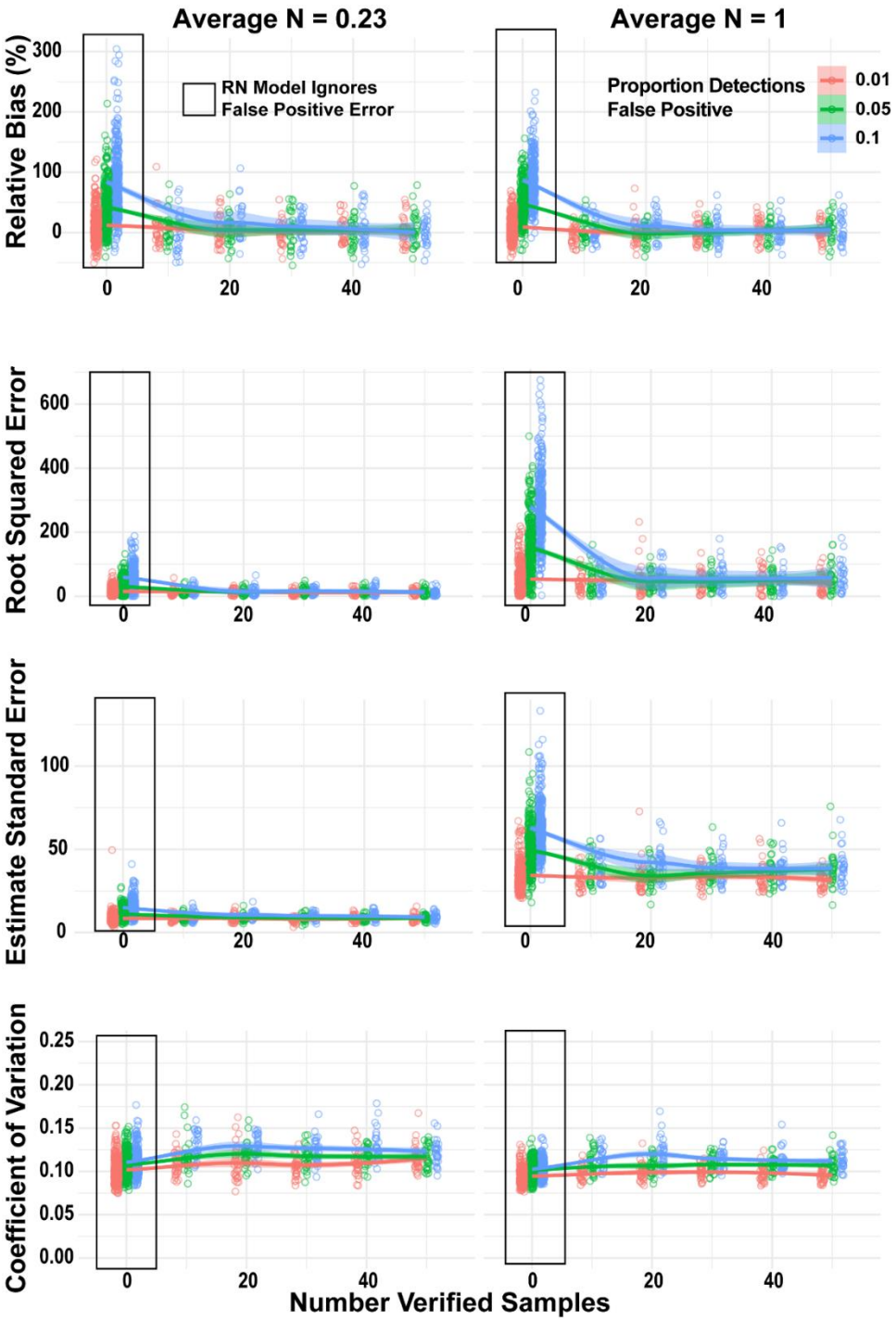

Figure 2. Performance of the Royle-Nichols model ignoring false positive error and of the model extension for false positive error with regard to finite-sample population size under varying levels of random false positive error (% of total detections = 1, 5, or 10) and verification effort (the number of sampling occasions in which all observations were verified). Standard error = standard deviation of the posterior distribution. The number of verified samples is truncated at 50 for visualization purposes (but see Appendix S2, Figure S1). Smoothers depict means across different verification levels.

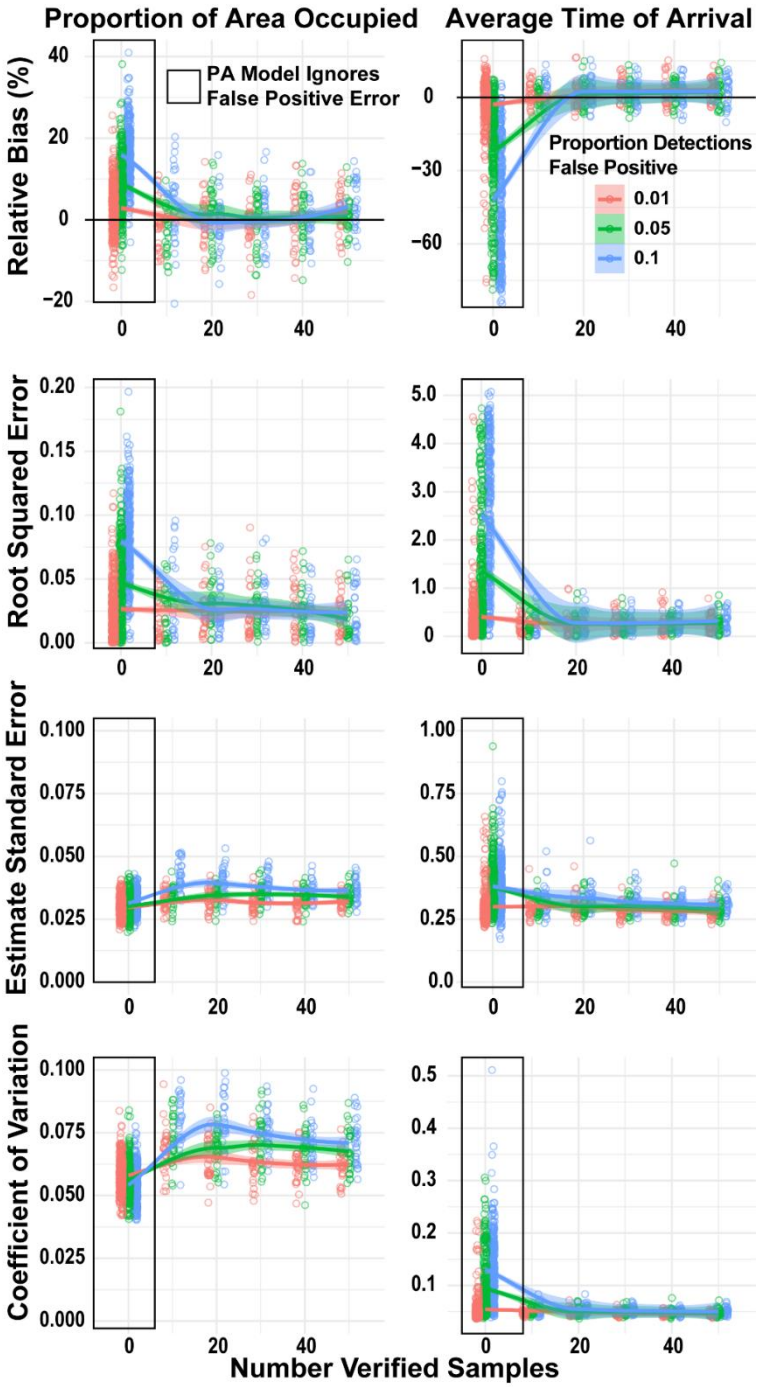

Figure 3. Performance of a standard phenological occupancy model ignoring false positive error and the model extension for false positive error with regard to proportion of area occupied and the time of arrival under varying levels of random false positive error (% of total detections = 1, 5, or 10) and verification effort (the number of sampling occasions in which all observations were verified). Standard error = standard deviation of the posterior distribution. The number of verified samples is truncated at 50 for visualization purposes (but see Appendix S2, Figure S2). Smoothers depict means across different verification levels.

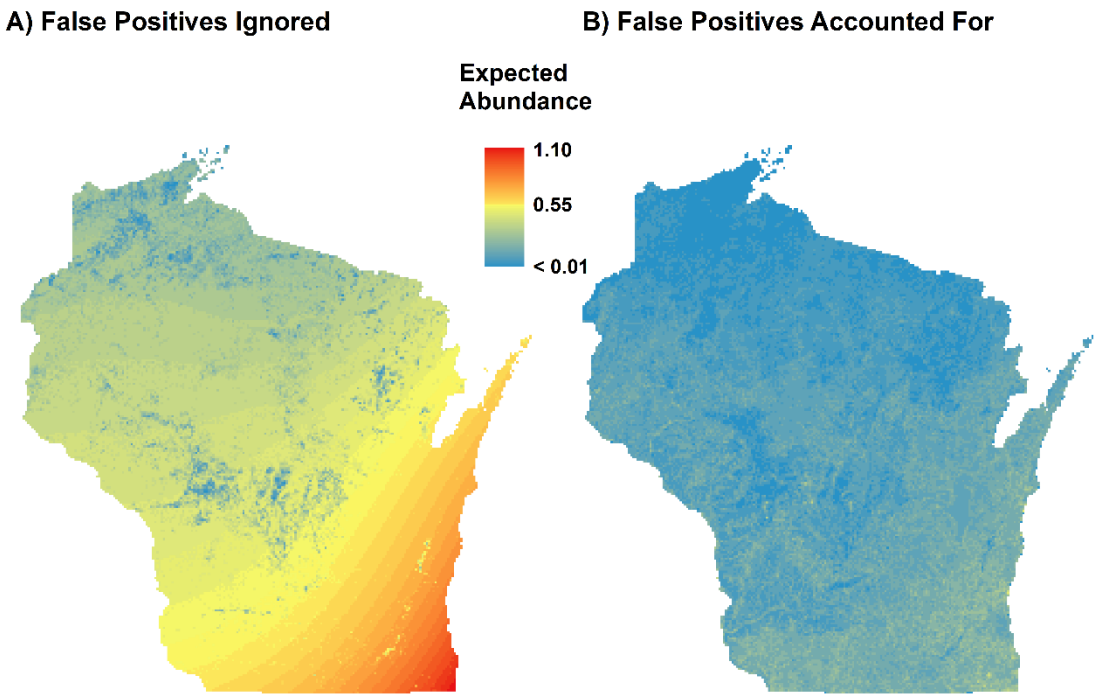

Figure 4. Predictions of gray fox abundance across Wisconsin, USA in 2017 using a Royle-Nichols model assuming no false positive error (left) and using an extension accounting for false positive error (right).

Appendix S1. **Extensions: other model types, other false positive protocols.**

Extension of other protocols.

The main text focuses on leveraging the observation-confirmation protocol to deal with false positive error. Figure S1 presents an overview of an implementation assuming observationally exclusive false positive error. However, it is important to note that other protocols used to estimate false positive parameters can also be effective across different model types.

We describe here alternative formulations following different described protocols for dealing with false positive error when estimating relative abundance using the model of Royle and Nichols (2003). The hiearchical likelihood for the base model is:

$$N_i \sim \text{Poisson}(\lambda)$$

$$p_i = 1 - (1 - r)^{N_i}$$

$$y_{i,j} \sim \text{Bernoulli}(p_i) \quad (1)$$

The likelihood for the observation-confirmation extension described in the main text is:

$$N_i \sim \text{Poisson}(\lambda)$$

$$\theta_{tp,i} = 1 - (1 - r)^{N_i}$$

$$\mathbf{\Omega_i} = [\{(1 - \theta_{tp,i}) \times (1 - \theta_{fp})\}\ \{\theta_{tp,i} \times (1 - \theta_{fp})\}\ \{\theta_{fp} \times (1 - \theta_{tp,i})\}\ \{\theta_{fp} \times \theta_{tp,i}\}]$$

$$v_{i,j} \sim \text{Categorical}(\mathbf{\Omega}_i)$$

$$y_{i,j} \sim \text{Bernoulli}(1 - \Omega_{i,i}) \quad (2)$$

Within most of the alternative models to account for false positive error, true and false positives are typically assumed to be conditionally exclusive. Throughout, we will denote the site-specific probability of a false positive (conditional on $\theta_{tp,i} = 0$) as p_{fp}, while continuing to present $\theta_{fp,i}$ as the

unconditional site-specific probability of detection. For the RN model, perhaps the most logical way to make θ_{fp} and θ_{tp} conditionally exclusive (such that false positives can only occur where true positives are impossible) is to constrain θ_{fp} to only occur at locations where abundance is zero. A formulation in the spirit of Royle and Link (2006) follows:

$$N_i \sim \text{Poisson}(\lambda)$$

$$\theta_{tp,i} = 1 - (1 - r)^{N_i}$$

$$\theta_{fp,i} = I(N_i = 0) \times p_{fp}$$

$$p_{fp} < r$$

$$y_{i,j} \sim \text{Bernoulli}(\theta_{tp,i} + \theta_{fp,i}) \quad (3)$$

Here, $I(N_i = 0)$ denotes an indicator function for whether the abundance of site $i = 0$, The constraint $p_{fp} < r$, where r is the probability of detecting a single present organism, mirrors the constraint in the original model for occupancy estimation that $p_{fp} < p_{tp}$.

Within the site-confirmation protocol (specifically, the multiple detection states version presented by Miller et al. 2011 and Chambert et al. 2015), the observed observations $y_{i,j}$ follow a categorical distribution pertaining to whether no species was observed ($y_{i,j} = 0$), a species was observed but not confirmed ($y_{i,j} = 1$), and a species was unambiguously observed ($y_{i,j} = 2$). The parameter b describes the conditional probability that a positive observation is unambiguous. The likelihood for a Royle-Nichols variant of the multiple detection states model can be described as:

$$N_i \sim \text{Poisson}(\lambda)$$

$$\theta_{tp,i} = 1 - (1 - r)^{N_i}$$

$$\theta_{fp,i} = I(N_i = 0) \times p_{fp}$$

$$\mathbf{\Omega_j} = [\{(1 - (\theta_{tp,j} + \theta_{fp,i})\}\ \{(\theta_{tp,j} \times (1 - b) + \theta_{fp,i})\}\ \{\theta_{tp,j} \times b\}]$$

$$y_{i,j} \sim \text{Categorical}\ (\mathbf{\Omega_j}) \quad (4)$$

Another variant of the site-confirmation protocol, the multiple detection methods model (Miller et al. 2011, Chambert et al. 2015), assumes that one sampling method (M1) generates detections $y_{i,j}$ that might either be true or false positives, and a second independent method (M2) operating during occasions s provides detections $w_{i,s}$ that can only be true positives. Let the distinct probabilities of truly detecting a single animal for M1 and M2 be denoted as r_1 and r_2. The hierarchical formulation is then:

$$N_i \sim \text{Poisson}\ (\lambda)$$

$$\theta_{tp,i,1} = 1 - (1 - r_1)^{N_i}$$

$$\theta_{tp,i,2} = 1 - (1 - r_2)^{N_i}$$

$$\theta_{fp,i} = I(N_i = 0) \times p_{fp}$$

$$y_{i,j} \sim \text{Bernoulli}\ (\theta_{tp,i,1} + \theta_{fp,i})$$

$$w_{i,s} \sim \text{Bernoulli}\ (\theta_{tp,i,2})\ ) \quad (5)$$

Finally, following the calibration protocol (Chambert et al. 2015), an investigator might have collected reference detection data under experimental conditions in which the state is known (or at least, the exclusive conditions for θ_{tp} and θ_{fp} are known): x_1 and x_0 respectively denote the total number of true and false positive detections in the reference data set, and n_1 and n_0 the number of trials in which true positive detections were possible or not—an example trial might be the presentation of a single image or recording in which the species of interest is present or not present. One way to specify the Royle-Nichols model likelihood hierarchically following the calibration protocol is:

$$N_i \sim \text{Poisson}\ (\lambda)$$

$$\theta_{tp,i} = 1 - (1 - r)^{N_i}$$

$$\theta_{fp,i} = I(N_i = 0) \times p_{fp}$$

$$x_1 \sim \text{Binomial}(n_1, r)$$

$$x_0 \sim \text{Binomial}(n_0, p_{fp})$$

$$y_{i,j} \sim \text{Bernoulli}\ (\theta_{tp,i} + \theta_{fp,i}) \quad (6)$$

The key to implementing using these different protocols within different model classes while assuming that true and false positives are conditionally exclusive is only slightly more complex than extending models assuming inclusive error. One must redefine $\theta_{tp,i}$ following the original model description, condition $\theta_{fp,i}$ so that a false positive is only possible under conditions when a true positive detection is impossible and alter the statements associated with the auxiliary data (if available) appropriately. For example, under Roth et al's. (2014) PA model, true positive detections are only possible at occupied sites once the species has arrived—in other words, $\theta_{tp,i,j} = z_i \times p \times I(j \geq x_i)$. If the probability of a false positive detection is conditional upon a true positive detection being impossible, then false positives can either occur at unoccupied sites ($z_i = 0$), or at occupied sites prior to the species arrival ($z_i = 1$, but $j < x_i$). That is, $\theta_{fp,i,,j} = ((1 - z_i) \times p_{fp}) + z_i \times \text{I}(j < x_i) \times p_{fp}$. Appendix S4 presents code used to fit the RN and PA models accounting for false positive error following the site confirmation (multiple detection states) protocol assuming that true and false positives are conditionally exclusive.

Importantly, it seems that the assumption that true and false positives are conditionally exclusive may not be strictly necessary for the full latent or calibration models. For example, the latent estimator we fit using informed priors (see main text, Appendix S2, and Appendix S4) relaxed the assumption that all false positives were site-level. We believe that fitting this model shares a similar constraint with the full estimator of Royle and Link (2006) in that the conditional probability of a true positive (if the species can be detected) must be greater than the probability of a false positive (here, unconditional), because all detections are of unknown quantity. Note, however, that we did not strictly enforce this constraint: rather,

our simulation settings typically gave rise to situations in which θ_{fp} was less than the average value of r or p, and thus our priors tended to reflect the constraint (more details in Appendix S4). Applied to the RN model, the likelihood looked like:

$$N_i \sim \text{Poisson}\ (\lambda)$$

$$\theta_{tp,i} = 1 - (1 - r)^{N_i}$$

$$\theta_{fp} < r$$

$$y_{i,j} \sim \text{Bernoulli}\ (\theta_{tp,i} + \theta_{fp} - [\theta_{tp,i} \times \theta_{fp}]) \quad (7)$$

Similarly, the calibration-based estimator described by Brost et al. (2018) assumes inclusive error. This model relies on auxiliary detections in controlled settings where true positive detections cannot occur. Brost et al. (2018) present a situation in which only false positives can be experimentally calibrated, which is a reasonable circumstance for laboratory studies, but we assume below one also might have calibrated true positives. For the RN likelihood, each positive trial reflects the ability to detect a single organism (i.e., it is implied N = 1). The likelihood is:

$$N_i \sim \text{Poisson}\ (\lambda)$$

$$\theta_{tp,i} = 1 - (1 - r)^{N_i}$$

$$x_1 \sim \text{Binomial}(n_1, r)$$

$$x_0 \sim \text{Binomial}(n_0, \theta_{fp})$$

$$y_{i,j} \sim \text{Bernoulli}\ (\theta_{tp,i} + \theta_{fp} - [\theta_{tp,i} \times \theta_{fp}]) \quad (8)$$

Based upon some exploratory analysis, extending the site confirmation models to accommodate inclusive or observation level false positives appears to require further constraints on some combination

of b, p_{fp}, and p_{tp} as there is some confounding between the probability that a true positive occurred but was not confirmed and the probability that a false positive detection occurred.

Observation Confirmation Protocol for the Spatial Royle-Nichols Model

The spatial Royle-Nichols model (Ramsay et al. 2015) uses z_i to denote whether individuals $i = 1,2…M$ exist within a geographic space $||S||$ with probability ψ. The state variable of interest, population size N in $||S||$, is estimated as $\widehat{N} = \sum_{i=1}^{M} z_i$, and population density is derived as $\widehat{N}/Area_{||S||}$. Individuals have distinct activity centers located within $||S||$ and the coordinates of these activity centers are denoted as s_i; individuals are detected at any of j detectors on given sampling occasions k with probability $p_{i,j.}$ The unconditional probability of detection is a function of whether an individual exists, the distance between an individual's latent activity center and the location of the detector, $d_{i,j}$, and the parameters g_0 and σ that respectively relate to the probability of individual detection at $d_{i,j} = 0$ and the rate at which individual encounter probability decays, and can be expressed as $p_{i,j} = g_0(–d_{i,j}/2\sigma^2) \times z_i$. Individuals are not distinguished, so these parameters are inferred by marginalizing across the latent individual encounter histories at a specific detector such that the unconditional probability of detection is described as $\theta_{tp,j} = 1 – \prod_{i=1}^{M}(1 - p_{i,j})$. The hierarchical likelihood is:

$$z_i \sim \text{Bernoulli } (\psi)$$

$$p_{i,j} = g_0(–d_{i,j}/2\sigma^2) \times z_i$$

$$\theta_{tp,j} = 1 – \prod_{i=1}^{M}(1 - p_{i,j})$$

$$\mathbf{\Omega_j} = [\{(1 – \theta_{tp,j}) \times (1 – \theta_{fp})\}\{\theta_{tp,j} \times (1– \theta_{fp})\}\ \{\ \theta_{fp} \times (1– \theta_{tp,j})\}\ \{\ \theta_{fp} \times \theta_{tp,j}\ \}]$$

Here, $v_{j,k} \sim$Categorical ($\mathbf{\Omega_j}$) and $y_{j,k} \sim$ Bernoulli $(1 – \Omega_{1,j})$. Critically, $\theta_{tp,j}$ has no support at 0, and as a result, without re-specifying the SRN model (e.g., by truncating the detection distance at a certain point), true and false positives cannot be conditionally exclusive. Although not directly considered within the manuscript, code required to simulate and fit extended SRN models is found in within Appendix S4. As a

general proof of concept that the model is sensitive to false positive error and an extended version can reduce bias, we present very limited simulation results here.

We simulated 100 replicate datasets to demonstrate proof of concept. Sampling parameters included a population size of 50 organisms; ||*S*|| defined as a 20 × 20 unit square; detection parameters g_0 = 0.15 and σ = 0.5; 196 detector locations within a 14 × 14 square grid with 1 unit spacing, and 20 sampling intervals: only the location of individual activity centers varied across simulation replicates. False positive observations (as 10% of all detections) and a verification sample were simulated following practices described in the main text; the size of the verification sample varied as {10 20 30 40 50} replicated 20 times each. We compared the standard model and the false positive extension on the basis of relative bias of $\widehat{N}$. The estimator ignoring false positive error was strongly biased (% bias = 81) at a 10% false positive rate; the extended model was not unbiased—perhaps due to small sample size considerations—but exhibited better performance (% bias = 25). Clearly, more research is needed to understand the general sensitivity of the SRN model to false positive error. Given the narrow range of parameter space within which the base model is unbiased even without false positive error (Ramsey et al. 2015), we expect that it is more likely to see usage as a component within integrated models (Sun et al. 2019), and we suggest that it may be more fruitful to explore sensitivity to false positives within this class of model.

Sometimes the assumed form (union) does not matter: spatiotemporal occupancy models

Finally, we wish to acknowledge that there are some models for which the assumed form for false positive error or the protocol employed to estimate false positive error should make little to no difference at all. An excellent example is the spatiotemporal occupancy model described by Hepler et al. (2018). This model leverages spatiotemporal autoregressive parameters, and the occupancy state at every site i may change across all time periods t; $z_{i,t}$ is assumed distributed as Bernoulli($\psi_{i,t}$) and detection/non-detection assumed distributed as Bernoulli($z_{i,t} \times p$), where each term's interpretation is consistent with its interpretation within a standard closed occupancy model. As noted in the main text, false positives

happening at the same time and place as true positives do not change the detection/non-detection data, and a model that assumes inclusive false positive error will provide exactly the same estimates of $\boldsymbol{\theta_{tp}}$ as a model that assumes observationally exclusive false positive error. Because the model of Hepler et al. (2018) operates under the assumption that the state variable is distinct for each site by occasion, a conditionally exclusive model form (e.g., false positives are only possible conditional on $z_{i,t} = 0$) ends up exactly equivalent to an observationally exclusive model form, and all three model forms will result in the same estimate of $\boldsymbol{\theta_{tp}}$.

Supporting References

Brost, B. M., B. A. Mosher, and K. A. Davenport. 2018. A model-based solution for observational errors in laboratory studies. Molecular Ecology Resources 18:580-589.

Chambert, T., D.A.W. Miller, and J. D. Nichols. 2015. Modeling false positive detections in species occurrence data under different study designs. Ecology 96:332-339.

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

Hepler, S. A., R. Erdhardt, and T. M. Anderson. 2018. Identifying spatial drivers of variation in occupancy with limited replication camera trap data. Ecology 99:2152-2158.

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

Ramsey, D. S. L., P. A. Caley, and A. Robley. 2015. Estimating population density from presence-absence data using a spatially explicit model. Journal of Wildlife Management 79:491-499.

Roth, T., N. Strebel, and V. Amrhein. 2014. Estimating unbiased phenological trends by adapting site-occupancy models. Ecology 95:2144-2154.

Royle, J. A., and J. D. Nichols. 2003. Estimating abundance from repeated presence-absence data or point counts. Ecology 84:777-790.

Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835-841.

Sun, C. C., A. K. Fuller, and J. A. Royle. 2019. Incorporating citizen science data in spatially explicit integrated population models. Ecology:e027777.

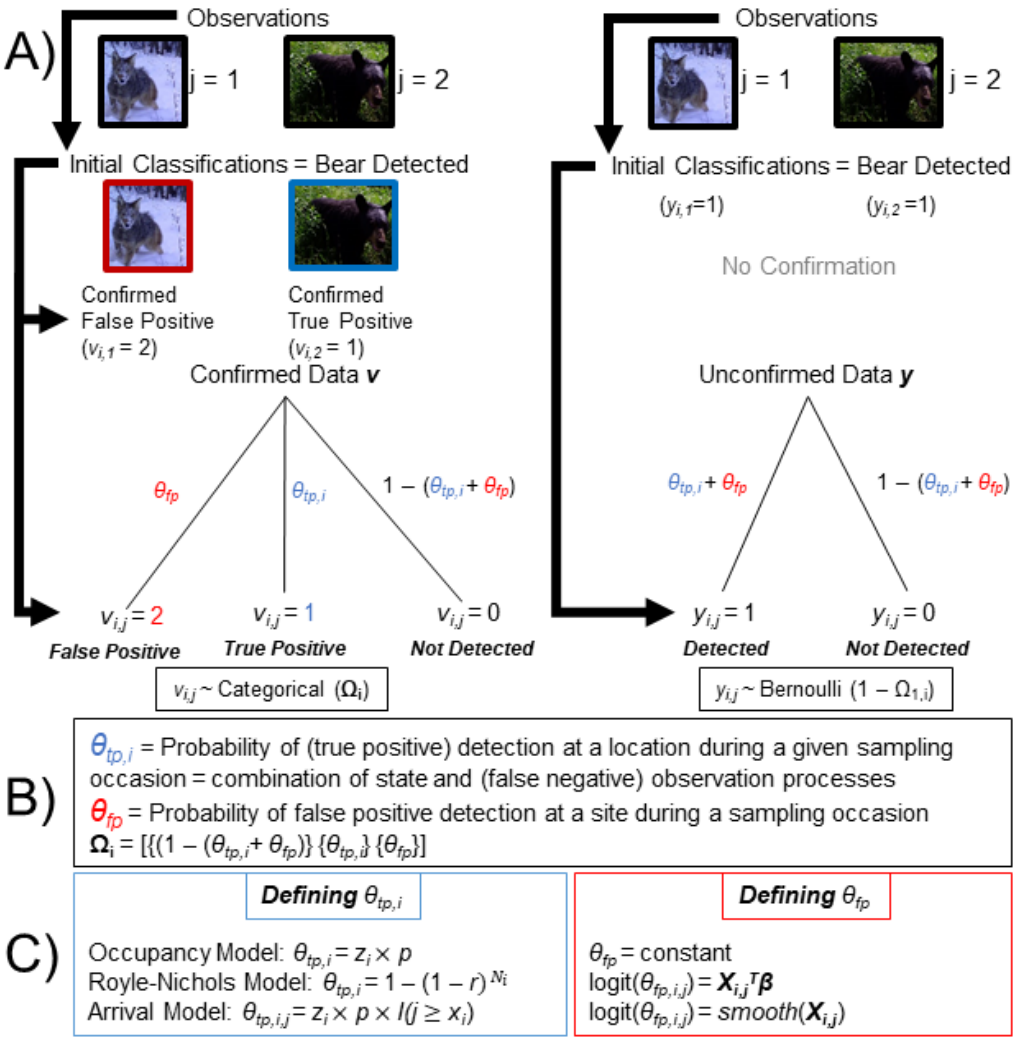

Figure S1. Schematic how the observation confirmation protocol can be implemented to deal with false positive error across several model classes *when assuming observationally exclusive false positive error* (c.f. Figure 1, and following eq. 3 in the main text): the figure largely mirrors Figure 1 in the main text. The example above presupposes that any detection at a site during a specific interval can be classified as either false positive or true positive, but not both (A, c.f. Figure 1, which presupposes that the observations within a specific interval at a specific site may be all true positives, all false positives, or a mixture of the two). At left, detections at a subset of sites and sampling occasions are confirmed *a posteriori:* the example here depicts that a detection on occasion 1 was confirmed as a false positive ($v_{i,1}$ = 2), and a detection at the same site on occasion 2 was confirmed as a true positive ($v_{i,2}$ = 1). Note that in contrast, Figure 1 depicts a set of observations occurring at the same site during a single sampling occasion (rather than multiple occasions). The unconfirmed data y_{ij} (example at right) is either classified as 0 (no detection) or 1(detected), and the example on the right shows the unconfirmed analogue to the data on the left (i.e, $y_{i,1}$ = 1, $y_{i,2}$ = 1). The probability that y_{ij} = 1 is equivalent to the union of the probabilities that a detection is a true positive (v_{ij} =1) or a false positive (v_{ij} = 2). The probabilities underpinning ***y*** and ***v*** reflect mixtures of the unconditional probabilities of true (θ_{tp}) and false positive (θ_{fp}) detection (B). C follows description from Figure 1. It would be perfectly reasonable to use the model structure depicted here to analyze the data in Figure 1 if one defined a false positive detection (v_{ij} = 2) as occurring when *all* observations in v_{ij} were confirmed as false positive, and if one defined a true positive detection (v_{ij} = 1) as occurring when >0 observations in v_{ij} were confirmed as true positive.

Appendix S2. Additional details associated with simulation study.

Most information here relates to tables and figures referred to within the main text. We briefly expand upon the description of the site-confirmation simulation described in the main text, elaborate upon the presumed offsetting biases observed when fitting the PA model of Roth et al. (2014), and also describe the results of a study focusing on model transferability and informed priors (with implications for fully latent estimators) here.

Enacting the site-confirmation protocol

Chambert et al. (2015) recognized that data collected under the observation confirmation protocol could be analyzed using a site confirmation approach if confirmed false positives were discarded and only confirmed true positives were used (the 'multiple detection state' model described by Miller et al. [2011]). Their simulation results demonstrate that both estimators are unbiased when false positives and true positives are simulated as being conditionally exclusive at the site level, but that the observation confirmation protocol is slightly more precise, and the likelihood is more unimodal.

We followed Chambert et al. (2015) in implementing the site-confirmation model referred to within the main text. From the base scenarios (constant false positive error) and simulation replicates described in the main text, we removed all confirmed false positives, and reclassified all observations within $\boldsymbol{y}$ to enforce consistency with the multiple detection states model so that $y_{i,j,sim} = 0$ if there was no detection, $y_{i,j,sim} = 1$ if there was a simulated detection but no simulated confirmation of at least 1 true positive, and $y_{i,j,sim} = 2$ if there was a simulated detection and a simulated confirmation that at least one observation was true positive. The likelihood used is described in Appendix S1, and sample code is presented in Appendix S4. Figures S3 and S4 depict results, and results are also included within the germane tables below. For each of the 6 RN scenarios, parameters associated with between 6 and 26 simulated datasets failed to converge (n provided in each table), even with prolonged run-times. This is potentially consistent with Chambert et al.'s (2015) observation that the site confirmation model had

multiple optima and was sensitive to starting values; the effect appeared more pronounced for scenarios 4- 6, in which observations were fairly sparse. The PA model exhibited no convergence problems. Note that all values in each table represent an average across all numbers of verified samples (non-converged simulated datasets are censored).

A note on the offsetting biases observed within the PA model.

In the main text, we note that PA models ignoring false positive error estimated arrival time with slightly less bias than models accounting for it when false positive error was considered impossible prior to occasion 5 and when average site-specific arrival was occasion 6. This was a somewhat surprising result. We fit each dataset simulated under this scenario to a standard PA model *after* getting rid of all simulated false positives: the standard PA model when there were no false positives was still slightly more biased than the standard PA model when false positives were included within the dataset (see Table S6). Given that the simulations performed within Roth et al. (2014) suggest the estimator is unbiased given a large sample and larger values of *p*, we believe this result arises because the small biases associated with false positives slightly prior to the time of arrival are being offset by small-sample biases associated with false negative error (i.e., with low *p*, the observed time of arrival is often much later than the actual arrival).

Evaluating the transferability of θ_{fp} using informed priors (and implications for fully latent models)

An appealing property of the generalized inclusive structure is that the unconditional θ_{fp} is a component of the model that does not necessarily need to be changed every time θ_{tp} is re-specified. This reflects the reality that for any specific detection/non-detection data, correctly classified data and misclassified data were generated from the same processes underlying the distribution (abundance, phenology, etc.) and perception of the focal organism relative to potential sources of misclassification. This suggests that if data or computational resources (and we note that incorporating data to help account for false positive error increases the number of nodes within the graphical model) are lacking, one might be able to use an informative prior for θ_{fp} given previous estimates of the parameter from a distinct (and more quickly fit)

model, or estimates from a comparable dataset. To briefly explore transferability across models, we fit the original observation confirmation occupancy model described by Chambert et al. (2015; i.e., $\theta_{tp,i} = z_i \times p_i$) to every simulated dataset described previously, and then fit a model with the correct (i.e., RN or PA) structure for $\theta_{tp,i}$ and for which θ_{fp} (or any variability in $\theta_{fp,i}$) was strictly informed by a prior distribution derived from the posterior distribution of the false positive parameter estimated by the Chambert et al. (2015) occupancy model (code in Appendix S4).

Using an informed prior for false positive error generally resulted in parameter estimates that were strongly correlated with estimates produced when confirmation results were directly incorporated into the likelihood, particularly for the RN model (Figure S8), and estimator performance scarcely differed (Appendix S2, Tables S1 – S9). Discrepancies did not appear to be related to the size of the verification sample; instead, relative to including confirmed observations in the likelihood, using an informed prior tended to result in slightly smaller estimates of proportion of area occupied and slightly larger estimates of arrival time (Figure S9).

Thus, results suggest that with reasonably well informed priors, a correctly-specified model for both the true positive and false positive process, and, we think, under the condition that the probability of a false positive detection is generally smaller than the probability for a true positive detection (as simulated here, values of θ_{fp} were far less than the constant r or p [or if varying across sites, their average values]), a fully latent estimator is extensible and effective for several model classes. It also suggests certain paths for increased efficiency if the model class in question is computationally demanding: one could estimate the unconditional probability of a false positive detection (and perhaps select an appropriate model structure for false positive error) within an occupancy model framework, and then use the results to set an informed prior within the model of interest. Although not depicted here, we note that estimates of θ_{fp} did change slightly across model structures–e.g., the posteriors from an RN model did not look exactly like the prior (itself, the posterior from an occupancy model), which suggests there is some waterbed effect associated with θ_{fp} and θ_{tp} such that slight changes to the specification of one will impact

the estimate of the other. The slight changes between the occupancy structure and the RN or PA model structures (again, using the same [generating] covariates) did not appear particularly consequential within the models considered. However, we caution that further exploration is warranted to understand the transferability of θ_{fp} estimates and how robust fully latent estimators are to poorly specified models.

Supporting References

Chambert, T., D.A.W. Miller, and J. D. Nichols. 2015. Modeling false positive detections in species occurrence data under different study designs. Ecology 96:332-339.

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

Roth, T., N. Strebel, and V. Amrhein. 2014. Estimating unbiased phenological trends by adapting site-occupancy models. Ecology 95:2144-2154.

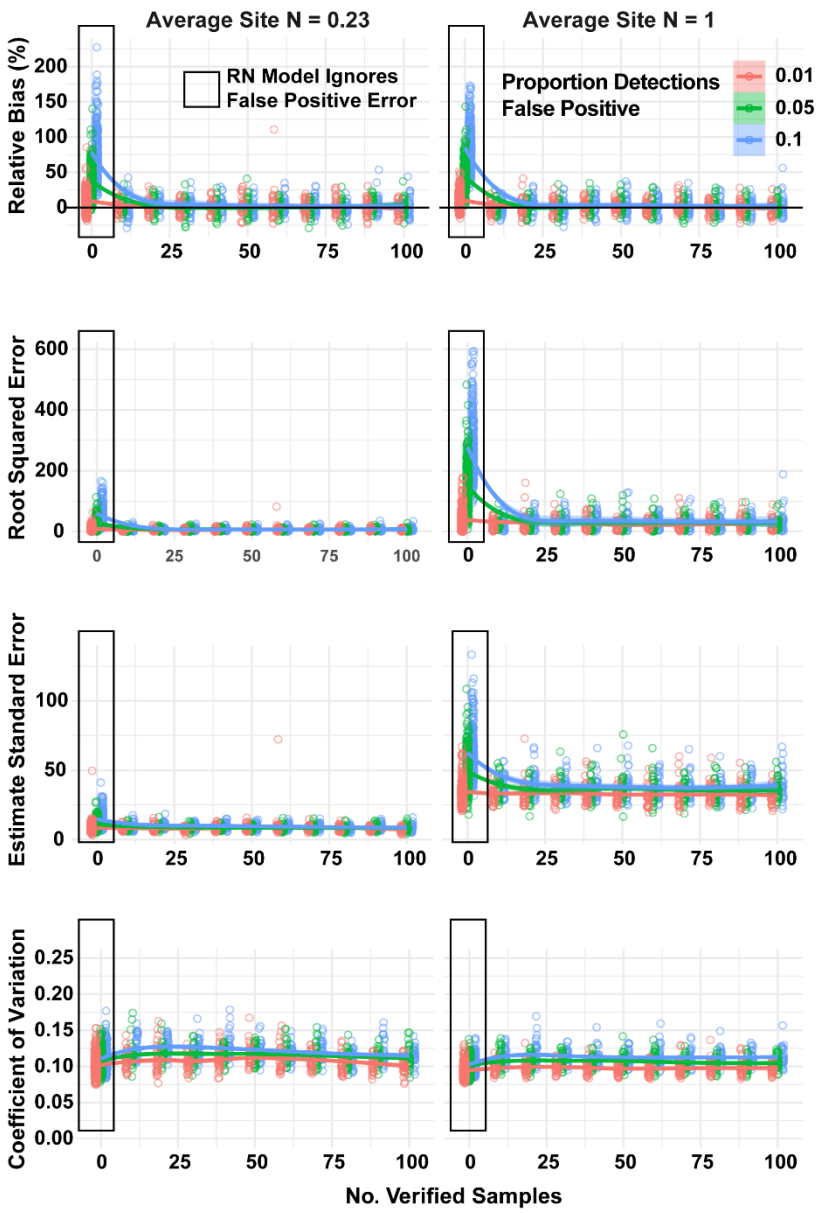

Figure S1. Performance of the Royle-Nichols model ignoring false positive error and the model extension for false positive error with regard to finite-sample population size under varying levels of random false positive error (% of total observations = 1, 5, or 10) and verification effort (# of verified samples, as in main text). Standard error = standard deviation of the posterior distribution. The sole distinction from Figure 2 in main text is that the number of verified samples is not truncated at 50. Smoothers fit to means.

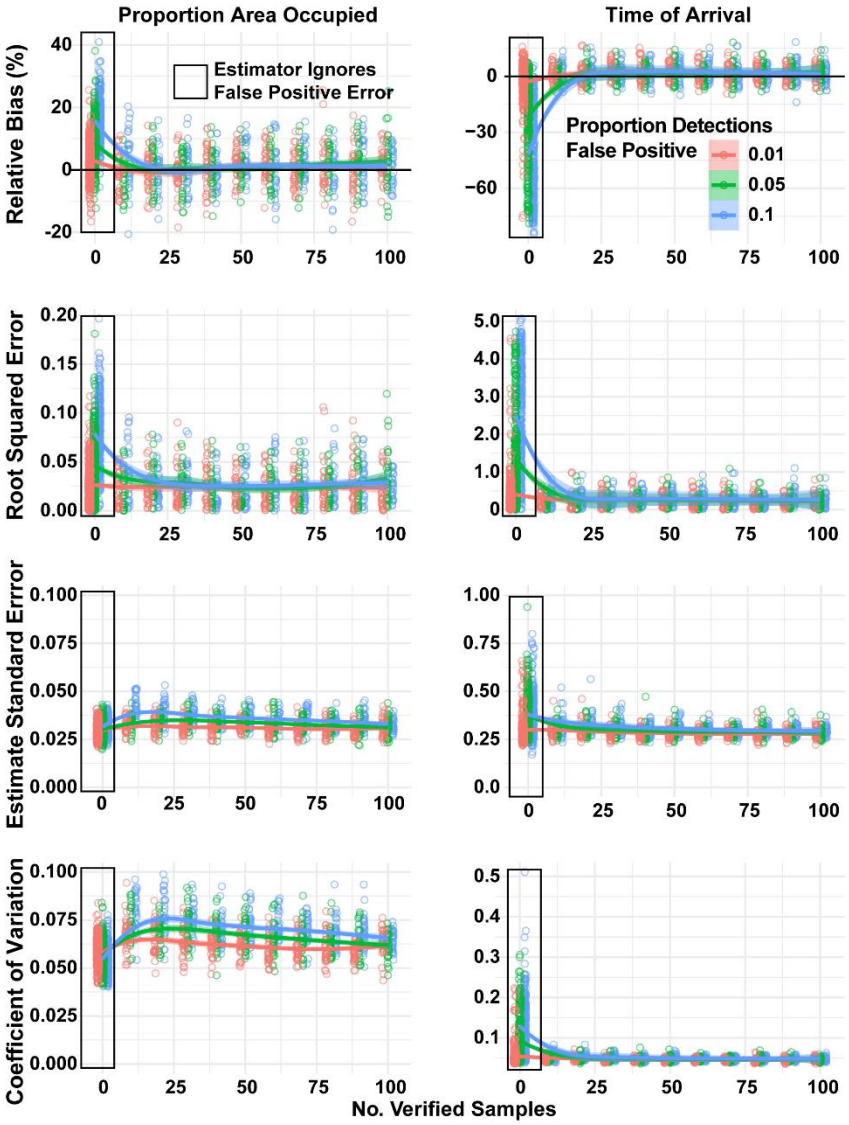

Figure S2. Performance of a standard phenological occupancy model ignoring false positive error and the model extension for false positive error with regard to proportion of area occupied and the time of arrival under varying levels of random false positive error (% of total observations = 1, 5, or 10) and verification effort (# of verified samples). Standard error = standard deviation of the posterior distribution. The sole distinction from Figure 3 in main text is that the number of verified samples is not truncated at 50. Smoothers depict mean values.

\

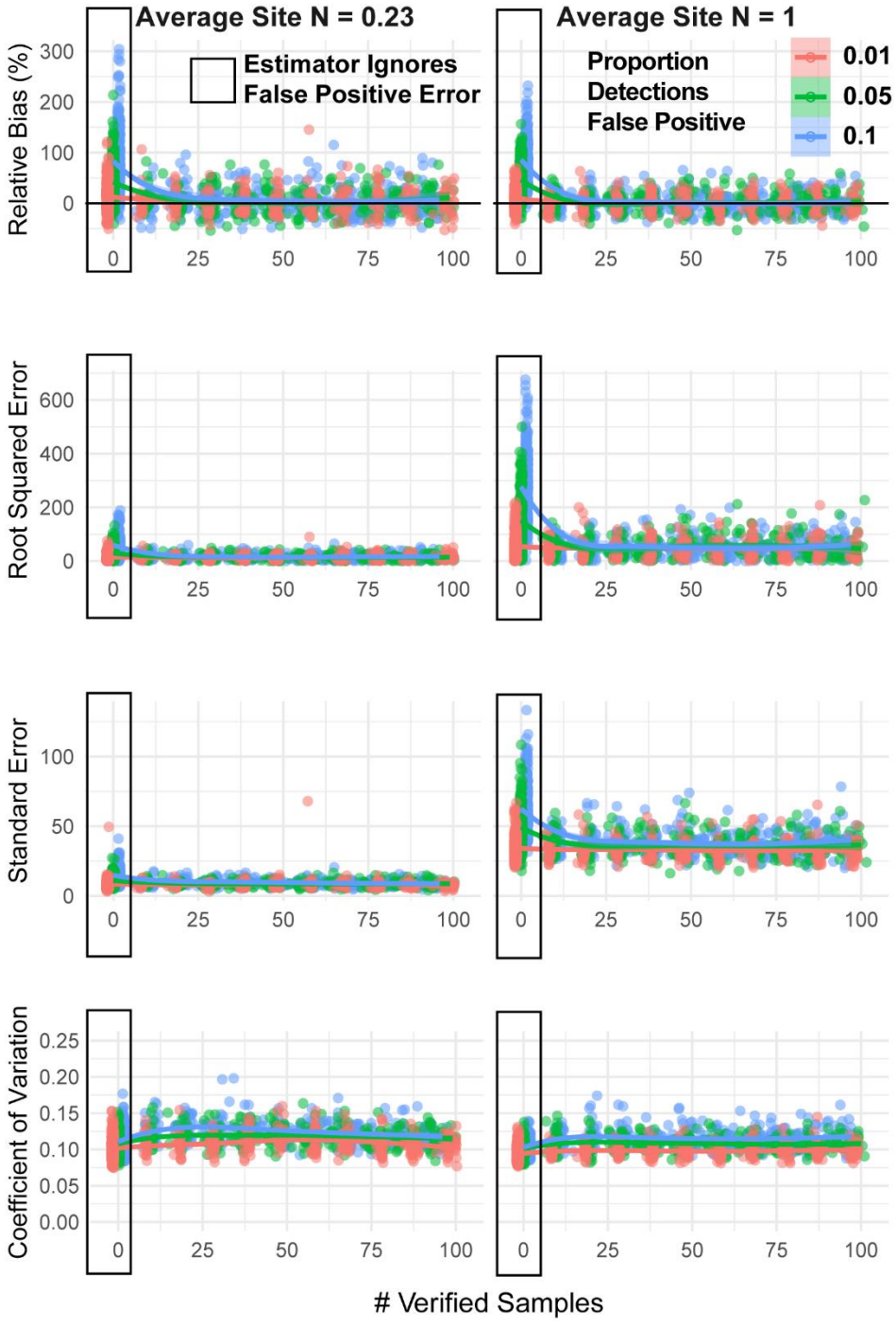

Figure S3. Performance of the Royle-Nichols model ignoring false positive error and the model extension for false positive error following the site confirmation protocol with regard to finite-sample population size under varying levels of random false positive error (% of total observations = 1, 5, or 10) and verification effort (# of verified samples). Standard error = standard deviation of the posterior distribution. Results based on the same simulation data used within Figure S1 and Figure 2 in the main text, but manipulated to remove confirmed false positives. Smoothers depict mean values.

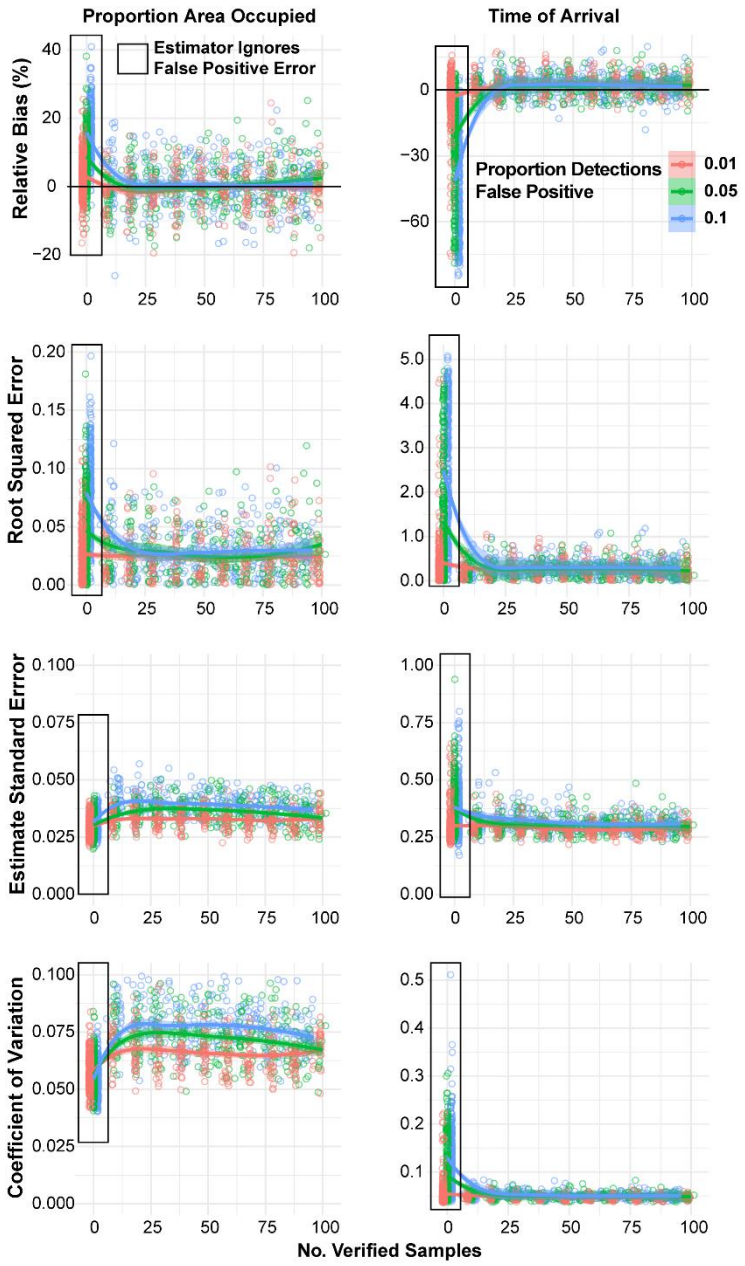

Figure S4. Performance of a standard phenological occupancy model ignoring false positive error and the model extension for false positive error following the site confirmation protocol with regard to proportion of area occupied and the time of arrival under varying levels of random false positive error (% of total detections = 1, 5, or 10) and verification effort (the # of sampling occasions in which all observations were verified). Standard error = standard deviation of the posterior distribution. Results based on the same simulation data used within Figure S2 and Figure 3 in the main text, but manipulated to remove confirmed false positives. Smoothers depict mean values.

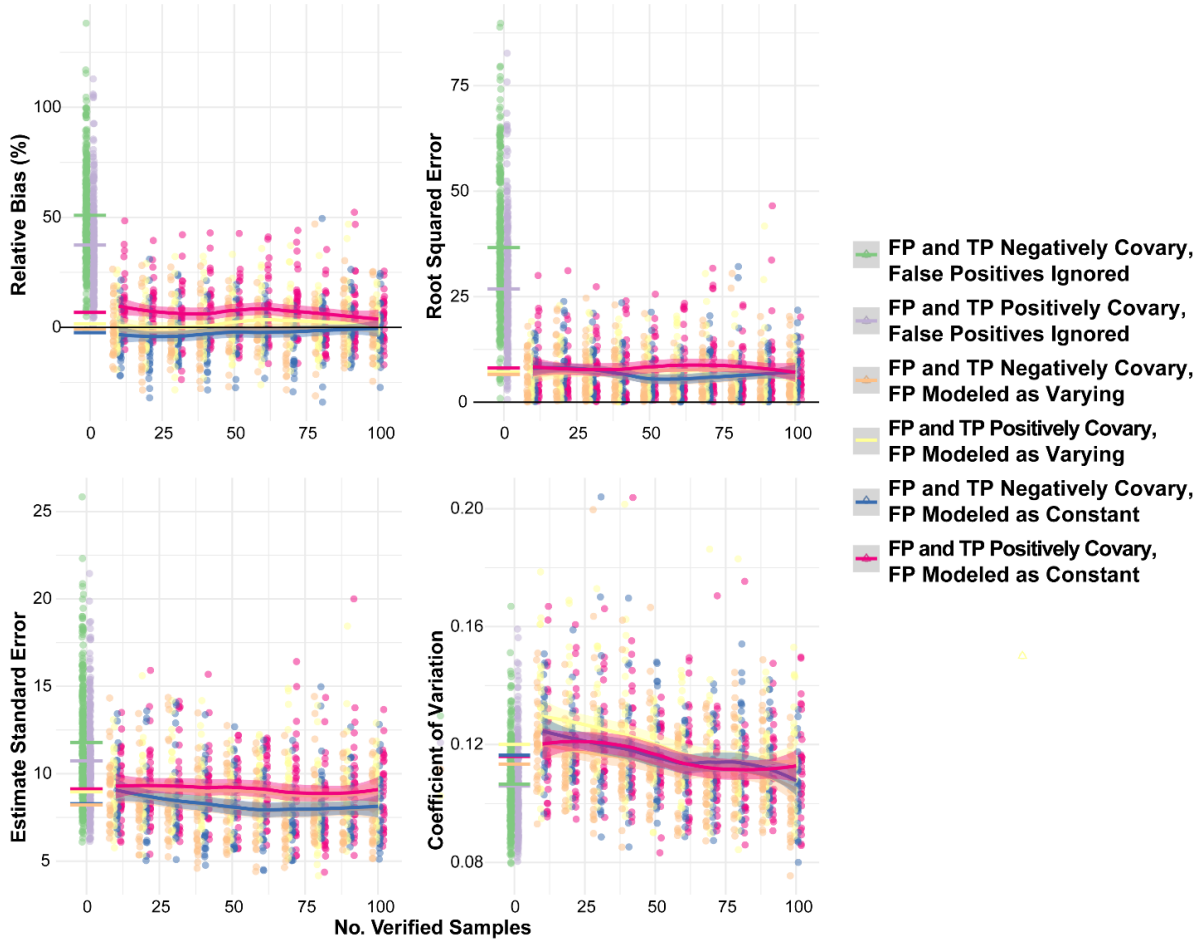

Figure S5. Performance of the standard Royle-Nichols model (when 0 observations are confirmed; green/gray) and model extensions for false positive error with regard to finite-sample population size under different directional covariance between true and false positive detections, and when different models for false positive error (either a misspecified constant model in blue/magenta or the generating model in yellow/orange) were fit. Standard Error = standard deviation of the posterior distribution. Data used includes scenarios 7 and 8 described below in Table S1. Horizontal bars at left depict the mean values for each model. Lines and shading (often overlapping) depict smoothed means and SE.

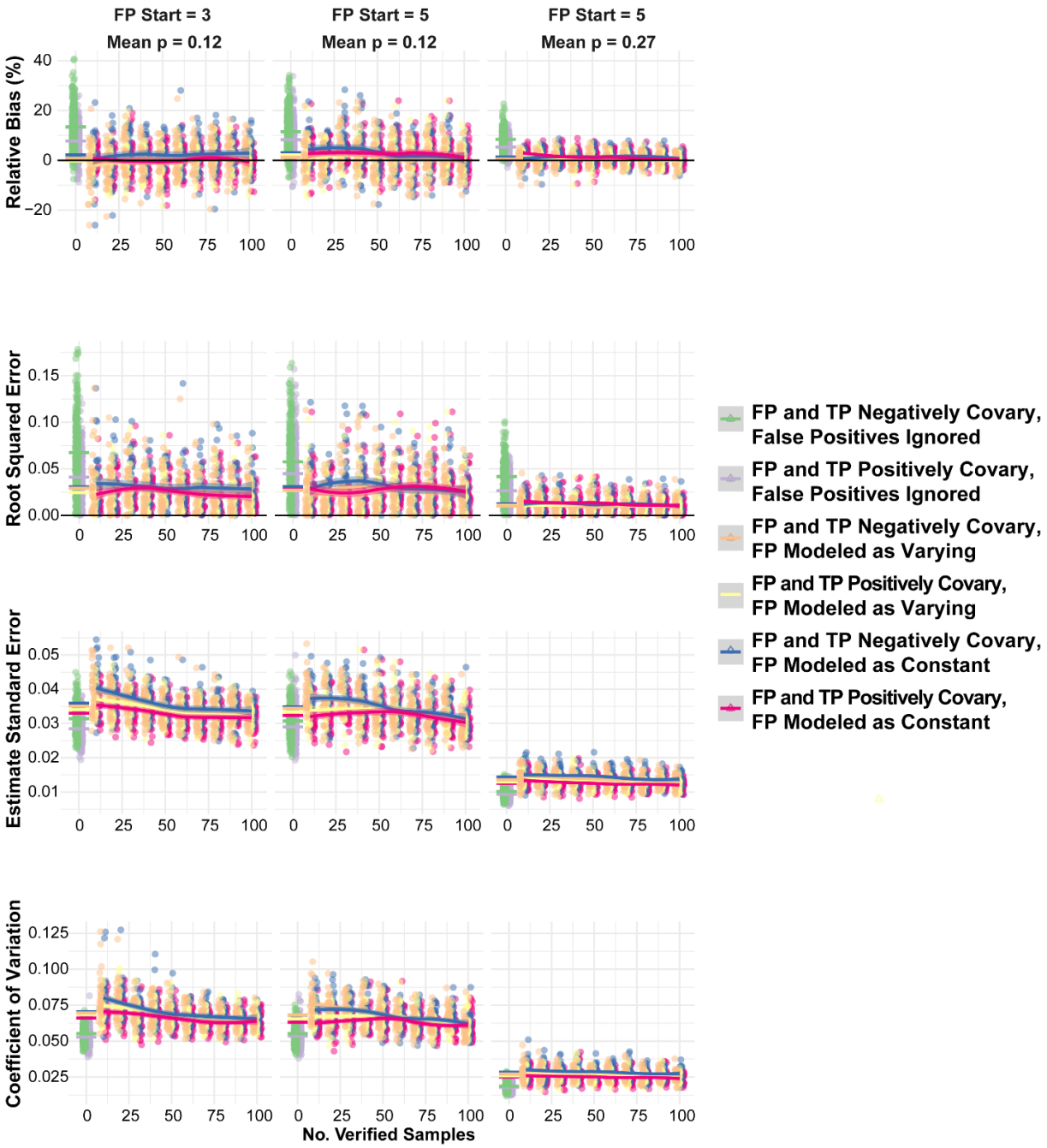

Figure S6. Performance of the standard phenological occupancy model (when 0 observations are confirmed, grey/green) and the model extension for false positive error with finite-sample population size with regard to proportion of area occupied under different directional covariance between true and false positive detections, and when different models for false positive error (either a misspecified constant model in blue/magenta or the generating model in yellow/orange) were fit. Standard Error = standard deviation of the posterior distribution. Data used includes scenarios 4-9 described below in Table S5. Horizontal bars at left depict the mean values for each model. Smoothers depict mean values at different levels of verification

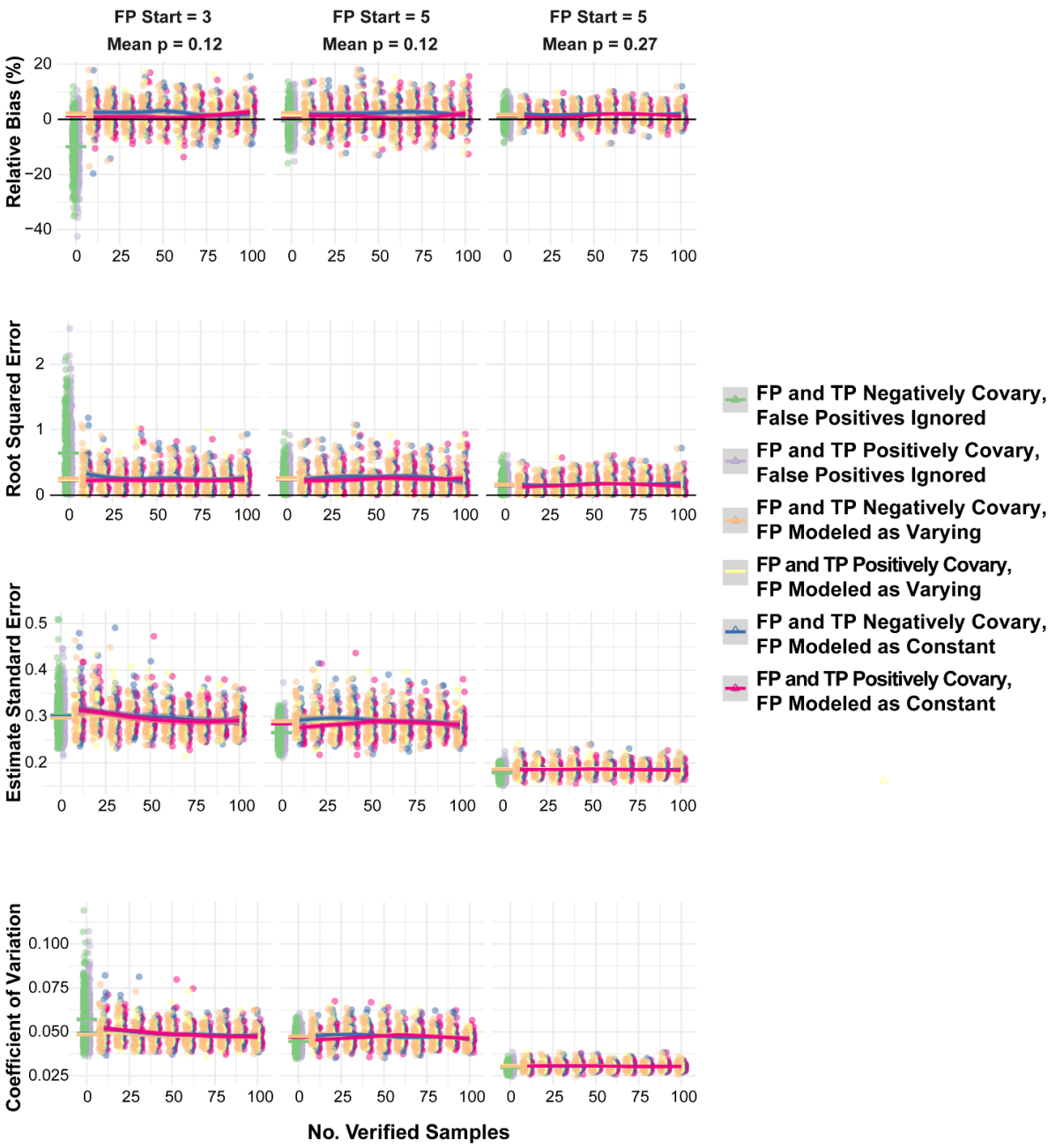

Figure S7 Performance of the standard phenological occupancy model (when 0 observations are confirmed, grey/green) and the model extension for false positive error with finite-sample population size with regard to expected arrival time under different directional covariance between true and false positive detections, and when different models for false positive error (either a misspecified constant model in blue/magenta or the generating model in yellow/orange) were fit. Standard Error = standard deviation of the posterior distribution. Data used includes scenarios 4-9 described below in Table S5. Horizontal bars at left depict the mean values for each model. Smoothers depict mean values at different levels of verification.

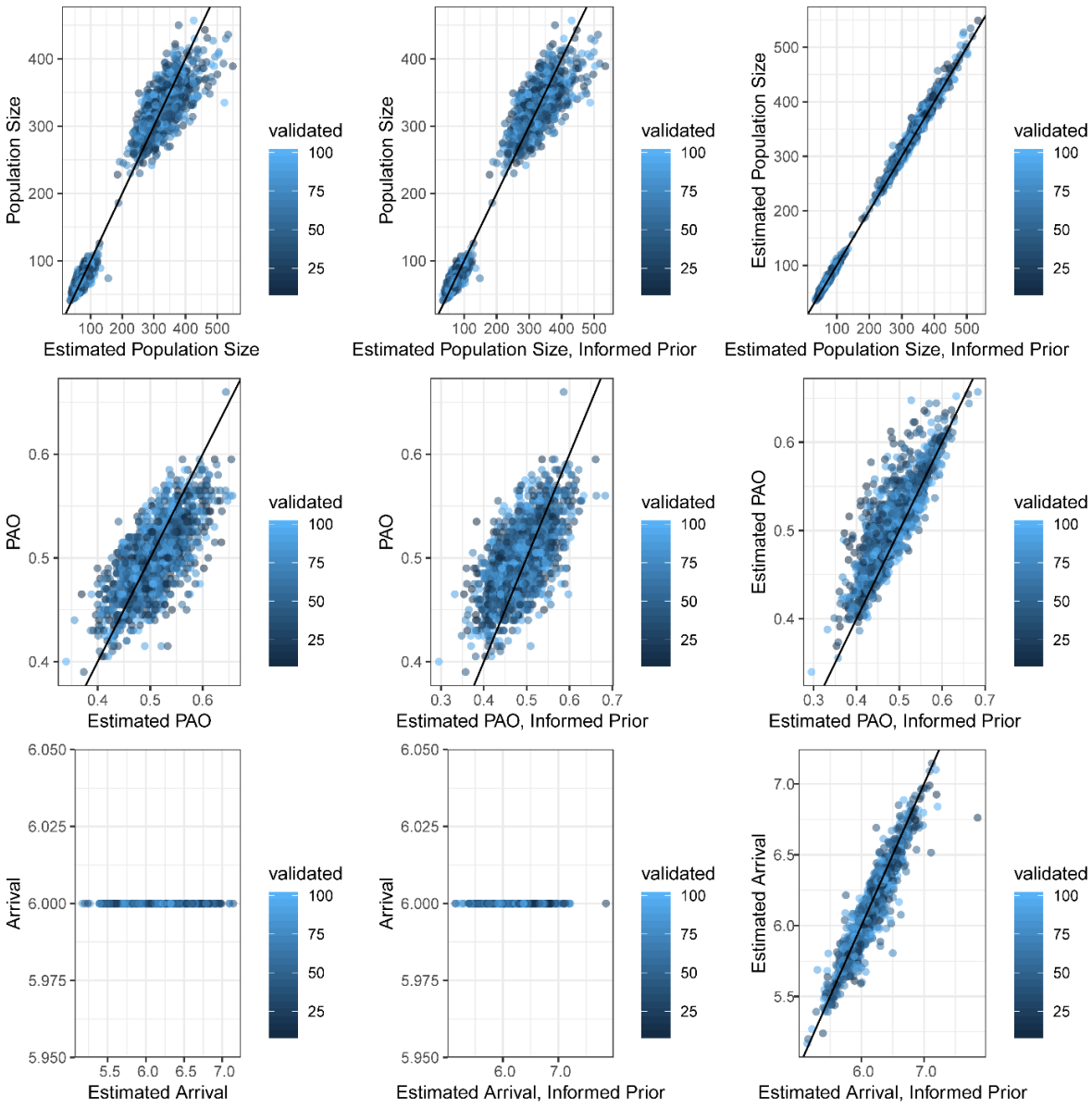

Figure S8. Correlation between true values, estimates made when incorporating confirmed observations, and estimates made when using an informed prior rather than incorporating confirmed observations in the likelihood.

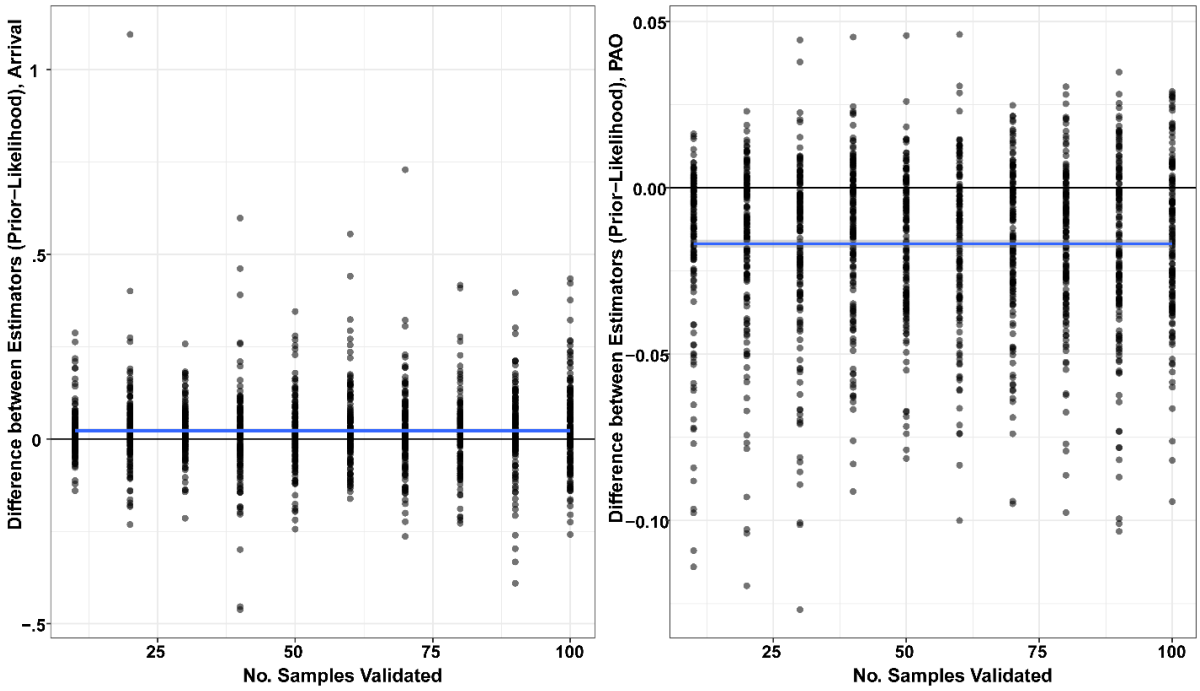

Figure S9. Noise in the correlations between estimates of PAO and arrival generated when including confirmed data in the likelihood vs. using an informed prior is not related to the size of the confirmation sample.

Table S1. Scenarios considered when evaluating the Royle-Nichols and extended models. **β** describes abundance terms, **α** describes detection terms. $\boldsymbol{X_1}$ is a vector of simulated covariates that influences abundance, and for some scenarios, the likelihood of false positive observation.

Scenario	β_0	β_1	α_0	α_1	False Positive Model
1	0	1	–1.73	1	$\theta_{fp} = 0.10$ of $\theta_{fp} + \theta_{tp}$
2	0	1	–1.73	1	$\theta_{fp} = 0.05$ of $\theta_{fp} + \theta_{tp}$
3	0	1	–1.73	1	$\theta_{fp} = 0.01$ of $\theta_{fp} + \theta_{tp}$
4	–1.50	1	–1.73	1	$\theta_{fp} = 0.10$ of $\theta_{fp} + \theta_{tp}$
5	–1.50	1	–1.73	1	$\theta_{fp} = 0.05$ of $\theta_{fp} + \theta_{tp}$
6	–1.50	1	–1.73	1	$\theta_{fp} = 0.01$ of $\theta_{fp} + \theta_{tp}$
7	–1.50	1	–1.73	1	$\text{logit}(\theta_{fp,i}) = -6 + 1X_{1,i}$
8	–1.50	1	–1.73	1	$\text{logit}(\theta_{fp,i}) = -6 - 1X_{1,i}$

Table S2. Mean error for parameters and relative bias of finite-sample population size for the standard and extended RN models across all scenarios and amounts of verification.

		Mean Error (Mean Absolute Bias)				Relative Bias (%)
Estimator	Scenario	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\widehat{N^*}$
Standard	1	0.71	-0.13	-0.58	-0.2	86
	2	0.42	-0.08	-0.35	-0.12	45
	3	0.1	-0.02	-0.08	-0.03	10
	4	0.71	-0.2	-0.48	-0.19	79
	5	0.37	-0.1	-0.28	-0.12	38
	6	0.08	-0.03	-0.07	-0.01	9
	7	0.28	0.00*	-0.23	-0.1	37
	8	0.61	-0.32	-0.35	-0.16	51
False Positive Extension	1	-0.01	0.01	-0.01	0.00	2
	2	-0.02	0.01	0.00	0.00	1
	3	-0.01	-0.01	0.00	0.00	1
	4	-0.02	-0.01	-0.01	0.01	2
	5	-0.06	0.02	0.00	0.00	1
	6	-0.03	0.01	-0.01	0.02	1
	7	-0.03	0.00	-0.01	0.01	2
	8	-0.06	0.00	0.00	0.01	-1
	1[A]	-0.04	0.02	0.00	0.01	0
	2[A]	-0.04	0.02	0.01	0.01	0
	3[A]	-0.02	0.00	0.01	0.00	0
	4[A]	-0.02	-0.01	-0.01	0.01	2
	5[A]	-0.06	0.02	0.00	0.00	1
	6[A]	-0.03	0.01	0.00	0.02	1
	7[A]	-0.03	-0.01	-0.01	0.01	1
	8[A]	-0.07	0.00	0.00	0.01	-1
	7[B]	-0.01	0.04	-0.02	0.00	6
	8[B]	-0.06	-0.03	0.01	0.01	-2
	1[C] (292) [D]	0.00	-0.02	0.08	-0.05	1
	2[C] (291) [D]	-0.01	-0.01	0.05	-0.03	1
	3[C] (293) [D]	0.00	-0.01	0.02	-0.01	0
	4 [C] (282)[D]	-0.01	-0.01	0.03	-0.01	5
	5 [C] (274)[D]	-0.06	0.01	0.02	-0.01	4
	6 [C] (277)[D]	-0.03	0.01	0.00	0.01	3

[A]Fit using an informed prior for θ_{fp} derived from estimates from an occupancy model following the observation confirmation protocol.
[B]A slightly mis-specified model that assumes θ_{fp} is constant.
[C]Fit following the site-confirmation protocol.
[D]The number of simulated datasets for which fitted models exhibited convergence (300 = max)
*Throughout, 0.00 used as a stand-in for absolute values < 0.01., or absolute % < 1

Table S3. Frequentist coverage of 95% CRI associated with parameters and finite-sample population size for the standard and extended RN models across the scenarios and amounts of verification considered.

		Coverage				
Estimator	Scenario	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\widehat{N^*}$
Standard	1	0	0.40	0.00	0.22	0.00
	2	0.093	0.76	0.18	0.56	0.08
	3	0.840	0.91	0.87	0.89	0.87
	4	0.066	0.56	0.09	0.52	0.01
	5	0.456	0.82	0.46	0.74	0.19
	6	0.906	0.93	0.90	0.88	0.90
	7	0.636	0.94	0.57	0.78	0.25
	8	0.066	0.29	0.24	0.63	0.06
False Positive Extension	1	0.98	0.95	0.95	0.93	0.95
	2	0.99	0.94	0.96	0.95	0.95
	3	0.98	0.95	0.97	0.96	0.96
	4	0.97	0.94	0.96	0.95	0.96
	5	0.97	0.96	0.96	0.92	0.96
	6	0.98	0.93	0.93	0.92	0.97
	7	0.97	0.93	0.94	0.95	0.95
	8	0.98	0.92	0.96	0.96	0.92
	1[A]	0.98	0.95	0.95	0.94	0.92
	2[A]	0.99	0.95	0.96	0.95	0.96
	3[A]	0.99	0.94	0.96	0.96	0.96
	4[A]	0.98	0.94	0.97	0.95	0.96
	5[A]	0.98	0.96	0.96	0.93	0.94
	6[A]	0.98	0.92	0.92	0.92	0.98
	7[A]	0.97	0.94	0.94	0.96	0.95
	8[A]	0.97	0.93	0.95	0.95	0.92
	7[B]	0.97	0.93	0.94	0.96	0.9
	8[B]	0.97	0.91	0.96	0.95	0.92
	1[C](292)[D]	0.91	0.93	0.88	0.89	0.74
	2[C](291)[D]	0.94	0.94	0.91	0.94	0.73
	3[C](293)[D]	0.95	0.95	0.96	0.96	0.73
	4[C](282)[D]	0.94	0.93	0.96	0.96	0.65
	5[C](274)[D]	0.92	0.95	0.95	0.93	0.67
	6[C](277)[D]	0.95	0.94	0.92	0.91	0.65

[A]Fit using an informed prior for θ_{fp} derived from estimates from an occupancy model following the observation confirmation protocol.
[B]A slightly mis-specified model that assumes θ_{fp} is constant.
[C]Fit following the site-confirmation protocol.
[D]The number of simulated datasets for which fitted models exhibited convergence (300 = max)
*Throughout, 0.00 used as a stand-in for absolute values < 0.01., or absolute % < 1

Table S4. Root mean squared error for parameters and relative bias of finite-sample population size for the standard and extended RN models across the scenarios and amounts of verification considered.

		RMSE				
Estimator	Scenario	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\widehat{N^*}$
Standard	1	0.71	0.14	0.58	0.2	280
	2	0.42	0.09	0.36	0.13	150
	3	0.1	0.07	0.11	0.07	39
	4	0.786	0.21	0.49	0.21	59
	5	0.377	0.14	0.28	0.16	28
	6	0.173	0.12	0.13	0.12	9.0
	7	0.29	0.11	0.24	0.15	27
	8	0.61	0.32	0.36	0.20	37
False Positive Extension	1	0.11	0.07	0.10	0.07	33
	2	0.11	0.06	0.10	0.07	29
	3	0.10	0.06	0.08	0.06	25
	4	0.17	0.12	0.12	0.11	7.2
	5	0.17	0.12	0.12	0.11	6.7
	6	0.16	0.12	0.12	0.11	6.0
	7	0.17	0.13	0.12	0.11	7.2
	8	0.19	0.13	0.12	0.11	6.6
	1[A]	0.12	0.07	0.10	0.07	32
	2[A]	0.11	0.06	0.10	0.07	28
	3[A]	0.10	0.06	0.09	0.06	25
	4[A]	0.18	0.12	0.12	0.11	7.4
	5[A]	0.17	0.12	0.12	0.11	6.9
	6[A]	0.16	0.12	0.12	0.11	6.5
	7[A]	0.17	0.13	0.13	0.11	7.3
	8[A]	0.19	0.13	0.12	0.11	6.9
	7[B]	0.17	0.13	0.12	0.10	8.1
	8[B]	0.19	0.14	0.12	0.11	6.9
	1[C] (292)[D]	0.12	0.07	0.13	0.08	51
	2[C] (291)[D]	0.11	0.06	0.11	0.06	52
	3[C] (293)[D]	0.09	0.07	0.09	0.06	47
	4[C] (282)[D]	0.18	0.12	0.12	0.11	16
	5[C] (274)[D]	0.18	0.12	0.12	0.11	14
	6[C] (277)[D]	0.16	0.12	0.12	0.11	13

[A]Fit using an informed prior for θ_{fp} derived from estimates from an occupancy model following the observation confirmation protocol.
[B]A slightly mis-specified model that assumes θ_{fp} is constant.
[C]Fit following the site-confirmation protocol. [D]The number of simulated datasets for which fitted models exhibited convergence (300 = max)
*Throughout, 0.00 used as a stand-in for absolute values < 0.01., or absolute % < 1

Table S5. Scenarios considered when evaluating the arrival model of Roth et al. (2014) and extended models. **β** describes occupancy terms, **α** describes detection terms. $\boldsymbol{X_1}$ is a vector of simulated covariates that influences abundance, and for some scenarios, the likelihood of false positive observation. φ denotes the simulated expected occasion of arrival. $\boldsymbol{X_1}$ is a vector of simulated covariates that influences the probability of occupancy, and for some scenarios, the likelihood of false positive observation.

Scenario	β_0	β_1	α_0	α_1	φ	Start of False Positives	False Positive Model
1	0	0.5	-2	0.5	6	>0	$\theta_{fp} = 0.10$ of $\theta_{fp} + \theta_{tp}$
2	0	0.5	-2	0.5	6	>0	$\theta_{fp} = 0.05$ of $\theta_{fp} + \theta_{tp}$
3	0	0.5	-2	0.5	6	>0	$\theta_{fp} = 0.01$ of $\theta_{fp} + \theta_{tp}$
4	0	0.5	-2	0.5	6	5	$\text{logit}(\theta_{fp,i}) = -6 + 1X_{1,i}$
5	0	0.5	-2	0.5	6	5	$\text{logit}(\theta_{fp,i}) = -6 - 1X_{1,i}$
6	0	0.5	-2	0.5	6	3	$\text{logit}(\theta_{fp,i}) = -6 + 1X_{1,i}$
7	0	0.5	-2	0.5	6	3	$\text{logit}(\theta_{fp,i}) = -6 - 1X_{1,i}$
8	0	0.5	-1	0.5	6	5	$\text{logit}(\theta_{fp,i}) = -6 + 1X_{1,i}$
9	0	0.5	-1	0.5	6	5	$\text{logit}(\theta_{fp,i}) = -6 - 1X_{1,i}$

Table S6. Estimator error and relative bias associated with parameters using the standard and extended arrival occupancy models across the scenarios in table S5.

		Mean Error (Mean Absolute Bias)				Relative Bias (%)	
Estimator	Scenario	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\widehat{PAO}$	$\hat{\varphi}$
Standard	1	0.36	0.03	-0.22	-0.04	16	-42
	2	0.18	0.02	-0.13	-0.02	9	-22
	3	0.07	0.03	-0.04	-0.01	3	-3
	4	0.19	0.17	-0.01	-0.02	8	-1
	5	0.24	-0.21	-0.04	-0.02	11	-1
	6	0.18	0.17	-0.04	-0.03	8	-1
	7	0.28	-0.24	-0.1	-0.03	13	-1
	8	0.11	0.13	-0.04	-0.01	5	0
	9	0.17	-0.15	-0.07	-0.01	8	1
False Positive Extension	1	0.03	0.07	-0.01	0.00*	0.01	2
	2	0	0.05	-0.01	0.00	0.01	2
	3	0.01	0.03	-0.02	-0.01	0.00	2
	4	0.05	0.09	-0.01	0.00	0.02	1
	5	0.02	-0.01	0.00	0.00	0.01	2
	6	0.00	0.06	0.00	0.00	0.00	2
	7	0.02	0.00	-0.11	0.00	0.01	2
	8	0.00	0.07	0.00	0.00	0.01	2
	9	0.01	0.02	-0.01	0.01	0.00	2
	1^A	0.03	0.08	-0.01	0.00	0.01	2
	2^A	-0.06	0.06	-0.01	0.01	-0.02	2
	3^A	-0.1	0.05	-0.01	0.00	-0.04	3
	4^A	0.06	0.11	-0.01	0.00	0.03	1
	5^A	0.05	0.04	0.01	0.01	0.04	3
	6^A	-0.02	0.06	-0.01	0.00	0.03	2
	7^A	0.06	0.05	-0.01	0.01	0.03	3
	8^A	0.04	0.04	-0.01	0.00	0.02	2
	9^A	0.03	0.02	0.02	0.00	-0.05	2
	4^B	0.06	0.13	0.00	0.00	0.03	1
	5^B	0.06	-0.1	-0.01	0.00	0.03	2
	6^B	0.01	0.13	0.01	0.00	0.00	1
	7^B	0.05	-0.11	-0.02	0.00	0.02	2
	8^B	0.02	0.10	0.00	0.00	0.01	1
	9^B	0.03	-0.03	-0.01	0.01	0.01	2
	1^C	0.02	0.08	0.04	-0.02	0.02	1
	2^C	-0.01	0.05	0.02	-0.01	0.02	0
	3^C	-0.01	0.04	0.00	-0.01	0.02	-1

Table S6 (Continued)

Estimator	Scenario	Mean Error (Mean Absolute Bias) $\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	Relative Bias (%) $\widehat{PAO}$	$\hat{\varphi}$
Standard Estimator, fit only using simulated true positives[D]	1	0.04	0.07	-0.01	0	0.02	2
	2	0.02	0.04	-0.01	0	0.02	2
	3	0.04	0.04	-0.02	0.01	0.02	2
	4	0.05	0.05	-0.02	0.01	0.02	1
	5	0.02	0.05	0	0	0.01	2
	6	0.03	0.04	0	0	0.01	2
	7	0.03	0.03	-0.02	0	0.01	2
	8	0	0.04	0	0	0	2
	9	0.01	0.04	-0.01	0.01	0	2
	1	0.04	0.07	-0.01	0	0.02	2

[A]Fit using an informed prior for θ_{fp} derived from estimates from an occupancy model following the observation confirmation protocol.
[B]A slightly misspecified model that assumes θ_{fp} is constant.
[C]Fit following the site confirmation protocol.
[D]This is the standard estimator fit only to simulated true positives, and as noted in the appendix text, there appears to be some small-sample bias associated estimates of arrival time.
*Throughout, 0 (or 0.00) used to denote an absolute value smaller than 0.01 after rounding (or < 1 if %).

Table S7. Frequentist coverage of 95% CRI associated with parameters using the standard and extended arrival occupancy models across the scenarios in table S5.

		Coverage					
Estimator	Scenario	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\widehat{PAO}$	$\hat{\varphi}$
Standard	1	0.54	0.94	0.41	0.9	0.25	0.1
	2	0.84	0.94	0.71	0.93	0.64	0.35
	3	0.94	0.93	0.94	0.95	0.93	0.82
	4	0.85	0.90	0.92	0.92	0.65	0.91
	5	0.81	0.76	0.91	0.94	0.53	0.93
	6	0.84	0.87	0.94	0.93	0.71	0.50
	7	0.67	0.74	0.85	0.97	0.40	0.53
	8	0.89	0.86	0.87	0.91	0.26	0.92
	9	0.82	0.84	0.78	0.90	0.08	0.95
False Positive Extension	1	0.96	0.93	0.95	0.96	0.97	0.93
	2	0.96	0.93	0.96	0.95	0.95	0.95
	3	0.95	0.92	0.97	0.95	0.96	0.94
	4	0.95	0.94	0.92	0.96	0.94	0.93
	5	0.96	0.95	0.93	0.95	0.94	0.93
	6	0.95	0.93	0.95	0.94	0.97	0.96
	7	0.94	0.94	0.95	0.97	0.96	0.94
	8	0.95	0.93	0.95	0.95	0.97	0.90
	9	0.96	0.95	0.93	0.96	0.97	0.94
	1[A]	0.97	0.93	0.97	0.95	0.97	0.94
	2[A]	0.95	0.93	0.96	0.94	0.94	0.94
	3[A]	0.91	0.93	0.97	0.96	0.91	0.92
	4[A]	0.94	0.94	0.93	0.95	0.96	0.94
	5[A]	0.92	0.94	0.95	0.94	0.96	0.92
	6[A]	0.93	0.95	0.96	0.93	0.95	0.95
	7[A]	0.94	0.94	0.95	0.94	0.96	0.94
	8[A]	0.92	0.94	0.94	0.96	0.96	0.90
	9[A]	0.94	0.93	0.94	0.95	0.96	0.93
	4[B]	0.94	0.9	0.92	0.95	0.93	0.92
	5[B]	0.94	0.9	0.93	0.95	0.92	0.93
	6[B]	0.95	0.92	0.94	0.94	0.98	0.95
	7[B]	0.93	0.9	0.95	0.97	0.93	0.93
	8[B]	0.93	0.91	0.95	0.95	0.93	0.91
	9[B]	0.96	0.94	0.93	0.95	0.97	0.92
	1[C]	0.96	0.93	0.93	0.95	0.98	0.93
	2[C]	0.96	0.94	0.95	0.94	0.96	0.94
	3[C]	0.97	0.94	0.97	0.95	0.96	0.93

[A]Fit using an informed prior for θ_{fp} derived from estimates from an occupancy model following the observation confirmation protocol.
[B]A slightly mis-specified model that assumes θ_{fp} is constant.
[C]Fit following the site confirmation protocol.

Table S8. Root mean squared error associated with parameters using the standard and extended arrival occupancy models across the scenarios in table S5.

		RMSE					
Estimator	Scenario	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\widehat{PAO}$	$\hat{\varphi}$
Standard	1	0.36	0.18	0.224	0.08	0.08	2.52
	2	0.23	0.17	0.15	0.08	0.05	1.36
	3	0.17	0.16	0.1	0.08	0.03	0.4
	4	0.23	0.21	0.09	0.08	0.04	0.23
	5	0.25	0.24	0.1	0.07	0.06	0.24
	6	0.22	0.22	0.09	0.08	0.04	0.64
	7	0.3	0.26	0.12	0.08	0.07	0.65
	8	0.15	0.18	0.06	0.06	0.03	0.15
	9	0.19	0.18	0.09	0.06	0.04	0.14
False Positive Extension	1	0.16	0.19	0.09	0.08	0.03	0.28
	2	0.16	0.17	0.09	0.08	0.03	0.25
	3	0.16	0.17	0.08	0.08	0.02	0.24
	4	0.18	0.18	0.1	0.08	0.03	0.25
	5	0.16	0.17	0.1	0.07	0.03	0.27
	6	0.17	0.18	0.09	0.08	0.02	0.23
	7	0.17	0.17	0.09	0.08	0.03	0.26
	8	0.13	0.15	0.06	0.06	0.01	0.15
	9	0.12	0.13	0.06	0.06	0.01	0.17
	1[A]	0.17	0.2	0.09	0.08	0.03	0.28
	2[A]	0.18	0.18	0.09	0.08	0.03	0.27
	3[A]	0.19	0.17	0.09	0.08	0.03	0.27
	4[A]	0.19	0.19	0.1	0.08	0.03	0.25
	5[A]	0.18	0.18	0.1	0.07	0.03	0.27
	6[A]	0.19	0.18	0.09	0.08	0.02	0.24
	7[A]	0.19	0.18	0.09	0.08	0.03	0.27
	8[A]	0.14	0.15	0.07	0.06	0.01	0.17
	9[A]	0.14	0.14	0.06	0.06	0.01	0.18
	4[B]	0.18	0.2	0.1	0.07	0.03	0.24
	5[B]	0.16	0.19	0.1	0.07	0.03	0.27
	6[B]	0.18	0.2	0.09	0.08	0.02	0.23
	7[B]	0.17	0.19	0.09	0.08	0.03	0.26
	8[B]	0.13	0.16	0.06	0.06	0.01	0.15
	9[B]	0.12	0.14	0.06	0.06	0.01	0.17
	1[C]	0.17	0.19	0.1	0.08	0.03	0.29
	2[C]	0.16	0.18	0.09	0.08	0.03	0.25
	3[C]	0.16	0.17	0.08	0.08	0.02	0.25

[A]Fit using an informed prior for θ_{fp} derived from estimates from an occupancy model following the observation confirmation protocol.
[B]A slightly mis-specified model that assumes θ_{fp} is constant.
[C]Fit following the site confirmation protocol.

Appendix S3. Details associated w/ case study.

The data used here were trail camera images classified via crowdsourcing from 91,276 24-h periods at 944 distinct locations across the state in 2017 between Julian days 150 and 320 (Figure S1). We defined sampling occasions as 24-hr periods, and reviewed all reported images reported as gray fox classifications (n =247 images) from 179 occasions at 127 locations; 90 other occasions at 55 distinct locations included putative but unconfirmed gray fox detections (Figure S1). Covariates for expected abundance are noted in Table S1 and were extracted from a circular buffer of 1 or 5 km radius surrounding the camera locations. Spatial smoothing was implemented via a 2-dimensional cubic spline (Guélat and Kéry 2018) across latitude and longitude with 20 knots placed across the state. The detection probability of an individual animal at different sites was modeled as varying in relation to whether the camera was placed on a maintained trail or not and as a quadratic function of the distance between the camera and the location the camera was targeting (as reported by volunteers), and false positive error probability was modeled as a logistic function of the proportion of cropland within a circular buffer with 5 km radius to account for what we expected to be increased prevalence of species confused with gray foxes (red fox, coyote) relative to foxes themselves. The prediction grain (a 2 x 2 km lattice) was chosen to approximate gray fox home range sizes in Wisconsin, which are believed to be slightly larger than the home ranges reported slightly further south (e.g., Haroldson and Fritzell 1984, Duell et al. 2017).

In Clare et al. (2019), we noted that perhaps the most useful predictor for classification error across a range of species was the degree of unanimity in the crowdsourced classifications (e.g., 100% of votes for gray fox = more likely to be gray fox than 50% votes). We do not use this term here to avoid the inelegance of having to define/impute values associated with crowdsourced agreement within sampling occasions in which the gray fox was not detected, although following the logic that greater confidence leads to lower probability of a false positive, imputing these values as 1.0 (i.e., 100% agreement) seems like a reasonable hack. Similar inelegancies associated with agreement arise for sampling occasions with multiple images; one way to model this that might be reasonable might be to define a covariate (or the

false positive probability) as $1 - \prod_{p=1}^{npictures}(1 - agreement_p)$. For example, in a case with 2 pictures in with 50 and 40 % agreement, the value of the operator = 0.7, in a case with 1 picture with 90% agreement, the value = 0.1, which seems to correctly imply a false positive is more likely in the first case because there are more images with less confidence. Of course, in situations with 0 pictures, the operator breaks down, and the value might need to be fixed at one.

Of the reviewed images, we confirmed 67% as correct classifications, with the rest either misclassifications of coyote (*Canis latrans*) or red fox (*Vulpes vulpes*). Once aggregated within sampling occasions, 60% (108) of the occasions consisted exclusively of true positives, and 40% (71) included exclusively false positives; no confirmed sampling intervals included a mixture of true and false positive observations. This might suggest some potential lack of independence between true and false positive outcomes, but we ignore that here because volunteers view images on the crowdsourcing platform at random, which makes it difficult to imagine that misperception has any serial structure or exhibits any other form of dependence. We note that both true and false positives were reported at 6 locations. Estimates are summarized in Tables S2 and S3.

References

Deuel, N.R., Conner, L.M., Miller, K.V., Chamberlain, M.J., Cherry, M.J. and Tannenbaum, L.V. 2017. Gray fox home range, spatial overlap, mated pair interactions and extra-territorial forays in southwestern Georgia, USA. Wildlife Biology: https://doi.org/10.2981/wlb.00326.

Guélat, J. and Kéry, M. 2018. Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods in Ecology and Evolution 9:1614-1625.

Haroldson, K.J. and Fritzell, E.K. 1984. Home ranges, activity, and habitat use by gray foxes in an oak-hickory forest. The Journal of Wildlife Management 48:222-227.

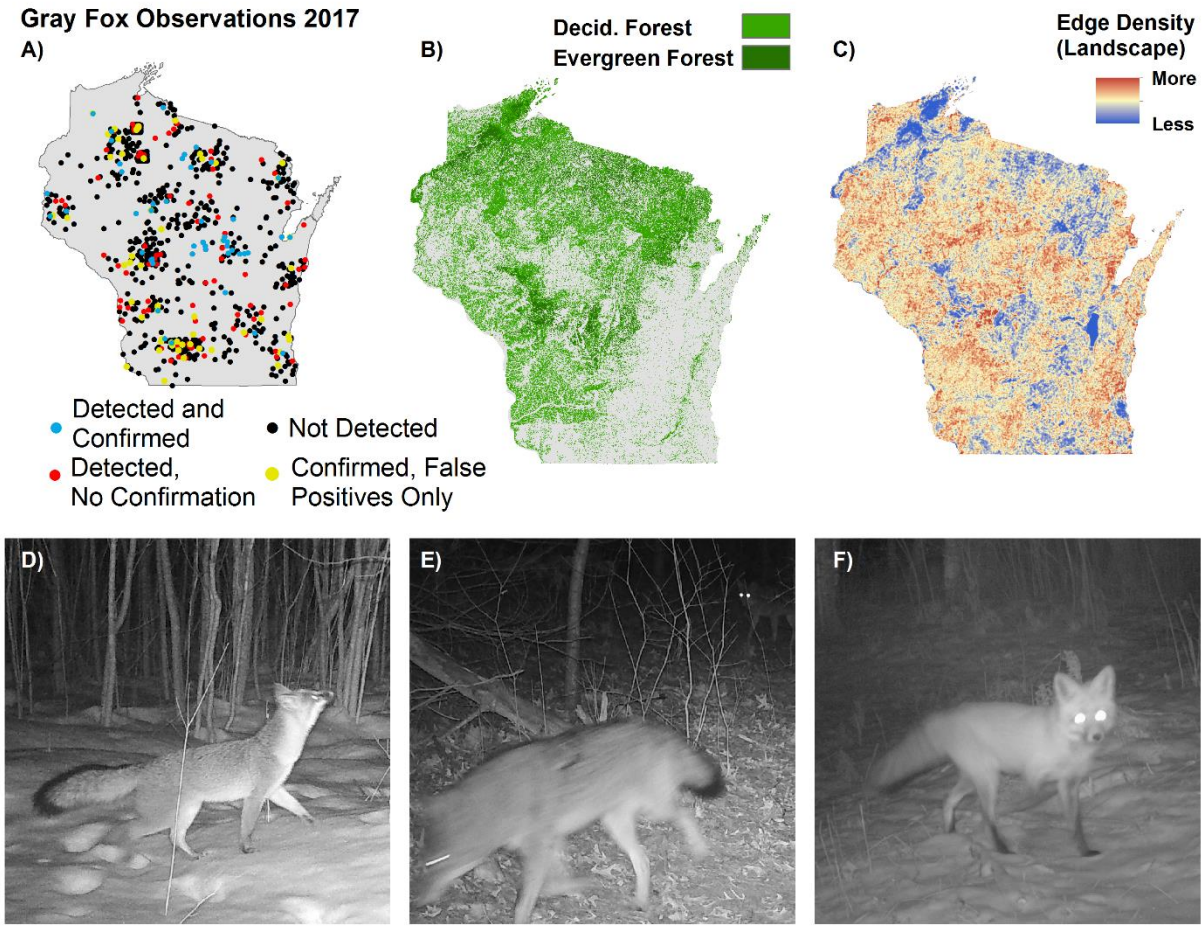

Figure S1. Location of sampling locations, confirmed observations, and unconfirmed detections used in the case study (A); covariates associated with model fitting (B and C). Reported gray fox (D) observations were commonly truly either coyote (E) or red fox (F).

Table S1. Covariates used in case study analysis.

Name	Description	Source	Real Parameter
Evergreen	% Evergreen Forest in 1 km radius buffer surrounding camera location	NLCD 2011[A]	λ
Snowdepth	Mean Snowdepth in 1 km radius buffer surrounding camera location between 2010 and 2017	SNODAS[B]	λ
Edge Density	(Landscape) edge density within 1 km radius buffer surrounding camera location	NLCD 2011	λ
Cropland	% Cropland in 1 km radius buffer surrounding camera location	NLCD 2011	λ, θ_{fp}
Grassland	% Grassland in 1 km radius buffer surrounding camera location	NLCD 2011	λ
Deciduous	% Deciduous forest in 1 km radius buffer surrounding camera location	NLCD 2011	λ
Trail	Binary; camera placed on maintained trail (vs. game trail or non-trail);		r
Distance	Distance between camera and target location (m); modeled as quadratic		r

[A]Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J. and Megown, K., 2015. Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. Photogrammetric Engineering & Remote Sensing 81:345-354.

[B]Barrett, A. 2003. National Operational Hydrologic Remote Sensing Center SNOw Data Assimilation System (SNODAS) Products at NSIDC. NSIDC Special Report 11. Boulder, CO, USA: National Snow and Ice Data Center. Digital media.

Table S2. Parameter estimates, variable inclusion probabilities (*w*), and 95% credible intervals (LCI, UCI) from a RN model for gray fox relative abundance ignoring false positive error.

Parameter	Estimate	LCI	UCI
Intercept, λ	-1.069	-2.573	1.107
Evergreen	-0.381	-2.988	2.148
Snowdepth	0.003	-6.171	6.197
Edge Density	-0.031	-6.218	6.088
Cropland	-0.009	-6.125	6.166
Grassland	0.011	-6.174	6.117
Deciduous	-0.011	-6.051	6.054
Evergreen, w^A	0.889	0	1
Snowdepth, w^A	0.034	0	1
Edge Density, w^A	0.028	0	1
Cropland, w^A	0.042	0	1
Grassland, w^A	0.068	0	1
Deciduous, w^A	0.048	0	1
Evergreen^B	-0.344	-0.653	0
Snowdepth^B	-0.001	0	0
Edge Density^B	0	0	0
Cropland^B	0.003	0	0.060
Grassland^B	0.008	0	0.153
Deciduous^B	-0.005	-0.091	0
Logit Intercept, r	-4.359	-4.587	-4.143
Trail	-0.166	-0.504	0.172
Distance	0.266	0.067	0.471
Distance^2	-0.133	-0.234	-0.045
Spline Coefficient[1]	0.010	-0.503	0.753
Spline Coefficient[2]	-0.023	-0.618	0.423
Spline Coefficient[3]	0.008	-0.267	0.648
Spline Coefficient[4]	-0.028	-0.490	0.218
Spline Coefficient[5]	0.001	-0.542	0.506
Spline Coefficient[6]	0.011	-0.505	0.608
Spline Coefficient[7]	-0.010	-0.559	0.519
Spline Coefficient[8]	0.034	-0.334	0.337
Spline Coefficient[9]	0.016	-0.478	0.602
Spline Coefficient[10]	0.037	-0.443	0.642
Spline Coefficient[11]	-0.111	-0.85	0.162
Spline Coefficient[12]	0.023	-0.482	0.624
Spline Coefficient[13]	-0.004	-0.488	0.400
Spline Coefficient[14]	0.003	-0.524	0.582
Spline Coefficient[15]	0.052	-0.353	0.702
Spline Coefficient[16]	0.011	-0.496	0.588
Spline Coefficient[17]	0.025	-0.444	0.649
Spline Coefficient[18]	0.010	-0.516	0.589
Spline Coefficient[19]	0.030	-0.443	0.694
Spline Coefficient[20]	0.040	-0.423	0.672

Table S2 (Continued)

[A]Inclusion probability—marginal posterior probability that term was included within the model
[B]Regularized coefficient. Unregularized coefficients include draws from the prior when the model term was not included within the model; regularized coefficients represent the iterative product of the raw coefficient and the indicator for inclusion.

Table S3. Parameter estimates, variable inclusion probabilities (*w*), and 95% credible intervals (LCI, UCI) from a RN model for gray fox relative abundance accounting for false positive error.

Parameter	Estimate	LCI	UCI
Log intercept, λ	-2.366	-4.415	1.022
Evergreen	-0.641	-4.228	4.132
Snowdepth	0.033	-6.062	6.118
Edge Density	0.217	-5.297	5.485
Cropland	-0.088	-5.914	5.913
Grassland	-0.011	-6.025	6.122
Deciduous	-0.272	-5.292	5.111
Logit Intercept, s_0	-6.486	-6.670	-6.311
Cropland, s_0	0.289	0.135	0.439
Evergreen, w^A	0.731	0	1
Snowdepth, w^A	0.102	0	1
Edge Density, w^A	0.451	0	1
Cropland, w^A	0.198	0	1
Grassland, w^A	0.088	0	1
Deciduous, w^A	0.493	0	1
Evergreen^B	-0.642	-1.664	0
Snowdepth^B	0.021	0	0.362
Edge Density^B	0.207	0	0.779
Cropland^B	-0.068	-0.590	0
Grassland^B	-0.017	-0.313	0
Deciduous^B	-0.244	-0.861	0
Logit Intercept, r	-3.514	-3.825	-3.237
Trail	-0.384	-0.983	0.195
Distance	0.668	0.344	1.019
Distance^2	-0.218	-0.495	0.191
Spline Coefficient[1]	0.008	-0.565	0.519
Spline Coefficient[2]	-0.030	-0.657	0.511
Spline Coefficient[3]	-0.114	-0.632	0.184
Spline Coefficient[4]	0.016	-0.285	0.366
Spline Coefficient[5]	0.003	-0.607	0.595
Spline Coefficient[6]	0.013	-0.596	0.637
Spline Coefficient[7]	-0.007	-0.618	0.598
Spline Coefficient[8]	0.005	-0.528	0.410
Spline Coefficient[9]	0.013	-0.591	0.635
Spline Coefficient[10]	0.032	-0.526	0.660
Spline Coefficient[11]	-0.079	-0.718	0.352
Spline Coefficient[12]	0.010	-0.606	0.622
Spline Coefficient[13]	-0.009	-0.538	0.534
Spline Coefficient[14]	-0.003	-0.620	0.605
Spline Coefficient[15]	0.004	-0.610	0.587
Spline Coefficient[16]	0.001	-0.604	0.610
Spline Coefficient[17]	-0.002	-0.628	0.593
Spline Coefficient[18]	0.002	-0.610	0.613

Table S3 (Continued)

Parameter	Estimate	LCI	UCI
Spline Coefficient[19]	-0.003	-0.629	0.590
Spline Coefficient[20]	0.008	-0.600	0.602

[A]Inclusion probability—marginal posterior probability that term was included within the model
[B]Regularized coefficient. Unregularized coefficients include draws from the prior when the model term was not included within the model; regularized coefficients represent the iterative product of the raw coefficient and the indicator for inclusion.

Chapter 3 – A phenology of fear? Static and seasonal predictors mediate white-tailed deer responses to metrics of predation risk from wolves.

Abstract

Quantifying non-consumptive predator effects upon prey, cascading effects upon vegetation, and elucidating general rules for their occurrence and strength is important for ecological applications ranging from forest or rangeland managagement to endangered species conservation. Despite substantial research effort along each front, a variety of challenges have complicated efforts to confront theory with empirical data, particularly with large mammalian predators and prey. Here, we assess the responses of white-tailed deer to wolves and other predators using a large and year-round network of trail cameras in a system where empirical research suggests wolves are triggering a non-consumptive trophic cascade that the existing body of theory suggests should be unlikely. We use a broad suite of risk and response metrics to consider support for multiple competing hypotheses regarding the state of the system and to assess the dynamics of habitat-mediated responses to predation risk. Deer responded to risk metrics in contrast with the expectations of a purely bottom-up system, although results indicate that predictors derived from satellite-based vegetation indices may poorly describe foraging resources. Deer responses to measures of predator occurrence were often environmentally mediated, suggesting proactive risk responses unexpected if wolves behaved as archetypical active predators. Deer activity decreased given the near-term occurrence of wolves, particularly in areas with greater surrounding forest cover, but wolves did not appear to use these areas more frequently following expectations of archtypical stalking predators. Deer allocated less time to foraging along linear features given near-term wolf occurrence (and wolves used these features more frequently), but were more likely to aggregate and forage in areas recently used by wolves given deeper snow where previous research suggests an inflated likelihood of mortality. Overall, results are consistent with the contention that wolves are triggering a non-consumptive trophic cascade in the Great Lakes region of North America, but suggest that predator hunting mode may be less important than environmental context when predicting responses to predation risk, and that the landscape of fear may follow a distinct phenology driven by temporal variation in foraging resources and the nutrional and behavioral states of both predators and prey.

Introduction

Predators impact prey in three primary ways: by killing and eating them, scaring them away from beneficial resources like food, or elevating chronic stress responses that reduce fitness (e.g., Brown and Kotler 2004, Sheriff et al. 2009, MacLeod et al. 2018, Creel 2018). The direct and indirect effects of predators on their prey may have further cascading effects on other ecosystem attributes such as vegetation structure (e.g., Estes et al. 2011). As such, understanding and quantifying how predators affect prey has important implications for understanding biotic interactions, ecosystem function, and how removing or restoring predator populations might influence managerial or conservation goals (Creel and Christianson 2008, Smith et al. 2010, Ritchie et al. 2012, Prugh et al. 2019, Gaynor et al. 2020).

A well-developed body of theory has sought to explain and predict how predators impact prey, how prey respond to predation risk, and how these processes impact ecosystem processes. Prey are expected to behaviorally respond to the threat of predation if 1) they can perceive risk; 2) a behavioral response reduces risk; 3) the risk varies over space or time; and 4) there are costs to constitutive defenses (Lima and Bednekoff 1999, Creel 2018). However, responses to predation risk exhibit variable costs and benefits. Predictable and surmountable risks may produce stronger proactive responses in prey, such as shifting space use, altering periods of activity, aggregating in larger groups, or increasing baseline vigilance (Schmitz et al. 2004, Creel et al. 2018). Some of these proactive responses, such as shifts in space use or reduced foraging, may carry substantial fitness costs and have cascading effects upon vegetation (Schmitz et al. 2004, Creel and Christianson 2008, Creel 2018). Other proactive responses (circadian shifts in activity) may induce smaller fitness costs and result in negligible effects upon vegetation (Kohl et al. 2018). If prey cannot predict risk or proactive responses are otherwise ineffective, prey are expected to respond to predators more proximally or reactively (i.e., after the initial encounter), with such responses broadly expected to have limited fitness costs or cascading effects (Lima and Bednekoff 1999, Schmitz 2004, Creel 2018).

Variation in risk and its predictability is primarily ascribed to a combination of three factors: predator hunting mode, predator habitat domain relative to prey, and the degree to which spatiotemporal environmental variability enhances (or dampens) risk by increasing predator lethality or the likelihood of encounter (Schmitz et al. 2004, Preisser et al. 2007, Palmer et al. 2017, Moll et al. 2017, Kohl et al. 2018). Sit-and-wait, sit-and-pursue, and other stalking hunting modes are believed to produce more predictable and variable risks that elicit stronger anti-predator responses (Schmitz et al. 2004, Preisser et al. 2007). These hunting modes often depend upon a narrow environmental domain where foraging or other resources are sufficient to attract prey to use it despite added predation risk (Sih 2005, Smith et al. 2020). As predators concentrate their space use in environmental features with greater prey resources (winning the so-called 'space-race'; Sih 2005), they produce more concentrated risk cues (Preisser et al. 2007, Miller et al. 2014). Prey respond to this predictable risk by reducing resource use or intake (Luttbeg et al. 2020), which is more likely to carry foraging costs (Brown and Kotler 2004, Creel 2018) and lead to cascading effects upon vegetation (Schmitz et al. 2004). In contrast, active, chase-and-pursue predators are not typically bound to a narrow environmental domain (although they may preferentially use or be more effective in some habitats, Kauffman et al. 2007) and are expected to leave more diffuse cues that prey should be less likely to proactively respond to (Lima and Bednekoff 1999, Preisser et al. 2007, Luttberg et al. 2020).

How predators affect prey, how prey respond to risk, and how prey responses to risk may in turn affect vegetation are questions of great importance for managing wildlife populations and other natural resources (Gaynor et al. 2020). For example, white-tailed deer (*Ocodeilius virginianus*) are a valuable game species that heavily influence their environments via herbivory or other activities associated with heavy use, and have been implicated in undesirable changes to forest vegetation across the northeastern United States (Horsely et al. 2003, Ferker et al. 2014, Bradshaw and Waller 2016, Sabo et al. 2017). Recent research within the Great Lakes region suggests that reduced deer foraging pressure in areas occupied or heavily used by wolves (*Canis lupus*) may be having cascading vegetation effects (Callan et

al. 2013, Flagel et al. 2016) that are desirable for both conservation and commercial objectives (Horsely et al. 2003, Rooney et al. 2004). For example, Callan et al. (2013) found that local forb species richness was greater in plots within wolf pack territories that had been occupied longer. Flagel et al. (2016) found that deer visitation rates and foraging behaviors were markedly reduced in areas of heavy wolf use, that sapling height and forb species richness were greater in areas of heavy wolf use, and that differences in sapling height and forb richness between control and exclosed plots disappeared in areas heavily used by wolves.

Despite being a major research focus for theoretical and applied ecologists, quantifying and predicting prey responses to predation risk remains challenging. This poses further challenges for the study of trophic cascades and the broader ecosystem impacts of predation (Ford and Goheen 2015). Experimental tests (e.g., Schmitz 1998, Miller et al. 2014, Luttbeg et al. 2020) have upheld theoretical predictions regarding how prey should respond to different forms of risk, but their transferability to natural systems has been questioned (Peers et al. 2018, Smith et al. 2019). Conversely, observational studies commonly produce conflicting results that may reflect either real variation between systems and limitations to the existing theory, methodological or scalar inconsistencies (Moll et al. 2017, Prugh et al. 2019), or may be a product of inability to directly assess certain key interactions (Ford and Goheen 2015). For example, the results of Callan et al. (2013) are consistent with a food-web driven from the bottom-up (Ford and Goheen 2015), while Flagel et al.'s (2016) suggestion that deer respond to predation risk from wolves because the latter operate more like stalking predators conflicts with existing beliefs related to wolf hunting mode within other systems (Kaufmann et al. 2010, Middleton et al. 2013, Kohl et al. 2018, Mumma et al. 2018, Dickie et al. 2020). Moroever, Flagel et al. (2016) focused upon an area smaller than a wolf pack territory, leading to uncertainty regarding whether these compelling patterns hold over a more meaningful spatial extent, and neither they nor Callan et al. (2013) assessed the effects of alternative predators.

A final complication is that the landscape of fear (and the reciprocal landscape of forage, Gallagher et al. 2016) is often dynamic (Palmer et al. 2017, Kohl et al. 2018). Most experimental studies sample responses to predation risk under experimental conditions where risk and resource trade-offs are essentially static; while most observational studies either limit sampling to a relatively narrow temporal domain or average over potentially important temporal variation in favor of assessing spatial variation in responses (Palmer et al. 2017, Kohl et al. 2018). It is believed that the foraging costs of effects of predation risk are highly context-dependent (Wirsing et al. 2020). Such context includes the behavioral and nutritional states of predators and prey, potentially emphemeral environmental factors that facilitate rates of encounter, lethality, or prey escape, and the foraging resources that prey must forgo when investing in anti-predator behavior (Brown and Kotler 2004, Moll et al. 2017, Wirsing et al. 2020). Of the potential temporal axes that might be used to describe the dynamics of fear, perhaps none is more germane to each of these components than seasonality, which may influence forage availability, rates of predator encounter and lethality, and strongly shapes the condition and life-history of predators and prey.

Here, we take advantage of occurrence and behavioral data collected across a large environmental gradient using a network of trail cameras maintained across a calendar year by volunteers to gain insights into how deer behaviorally respond to wolves or other predators, which carries implications for assessing existing theoretical predictions and inferring whether wolves are causing a trophic cascade. We ground our study in competing (but not mutually-exclusive) hypotheses (Figure 1). First, observed associations between wolves and vegetation may be driven by bottom-up processes (Figure 1A, Callan et al. 2013, Ford and Goheen 2015). Alternatively, associations between wolves and vegetation may result from a trait-mediated (behavioral) trophic cascade. This cascade may occur because: a) despite wolves operating as unpredictable and generalist active predators, even the weak cascading effects expected in these circumstances are detectable (Figure 1B); b) deer respond strongly to wolf risk cues because wolves behave as stalking predators in the region (Flagel et al. 2016, Figure 1C); c) some other component(s) of wolf hunting mode—e.g., attraction to travel corridors that increase the likelihood of encountering prey

(Dickie et al. 2020)—creates strong and predictable risk cues that deer respond to (Figure 1D). A further possibility is that the patterns observed by Callan et al. (2013) and Flagel et al. (2016) result from a consumptive (density-mediated) trophic cascade (i.e., predation reduces deer densities and leads to vegetation release). Generalist active predators are perhaps more likely to generate consumptive trophic cascades (Schmitz et al. 2004, Middleton et al. 2013), but as we shortly describe, this seems unlikely in this system.

Indeed, that none of the hypotheses is clearly favored by existing evidence, theory, or general understanding of the system and species in question is a primary motivation for the study with clear implications for cascading effects on ecosystem function. Under a bottom-up driven system, one would expect a positive association between predator space use and prey resources. However, this sort of indirect resource matching is also consistent with the expectations of a predator winning the predator-prey space race (Sih 2005), a strategy more strongly associated with predators that employ sit-and-pursue, stalking, or other more sedentary hunting modes (Luttbeg et al. 2020, Smith et al. 2020). If wolves operate as ambush predators in more densely forested regions, one would expect that wolves would be more likely occur within forested areas that increase the lethality of their attacks, and that deer would exhibit proactive risk reduction responses (e.g., avoidance) to areas of greater forest cover within the range of the predator species but not outside of its range, and that deer would respond more strongly to proximal wolf risk cues within forested areas than outside of them (Lima and Dill 1990, Moll et al. 2017, Creel 2018). Although wolf-killed deer in some regions are concentrated within denser vegetation (e.g., Kunkel and Pletscher 2001), differences between wolf depredation locations and random locations or locations used by deer in northern Wisconsin are primarily explained by snow depth rather than land cover (Olson 2019). If wolves behave consistently with the archetypical expectations for active predators with a broad habitat domain, then few marked environmental patterns in wolf space use or deer responses to wolf risk metrics are expected (Schmitz et al. 2004). Finally, wolves might be active predators with foraging strategies that center on using linear features to increase rates of movement and the likelihood of

encountering prey (Whittington et al. 2011, Dickie et al. 2017, Mumma et al. 2018, Dickie et al. 2020). Such behavior has not been formally integrated into the broader theory, but suggests that active predators could also be capable of generating concentrated risk cues expected to trigger stronger prey responses, and there is evidence that prey may perceive and respond to these cues (Demars and Boutin 2018, Dickie et al. 2020). As an orthogonal line of inquiry, we were interested in the degree to which predator-deer interactions exhibited seasonal variation, as would be expected given the seasonality of many risk and forage metrics in this system.

Methods

Study Area and Ecological Context

We focus on deer responses to metrics of predation risk associated with three predators across Wisconsin and a small portion of Michigan, USA. Deer across the Great Lakes region of North America are widely hunted and contend with 4 mammalian predators: wolves, coyotes (*C. latrans*), black bears (*Ursus americanus*), and bobcats (*Lynx rufus*). Harvest is the primary cause of adult mortality across most of the region (DelGiudice et al. 2002, Norton 2015). The combined predation from coyotes and wolves during particularly snowy years in regions with above average snowfall may induce nearly as much mortality as human harvest (DelGiudice et al. 2002), in most parts of Wisconsin during most years, predation is a minor cause of mortality relative to hunting, and overall rates may be overstated due to difficulty distinguishing between predation and scavenging events (Norton 2015). Even in the harshest climatic zones within the region, deer population growth rates are better explained by winter severity than wolf population size (Post and Stenseth 1998), and the size of Wisconsin's deer herd within the part of the state most heavily occupied by wolves exhibits no clear trend over the previous 20 years despite marked increases in wolf population size (Wojcik and Stenglein 2020, Wiedenhoft et al. 2020). In the northern (more forested and snowier) parts of the region, predation is a major cause of fawn mortality, but is primarily attributable to black bears, coyotes, and bobcats, while in milder and more agricultural areas,

other natural causes of mortality such as malnutrition appear to be more common (Carstensen et al. 2009, Warbington et al 2017, Kautz et al. 2019). Consequently, it is believed that wolves have only limited consumptive effects upon deer populations and has been suggested that any cascading effects within the region would be more likely to arise from non-consumptive effects (Rooney and Waller 2009, Flagel et al. 2016).

Data Collection and Analysis

We used images collected across the 2017 calendar year by Snapshot Wisconsin, a state-wide trail camera monitoring program managed by the Wisconsin Department of Natural Resources (Locke et al. 2019, Townsend et al. 2020), to analyze patterns in predator occurrence, and deer occurrence, counts, and behaviors. The initial dataset included images sampled across 216,774 camera days at 1,213 locations. We thinned locations so that they had 1.5 km nearest neighbor distance to minimize spatiotemporal dependence that would be difficult to account for. A major component of our analysis focused on estimating autoregressive occurrence effects, and we censored data without a previous time-series step (24-hr period) rather than use imputation techniques. This left 816 camera locations sampling over 151,980 camera days for most analyses.

Project cameras are set to take an image triplicate (hereafter, sequences) when a triggering event occurs, with a minimum 15 s delay between triggers. Species in image sequences were classified by a combination of trail camera hosts using an agency-developed platform and on a crowdsourcing platform hosted by the Zooniverse. Species classification accuracy of deer and black bears across both platforms is approximately 99%, and we ignored any potential misclassification within the analyses here because simulation has suggested it is not likely to influence results (Clare et al. 2019, 2020). Coyotes and wolves are classified less accurately, and each putative image of these two species (n = 11,000) was given further expert review: the few images that could not be reconciled (n < 50) were censured from analysis.

Volunteers using Snapshot Wisconsin's crowdsourcing platform can optionally classify image sequences with tags denoting the image contains foraging, vigilant, moving, resting, or interacting deer. The tag(s) apply to the image sequence rather than to specific animals within the image. As has been previously noted, such data is analogous to traditional behavioral observation methods such as scan sampling (Gallo et al. 2018). The crowdsourcing platform provides guidelines for behavioral classification largely following the process of Olson et al. (2019), and independent assessments have strongly agreed (> 90%) with the crowdsourced classification (Townsend et al. 2020, Appendix S6). Because behavioral classification of deer within images only occurs on the crowd-sourcing platform, and these analyses had smaller overall sampling parameters (27,345 camera days at 719 camera locations). We strictly focus on images of a single deer (n = 113,654 image sequences) given that it was challenging to assign specific behaviors to separate individuals in an image sequence, and because we expected that behavior would be dependent upon group size (Olson et al. 2019).

Risk Metrics

Following the definitions of Moll et al. (2017), we identified metrics of risk describing "risky places" (predator occupancy), "risky times" (near-term predator detection), "risky habitats" that either enhanced predator lethality or encounter rate (forest cover, snow depth, camera placement along linear features), and measures of productivity or land cover diversity that might capture bottom-up associations. Responses to risky places or habitats are indicative of more persistent risk cues, and responses to these factors are expected to carry larger foraging costs and result in stronger cascading effects upon vegetation (Schmitz et al. 2004, Moll et al. 2017, Creel 2018). However, prey responses to risky habitat should also depend on the degree to which prey can effectively surmount associated risk: it may be easier to avoid a static land cover feature than an ephemeral environmental feature like snow cover. Risky times capture more proximal risk (Moll et al. 2017) to which responses should have smaller foraging costs and weaker cascading effects.

We focused on three predator species: wolves, coyotes, and black bears, ignoring bobcats to reduce model complexity as they appear to be the least significant deer predator in Wisconsin (Norton 2015). We did not expect deer to respond equally to each predator, instead predicting deer would respond most strongly to wolves as the most dangerous but least frequently encountered due to low density (Lima and Bednekoff 1999), next strongly to coyotes given the threat they pose to adults but also the frequency of encounter, and least strongly to bears that essentially only kill fawns during a 'hiding' stage (Warbington et al. 2017) when the latter are not readily detected on camera.

We assembled covariates (Table S1) measured directly at the camera location by camera hosts (placement along a linear feature or not; Trail), or extracted using the 'raster' package from the cell containing the camera location (snow depth and varied vegetation greenness measures) or within a circular buffer (250 m or 5 km) surrounding the camera location (land cover covariates). These included daily snow depth data from SNODAS (Snow; National Operational Hydrologic Remote Sensing Center 2004), and measures of wooded land cover (Forest) and land cover richness (Richness) from the 2016 National Land Cover Database (Homer et al. 2018), and varied measures of the enhanced vegetation derived from 16-day MODIS reflectance measurements at 500 m resolution (product MCD43A4).

We derived interpolated daily EVI (DailyEVI) estimates using a double-logistic smooth (Beck et al. 2006). From the daily EVI estimates, we derived an estimate of annually integrated EVI (IntEVI, for 2016) by summing the daily estimates between the estimated start and end of the growing season, estimates of the daily EVI of a given pixel relative to the daily EVI of pixels within its queen's neighborhood (RelEVI), and estimates of the daily rate of change in EVI (DeltaEVI) as the first derivative of the smoothed daily EVI curve. Changes in EVI were further transformed into estimates of the daily instantaneous rate of green-up (IRG = DeltaEVI if DeltaEVI > 0, or else 0).

We assume that EVI variables broadly approximate patterns in plant productivity and food resources, and that predator and deer occurrence and behavior are associated with annual measures of productivity (IntEVI), measures of greenness on the day of observation (daily EVI) or greenness relative

to other proximal locations (RelEVI), and the rate of change in vegetation greenness (DeltaEVI). We purposely chose a limited set of spatial or temporal scales for these covariates that either reflect the native measurement scale (e.g., 500 m x 500 m MODIS pixels or *in situ* camera metadata), align with the scale of a particular response (e.g., a 5 km radius buffer approximating the home range of a bear or wolf pack or a 250 m radius buffer approximating the scale at which deer might perceive predators), or represent a plausible scale for management manipulations (e.g., forest cover within a 250 m radius buffer). We did not intend our scales to be an exhaustive search for the scale-of-effect.

Analysis of Predator Space Use

We fit zero-inflated binomial (occupancy) models (MacKenzie et al. 2002) to estimate occupancy and the daily probability of detection (occurrence) of wolves, black bears, and coyote using Hamiltonian Markov Chain Monte Carlo simulation fit with Stan using the library 'rstan' (Carpenter et al. 2017, Stan Development Team 2018). In most applications, occupancy models are applied to rigorously estimate a binary occupancy state (z_i), a probability of occupancy (ψ_i), and important predictors across sites $i = 1, 2, \ldots R$ while accounting for imperfect detection; we employed the models largely to make inference about finer-grained patterns of occurrence while accounting for the fact that some cameras fall outside the species ranges (i.e., $p|z_i$). Because our interest was inference rather than minimizing predictive error, we fit a single model for each species roughly following a degrees-of-freedom spending approach to determine a tenable model complexity for the organism in question (Giudice et al. 2012). We defined sampling occasions as single days because we expected that the patterns in occurrence and co-occurrence between predators and prey would be most meaningful at fine temporal grains (Valeix et al. 2009). Analysis of diel activity patterns, following Rowcliffe et al. (2014), indicated that each species was generally crepuscular, and so there was little benefit to realigning days to reflect different 24-hour intervals. As the occupancy of a camera viewshed is not strictly closed, we interpret the latent occupancy state as an approximation of whether a camera location falls within the home range of an individual organism (MacKenzie and Royle 2005).

We modeled bear and wolf occupancy as logit ($\psi_{i,species}$) = $\beta_{0,species}$+ $\beta_{1,species}$Forest5km$_i$+ $\beta_{2,species}$IntEVI$_i$+f_1(x$_i$,y$_i$) to test the hypothesis that these predators were more likely to occupy more productive areas (as might be expected if the system was purely bottom-up), while accounting for spatial structure in their distributions related to forest cover and broader spatial effects: f_1(x$_i$,y$_i$) denotes a spatial B-spline smooth with 20 basis functions. We modeled coyote occupancy as logit (ψ_i) = β_0+ β_1IntEVI$_i$, as coyotes were observed across a far greater geographic range of locations exhibiting less obvious spatial structure, and previous modeling efforts have found few useful predictors of coyote occupancy (e.g., Clare et al. 2016). We derived the posterior mean of the latent occupancy states, or pr($z_i|y_i$), for each species within the associated Stan program following formulas provided by MacKenzie et al. (2006, p. 124).

The detection models for these species were our primary focus as tools for making inference about finer-grained space use strategies. In particular, we were interested in the degree to which bear, wolf, and coyote occurrence might suggest their use of features that: facilitated predation (daily snow depth or forest cover); elevated the probability of encounter with deer (linear features); were associated with deer use (whether deer were recently detected) or resources targeted by deer (increased plant productivity). Given pronounced differences in the approximate effective sample size for each species (ranging from 229 observed daily wolf occurrences to 5969 observed daily coyote occurrences), we varied model complexity across species.

We specified the model for bear detection, given occupancy, as:

$$\text{logit}(p_{i,j,bear}) = \alpha_0 + \alpha_1\text{Forest}_i + \alpha_2\text{Trail}_i + \alpha_3\text{YBear}_{i,j-1} + \alpha_4\text{Richness}_i + \alpha_5\text{DailyEVI}_{i,j} + \alpha_6\text{RelEVI}_{i,j} + \alpha_7\text{IRG}_{i,j} + \alpha_8\text{Day}_j + \alpha_9\text{Day}_j^2 + \alpha_{10}\text{Deer}_{i,j} + \alpha_{11}\text{Deer}_{i,j}\text{Forest}_i + \alpha_{12}\text{Deer}_{i,j}\,\text{Trail}_i + \alpha_{13}\text{Deer}_{i,j}\text{IRG}_{i,j} + \varepsilon_i$$

For coyotes:

$$\text{logit}(p_{i,j,coyote}) = \alpha_0 + \alpha_1 Forest_i + \alpha_2 Trail_i + \alpha_3 YBear_{i,j-1} + \alpha_4 Richness_i + \alpha_5 DailyEVI_{i,j} + \alpha_6 RelEVI_{i,j} + \alpha_8 Snow_{i,j} + \alpha_9 Day_j + \alpha_{10} Day_j^2 + \alpha_{11} Deer_{i,j} + \alpha_{12} Deer_{i,j} Forest_i + \alpha_{13} Deer_{i,j} Trail_i + \alpha_{14} Deer_{i,j} Snow_{i,j} + \alpha_{15} Deer_{i,j} IRG_{i,j} + \varepsilon_i$$

For wolves:

$$\text{logit}(p_{i,j,wolf}) = \alpha_0 + \alpha_1 Forest_i + \alpha_2 Trail_i + \alpha_3 YWolf_{i,j-1} + \alpha_4 Richness_i + \alpha_5 DailyEVI_{i,j} + \alpha_6 Snow_{i,j} + \alpha_7 Deer_{i,j}$$

Above, (e.g.) $YBear_{i,j-1}$ denotes whether a bear (coyote, wolf) was detected during the previous occasion, which we include to account for potential temporal autocorrelation in detection between subsequent 24-hr periods, and because we posited that predators using space in an autoregressive fashion might be more likely to deposit cues to which deer might respond. $Deer_{i,j}$ denotes whether deer were detected on the same occasion and or the previous occasion (i.e., a count of 0, 1, or 2), which we consider a reasonable proxy for contemporaneous deer occurrence in close vicinity to the camera location. Although we initially considered estimating separate terms for deer detection on the same occasion or previous occasion, we pooled the detection totals to simplify the model structures. Term ε_i denotes a logit-normal random effect at the camera level: $\varepsilon_i \sim \text{Normal}(0, \sigma)$. Interactions between near-term deer occurrence and the instantaneous rate of green-up consider the hypothesis that bears or coyotes might more closely track deer during fawning periods, while other interactions were intended to assess the degree to which predators might track deer occurrence more or less strongly across different environmental contexts.

Deer Response Metrics

We considered four potential metrics describing deer responses to risk: a probability of daily occurrence, and conditional on occurrence, the expected number of image sequences per day containing deer (hereafter 'counts'), and the probability that deer within image sequences were either foraging or exhibiting vigilance. Although changes in occupancy are sometimes used to infer species interactions, deer were detected at 804 out of 816 camera locations, and we did not view it as likely that the species

was structurally absent from any location nor meaningful to try to model site-structured zero inflation within the detection history as a function of risk. We modeled occurrence and counts (given > 0 sequences) separately analogous to a hurdle model (Swanson et al. 2016), given some challenges finding a reasonably-fitting combined count model and reflecting that occurrence and count metrics capture slightly different elements of deer space use. The frequency of binary occurrence may relate more to how regularly move through the camera viewshed (Stewart et al. 2018), while counts better describe the intensity of viewshed use; we found that large image counts were often associated with prolonged periods of relatively sedentary behaviors within the viewshed that would be expected to potentially impact vegetation (Sabo et al. 2017). Although previous efforts have sought to directly quantify the duration of an individual's residency directly (Flagel et al. 2016), we found that this was often impossible to do without arbitrarily establishing rules for what constituted a new 'event' or 'encounter' or making tenuous assumptions about individual identify. Observations of foraging and vigilance behaviors are conditional on visitation and image counts, and so these metrics describe the allocation of different activity types within the camera viewshed.

We assumed daily deer occurrence was a Bernoulli random variable with probability $p_{i,j,deer}$, and specified the model for occurrence probability as:

$$\text{logit}\,(p_{i,j,deer}) = \alpha_0+\alpha_1 \text{YDeer}_{i,j-1}+ f_1(\text{Long}_i,\text{Lat}_i) + f_2(\text{Day}_j)+ f_3(\text{Day}_j\text{Long}_i\text{yLat}_i) + \varepsilon_i + \mathbf{\Delta_{i,j}}$$

Term $f_1(\text{Long}_i,\text{Lat}_i)$ denotes a marginal smoothing terms over space (the tensor product of two cubic splines with 5 basis function each), $f_2(\text{Day}_j)$ denotes a marginal smooth over time (a cyclical cubic regression spline function with 15 basis functions), and $f_3(\text{Day}_j\text{Long}_i\text{Lat}_i)$ was the tensor interaction of the two marginal smooths. These functions provided the model an explicit spatiotemporal structure to account for broad sources of variation not captured by the predictors, while the camera-specific random effect ε_i was used to account for unmeasurable fine-scale variation at the camera location. $\text{YDeer}_{i,j-1}$ is a first order autoregressive term describing deer observed occurrence on the previous day. The term $\mathbf{\Delta_{i,j}}$ denotes the vector product of coefficients and a set of variables including the proximal occurrence of

coyotes, bears, and wolves on the same or previous day, the estimated latent occupancy state of wolves ($WolfOccupancy_i$), the previously described environmental variables, and a set of interactions between environmental variables or environmental variables and metrics of predator occurrence or occupancy (fully listed in Table S2). The environmental variables we considered to potentially interact with predator occurrence or occupancy were Forest, Trail, DailyEVI, and Snow. We ignored the latent occupancy states of coyotes and bears because the former was nearly ubiquitous, and because the latter was spatially structured in a manner that was largely redundant with the spatial smoothing effects (Figure S1).

Again, the model specification was meant to allow us to broadly test all components of our hypotheses of interest (Figure 1) and other factors rather than minimize predictive error. Under a purely bottom-up system, deer would be expected to be more likely to occur, aggregate, and forage in areas with greater resource availability irrespective of any predation risk. Assuming the metrics of forage availability derived from EVI adequately describe forage resources, we would consequently expect deer to respond to their main effects or interactions between these variables. Assuming the non-consumptive effects in the system here reflect the classical assumptions of predator-prey systems with chase and pursue predators where both species have broad habitat domains, deer would be primarily expected to primarily respond to very proximal predation risk metrics that might be captured as the main effects of near-term occurrence of the individual predator species or might be too proximal to be measured here. If predators actually exhibited stalking behaviors and preferentially used areas with cover that facilitated hunting, we would expect deer to respond to interactions between predator use or occupancy and forest cover. Finally, if predator space use created concentrated environmental risk cues in other environmental contexts associated with greater lethality (snow), likelihood of encounter (linear features), or with different foraging benefits (greater EVI), we expected that deer would respond to interactions between these variables and occurrence or occupancy.

Our model for deer counts followed the structure of the model for deer occurrence, except that: $\alpha_1 YDeer_{i,j-1}$ was replaced with $\alpha_1 CDeer_{i,j-1}$, where $CDeer_{i,j-1}$ was the standardized count of deer at the

camera location on the previous occasion; and we estimated expected deer count using a log-link and assuming a Negative Binomial response distribution. We modeled foraging and vigilance tags as quasi-binomial counts using the logit-link following Eq. 4 such that $y_{foraging,i,j}$ ~ Quasi-binomial (Numberimages$_{i,j}$, $p_{foraging,i,j}$). Model terms followed the occurrence and count models, although here we did not include an autoregressive term. All deer models were fit in R using the library "mgcv" (Wood 2011) with the 'bam' function (Wood 2017) and goodness of fit was assessed using diagnostics and charts provided by the 'gam.check' function.

Results

Predator Space Use

As expected, bears ($\hat{\beta}$ = 0.90, 0.24 – 1.64) and wolves ($\hat{\beta}$ = 0.80, 0.20 – 1.40) were more likely to occupy locations with a high proportion of surrounding forest cover. Integrated EVI had mixed effects on predator occupancy: bears were less likely to occupy more productive sites ($\hat{\beta}$ = -0.38, -0.80 – 0.00), coyotes were more likely ($\hat{\beta}$ = 0.37, 0.03 – 0.48), and wolves exhibited little association ($\hat{\beta}$ = 0.14, -0.17 – 0.43), suggesting that as a collective, predator species were not strongly associated with plant productivity as would be expected under a bottom-up driven system. Patterns in camera-specific latent occupancy states largely reflect existing understanding of the species distributions (bears and wolves more likely to occur in northern Wisconsin, with coyotes more ubiquitous, Fig. S1), although gaps in estimates of wolf occupancy that fall within what is generally understood to be their range across much of the northern part of the state suggest that there may be many locations within their broader range that wolves do not use.

Wolves were more likely to be detected on trails ($\hat{\beta}$ = 1.10, 0.78 – 1.43) and less likely to be detected as the concurrent daily EVI increased ($\hat{\beta}$ = -0.29, -0.45 – -0.13), with weaker evidence that detection was more likely given detection during the previous occasion ($\hat{\beta}$ = 0.70, -0.30 – 1.53). There was weak indication that wolf detection was more likely if deer were recently detected ($\hat{\beta}$ = 0.15, -0.03 –

0.33), while land cover richness, forest cover, and snow depth had little effect (Table S3). Thus, there was no evidence that wolves targeted forested areas purported to either facilitate stalking or prey capture, nor that wolf occurrence was driven by bottom-up productivity at fine temporal scales, as they appeared to move less frequently during more productive times of year and/or avoid more productive locations. Rather, wolves primarily appeared to target linear features, perhaps moving serially (i.e., exhibiting repeated use of certain locations) and perhaps weakly tracking deer occurrence.

Similarly, coyotes were more likely to be detected given previous occurrence ($\hat{\beta}$ = 0.83, 0.75 – 0.91) and along linear features ($\hat{\beta}$ = 0.83, 0.61 – 1.04). To a lesser degree, their occurrence was positively associated with near-term deer occurrence ($\hat{\beta}$ = 0.26, 0.21 – 0.30). Coyotes occurrence was negatively associated with greater forest ($\hat{\beta}$ = -0.17, -0.27 – - 0.06), and they were observed slightly less frequently as snow depth, vegetation greenness, or the instantaneous rate of green-up increased: in general, they appeared to be more active during the beginning and end of the year (Table S4). No interaction with deer occurrence appeared meaningful.

Bear detection was similarly serial ($\hat{\beta}$ = 0.91, 0.73 – 1.00), and otherwise most strongly driven by vegetation greenness ($\hat{\beta}$ = 0.87, 0.78 – 0.96) and a quadratic effect of day of year (Table S4). Bears were also more likely to be detected as local-scale (250 m) forest cover increased ($\hat{\beta}$ = 0.32, 0.11 – 0.56), on-trail ($\hat{\beta}$ = 0.37, 0.02 – 0.72), and if deer were detected on the same or previous day ($\hat{\beta}$ = 0.17, 0.06 – 0.28). Other parameters of interest were estimated as having weak and uncertain effects (Table S5).

Deer Responses

Deer occurrence and counts generally exhibited strong spatiotemporal structure with marginal smooths suggesting a broad spatial gradient in decreasing occurrence and counts running from SW to NE Wisconsin (Figure S2 A and B), a pattern broadly consistent with estimates of deer abundance derived from harvest-based techniques (Townsend et al. 2020). Occurrence peaked in mid spring and again around the time of the rut, reaching a minimum during mid-to-late winter; counts peaked during the

winter, suggesting that deer generally used less space more intensively during this period (Figure S3, A and B). As expected, both deer occurrence ($\hat{\beta}$ = 0.86, 0.84 – 0.89) and deer counts ($\hat{\beta}$ = 0.17, 0.17 – 0.18) were strongly positively associated with the occurrence or count on the previous day. The likelihood of deer foraging within an image sequence tended to increase throughout the winter months from a low near the beginning of the year until a peak in later April, with a second peak in early September, when hard mast such as acorns typically begin to fall across the state (Figure S3C).

Deer occurrence, intensity of use, and behaviors were influenced by attributes of plant phenology and greenness, although effects were typically modest (Figure 3, Tables S5 – S8). Deer were more likely to occur as the concurrent daily greenness increased, particularly at locations with greater overall annual productivity, with similar but weaker patterns observed for deer counts. Similarly, the probability of deer foraging increased with higher concurrent greenness at locations that also exhibited greater integrated EVI, but the association reversed at locations with lower integrated productivity. The influence of relative EVI on deer occurrence and counts was weaker and mediated by other productivity variables, tending to have a more positive influence when the concurrent or integrated EVI was less. Thus, deer occurred more, used more heavily, and spent more time foraging in the most annually productive locations (typically deciduous forests) during periods of peak productivity, with dampened temporal patterns in less annually productive areas such as coniferous forests. Deer vigilance was more likely as the greenness on the day of observation increased, less likely as annually integrated vegetation increased, and both associations strengthened as the greenness on the day of observation relative to surrounding pixels increased. Thus, deer were most likely to exhibit vigilance during periods of peak greenness in less annually productive areas, particularly in locations exhibiting more concurrent greenness than surrounding pixels.

Deer responses to metrics of predator risk were variable. Deer detection was more likely given the near-term occurrence of all predators (Figure 4A). Other responses to bear or coyote-related metrics or interactions were typically weak (Figure 4, Tables S5-S8). There was some indication that both deer counts and deer vigilance decreased given near-term bear occurrence, particularly in locations with higher

surrounding forest cover (Fig 4 B and D). As well, deer vigilance appeared to increase given near-term coyote occurrence at off-trail locations (Figure 4D), suggesting that deer tended to increase their vigilance in response to coyote or bear occurrence in locations that these predators generally use less frequently. Although deer foraging probability was statistically significantly affected by interactions between coyote occurrence and both daily EVI and snow depth, the strength of the estimated effects was extremely weak (Figure 4C).

Wolf-related risk metrics tended to have larger influence on deer responses (Figure 4, Tables S5-S8). Near-term wolf occurrence reduced the probability of foraging, particularly on-trail (Figure 4C), while the effects of near-term occurrence upon deer counts became increasingly negative as EVI and snow depth decreased, and as forest cover increased (Figure 4 B). This combination of terms suggests that the effect of proximal wolf occurrence upon deer counts exhibits a complex "phenology" depending upon environmental attributes (Figure 5 A-C, F). At camera locations 'occupied' by wolves, there was an increased negative effect of deeper snow on deer counts and a weaker positive effect of EVI (Figure 4B): locations occupied by wolves tended to exhibit greatly reduced deer counts during the winter relative to locations unoccupied by wolves (Figure 5 D-E). Moreover, wolf occupancy flipped an otherwise positive association between deer vigilance and daily EVI, and dampened the negative influence of snow depth on foraging likelihood (Figure 4 C and D).

Discussion

We found little evidence that wolves were associated with patterns in plant productivity as might be expected under a bottom-up driven system (Ford and Goheen 2015). Instead, our results broadly suggest that white-tailed deer respond to both proximal and more persistent metrics of predation risk from predators—primarily wolves—by locally reducing their activity or foraging (given usage), factors previously implicated as drivers of vegetation properties in the region (Ferker et al. 2014, Sabo et al. 2017). This is consistent with the contention that wolf recovery has the potential to trigger a non-

consumptive tri-trophic cascade across parts of the Great Lakes (Callan et al. 2013, Flagel et al. 2016), and suggests that said recovery could have some desirable consequences for forest management.

We first acknowledge key uncertainties and limitations. Cameras sample small viewsheds, and the spatial scale of any avoidance or other behavioral responses is unclear. Moreover, because we did not directly sample vegetation itself, the extent of the vegetation responses reported by previous studies was not directly tested (Callan et al. 2013, Flagel et al. 2016). Such research is needed to assert whether a trophic cascade is occurring and quantify its strength (Ford and Goheen 2015). The nature and duration of our sampling further preclude any direct assessment of the actual foraging/fitness costs of these behaviors or broader population consequences. Limited association between the metrics of deer behavior and foraging resources considered suggest our study may have poorly quantified energetic considerations. Because the evidence here suggests that deer do respond to predation risk in ways that would lead to risk-related non-consumptive cascading effects, we are less concerned about these shortcomings, but wish to highlight that satellite-derived vegetation indices exhibit three potentially surmountable limitations related to viewpoint and spatial and biological resolution. The nadir viewpoint employed in satellite imagery can conflate canopy and sub-canopy measurements, the spatial or temporal resolution of sampling is typically coarse, and vegetation greenness is a rough proxy for more nuanced chemical and nutritional metrics that might better describe foraging value. The use of hybrid products blending spatiotemporal resolution strengths, sub-canopy measures (potentially derived from trail cameras (Liu et al. *in review*), and more nuanced proxies for foliar chemistry derived from a greater number of spectra (Wang et al. 2020) could help clarify foraging resources and trade-offs. In turn, direct measures of deer condition (*sensu* Middleton et al. 2013) and longer term population monitoring might clarify the population consequences of these trade-offs, although we note that the evidence to date suggests that non-consumptive effects exert little influence on prey populations (Sherriff et al. 2020). Such information will be critical for assessing the role of non-consumptive effects within management decision-making focused on deer, predators, and forests.Regardless, our results do not contradict previous evidence for a trophic cascade (Callan et al.

2013, Flagel et al. 2016), and consequently, our objective to evaluate hypotheses pertaining to *how* such a cascade might be occurring remains pertinent. Theory predicts that non-consumptive cascading effects should be weaker given unpredictable risk from predators hunting actively over a broad habitat domain, because prey should exhibit more reactive responses to immediate risk (i.e., immediate predator presence) that are not concentrated in any specific habitat type (Schmitz et al. 2004, Creel 2018). All predators considered here exhibited varied detection attributes—namely, associations with linear features or forest cover, and autoregressive patterns that suggest serial space use—that could be predictable from the standpoint of prey, and the strongest deer responses were associated with interactions between predator (specifically, wolf) occurrence and environmental features rather than responses to predator occurrence irrespective of environmental context. Thus, predator-prey interactions here do not appear to follow the archetypical expectations of active predator-prey systems where each player has a broad habitat domain. Instead, results were more consistent with the hypothesis that wolf foraging strategies centered around targeting linear features expected to increase travel speed and the likelihood of encountering prey or other food resources per unit time (Avgar et al. 2011, Dickie et al. 2016). Deer tended to aggregate and forage less along these features, particularly given near-term wolf occurrence, evidence for a behavioral response to more concentrated risk cues along these features. This consistent with patterns of avoidance practiced by other prey in similar systems (Dickie et al. 2020) and the idea that prey should respond most strongly to spatiotemporal spikes in risk embedded within low baseline risk (Lima and Bednefoff 1999, Creel et al. 2008). An important remaining challenge involves improving characterization or description of the linear features used or avoided by predators and prey. Here, we relied upon a binary classification provided by volunteer scientists maintaining specific camera locations because many cameras were deployed on private land and privately maintained trails are not well captured by existing GIS layers, and also because volunteer characterization of site descriptors may be more accurate when the task is simpler (Kallamansis et al. 2017).

Despite the suggestion by Flagel et al. (2016) that wolves might operate more as stalking predators in forested landscapes, there was no indication that wolves preferentially use forests in ways that might facilitate this. Although their occupancy was associated with broader surrounding forest cover, this likely reflects avoidance of or exploitation within areas in closer proximity to human settlements (Stenglein et al. 2015), given that their frequency of detection was not strongly associated with forests at a local scale. However, deer responded more strongly to near-term wolf occurrence as local forest cover increased. This may represent a response to increased wolf lethality in areas with more cover (Kunkel and Pletscher 2001, Hebblewhite et al. 2005), potentially because such cover inhibits escape (Gervasi et al. 2013) or reduces visibility and the distance at which predators are perceived and encounters initiate. Although Olson (2019) found no forest-related patterns in the location of wolf-killed deer, this is expected if prey accurately perceive risk and have the capacity to respond to it (Gaynor et al. 2019, Smith et al. 2020): the landscape of mortality should be distinct from the landscape of risk if prey mitigate risk via their responses (Moll et al. 2017). Indeed, it is possible that the landscape of mortality describes locations where prey are *least* afraid. That the landscape of deer mortality in this system is associated with snow depth (e.g., Olson 2019) may suggest that any added risk associated with deeper snow is more difficult or less worthwhile to mitigate, or that the location of mortality is otherwised decoupled from the landscape of fear. We return to these ideas shortly.

That wolves, which are considered archetypically active and generalist predators, may be capable of generating acute risk cues in certain environments suggests a need to weigh environmental context more heavily within the study of predator-prey interactions (Wirsing et al. 2020). We pose two considerations. The first is that predator hunting modes and habitat domains may require more careful definitions. Predation risk arises from some combination of the per-capita likelihood and lethality of encounter and the duration of exposure to encounter (Lima and Dill 1990, Moll et al. 2017), and we suggest that environmental variation in these factors are what defines a predator's hunting mode and habitat domain. In other words, habitat is definied by attributes that make a given hunting mode more

effective by increasing encounter likelihood or lethality. For sit and pursue or stalking predators, areas that increase the likelihood of encounter and lethality are often the same, because the strategy typically restricts predators to a narrow habitat domain and is only viable if prey have some attraction to locations where attacks are likely to be successful (Smith et al. 2019, Smith et al. 2020). This may make habitat domain easier to define. For actively hunting species, areas where the likelihood of enocounter is greater or lethality is greater may be distinct, and predators may make bifurcating selection decisions to maximize one component of habitat domain or another (Kittle et al. 2017). While wolves in our system may potentially be more dangerous to deer in forests, patterns in their occurrence suggest a hunting strategy that hinges on increasing the likelihood of encounter along finer-grained travel corridors. From this perspective, despite their broad geographic range, the habitat domain of wolves here might be fairly narrow.

The second consideration is merely that predator hunting mode, per se, may not predict behavioral responses as well as the combination of hunting mode and environmental settting. In fact, prey responses to predators with different hunting modes can either diverge or converge depending upon the surrounding environment (Wirsing et al. 2010, Wirsing et al. 2020). Although recent research in the Greater Yellowstone Ecosystem has coalesced around the idea that wolves should not generally be expected to induce strong effects upon prey given their hunting mode (Middleton et al. 2013, Kohl et al. 2018), results here more closely align with those from other forested regions in boreal Canada (e.g., Leblond et al. 2016, Dickie et al. 2020). Thus, environmental similarity may play a key role in the transferability of predator-prey interactions across systems.

The season of study may be a key type of environmental similarity to consider, as seasonality influences the physical and behavioral states of predators and prey in several ways that might impact risk perception and response. Indeed, although not the primary focus of our study, deer responses to proximal and longer-term wolf risk cues were mediated by dynamic environmental variables in ways that a) suggest deer were able to reliably discern seasonal variation in risk and their ability to control risk, and b)

implies that studies with a narrow temporal domain may not effectively describe predator-prey interactions. Deer counts were most strongly negatively impacted by near-term wolf occurrence towards the tails of the growing season. This is consistent with the forb responses that Callan et al. (2013) and Flagel et al. (2016) attribute to wolf-related effects: emergent and evergreen forbs are important parts of deer spring diets, while late-senescing forbs (primarily asters) are widely consumed by deer in autumn (McCaffery et al. 1974). Less negative associations between deer counts and wolf occurrence as the concurrent EVI increased suggest that deer perceived wolves as less dangerous during summer months, when deer are in better physical condition and when wolves may be less lethal because they often travel individually (Peterson et al. 1984). In contrast, although deer mortality risk increases in areas with or during periods of deeper snow (Post and Stenseth 1998, Norton 2015, Olson 2019), deer counts and wolf occurrence were increasingly positively associated as snow depth increased. This may reflect energetic constraints associated with different risk reduction strategies, and more broadly suggests that wolf encounter during periods of deeper snow may be a risk that is difficult to control behaviorally (Gallagher et al. 2016, Creel 2018, Wirsing et al. 2020). One reason may be that deep snow simply shrinks the habitat domain of deer more than wolves, by more severely limiting deer movement and making near-term avoidance of wolves more costly, particularly given that this is a period of nutritional stress. That deer use of locations *occupied* by wolves *decreased* with increasing snow-depth suggests a bifurcation of strategies. Some deer broadly relocate to areas relatively unused by wolves during times of deeper snow (Nelson and Mech 1984, Nelson and Mech 1991). Those that do not proactively avoid areas of wolf use during periods of low vegetation productivity and deeper snow may compensate for increased predation risk by using strategies that minimize energetic deficits (Gallagher et al. 2016), as deer using locations occupied by wolves spent more time foraging as snow depth increased relative to deer in locations unoccupied by wolves. The cost of doing so may be to become easier for wolves to both find and kill. It is also possible that deer experience the risks associated with deeper snow after the initial encounter with the predator: for example, if wolves kill deer in winter primarily by chasing them until deer encounter

pockets of deeper snow. This would make it difficult for deer to proactively respond to such risk (Creel 2018).

Ultimately, deer responses to metrics of potential predation risk here suggest some possibility that wolves may be an enacting a non-consumptive trophic cascade across parts of the Great Lakes region, but also raise new questions and pose new possibilities. Seasonal variation in deer responses suggest that our system may exhibit a distinct 'phenology of fear', where consumptive effects may have primacy during the winter in snowy areas, and non-consumptive effects may play a stronger role during other times of the year. Similar seasonal variation in encounter rate and avoidance has been observed previously between wolves and caribou (Whittington et al. 2011). Because resource-limited organisms are expected to make smaller investments in anti-predator behaviors (Bolnick and Preisser 2005, Wirsing et al. 2020), we posit such a fear phenology may be a general rule across environments with strong seasonal variation in forage availability. This raises questions about what can be inferred from studies of fear landscapes during specific seasonal periods. With fear dynamics seeing growing research interest (Palmer et al. 2017, Kohl et al. 2018), there is a need to develop overarching hypothesis to explain such patterns: seasonal variation associated with a changing energy landscape could be a fruitful starting point.

Acknowledgments

Funding was provided by NASA Ecological Forecasting grant #NNX14AC36G and Earth and Space Science Fellowship #NNX16A061H, the University of Wisconsin Cooperative Extension, and a grant from the Federal Aid in Wildlife Restoration act awarded to WDNR. This publication uses data generated via the Zooniverse.org platform, funded by in part by a grant from the Alfred P. Sloan Foundation and a Global Impact Award from Google.

References

Avgar, T., D. Keufler, and J. M. Fryxell. 2011. Linking rates of diffusion and consumption in relation to resources. Americna Naturalist 178:182-190.

Barrett, A. 2003. National Operational Hydrologic Remote Sensing Center SNOw Data Assimilation

System (SNODAS) Products at NSIDC. NSIDC Special Report 11. Boulder, CO, USA: National Snow and Ice Data Center. Digital media.

Bolnick, D. I, and E. L. Preisser. 2005. Resource competition modifies the strength of trait-mediated predator-prey interactions: a meta-analysis. Ecology 86:2771-2779.

Bradshaw, L., and D. M. Waller. 2016. Impacts of white-tailed deer on regional patterns of forest tree recruitment. Forest Ecology and Management 375:1-11.

Brown, J. S., and B. P. Kotler. 2004. Hazardous duty pay and the foraging cost of predation. Ecology Letters 7:999-1014.

Callan, R., N. P. Nibbelink, T. P. Rooney, J. E. Wiedenhoeft, and A. P. Wydeven. 2013. Recolonizing wolves trigger a trophic cascade in Wisconsin (USA). Journal of Ecology 101:837-845.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P, Li, and A. Riddell. 2017. Stan: a probabilistic programming language. Journal of Statistical Software 76:1-32.

Carstensen, M., G. D. DelGuidice, B. A. Sampson, and D. W. Kuehn. 2009. Survival, birth characteristics, and cause-specific mortality of white-tailed deer neonates. Journal of Wildlife Management 73:175-183.

Clare, J. D. J., D. W. Linden, E. M. Anderson, and D. M. MacFarland. 2016. Do the anti-predator strategies of shared prey mediate intraguild predation and mesopredator suppression? Ecology and Evolution 6:3884-3897.

Creel, S. 2018. The control of risk hypothesis: reactive vs. proactive antipredator responses and stress-mediated vs. food-mediated costs of response. Ecology Letters 21:947-956.

Creel, S., J. A. Winnie Jr., D. Christianson, and S. Liley. 2008. Time and space in general models of antipredator response: tests with wolves and elk. Animal Behavior 76:1139-1146.

Creel, S., J. A. Winnie, Jr., and D. Christianson. 2009. Glucocorticoid stress hormones and the effect of predation risk on elk reproduction. PNAS 106:12388-12393.

DelGiudice, G. D., M. R. Riggs, P. Joly, and W. Pan. 2002. Winter severity, survival, and cause-specific mortality of female white-tailed deer in north-central Minnesota. Journal of Wildlife Management 66:698-717.

DeMars, C. A., and S. Boutin. 2018. Nowhere to hide: effects of linear features on predator-prey dynamics in a large mammal system. Journal of Animal Ecology 87:274-284.

Dickie, M., R. Serrouya, R. S. McNay, and S. Boutin. 2017. Faster and farther: wolf movement on linear features and implications for hunting behavior. Journal of Applied Ecology 54:253-263.

Dickie, M., S. R. McNay, G. D. Sutherland, M. Cody, and T. Avgar. 2020. Corridors or risk? Movement along, and use of, linear features varies predictably among large mammal predator and prey species. Journal of Animal Ecology 89:623-634.

Gallagher, A. J, S. Creel, R. P. Wilson, and S. J. Cooke. 2017. Energy landscapes and the landscape of fear. Trends in Ecology and Evolution 32:88-96.

Gallo, T., M. Fidino, E. W. Lehrer, and S. Magle. 2019. Urbanization alters predator-avoidance behaviors. Journal of Animal Ecology 88:793-803.

Gaynor, K. M., J. S. Brown, A. D. Middleton, M. E. Power, and J. S. Brashares. 2019. Landscapes of fear: spatial patterns of risk perception and response. Trends in Ecology and Evolution 34: 355-368.

Gaynor, K. M., M. J. Cherry, S. L. Gilbert, M. T. Kohl, C. L. Larson, T. M. Newsome, L. R. Prugh, J. P. Suraci, J. K. Young, and J. A. Smith. 2020. An applied ecology of fear framework: linking theory to conservation practice. Animal Conservation doi:10.1111/acv.12629.

Gervasi, V., H. Sand, B. Zimmerman, J. Mattisson, P. Wabakken, and J. D. C. Linnell. 2013. Decomposing risk: landscape structure and wolf behavior generate different predation patterns in two sympatric ungulates. Ecological Applications 23:1722-1734.

Giudice, J. H., J. R. Fieberg, and M. S. Lenarz. 2012. Spending degrees of freedom in a poor economy: a case study of building a sightability model for moose in northeastern Minnesota. Journal of Wildlife Management 76:75-87.

Flagel, D. G., G. E. Belovsky, and D. E. Beyer, Jr. 2016. Natural and experimental tests of trophic cascades: gray wolves and white-tailed deer in a Great Lakes forest. Oecologia: 180:1183-1194.

Ford, A. T., and J. R. Goheen. 2015. Trophic cascades by large carnivores: a case for strong inference and mechanism. Trends in Ecology and Evolution 30:725-735.

Ferker, K., A. Sabo, and D. Waller. 2014. Long-term regional shifts in plant community composition are largely explained by local deer impact experiments. PLoS One 9:e115843.

Hebblewhite, M., E. H. Merrill, and T. L. McDonald. 2005. Spatial decomposition of predation risk using resource selection functions: an example in a wolf-elk predator-prey system. Oikos 111:101-111.

Hopcraft, J. G. C., A. R. E. Sinclair, and C. Packer. 2005. Planning for success: Serengeti lions seek prey accessibility rather than abundance. Journal of Animal Ecology 74:559-5666.

Horsely, S. B., S. L. Stout, and D. S. deCalesta. 2003. White-tailed deer impact on the vegetation dynamics of a northern hardwood forest. Ecological Applications 13:98-118.

Kittle, A. M., M. Anderson, T. Avgar, J. A. Baker, G. S. Brown, J. Hagens, E. Iwachewski, S. Moffat, A. Mosser, B. R. Patterson, D. E. B. Reid, A. R. Rodgers, J. Shuter, G. M. Street, I. D. Thompson, L. M. Vander Vennen, and J. M. Fryxell. 2017. Landscape-level wolf space use is correlated with prey abundance, ease of mobility, and the distribution of prey habitat. Ecosphere 8:e01783.

Kohl, M. T., D. R. Stahler, M. C. Metz, J. D. Forester, M. J. Kauffman, N. Varley, P. J. White, D. W. Smith, and D. R. MacNulty. 2018. Diel predator activity drives a dynamic landscape of fear. Ecological Monographs 88:638-652.

Kunkel, K. E., and D. H. Pletscher. 2001. Winter hunting patterns of wolves in and near Glacier

National Park, Montana. Journal of Wildlife Management 65:520-530.

Leblond, M., C. Dussault, J. Oullet, and M. St-Laurent. 2016. Caribou avoiding wolves faced increased predation by bears – caught between Scylla and Charybdis. Journal of Applied Ecology 53:1078-1087.

Lima, S. L., and L. M. Dill. 1990. Behavioral decisions made under the risk of predation: a review and prospectus. Canadian Journal of Zoology 68:619-640.

Lima, S. L. and P. A. Bednekoff. 1999. Temporal variation in danger drives antipredator behavior: the predation risk allocation hypothesis. American Naturalist 153:649-659.

Liu, N., M. E. Garcia, A. Singh, J. D. J. Clare, J. L. Stenglein, B. Zuckerberg, E. L. Kruger, and P. A. Townsend. *In review.* Trail camera networks provide insights into satellite phenology for ecological studies.

Locke, C. M., C. M. Anhalt-Depies, S. Frett, J. L. Stenglein, S. Cameron, V. Malleshappa, T. Peltier, B. Zuckerberg, and P. A. Townsend. 2019. Managing a large citizen science project to monitor wildlife. Wildlife Society Bulletin 43:4-10.

Luttbeg, B., J. I. Hammond, T. Brodin, and A. Sih. 2020. Predator hunting modes and predator-prey space games. Ethology 126:476-485.

Kallimansis, A. S., M. Panitsa, and P. Dimopoulos. 2017. Quality of non-expert citizen science data collected for habitat type conservation status assessment in Natura 2000 protected areas. Scientific Reports 7:8873.

Kauffman, M. J., N. Varley, D. W. Smith, D. R. Stahler, D. R. MacNulty, and M. S. Boyce. 2007. Landscape heterogeneity shapes predation in a newly restored predator-prey system. Ecology Letters 10:690-7000.

Kauffman, M. J., J. F. Brodie, and E. S. Jules. 2010. Are wolves saving Yellowstone's aspen? A landscape-level test of a behaviorally mediated trophic cascade. Ecology 91:2742-2755.

Kautz, T. M., J. L. Belant, D. E. Beyer Jr., B. K. Strickland, T. R. Petroelje, and R. Sollmann. 2019. Predator densities and white-tailed deer fawn survival. Journal of Wildlife Management:1261-1270.

MacLeod, K. J., C. J. Krebs, R. Boonstra, and M. J. Sherriff. 2018. Fear and lethality in snowshoe hares: the deadly effects of non-consumptive predation risk. Oikos 127:375-380.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248-2255.

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

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2006. Occupancy estimation and modeling. Academic Press, London, UK.

McCaffery, K. R., J. Tranetzki, and J. Peichura, Jr. 1974. Summer foods of deer in northern Wisconsin. Journal of Wildlife Management 38:215-219.

Middleton, A. D., M. J. Kauffman, D. E. McWhirter, M. D. Jimenez, R. C. Cook, J. G. Cook, S. E. Albeke, H. Sawyer, and P. J. White. 2013. Linking anti-predator behavior to prey demography reveals limited risk effects of an actively hunting large carnivore. Ecology Letters 16:1023-1030.

Miller, J. R. B., J. M. Ament, and O. J. Schmitz. 2014. Fear on the move: predator hunting mode predicts variation n prey mortality and plasticity in prey spatial response. Journal of Animal Ecology 83:214-222.

Moll, R. J., K. M. Redilla, T. Mudumba, A. B. Muneza, S. M. Gray, L. Abade, M. W. Hayward, J. J. Millspaugh and R. A. Montgomery. 2017. The many faces of fear: a synthesis of the methodological variation in characterizing predation risk. Journal of Animal Ecology 86:749-765.

Montgomery, R. A., R. J. Moll, E. Say-Sallaz, M. Valeix, and L. R. Prugh. 2019. A tendency to simplify complex systems. Biological Conservation 233:1-11.

Mumma, M. A., M. P. Gillingham, K. L. Parker, C. J. Johnson, and M. Watters. Predation risk for boreal woodland caribou in human-modified landscapes: evidence of wolf spatial responses independent of apparent competition. Biological Conservation 228:215-223.

Nelson, M. E., and L. D. Mech. 1981. Deer social organization and wolf predation in northeastern Minnesota. Wildlife Monographs 77:3-53.

Nelson, M. E., and L. D. Mech. 1984. Home-range formation and dispersal of deer in northeastern Minnesota. Journal of Mammalogy 65:567-575.

Norton, A. S. 2015. Integration of harvest and time-to-event data used to estimate demographic parameters for white-tailed deer. Dissertation, University of Wisconsin, Madison, WI.

Olson, E. R., T. R. Van Deelen, and S. J. Ventura. 2019. Variation in anti-predator behaviors of white-tailed deer in a multi-predator system. Canadian Journal of Zoology 97:1030-1041.

Olson, L.O. Physical factors affect the distributions of two Wisconsin ungulates. Thesis. University of Wisconsin, Madison, WI.

Palmer, M. S., J. Fieberg, A. Swanson, M. Kosmala, and C. Packer. 2017. A ‘dynamic’ landscape of fear: prey responses to spatiotemporal variation in predation risk across the lunar cycle. Ecology Letters 20:1364-1373.

Peers, M. J. L., Y. N. Majchrzak, E. Neilson, C. T. Lamb, A. Hamalainen, J. A. Haines, L. Garland, D. Doran-Myers, K. Broadley, R. Boonstra, and S. Boutin. 2018. Quantifying fear effects on prey demography in nature. Ecology 99:1716-1723.

Peterson, R. O., J. D. Woolington, and T. N. Bailey. 1984. Wolves of the Kenai peninsula, Alaska. Wildlife Monographs 88:3-52.

Post, E., and E. C. Stenseth. 1998. Large-scale climatic fluctuation and population dynamics of moose and white-tailed deer. Journal of Animal Ecology 67:537-543.

Preisser, E. L., J. L. Orrock, and O. J. Schmitz. 2007. Predator hunting mode and habitat domain alter nonconsumptive effects in predator-prey interactions. Ecology 88:2744-2751.

Prugh, L. R., K. J. Sivy, P. J. Mahoney, T. R. Ganz, M. A. Ditmer, M. van de Kerk, S. L. Gilbert, and R. A. Montgomery. 2019. Designing studies of predation risk for improved inference in carnivore-ungulate systems. Biological Conservation 232:194-207.

R Core Team. 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ritchie, E G., B. Elmhagen, A. S. Glen, M. Letnic, G. Ludwig, and R.A. McDonald. 2012. Ecosystem restoration with teeth: what role for predators? Trends in Ecology and Evolution 27: 265-271.

Rooney, T. P., S. M. Wiegmann, D. A. Rogers, and D. M. Waller. 2004. Biotic impoverishment and homogenization in unfragmented forest understory communities. Conservation Biology 18:787-798.

Rooney, T. P., and D. P. Anderson. 2009. Are wolf-mediated trophic cascades boosting biodiversity in the Great Lakes region? Pages 205-216 in Wydeven, A. P., T. R. Van Deelen, and E. J. Heske, editors. Recovery of gray wolves in the Great Lakes region of the United States: an endangered species success story. Springer, New York, USA.

Rowcliffe, J. M., R. Kays, B. Kranstauber, C. Carbone, and P. A. Jansen. 2014. Quantifying levels of animal activity using camera trap data. Methods in Ecology and Evolution 5:1170-1179.

Sabo, A. E., K. L. Frerker, D. M. Waller, and E. L. Kruger. 2017. Deer-mediated changes in environment compound the direct impacts of herbivory on understorey plant communities. Journal of Ecology 105:1386-1398.

Schmitz, O. J. 1998. Direct and indirect effects of predation and predation risk in old-field interaction webs. American Naturalist 151:327-342

Schmitz, O. J., V. Krivan, and O. Ovadia. 2004. Trophic cascades: the primacy of trait-mediated indirect interactions. Ecology Letters 7:153-163.

Sherriff, M. J., C. J. Krebs, and R. Boonstra. 2009. The sensitive hare: sublethal effects of predator stress on reproduction in snowshoe hares. Journal of Animal Ecology 78:1249-1258.

Sherriff, M. J., S. D. Peacor, D. Hawlena, and M. Thanker. 2020. Non-consumptive predator effects on prey population size: a dearth of evidence. Journal of Animal Ecology 89:1302-1316.

Sih, A. 2005. Predator-prey space use as an emergent outcome of a behavioral response race. Pages 240-255 in Barbosa, P. and I. Castellanos, editors. Ecology of Predator-Prey Interactions. Oxford University Press, Oxford, UK.

Smith, R. K., Pullin, A. S., Stewart, G. B., & Sutherland, W. J. (2010). Effectiveness of predator removal for enhancing bird populations. *Conservation Biology*, *24*(3), 820-829.

Smith, J. A., E. Donadio, J. N. Pauli, M. J. Sherriff, and A. D. Middleton. Habitat complexity

mediates the predator-prey space race. Ecology 100:e02724.

Smith, J. A., E. Donadio, O. R. Bidder, J. N. Pauli, M. J. Sheriff, P. L. Perrig, and A. D. Middleton. 2020. Where and when to hunt? Decomposing predation success of an ambush carnivore. Ecology: e03172.

Stan Development Team. 2018. RStan: the R interface to Stan. R package version 2.18.2, http://mc-stan.org/.

Stenglein, J. L., J. H. Gilbert, A. P. Wydeven, and T. R. Van Deelen. 2015. An individual-based model for southern Lake Superior wolves: a tool to explore the effect of human-caused mortality on a landscape of risk. Ecological Modeling 302:13-24.

Stewart, F. E. C., J. T. Fisher, A. C. Burton, and J. P. Volpe. 2018. Species occurrence data reflect the magnitude of animal movements better than the intensity of animal space use. Ecosphere:e02112.

Townsend, P. A., J. Clare, N. Liu, J. L. Stenglein, C. Anhalt-Depies, T. R. Van Deelen, N. A. Gilbert, A. Singh, K. J. Martin, and B. Zuckerberg. Integrating remote sensing within jurisdictional observation networks to improve the resolution of ecological management. doi: 10.1101/2020.06.08.140848.

Valeix, M., A. J. Loveridge, S. Chamaille-Jammes, Z. Davidson, F. Murindagomo, H. Fritz, and D. W. Macdonald. 2009. Behavioral adjustments of African herbivores to predation risk by lions: spatiotemporal variations influence habitat use. Ecology 90:23-30.

Warbington, C. H., T. R. Van Deelen, A. S. Norton, J. L. Stengelin, D. J. Storm, and K. J. Martin. 2017 Cause-specific neonatal mortality of white-tailed deer in Wisconsin, USA. Journal of Wildlife Management 81:824-833.

Wang, Z., A. Chlus, R. Geygan, Z. Ye., A. Singh, J. J. Couture, J. Cavender-Bares, E. L. Kruger, and P. A. Townsend. Foliar functional traits from imaging spectroscopy across biomes in eastern North America. New Phytologist 228:494-511.

Whittington, J., M. Hebblewhite, N. J. DeCesare, L. Neufield, M. Bradley, J. Wilmshurst, and M. Musiani. 2011. Caribou encounters with wolves increase near roads and trails: a time-to-event approach. Journal of Applied Ecology 48:1535-1542.

White, P. J., and R. A. Garrott. 2013. Predation: wolf restoration and the transition of Yellowstone elk. eds. White, P. J., R. A. Garrott, and G. E. Plumb, pages 69-93. Yellowstone's Wildlife in Transition. Harvard University Press, Cambridge, MA.

Wiedenhoeft, J. E., S. Walter, M. Gross, N Kluge, S. McNamara, G. Stauffer, J. Price-Tack, and R. Johnson. 2020. Wisconsin Gray Wolf Monitoring Report 15 April 2019 through 14 April 2020. Wisconsin Department of Natural Resources, Madison, Wisconsin.

Wirsing, A., K. E. Cameron, and M. R. Heithaus. 2010. Spatial responses to predators vary with prey escape mode. Animal Behaviour 79:531-537.

Wirsing A J., M. R. Heithaus, J. S. Brown, B. P. Kotler, and O. J. Schmitz. 2020. The context dependence of non-consumptive predator effects. Ecology Letters. doi:10.1111/ele.13614.

Wojcik, B., and J. Stenglein. 2020. White-tailed deer population status 2019. Wisconsin Department of Natural Resources. https://dnr.wi.gov/topic/WildlifeHabitat/documents/reports/wtaildeerpop.pdf

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society B 73:3-36.

Wood, S. N., Z. Li, G. Shaddick, and N. H. Augustin. 2017. Generalized additive models for gigadata: modelling the UK black smoke network daily data. Journal of the American Statistical Association 112:1199-1210.

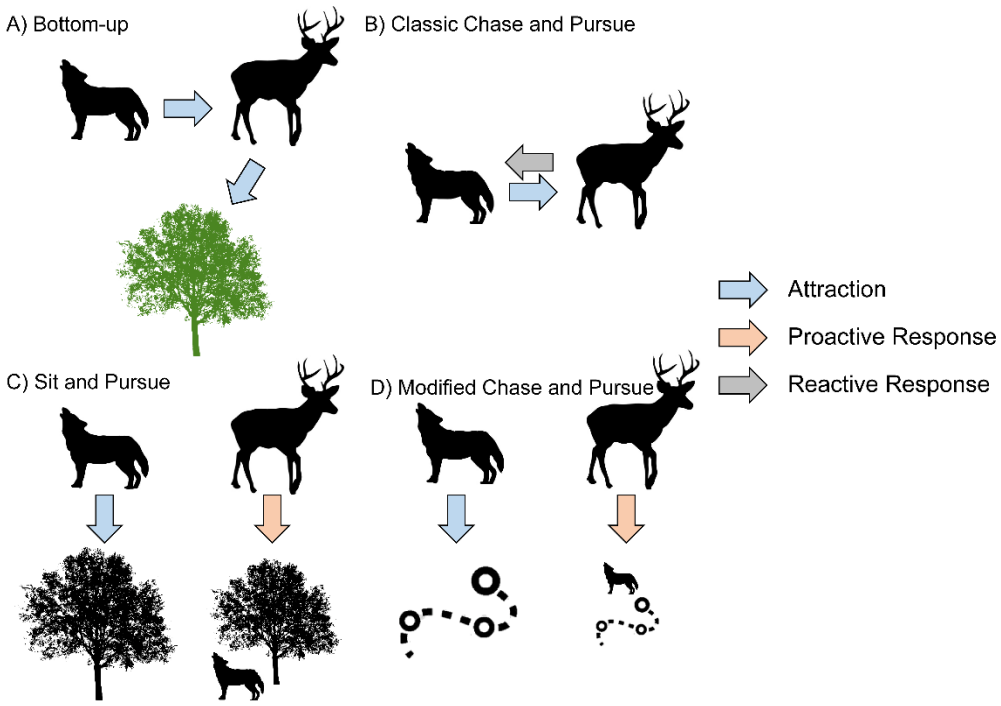

Figure 1. Overview of the motivating hypothesis for the study. In A), the system is bottom-up driven, such that deer tend to be attracted to areas with greater resource availability and wolves are attracted to greater deer accessibility, and indirectly, regions with increased foraging resources for deer. In B), the system follows classical expectations for chase and pursue (active) predators: wolves seek out deer, and deer respond to wolves proximally (i.e., after perceiving them) and relatively uniformly across environmental conditions. In C) wolves operate as sit and pursue (stalking) predators within a forested environmental context, selecting for landscape attributes that provide stalking cover and increase the lethality of their attacks. In turn, deer respond proactively to concentrated risk cues associated with wolf selection for these attributes. In D), wolves are active predators, but their selection patterns (e.g., for landscape attributes that amplify the probability of encounter) leave concentrated risk cues that trigger stronger deer responses.

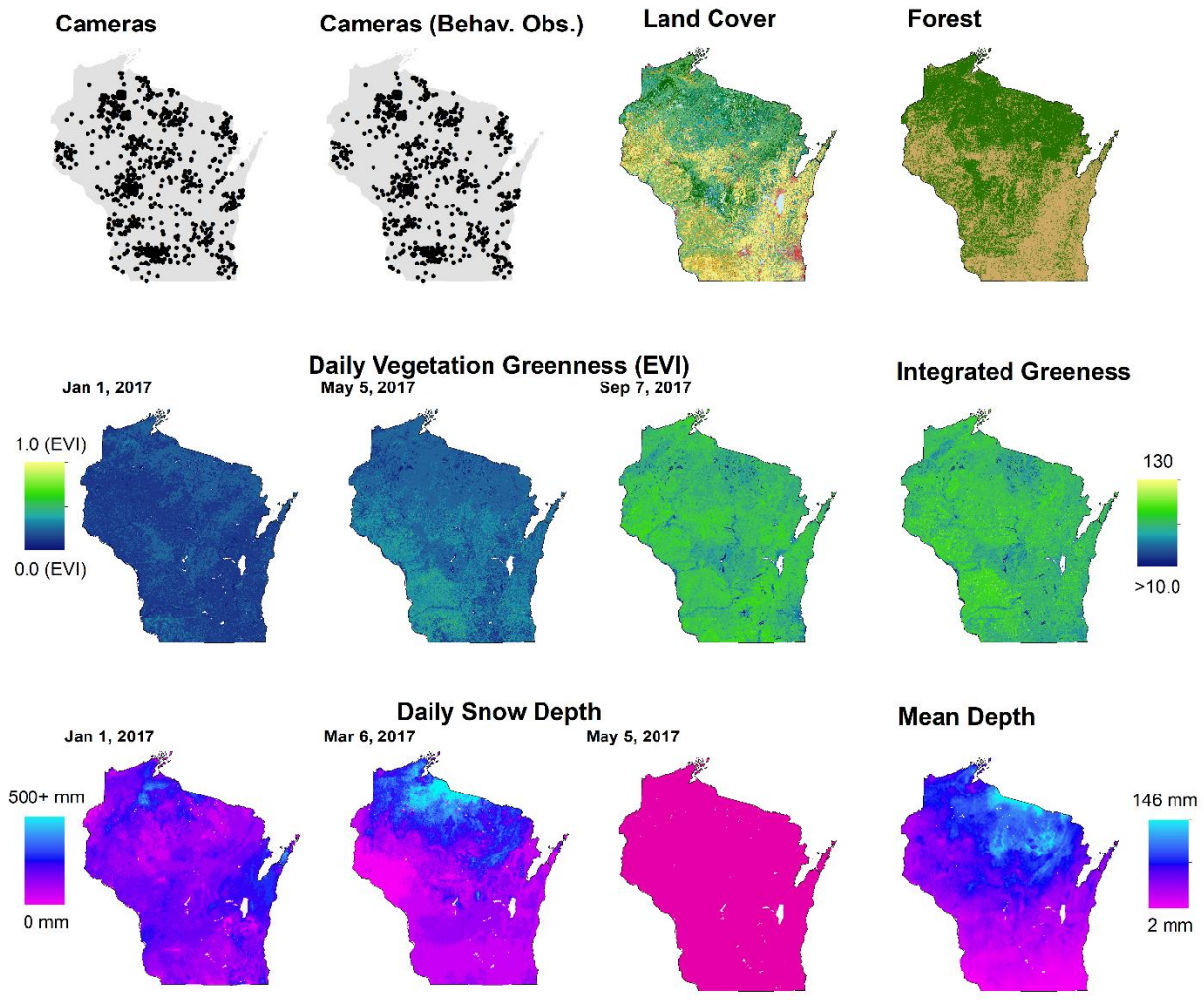

Figure 2. Camera locations and geospatial predictor variables used in the analyses. Land cover richness and forest proportion are static covariates, while the enhanced vegetation index (EVI) and snow depth are dynamic (daily) covariates from which other predictors were derived.

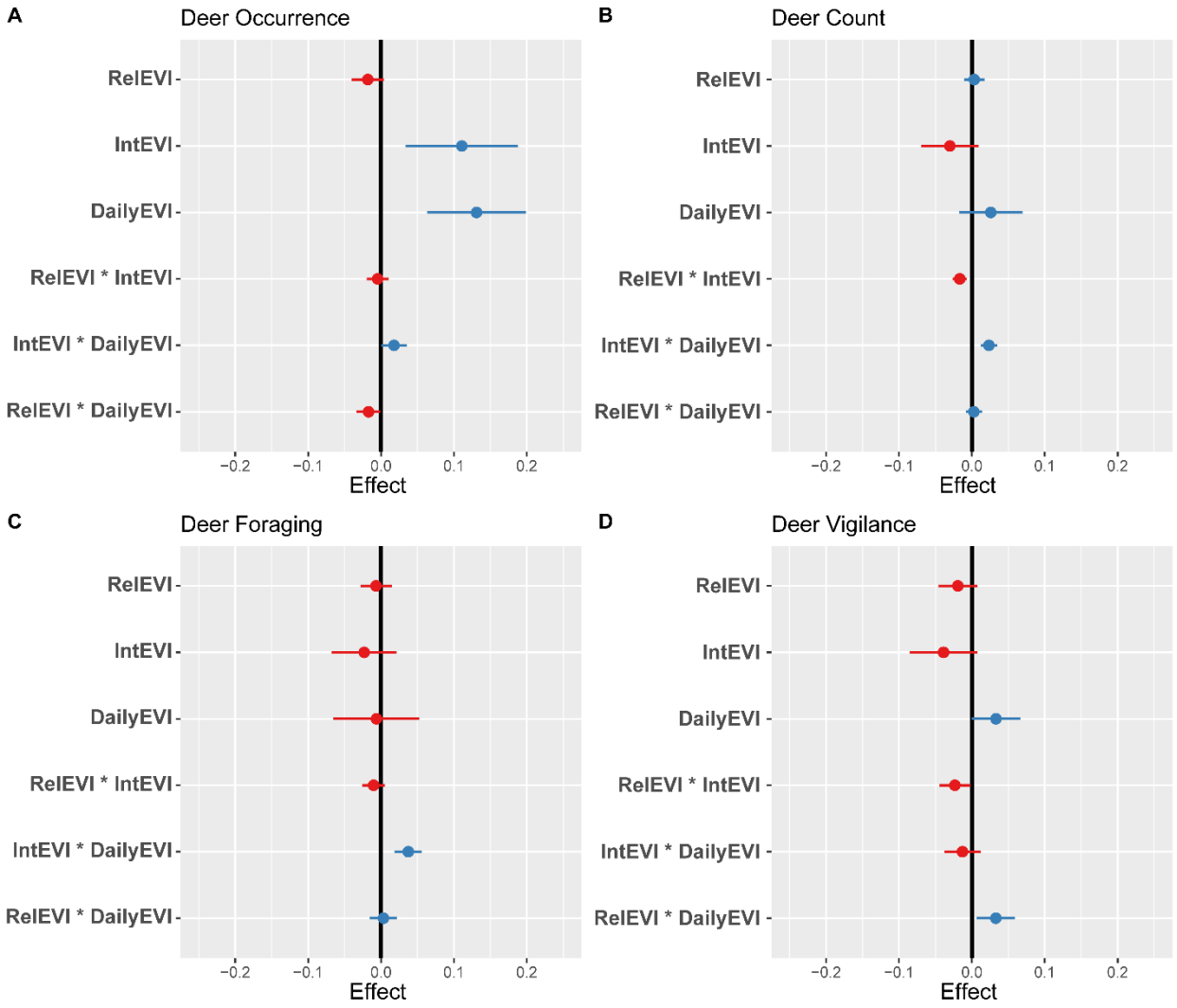

Figure 3. The effects of satellite-derived estimates of vegetation greenness derived concurrently within the pixel containing a camera location (DailyEVI), derived concurrently within a pixel containing a camera location relative to the queen's neighborhood (RelEVI), and derived based upon the integrated daily enhanced vegetation index over the previous year between the estimated start and end of the growing season (IntEVI) upon the probability of daily deer occurrence/detection at a camera location (A), the expected daily count of image sequences at a camera locations (B), the probability that a single deer in a sequence was tagged as exhibiting foraging behavior (C), and the probability that a single deer within a sequence was tagged as exhibiting vigilance (D). Blue and red colors denote positive and negative estimates of effect, respectively.

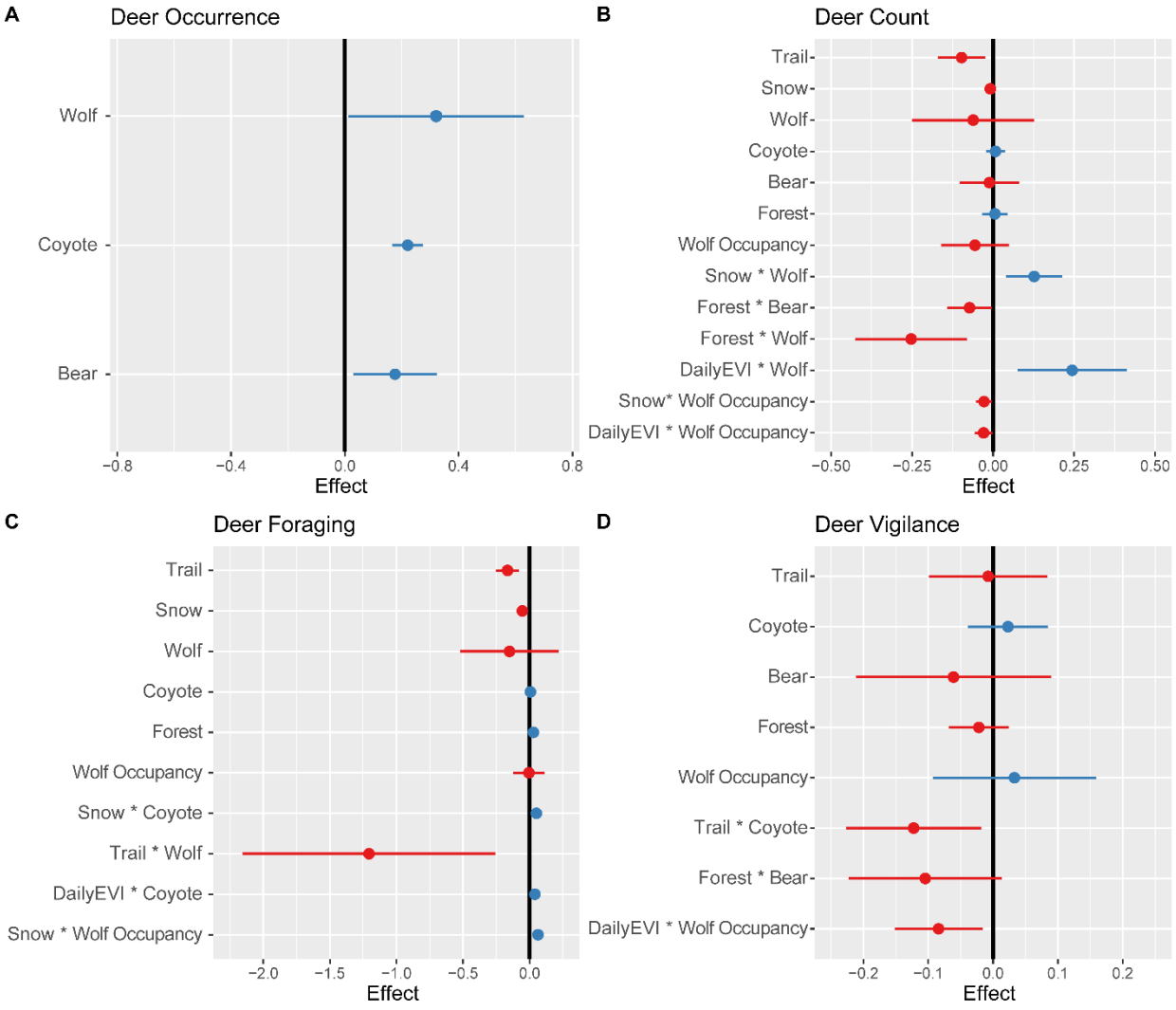

Figure 4. The effects of varied risk metrics (Table S1 and S2) on the probability of daily deer occurrence/detection at a camera location (A), the expected daily count of image sequences at a camera locations (B), the probability that a single deer in a sequence was tagged as exhibiting foraging behavior (C), and the probability that a single deer within a sequence was tagged as exhibiting vigilance (D). Blue and red colors denote positive and negative estimates of effect, respectively. For aesthetic purposes, we only depict terms statistically significant at $\alpha = 0.05$, although we present main effects if interaction terms did not overlap 0.

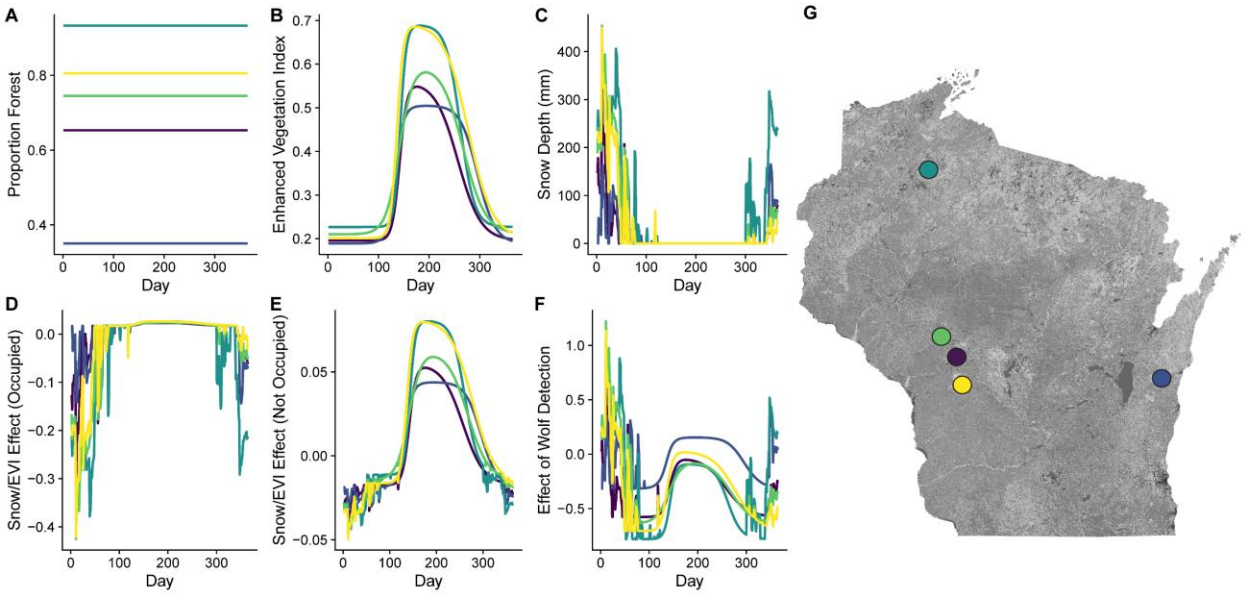

Figure 5. Interactions between wolf occupancy/detection and dynamic environmental covariates create distinct 'phenology' effects on deer counts for five sample camera locations from the data (different line colors). Environmental covariates depicted across the year include (A) forest cover, (B) daily enhanced vegetation index, and (C) snow depth. (D) In sites occupied by wolves, deer detection counts given occurrence are expected to markedly decrease during periods of low EVI and deeper snow. (E) However, this pattern is much less pronounced in sites unoccupied by wolves (note different scale). (F) Conversely, the effect of proximal wolf occurrence upon expected deer counts is most strongly positive in relatively unforested areas during periods of deep snow or in unforested areas during the peak of the growing season, and is typically most negative in forested areas near the start and end of the growing seasons. Location of sample cameras depicted in in right panel (G)—note that wolves are not necessarily present at all sample locations, and the presented effects in panels D-F ignore site-specific random effects.

Appendix S1. Supporting Figures and Tables.

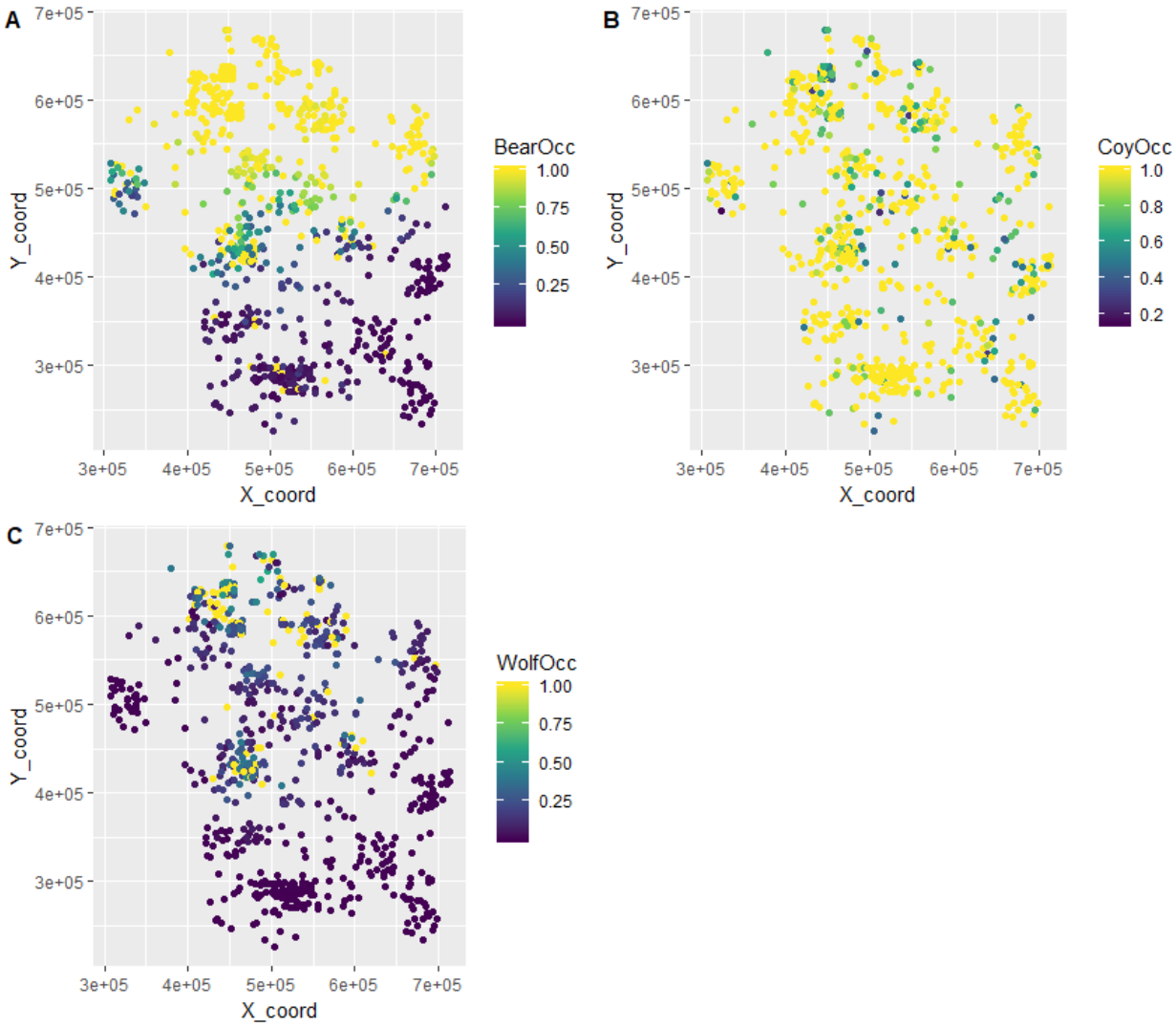

Figure S1. Finite sample estimates of realized occupancy of (A) black bears, (B) coyotes, and (C) wolves at camera locations used within the study.

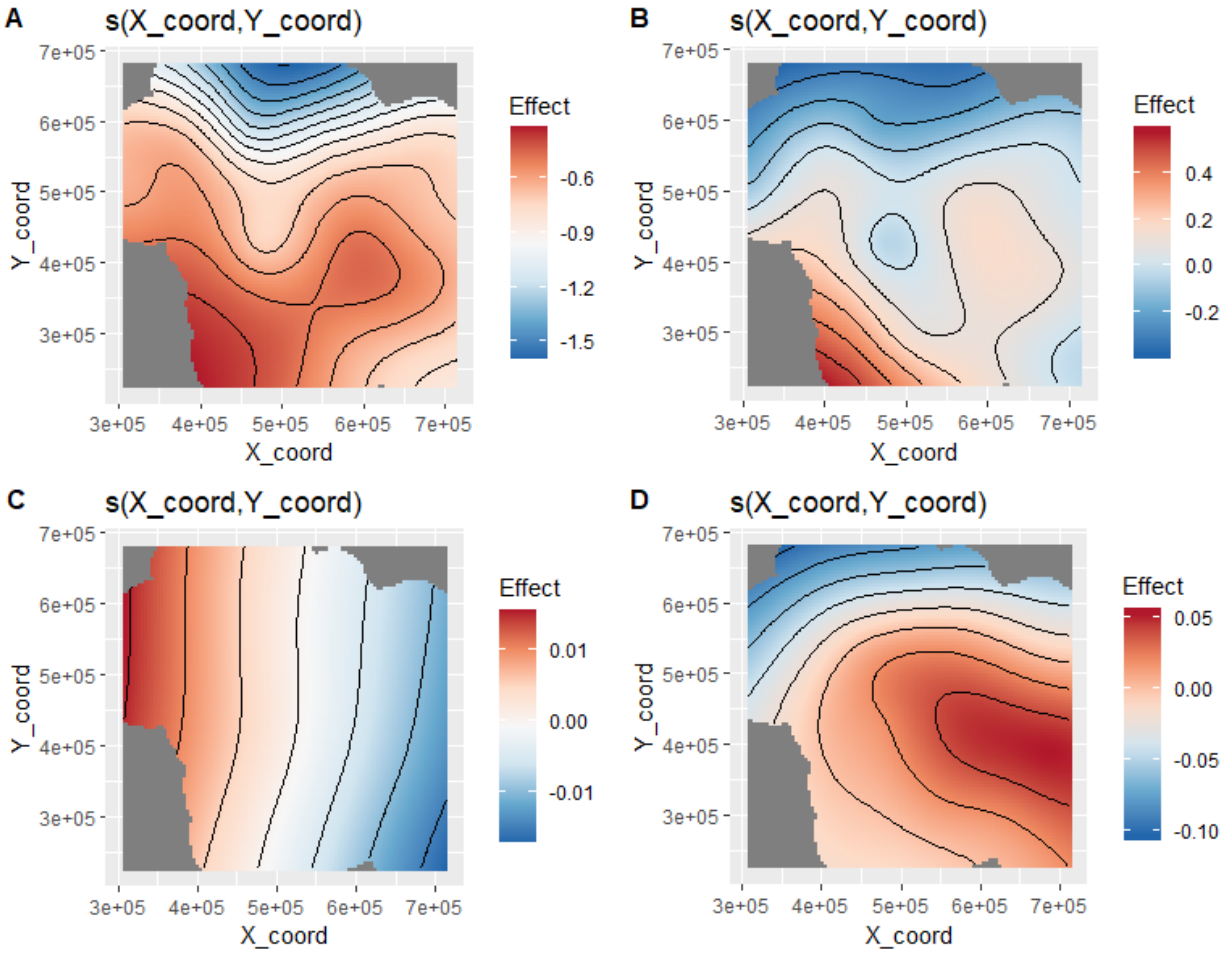

Figure S2. Marginal spatial smoother effects for deer occurrence (A), trigger counts (B), probability of foraging (C), and probability of vigilance (D).

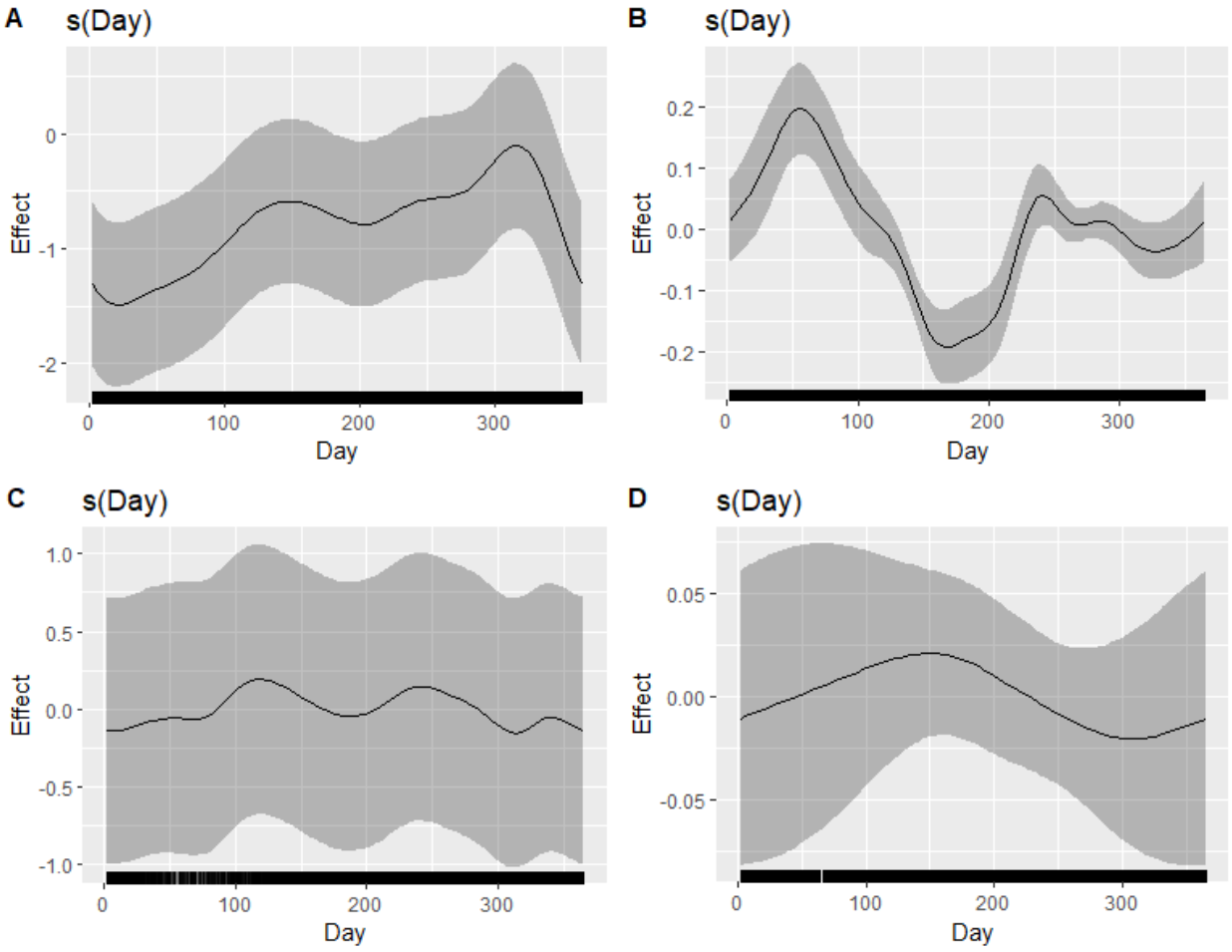

Figure S3. Marginal temporal smoother effects for deer occurrence (A), trigger counts (B), probability of foraging (C), and probability of vigilance (D).

Table S1. Variables used as predictors within the analysis, description of their derivation, and which models they were employed within.

Term	Description	Analyses used
YDeer	Binary, whether a deer was detected at a given camera location on the previous day.	Deer occurrence
YWolf	Binary, whether a deer was detected at a given camera location on the previous day.	Wolf occupancy/detection
YBear	Binary, whether a deer was detected at a given camera location on the previous day.	Bear occupancy/detection
YCoyote	Binary, whether a deer was detected at a given camera location on the previous day.	Coyote occupancy/detection
CDeer	Standardized count of deer image sequences at given camera location on the previous day.	Deer counts
Trail	Binary, whether a camera was placed along a maintained trail/road or not.	All
Snow	Daily concurrent snow depth derived from Snodas pixel containing camera after resampling to 500 m resolution.	All
Richness	Land cover richness (derived from 2016 NLCD) within 250 m radius buffer surrounding camera location.	All
DailyEVI	Daily enhanced vegetation index within MODIS pixel containing camera. Derived from MODIS reflectance (product MCD43A4), and smoothed using double-logistic function.	All
IntEVI	Summation of the daily smoothed enhanced vegetation index between the start and end of growing season (as estimated using double-logistic smooth) at pixel containing camera during the previous year (2016).	All
RelEVI	Daily enhanced vegetation index within MODIS pixel containing camera relative to the mean daily EVI of pixels within the Queen's neighborhood.	All Deer, Bear and Coyote occupancy/detection
IRG	Instantaneous rate of green-up derived from first derivative of daily EVI (either positive, or 0).	Bear, Coyote occupancy/detection
Deer	Binomial count (0, 1, 2) of deer occurrence at camera location on the same day or previous day.	Bear, Coyote, Wolf occupancy/detection
Wolf	Binomial count (0, 1, 2) of wolf occurrence at camera location on the same day or previous day.	All deer models
Coyote	Binomial count (0, 1, 2) of coyote occurrence at camera location on the same day or previous day.	All deer models
Bear	Binomial count (0, 1, 2) of bear occurrence at camera location on the same day or previous day.	All deer models
Wolf Occupancy	Estimate of wolf occupancy state at camera location.	All deer models
Forest	% Forest cover (classes deciduous forest, mixed forest, evergreen forest, wooded wetland, and shrubland) in circular buffer of 250 m (Forest) or 5km (Forest5km) radius.	Forest: All Forest5km: Bear, Coyote, Wolf occupancy/detection

Table S2. Parametric terms estimated within deer models and underlying hypothetical justification.

Term	Justification
YDeer/CDeer	Expected deer occurrence/counts to exhibit temporal autocorrelation.
Trail	Proposed indicator of risk, but also might facilitate movement and increase detection rates.
Snow	Hindrance to deer movement and indicator of periods with low food availability and greater risk, anticipated to effect occurrence, counts, and foraging/vigilance behavior.
Richness	Indicator of diversity in foraging and shelter resources.
RelEVI	Indicator of resources (during summer, food, and during winter, evergreen shelter) relative to other patches deer might conceivably use.
IntEVI	Indicator of deer condition entering the study (note, derived from previous year's phenology) and also overall annual productivity of given location.
DailyEVI	Indicator of seasonal forage availability.
Wolf	Proximal risk from wolves.
Coyote	Proximal risk from coyotes.
Bear	Proximal risk from black bears.
Forest	Previously proposed indicator of risk.
Wolf Occupancy	Longer-term risk from wolves.
Richness x IntEVI	Diversity of foraging resources (and possible phenological diversity associated with land cover diversity) may be more important in areas of lower overall productivity.
RelEVI x IntEVI	Resource availability relative to neighborhood may be less important if annually integrated productivity is relatively high.
IntEVI x DailyEVI	Expected that responses to temporal variability in resources might be more muted or amplified in areas of greater and lesser annually integrated resource availability.
RelEVI x DailyEVI	Resource availability relative to neighborhood may be more important during periods of overall resource scarcity.
Snow x Wolf Snow x Coyote Snow x Wolf Occupancy	Because snow is strong predictor of deer mortality, expected deer to respond more strongly to risk proxies as snow depth increased
Trail x Bear Trail x Wolf Trail x Coyote Trail x Wolf Occupancy	Following hypothesis in main text: if active predators create concentrated cues along linear features, deer should respond to these cues when predators are present.
Forest x Bear Forest x Wolf Forest x Coyote Forest x Wolf Occupancy	Following hypothesis in main text: if predators are more lethal in areas with more cover, deer should respond to these cues when predators are present.
DailyEVI x Bear DailyEVI x Wolf DailyEVI x Coyote DailyEVI x Wolf Occupancy	Expected resource availability/site foraging value would mediate deer reponses to proxies of predation risk

Table S3. Coefficient estimates and uncertainty intervals associated with wolf occupancy/detection model.

Variable	Process	Mean	Monte Carlo Error	LCI	UCI
Intercept	p	-5.063	0.003	-5.316	-4.818
Trail	p	1.096	0.003	0.780	1.425
Y_{Wolf}	p	0.703	0.008	-0.304	1.527
Richness	p	-0.037	0.001	-0.208	0.135
Forest	p	0.080	0.002	-0.128	0.298
Deer	p	0.149	0.002	-0.029	0.335
Snow	p	-0.040	0.001	-0.168	0.082
DailyEVI	p	-0.286	0.001	-0.445	-0.135
Intercept	ψ	-2.871	0.009	-3.812	-2.134
Forest (5km)	ψ	0.799	0.005	0.222	1.430
IntEVI	ψ	0.138	0.002	-0.170	0.425

Table S4. Coefficient estimates and uncertainty intervals associated with coyote occupancy/detection model.

Variable	Process	Mean	Monte Carlo Error	LCI	UCI
Intercept	p	-3.775	0.005	-3.904	-3.645
Forest	p	-0.159	0.003	-0.263	-0.058
Trail	p	0.821	0.008	0.565	1.035
Y_{Coyote}	p	0.827	0.001	0.752	0.906
IRG	p	-0.095	0.001	-0.152	-0.039
Richness	p	-0.101	0.003	-0.198	-0.006
Snow	p	-0.128	0.001	-0.187	-0.070
DailyEVI	p	0.073	0.001	0.008	0.143
RelEVI	p	0.009	<0.001	-0.027	0.047
Day of Year	p	0.043	0.001	0.000	0.086
Day of Year^2	p	0.256	0.001	0.192	0.320
Deer	p	0.258	0.000	0.217	0.300
Trail x Deer	p	0.005	0.001	-0.073	0.084
Snow x Deer	p	-0.021	0.001	-0.077	0.035
Forest x Deer	p	0.005	0.001	-0.033	0.041
IRG x Deer	p	0.042	<0.001	0.002	0.083
DailyEVI x Deer	p	-0.006	<0.001	-0.043	0.030
σ	p	1.083	0.003	0.999	1.183
Intercept	ψ	2.529	0.015	2.054	3.217
IntEVI	ψ	0.369	0.005	0.031	0.713

Table S5. Coefficient estimates and uncertainty intervals associated with black bear occupancy/detection model.

Variable	Process	Mean	Monte Carlo Error	LCI	UCI
Intercept	p	-4.966	0.006	-5.233	-4.698
Forest	p	0.317	0.005	0.09	0.536
Trail	p	0.379	0.007	0.035	0.727
Y_{BEAR}	p	0.913	0.002	0.735	1.091
IRG	p	-0.026	0.001	-0.100	0.043
Richness	p	-0.064	0.003	-0.220	0.081
Deer	p	0.174	0.002	0.066	0.284
DailyEVI	p	0.872	0.001	0.788	0.959
RelEVI	p	-0.087	0.001	-0.163	-0.008
Day of Year	p	-0.194	0.001	-0.318	-0.067
Day of Year^2	p	-0.709	0.003	-0.947	-0.506
Forest x Deer	p	0.010	0.002	-0.103	0.124
Trail x Deer	p	-0.053	0.002	-0.224	0.129
IRG x Deer	p	0.000	0.001	-0.056	0.053
σ	p	1.038	0.003	0.922	1.165
Intercept	ψ	1.398	0.027	0.548	2.651
Forest (5km)	ψ	1.015	0.009	0.391	1.671
IntEVI	ψ	-0.335	0.004	-0.726	0.027

Table S6. Coefficient estimates, standard error, Z score, and p-values for deer occurrence model.

Variable	Estimate	SE	Z	Pr(>|z|)
Intercept	-0.776	0.038	-20.537	< 2e-16
Y_{Deer}	0.843	0.013	65.941	< 2e-16
Trail	0.095	0.078	1.221	0.222
Snow	0.019	0.013	1.427	0.154
Richness	0.019	0.035	0.530	0.596
RelEVI	-0.018	0.011	-1.603	0.109
IntEVI	0.111	0.039	2.820	0.005
DailyEVI	0.131	0.035	3.802	0.000
Wolf	0.322	0.157	2.050	0.040
Coyote	0.221	0.027	8.049	0.000
Bear	0.176	0.075	2.354	0.019
Forest	0.004	0.042	0.085	0.933
Wolf Occupancy	-0.124	0.110	-1.125	0.261
Richness x IntEVI	0.024	0.030	0.819	0.413
RelEVI x IntEVI	-0.005	0.008	-0.618	0.537
IntEVI x DailyEVI	0.018	0.009	2.010	0.044
RelEVI x DailyEVI	-0.017	0.008	-2.059	0.039
Snow x Wolf	0.113	0.095	1.183	0.237
Snow x Coyote	-0.006	0.029	-0.216	0.829
Trail x Bear	0.025	0.097	0.255	0.799
Trail x Wolf	-0.149	0.223	-0.668	0.504
Trail x Coyote	0.012	0.044	0.280	0.779
Forest x Bear	0.022	0.057	0.397	0.691
Forest x Wolf	-0.119	0.145	-0.826	0.409
Forest x Coyote	-0.005	0.021	-0.255	0.799
DailyEVI x Bear	-0.040	0.048	-0.825	0.409
DailyEVI x Wolf	0.035	0.131	0.266	0.790
DailyEVI x Coyote	0.034	0.021	1.618	0.106
Snow x Wolf Occupancy	-0.029	0.024	-1.213	0.225
Trail x Wolf Occupancy	0.055	0.240	0.230	0.818
Forest x Wolf Occupancy	-0.077	0.095	-0.817	0.414
DailyEVI x Wolf Occupancy	0.025	0.021	1.196	0.232

Table S7. Coefficient estimates, standard error, Z score, and p-values for deer count model.

Variable	Estimate	SE	Z	Pr(>|z|)
Intercept	1.454	0.019	75.087	< 2e-16
Y_{Deer}	0.177	0.004	40.204	< 2e-16
Trail	-0.097	0.037	-2.597	0.009
Snow	-0.009	0.007	-1.190	0.234
Richness	-0.021	0.017	-1.244	0.213
RelEVI	0.003	0.007	0.449	0.654
IntEVI	-0.030	0.020	-1.499	0.134
DailyEVI	0.026	0.022	1.165	0.244
Wolf	-0.062	0.096	-0.640	0.522
Coyote	0.007	0.015	0.496	0.620
Bear	-0.011	0.047	-0.239	0.811
Forest	0.006	0.020	0.293	0.770
Wolf Occupancy	-0.056	0.053	-1.046	0.295
Richness x IntEVI	0.006	0.015	0.406	0.685
RelEVI x IntEVI	-0.017	0.005	-3.323	0.001
IntEVI x DailyEVI	0.023	0.006	3.988	0.000
RelEVI x DailyEVI	0.003	0.006	0.477	0.633
Snow x Wolf	0.127	0.045	2.838	0.005
Snow x Coyote	0.011	0.013	0.820	0.412
Trail x Bear	-0.048	0.063	-0.759	0.448
Trail x Wolf	0.152	0.164	0.928	0.353
Trail x Coyote	0.026	0.024	1.092	0.275
Forest x Bear	-0.073	0.035	-2.067	0.039
Forest x Wolf	-0.253	0.088	-2.862	0.004
Forest x Coyote	-0.019	0.011	-1.680	0.093
DailyEVI x Bear	0.008	0.031	0.252	0.801
DailyEVI x Wolf	0.244	0.086	2.828	0.005
DailyEVI x Coyote	0.017	0.012	1.413	0.158
Snow x Wolf Occupancy	-0.028	0.013	-2.185	0.029
Trail x Wolf Occupancy	-0.028	0.116	-0.241	0.810
Forest x Wolf Occupancy	0.016	0.046	0.351	0.726
DailyEVI x Wolf Occupancy	-0.029	0.014	-2.116	0.034

Table S8. Coefficient estimates, standard error, Z score, and p-values for deer foraging model.

Variable	Estimate	SE	Z	Pr(>|z|)
Intercept	-1.373	0.038	-35.903	< 2e-16
Trail	-0.166	0.044	-3.794	0.000
Snow	-0.055	0.013	-4.181	0.000
Richness	0.004	0.020	0.183	0.855
RelEVI	-0.007	0.011	-0.627	0.530
IntEVI	-0.023	0.023	-1.021	0.307
DailyEVI	-0.006	0.030	-0.210	0.834
Wolf	-0.151	0.189	-0.799	0.424
Coyote	0.006	0.020	0.311	0.756
Bear	0.015	0.047	0.316	0.752
Forest	0.029	0.022	1.330	0.184
Wolf Occupancy	-0.005	0.060	-0.082	0.934
Richness x IntEVI	0.019	0.018	1.014	0.311
RelEVI x IntEVI	-0.010	0.008	-1.293	0.196
IntEVI x DailyEVI	0.037	0.009	3.940	0.000
RelEVI x DailyEVI	0.003	0.009	0.356	0.722
Snow x Wolf	0.113	0.104	1.089	0.276
Snow x Coyote	0.051	0.023	2.265	0.024
Trail x Bear	-0.054	0.076	-0.709	0.478
Trail x Wolf	-1.206	0.484	-2.491	0.013
Trail x Coyote	-0.051	0.034	-1.498	0.134
Forest x Bear	0.021	0.041	0.501	0.616
Forest x Wolf	0.178	0.226	0.789	0.430
Forest x Coyote	-0.021	0.016	-1.309	0.190
DailyEVI x Bear	-0.014	0.038	-0.371	0.711
DailyEVI x Wolf	0.036	0.177	0.205	0.838
DailyEVI x Coyote	0.040	0.017	2.393	0.017
Snow x Wolf Occupancy	0.063	0.019	3.323	0.001
Trail x Wolf Occupancy	-0.064	0.136	-0.471	0.638
Forest x Wolf Occupancy	0.004	0.053	0.076	0.940
DailyEVI x Wolf Occupancy	-0.037	0.023	-1.578	0.115

Table S9. Coefficient estimates, standard error, Z score, and p-values for deer vigilance model.

Variable	Estimate	SE	Z	Pr(>|z|)
Intercept	-2.454	0.026	-93.732	<2e-16
Trail	-0.008	0.046	-0.165	0.869
Snow	-0.011	0.013	-0.829	0.407
Richness	-0.007	0.021	-0.336	0.737
RelEVI	-0.019	0.014	-1.409	0.159
IntEVI	-0.039	0.024	-1.639	0.101
DailyEVI	0.033	0.017	1.922	0.055
Wolf	-0.003	0.279	-0.010	0.992
Coyote	0.023	0.031	0.733	0.464
Bear	-0.061	0.077	-0.795	0.427
Forest	-0.022	0.024	-0.932	0.351
Wolf Occupancy	0.033	0.064	0.513	0.608
Richness x IntEVI	-0.003	0.019	-0.138	0.891
RelEVI x IntEVI	-0.023	0.011	-2.137	0.033
IntEVI x DailyEVI	-0.013	0.013	-1.021	0.307
RelEVI x DailyEVI	0.033	0.013	2.456	0.014
Snow x Wolf	0.053	0.188	0.283	0.777
Snow x Coyote	0.039	0.031	1.252	0.211
Trail x Bear	0.005	0.125	0.042	0.966
Trail x Wolf	-0.089	0.458	-0.194	0.846
Trail x Coyote	-0.123	0.053	-2.298	0.022
Forest x Bear	-0.105	0.060	-1.741	0.082
Forest x Wolf	0.264	0.316	0.833	0.405
Forest x Coyote	-0.005	0.025	-0.184	0.854
DailyEVI x Bear	0.059	0.066	0.896	0.370
DailyEVI x Wolf	0.089	0.257	0.345	0.730
DailyEVI x Coyote	0.007	0.025	0.283	0.777
Snow x Wolf Occupancy	-0.028	0.029	-0.989	0.323
Trail x Wolf Occupancy	-0.102	0.142	-0.722	0.471
Forest x Wolf Occupancy	0.011	0.057	0.187	0.852
DailyEVI x Wolf Occupancy	-0.084	0.035	-2.433	0.015

Chapter 4 – Snow and seasonality structure wildlife communities over the annual cycle.

Abstract

Modeling and mapping species distributions and assemblages is a ubiquitous and core need for a variety of ecological applications. Given increased appreciation that wildlife distributions—and by extension, communities—can be highly dynamic, there is growing impetus to characterize these dynamics to gain a more complete understanding of species phenology and important environmental drivers. Here, we apply a spatiotemporal multi-species occupancy model to a community of mammals and gallinaceous birds sampled across the state of Wisconsin over a calendar year. Results suggest that although the species considered are essentially non-migratory, there are pronounced shifts in occupancy dynamics associated with changes in activity and movement. Moreover, dynamic environmental variables associated with plant vigor and snow depth better explained species distributions over the year than static variables. Seasonal variation in species richness was higher in regions with greater overall richness, suggesting that seasonal space use represents a potentially important niche axis. Results suggest that seasonal variation plays a strong role in structuring communities, and that expected changes to seasonal patterns in plant and snow phenology may reshape Wisconsin's biodiversity more powerfully than other broad types of environmental change such as land cover conversion or increased urbanization.

Introduction

Understanding and delineating patterns in species distributions and broader biodiversity attributes such as alpha and beta diversity is important for both basic and applied ecology (Elith and Leathwick 2009). Species occurrence, richness, the distinctiveness of species assemblages, and the environmental correlates of each are commonly used to guide or prioritize conservation efforts (Leathwick et al. 2010, Zipkin et al. 2010). However, it is increasingly appreciated that wildlife distributions are fluid entities (Fink et al. 2010, Conn et al. 2015, Zuckerberg et al. 2016), as individual organisms face and respond to intra-annually varying environmental contexts and pressures. This has spurred calls to expand the scope of ecological studies to better describe patterns and processes over the full annual cycle (Marra et al. 2015). Empowered by tagging and sensing datastreams that provide continuous information at broad scales and growing commitment to data-sharing across distinct locations (Sullivan et al. 2009, Kays et al. 2015, Steenweg et al. 2017), ecologists are increasingly rising to the challenge and estimating seasonal variation in demographic rates, occurrence, and abundance.

Characterizing spatiotemporal variation in distributions and vital rates is often needed to prioritize actions that target different life-history stages (Rushing et al. 2017, Hardy et al. 2020) or different regions and habitat types both permanently or ephemerally (Johnston et al. 2015, Zuckerberg et al. 2016, Schuster et al. 2019). To date, the vast majority of this research (but see Hardy et al. 2020) has focused on migratory avian species for which the spatial scale of individual movement across the year often vastly exceeds individual management jurisdictions. Communities of mid-sized to large mammals pose similar challenges in that many also seasonally range beyond protected areas or across jurisdictional boundaries (Newmark 2008, Bischof et al. 2020). However, most efforts to monitor mammal distributions year-round rely on tagging technologies and focus on a relatively small number of species and individuals. Although the importance of monitoring mammal communities rather than individual species to prioritize management actions is increasingly recognized (Rich et al. 2016), most community monitoring (indeed, most mammalian monitoring in general; Marra et al. 2015) continues to rely on ‘snapshot’ sampling

efforts with limited capacity to characterize intra-annual variation in species distributions or the composition of the broader community.

Because of these limitations, conservation and management decision-making predicated upon assessments of mammalian biodiversity patterns should attempt to assess (or address) seasonal variation. A major component of this is to consider biodiversity responses to both static and dynamic elements of the environment. Across temperate latitudes, snow cover and plant vigor are key indicators of seasonal varation and species' phenologies. Although the triggering cues vary, nearly all wildlife species in such environments have evolved to coarsely time emergence from hibernation, migration, and birthing with the initiation of spring vegetation green-up (Visser et al. 1998, Inouye et al. 2000, Merkle et al. 2016, Aikens et al. 2017). At at finer grains, the the daily movements or activity of many species is linked to spatiotemporal patterns in greenness (Mueller et al. 2011, van Moorter et al. 2013). In turn, snow can provide a critical ephemeral habitat in its own right (Pauli et al. 2013) that strongly influences species behaviors and interactions (Post et al. 1999). It is possible that variation in these dynamic environmental variables—which, when considered, are typically aggregated into climatic averages—may describe species' habitat requirements as well or more effectively than land cover composition and configuration.

From an applied perspective, assessing seasonal variation in communities in relation to both dynamic and relatively static predictors may provide two major benefits. First, seasonal and annually-integrated spatial patterns in community composition provide a basis to spatially or spatiotemporally prioritize conservation and management actions (Johnston et al. 2015, Schuster et al. 2019). Secondly, understanding the relative importance of different environmental factors upon species distributions can help agencies assess and rank the importance of different broad-scale drivers of biodiversity change ranging from land cover conversion to climate change (Sultaire et al. 2016).

Understanding intra-annual variability in community composition also has more fundamental value. The study of phenology has largely focused upon cataloging shifts in the timing of specific activities and mismatches between species, as such changes are expected to eventually drive shifts in

community composition (Cohen et al. 2018, Simmonds et al. 2019). However, community composition should exhibit a regular phenology driven by species' life histories and predictable seasonal environmental variation. Interspecific variation in organismal and environmental phenology is expected play a major role in structuring communities, as time (more specifically here, seasonality) constitutes a pre-eminent niche axis (Chesson 2000, Wolkovich and Cleland 2011). Presumably, greater interspecific partitioning in the seasonal use of certain locations is associated with greater annually integrated species richness.

Here, we use data from a broad-scale monitoring program operating over a full calendar year in a strongly seasonal environment to gain insights into the intra-annual phenology of mammal communities. We use multi-species occupancy models with a spatiotemporal structure to i) estimate species occurrence probabilities over the year, ii) identify important environmental and anthropogenic predictors of both individual species and summarize broader community effects, and iii) derive and visualize community summary statistics to help inform initial biodiversity assessments.

Methods

We use data produced as part of Snapshot Wisconsin (Locke et al. 2019, Townsend et al. 2020) across the 2017 calendar year. While effort steadily increased over the course of the year as new cameras were deployed, deployment did not follow any pronounced spatial trend that might cause confounding between sampling effort and the environmental variables of interest. We focus on a community of 22 species (in a few cases, species combinations) that were regularly detected by the project (Table S1). All triggering sequences of species known to be misclassified fairly commonly (i.e., incidence > 3-5%) were subjected to a complete post-hoc review, and for other species, we assume that the classifications provided by volunteer camera hosts or via crowdsourced consensus are accurate (Clare et al. 2019, subsequent unpublished data) aside from reviewing clear spatial or temporal outliers. Prior to analysis, we thinned camera locations that were within 1 km proximity to one another, leaving a total sampling effort of 170,194 24-hr periods across 953 camera locations.

Analysis

We analyzed the data by fitting a dynamic multi-species occupancy model (Dorazio et al. 2010). We defined sites (i = 1, 2, 3…R) by overlying a grid of 3 by 3 km cells across the state; note that a site (R = 757) may include multiple camera locations c ($c_{[i]}$ ranged from 1-6, Figure 1). We defined primary periods (t = 1, 2, 3…T) as 28-d intervals (T = 13), each containing 7 secondary 4-d periods j. We restricted inference to the set of observed species (rather than augmenting the species pool with unobserved species *sensu* Dorazio et al. 2010) because the additional set of completely unobserved or unanalyzed species is poorly defined here (Guillera-Arroita et al. 2019).

Following Rushing et al. (2019), for each species s at each site during each primary period, the binary occupancy state $z_{s,i,t}$ is assumed to be a Bernoulli random variable with probability $\psi_{s,i,t}$:

$$z_{s,i,t} \sim \text{Bernoulli}\ (\psi_{s,i,t})$$

$$\text{logit}\ (\psi_{s,i,t}) = f_{t,s}(longitude_i, latitude_i) + \boldsymbol{\beta}_s \mathbf{X}_{i,t}$$

Above, $\boldsymbol{\beta}_s$ denotes a vector of species-specific coefficients describing responses to the vector of environmental covariates $\mathbf{X}_{i,t}$. Following custom for this model class, we assumed $\boldsymbol{\beta}_s \sim \text{Normal}\ (\boldsymbol{\mu}_\beta, \boldsymbol{\sigma}_\beta)$, where $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\sigma}_\beta$ were vectors describing the mean and standard deviation of species associations with specific covariates. Mean parameters were assigned prior distributions as Normal (0, 1.5), and standard deviations were assigned half-normal priors with scale 1.5. Covariates included a mix of time-varying and static predictors. Mean snow depth was calculated across the 3 km cell and 28-d primary period derived from SNODAS (Barrett 2003, $Snow_{i,t}$) and mean enhanced vegetation index across the same spatial and temporal window was derived from 16-d MODIS reflectance data (product MCD43A4) and smoothed using a double-logistic function (Beck et al. 2006, $EVI_{i,t}$)—and static predictors. Static predictors included the proportion of forested land-cover types (classes evergreen forest, deciduous forest, mixed forest, wooded wetland, and shrubland) derived from the 2016 National Land Cover Database (Homer et al. 2020, $Forest_i$), the inverse Simpson index of land cover types ($Simpson_i$), and the annual mean of night-

time light intensity ($Lights_i$, Roman et al. 2018). These covariates were selected to assess specific hypotheses related to the relative importance of vegetation structure, landscape heterogeneity, intensity of human development, and vegetation/snow seasonality upon species occurrence and community composition. Spatial analysis was performed in the R computing environment (R Core Team 2019) using libraries 'raster' and 'velox'.

Term $f_{t,s}(longitude_i, latitude_i)$ denotes a species-specific evolving spatial smoother employed to capture patterns in species occurrence not well-described by the environmental covariates (Rushing et al. 2019). The smoothing function is the dot product of K basis functions g and coefficients v (Wood et al. 2017):

$$f_{t,s}(longitude_i, latitude_i) = \sum_{k=1}^{30} g_k(longitude_i, latitude_i) v_{k,t,s}$$

Here, we employed a smoother with 30 degrees of freedom (K = 30), and radial cubic spline basis functions following Guelat and Kery (2018). Knot locations were selected using a space-filling algorithm using the library "fields". During the first primary period, we assumed coefficients $v_{k,1,s}$ ~Normal (0, σ_s^v), where σ_s^v~Half-normal (σ_v) and σ_v was assigned a prior distribution of Uniform (0, 3). This reflects previous implementations (Crainiceanu et al. 2005, Guelat and Kery 2018) save that there is a partial pooling of spline coefficient standard deviations across species. Following Rushing et al. (2019), we assumed that during subsequent primary periods (t = 2, 3, 4…13) $v_{k,t,s}$ ~Normal($v_{k,t-1,s}$, σ_s^t), again enacting partial-pooling by assuming σ_s^t~Half-normal (σ_t) and assigning σ_t a prior distribution of Uniform (0, 3).

Observations $y_{s,c,t}$ were entered at the camera-level (c) as binomial counts such that $y_{s,c,t}$ ~ Binomial($N_{c,t}$, $p_{s,c} \times z_{s,i[c],t}$), where $N_{c,t}$ denotes the number of secondary periods that camera c was active during primary period t, $i[c]$ denotes the site containing camera location c, and $p_{s,c}$ denotes the probability of detecting species s at camera c if $z_{s,i[c],t}$ = 1. We modeled variation in $p_{s,c}$ as logit($p_{s,c}$) = $\alpha_{0,s}$+ $\alpha_{1,s}$Trail$_c$+ $\alpha_{2,s}$Water$_c$+ $\varepsilon_{s,c}$. We assumed species-specific detection coefficients $\boldsymbol{\alpha_s}$ were distributed as Normal ($\boldsymbol{\mu_\alpha}$, $\boldsymbol{\sigma_\alpha}$) and employed the same hyper-priors used for occupancy coefficients. Term $\varepsilon_{s,c}$ denotes logit-normal error

for each species at each camera location distributed as Normal (0, σ_s^p). Priors and hyper-priors associated with σ_s^p follow previous descriptions.

We fit models using Markov Chain Monte Carlo simulation using Stan (Carpenter et al. 2017) via the R library 'rstan' (Stan Development team 2018), marginalizing over the latent occupancy state following MacKenzie et al. (2002). We fit 4 chains with a burn-in of 1000 iterations and a sampling duration of 1000 iterations, and assessed convergence by visually inspecting chains and using standard statistical tests.

Post fitting processing

We derived the expected occupancy state of each species for each primary period across the state-wide grid using posterior prediction, and derived annually-integrated occupancy (i.e., the probability of occupancy during any of the primary periods) for each species as $\psi_{s,i,annual} = 1 - \prod_{t=1}^{13} 1 - \psi_{s,i,t}$. We used these derivations to further derive annual and primary-specific predictions of species richness across the full network of sites. We used functional principal components analysis (Ramsay and Silverman 2005) with a B-spline basis function to partition variation in the collection of predicted richness time-series across sites using R library 'fda'.

A note on the interpretation of occupancy

Camera-trap observations arise from a combination of spatiotemporal variation in species abundance, movement, and camera perception (Burton et al. 2015). The interpretation of occupancy estimates derived from camera-trap sampling is challenging, particularly when multiple species are considered (e.g., MacKenzie and Royle 2005, Efford and Dawson 2012, Burton et al. 2015, Neilsen et al. 2018, Steenweg et al. 2018). Here, we defined sites as areal units sampled using (potentially multiple) point-detectors. It is tempting to imagine that occupancy in this context denotes whether an organism ever occurred within a given areal unit, but this is not strictly true: the statistical interpretation is whether ($z_i = 1$) or not ($z_i = 0$) there was some non-zero possibility of detecting the organism at cameras placed within the areal unit.

Thus, species exhibiting prolonged periods of inactivity may exist within a cell without 'occupying' it in the narrower statistical sense of the term. In addition, differences in movement extent and density across species, space, and time pose estimation challenges (Efford and Dawson 2012, Neilsen et al. 2018, Steenweg et al. 2018): a fast-moving and wide-ranging organism will encounter more point detectors than a relatively sedentary organism but, on a per-capita basis, will spend less time in front of any given detector. Thus, even if the two organisms occupy the same number of areal units at equal density, the estimated probability of occupancy will tend to be larger for the wider-ranging organism. Accordingly, differences in the estimated occupancy probabilities of cells across species, sites, or primary periods may reflect differences in within-cell space use: deep snow might greatly constrict the movement of certain species, making them appear to occupy fewer cells when they truly use less space within cells.

Our interest here was primarily related to understanding (and visualizing) intra-annual patterns in species activity/distributions and making inference about species that are active players within the community. We consequently use an analytical approach that facilitates these comparisons. The cost of doing so is that the state variable of interest is not consistently interpretable. Steps to ensure more consistent interpretability—such as defining a site as a camera viewshed (Efford and Dawson 2012, Steenweg et al. 2018) or using dynamic detection covariates to try to account for periods of reduced activity or other factors—make visualization challenging and otherwise mask the patterns we were interested in uncovering. Note that as the sampling duration increases, there is some evidence that the true 'asymptotic' occupancy state is better recovered (Steenweg et al. 2018), so it is likely that our annually-integrated estimates of occupancy better describe species' ranges.

Results

The fitted model exhibited adequate convergence based on visual and statistical assessments. Across the set of covariates considered, snow depth had the largest (and consistently negative, although most variable) effect upon species occupancy probability across the community (Figure 2, community-level hyper-parameters are tabulated in Table S2). More granularly, the (negative) effects of snow-depth

were less pronounced for species recognized as snow-adapted (e.g., snowshoe hare and ruffed grouse) or for canid and mustelid predator species (Figure 3, parameter effects are presented in Tables S3 – S13). Unsurprisingly, snow depth had a more pronounced effect on species undergoing periods of prolonged winter inactivity (e.g., black bear, *Ursus americanus*), although several species that remain active through winter (white-tailed deer, *Ocodeilus virginianus*) were also negatively affected.

The enhanced vegetation index also had negative effects, on average, across the community, but smaller interspecific variation in response: bears were uniquely positively associated with EVI, while most species were weakly negatively associated with site and primary-specific vegetation greenness. The proportion of forest cover had near net-zero community effect but relatively large interspecific variability (Figures 2 and 3): conforming to expectations, snowshoe hare (*Lepus americanus*) and porcupine (*Erithizon dorsatum*) were more likely to occupy site with increased increased forest cover while domestic cats *(Felis catus*) and red foxes (*Vulpes vulpes*) were more likely to occupy sites with less forest cover. The overall effects of nighttime light intensity and land cover diversity across the community were more muted and less variable. Gray fox (*Urocyon cinereoargenteus*) were uniquely positively associated with greater nighttime light intensity: all others species were negatively or not associated. On average, land cover diversity had positive effect, although all specific effect sizes were small.

Post-hoc, we were interested in understanding covariance between environmental effects (e.g., the degree to the effect of one variable was associated with the effect of another). We estimated correlation coefficients for all possible pairwise combinations of covariate-specific effects across each posterior iteration in order to derive a posterior distribution for effect correlations. The most strongly correlated effects were EVI and snow depth (Pearson r = –0.64, 95 % CRI = -0.80 – -0.44, Figure S1A), and forest cover and night-time light intensity (Pearson r = –0.39, 95 % CRI = -0.61 – -0.11, Figure S1B). Other pairwise combinations of coefficients exhibited weaker correlations (Table S14). Thus, species with strong negative responses to snow (forest cover) were likely to have less negative or positive associations with vegetation greenness (night-time lights). Overall patterns in predicted species distributions integrated

over the year reflect this, exhibiting a mix of northerly and southerly distributed species (Figure 4), although species richness appeared to be greater in the central and northern part of the state.

Functional analysis of the site-specific predicted species richness time-series suggested that the primary variation in richness phenology could be partitioned into a function describing whether species richness was consistently greater than expected (explaining roughly 80% of the variance), and a function describing the degree to which the phenology of richness was more hump-shaped or flat (explaining roughly 14% of the variance, Figure 5A). The general phenology of species richness across the complete spatial domain suggests distinct peaks in spring and autumn. This pattern did not appear to arise from broad-scale species movements (no species occupancy patterns appeared to drastically shift spatially), but rather from phenological shifts in activity and smaller-scale movement across the year: species primarily differentiated with respect to whether their occurrence peaked during the growing season or along its shoulders (Figure S2). Visualized across space, the boundary between the state's northern forests and other regions primarily demarcates shifts between more seasonally variable species richness and more consistently rich communities (Figure 5B). Greater seasonal variation in species richness (PCA function 2) was positively correlated with annually-integrated species richness (post-hoc Spearman rank correlation = 0.36, $P < 0.01$), while more consistently great richness was negatively correlated with annually integrated species richness (post-hoc Spearman $\rho = -0.44$, $P < 0.01$), suggesting areas with more pronounced differences between growing season and winter species richness tended to be used by a greater number of species over the course of a year, although the region of peak annually-integrated richness fell along a boundary between the two regions.

Discussion

We demonstrate here the capacity of modern ecological data-collection efforts empowered by community scientists and sensors to capture the phenology of community dynamics. The capacity to predict distributional variation over a broad swath of species across the year greatly expands the scope of conservation and management capacity to influence biodiversity outcomes. The environmental effects we

considered here were broadly intended to capture the presence of vertical structure (Forest), exposure to human activity (Lights), land cover diversity (Simpson), and both snow and vegetation phenology. Each factor has seen pronounced global change in recent years, as humans clear forests (Hansen et al. 2010), expand cities and increase artificial light (Longcore and Rich 2004), simplify landscapes via land use change (Homer et al. 2020), and alter plant and snow phenology (Cleland et al. 2007, Thompson et al. in press). Each of these environmental factors poses particular challenges for conservation and management organizations because they often have limited capacity to easily manipulate, for example, climate or land use (Townsend et al. 2020). Our results—namely, the primacy of seasonal predictors, and the limited correlation between patterns in species richness across different temporal extents—suggest that seasonal environmental factors and variation in species phenology may play important roles in shaping community composition, and that both seasonal effects and the temporal sampling frame may deserve greater consideration from practitioners seeking to make inference about wildlife communities.

Snow depth was the strongest predictor of species occurrence across the annual cycle, and the combination of snow depth and EVI appears to provide the most useful bivariate axis for describing species environmental associations. This is remarkable for several reasons. First, both variables exhibit relatively strong spatiotemporal autocorrelation that is not expected to result in broad-scale movement dynamics (Mueller et al. 2011, Van Moorter et al. 2013), and because the coarse spatiotemporal smoothing we employed likely explains variance that might otherwise be captured by synoptic patterns in snow and vegetation greenness—our smoothing functions almost certainly capture variation that might be more proximally explained by longer-term patterns in plant productivity and winter severity. Second, our study was not restricted to species recognized as requiring or being sensitive to seasonal environmental variation such as emphemeral snow cover (Pauli et al. 2013, Sultaire et al. 2016), but instead largely describes a set of generalist species that exist across a far broader range of environmental conditions than considered here. Camera traps primarily sample animal movement integrated over the temporal extent of a study (Burton et al. 2015). While snow certainly provides a distinct habitat domain for winter specialists

that is poorly described by surface sampling, our results also broadly suggest that snow plays an important role in constricting the usable space for many species by limiting movement: in essence, the surface species pool may seasonally cotntract because there is less physical space to partition.

We expect the effects of other environmental variables are best interpreted through this lens. Wildlife move less frequently and compress their activity into a smaller diel window in areas of greater human development (Tucker et al. 2018, Gaynor et al. 2019), which likely explains some portion of reduced richness estimates in areas with more artificial light. The effect of forest cover, which presumably captures elements of the vertical/structural niche-space, may be less important for the community of mammals considered here that are largely ground-bound. Instead, within-cell structural diversity may have been better captured by land cover diversity, a well-established positive driver of species richness that is believed to describe many potential niche axes (Stein et al. 2014). One important axis that may be captured by land cover diversity is phenological diversity: greater spatial heterogeneity in seasonal resource variation may allow individual consumers to accumulate more energy by moving less, and may also allow consumers with different specialties to better partition space over time (Armstrong et al. 2016). Notably, while the effect of land cover diversity on occupancy probability was positive, on average, specific effect sizes were generally small, suggesting that no species strongly depended upon or gained significant competitive advantage from increases in diversity, and that effects upon richness may have arisen from the increased likelihood of coexistence in these areas (Chesson 2000)

In contrast, on average, species occurrence was negatively associated with plant greenness, a widely considered proxy for resource availability and ecosystem productivity, and given that nearly all individual species exhibited negative associations, there was little evidence that some subset of ‘competitive’ species were better at exploiting available resources (sensu Mittelbach et al. 2001). Moreover, on average, temporal patterns in species richness dipped during periods of peak productivity. Here, consideration of the response variable (i.e., as a function of movement) and interspecific heterogeneity is warranted. Many species considered exhibited dips in occupancy probability during

peak-summer, which may both reflect a need to move less during periods of peak resource availability, or life history characteristics that constrain movement (i.e., offspring with limited mobility or constraint to a central foraging location, Sibly and Brown 2009).

Because there are relatively fewer species that exhibit this phenology, it is interesting that annually-integrated species richness tended to be associated with a more pronounced summer peak in primary-specific richness, and that there was no clear indication. We believe a primary reason is that the densities of species primarily found in northern Wisconsin tend to be smaller and their annual space use larger than those in southern Wisconsin, on average. This is generally consistent with the idea that seasonal space use represents an important niche axis that might be partitioned among a smaller set of more abundant species, or a larger set of less abundant species (Chesson 2000, Hurlburt 2004).

At a more synoptic grain than directly considered here, patterns in both annually-integrated richness and in seasonal richness variation appear to negatively align with patterns in anthropogenic modification to ecosystems via land use change and development (i.e., the human footprint index). It is tempting to ascribe community differences to human-driven reductions in the niche space (here, seasonal) available to partition (Tucker et al. 2018, Gaynor et al. 2019, Manlick and Pauli 2020). However, there are several confounding factors that complicate interpretation. Most broadly, many species distributions exhibit legacy effects associated with patterns in historical extirpation (or reintroduction), and it is unclear whether variation in space/time/dietary use results from human-modified systems or from compositional differences in the regional species pool: our results suggest that after controlling for broader spatial structure, there may be some benefit to moderate levels of land cover conversion. More specific to seasonality, the degree to which humans modulate resource phenology via land-use change, creating urban heat islands, or providing annually available resource subsidies vs. simply settling in more moderate systems is similarly unclear. These issues are symptoms of a broader challenge that pervades most attempts to characterize species distributions, niches, and niche axes. Studies with a narrow or poorly defined spatial or temporal domain will have difficulty distinguishing between equally viable

competing mechanisms. Although the temporal extent of our own study limits our own inferences, we believe our focus on more granular dynamics represents a step in the right direction.

Acknowledgments

Funding was provided by NASA Ecological Forecasting grant #NNX14AC36G and Earth and Space Science Fellowship #NNX16A061H, the University of Wisconsin Cooperative Extension, and a grant from the Federal Aid in Wildlife Restoration act awarded to WDNR. This publication uses data generated via the Zooniverse.org platform, funded by in part by a grant from the Alfred P. Sloan Foundation and a Global Impact Award from Google.

References

Aikens, E. O., M. J. Kauffman, J. A. Merkle, S. P. H. Dwinnell, G. L. Fralick, and K. L. Monteith. 2017. The greenscape shapes surfing of resource waves in a large migratory herbivore. Ecology Letters 20:741-750.

Armstrong, J. B., G. Takimoto, D. E. Schindler, M. M. Hayes, and M. J. Kauffman. 2016. Resource waves: phenological diversity enhances foraging opportunities for mobile consumers. Ecology 97:1099-1112.

Beck, P. S. A., C. Atzberger, K. A. Hodga, B. Johansen, and A. Skidmore. 2006. Improved monitoring of vegetation dynamics at very high latitudes: a new method using MODIS NDVI. Remote Sensing of Environment 100:321-334.

Bischof, R., C. Milleret, P. Dupont, J. C. Chipperfield, M. Tourani, A. Ordiz, P. de Valpine, D. Turek, J. A. Royle, O. Gimenez, O. Flagstand, M. Akesson, L. Svensson, H. Broseth, and J. Kindberg. 2020. Estimating and forecasting spatial population dynamics of apex predators using transnational genetic monitoring. Procedeedings of the National Academy of Sciences doi:10.1073/pnas.2011383117.

Burton, A. C., E. Neilson, D. Moreira, A. Ladle, R. Steenweg, J. T. Fisher, E. Bayne, and S. Boutin. 2015. Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. Journal of Applied Ecology 52:675-685.

Cohen, J., M. Lajeunesse, and J. Rohr. 2018. A global synthesis of animal phenological responses to climate change. Nature Climate Change 8:224-228.

Chesson, P. 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecology and Systematics 31:343-366.

Cleland, E. E., I. Chuine, A. Menzel, H. A. Mooney, and M. D. Schwartz. 2007. Shifting plant phenology in response to global change. Trends in Ecology and Evolution 22:357-365.

Conn, P. B., D. S. Johnson, J. M. Ver Hoef, M. B. Hooten, J. M. London, and P. L. Boveng. Using spatiotemporal statistical models to estimate animal abundance and infer ecological dynamics from survey counts. Ecological Monographs 85:235-252.

Crainceanu, C., D. Ruppert, and M. P. Wand. 2005. Bayesian analysis for penalized spline regression using WinBUGS. Journal of Statistical Software 14:1-24.

Dorazio, R. M., M. Kery, J. A. Royle, and M. Plattner. 2010. Models for inference in dynamic metacommunity systems. Ecology 91:2466-2475.

Efford, M. G., and D. K. Dawson. 2012. Occupancy in continuous habitat. Ecosphere 3:1-15.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution and Systematics 40:677-697.

Fink, D., W. M. Hochachka, B. Zuckerber, D. W. Winkler, B. Shaby, M. Arthur Munson, G. Hooker, M. Riedewald, D. Sheldon, and S. Kelling. 2010. Spatiotemporal exploratory models for broad-scale survey data. Ecological Applications 20:2131-2147.

Gaynor, K. M., C. E. Hojnowski, N. H. Carter, and J. S. Brashares. 2019. The influence of human disturbance on wildlife nocturnality. Science 360:1232-1235.

Guillera-Arroita, G., M. Kery, and J. J. Lahoz-Monfort. 2019. Inferring species richness using multispecies occupancy modeling: estimation performance and interpretation. Ecology and Evolution 9:780-792.

Guélat, J. and Kéry, M. 2018. Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods in Ecology and Evolution 9:1614-1625.

Hansen, M. C., S. V. Stehman, and P. V. Potapov. 2010. Quantification of global gross forest cover loss. Proceedings of the National Academy of Sciences 107:8650-8655.

Hardy, M.A., M.S. Broadway, C.D. Pollentier, J.D. Riddle, S.D. Hull, and B. Zuckerberg. 2020. Responses to land cover and grassland management vary across life-history stages for a grassland specialist. Ecology and Evolution DOI:10.1002/ece3.6805.

Homer, C., J. Dewitz, S. Jin, G. Xian, C. Costello, P. Danielson, L. Gass, M. Funk, J. Wickham, S. Stehman, R. Auch, and K. Ritters. 2020. Conterminous United States land cover change patterns 2001-2016 from the 2016 National Land Cover Database. ISPRS Journal of Photogrammetry and Remote Sensing 162:184-199.

Hurlbert, A. H. 2004. Species-energy relationships and habitat complexity in bird communities. Ecology Letters 7:714-720.

Inouye, D. W., B. Barr, K. B. Armitage, and B. D. Inouye. 2000. Climate change is affecting altitudinal migrants and hibernating species. Procededings of the National Academy of Sciences 97:1630-1633.

Johnston, A., D. Fink, M. D. Reynolds, W. M. Hochachka, B. L. Sullivan , N. E. Bruns, E. Hallstein, M. S. Merrifield, S. Matsumoto, and S. Kelling. Abundance models improve spatial and temporal prioritization of conservation resources. Ecological Applications 25:1749-1756.

Kays, R., M. C. Crofoot, W. Jetz, and M. Wikelski. 2015. Terrestrial animal tracking as an eye on life and planet. Science 348:aaa2478.

Leathwick, J. R., A. Moilanen, S. Ferrier, and K. Julian. 2010. Complentarity-based conservation prioritization using a community classification, and its application to riverine ecosystems. Biological Conservation 143:984-991.

Locke, C. M., C. M. Anhalt-Depies, S. Frett, J. L. Stenglein, S. Cameron, V. Malleshappa, T. Peltier, B. Zuckerberg, and P. A. Townsend. 2019. Managing a large citizen science project to monitor wildlife. Wildlife Society Bulletin 43:4-10.

Longcore, T., and C. Rich. 2004. Ecological light pollution. Frontiers in Ecology and the Environment 2:191-198.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248-2255.

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

Manlick, P. J., and J. N. Pauli. 2020. Human disturbance increases trophic niche overlap in terrestrial carnivore communities. Proceedings of the National Academy of Sciences 117:26842-26848.

Marra, P. P., E. B. Cohen, S. R. Loss, J. E. Rutter, and C. M. Tonra. 2015. A call for full annual cycle research in animal ecology. Biology Letters 11:20150552.

Merkle, J. A., K. L. Monteith, E. O. Aikens, M. M. Hayes, K. R. Hersey, A. D. Middleton, B. A. Oates, H. Sawyer, B. M. Scurlock, and M. J. Kauffman. 2016. Large herbivores surf waves of green-up during spring. Proceedings of the Royal Society B 283:20160456.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381-2396.

Mueller, T., K. A. Olson, G. Dressler, P. Leimgruber, T. K. Fuller, C. Nicolson, A. J. Novaro, M. J. Bolgeri, D. Wattles, S. DeStefano, J. M. Calabrese, and W. F. Fagan. 2011. How landscape dynamics link individual-to population-level movement patterns: a multispecies comparison of ungulate recolocation data. Global Ecology and Biogeography 20:683-694.

Neilson, E. W., T. Avgar, A. C. Burton, K. Broadley, and S. Boutin. 2018. Animal movement affects interpretation of occupancy models from camera-trap surveys of unmarked animals. Ecosphere 9:e02092.

Newmark, W. D. 2008. Isolation of African protected areas. Frontiers in Ecology and the Environment 6:321-328.

Pauli, J. N., B. Zuckerberg, J. P. Whiteman, and W. Porter. 2013. The subnivium: a deteriorating seasonal refugium. Frontiers in Ecology and the Environment 11:260-267.

Post, E., R. O. Peterson, N. C. Stenseth, and B. E. McLaren. 1999. Ecosystem consequences of wolf behavioural response to climate. Nature 401:905-907.

Ramsay, J., B. W. Silverman. 2005. Functional data analysis. Springer, New York, USA.

Rich, L. N., D. A. W. Miller, H. S. Robinson, J. W. McNutt, and M. J. Kelly. 2016. Using camera trapping and hierarchical occupancy modelling to evaluate the spatial ecology of an African mammal community. Journal of Applied Ecology 53:1225-1235.

Roman, M. O. Z. Wang, Q. Sun, V. Kalb, S. D. Miller, A. Molthan, L. Schultz, J. Bell, E. C. Stokes, B. Pandey, K. C. Seto, D. Hall, T. Oda, R. E. Wolf, G. Lin, N. Golpayegani, S. Devadiga, C. Davidson, and E. J. Masuoka. 2018. NASA's Black Marble nighttime lights product suite. Remote Sensing of Environment 210:113-143.

Rushing, C. S., J. A. Hostetler, T. S. Sillett, P. P. Marra, J. A. Rotenberg, and T. B. Ryder. 2017. Spatial and temporal drivers of avian population dynamics across the annual cycle. Ecology 98:2837-2850.

Rushing, C. S., J. A. Royle, D. J. Ziolkowski, and K. L. Pardieck. 2019. Modeling spatially and temporally complex range dynamics when detection is imperfect. Scientific Reports 9:12805.

Schuster, R., S. Wilson, A. D. Rodewald, P. Arcese, D. Fink, T. Auer, and J. R. Bennett. 2019. Optimizing the conservation of migratory species over their full annual cycle. Nature Communications 10:1754.

Sibly, R. M., and J. H. Brown. 2009. Mammal reproductive strategies driven by offspring mortality-size relationships. American Naturalist 173:E185-E199.

Simmonds, E. G., E. F. Cole, B. C. Sheldon, and T. Coulson. 2020. Phenological asynchrony: a ticking time-bomb for seemingly stable populations? Ecology Letters 23:1766-1775.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: a citizen-based observation network in the biological sciences. Biological Conservation 142:2282-2292.

Sultaire, S. M., J. N. Pauli, K. J. Martin, M. W. Meyer, M. Notaro, and B. Zuckerberg. 2016. Climate change surpasses land-use change in the contracting range boundary of a winter-adapted mammal. Proceedings of the Royal Society B 283:20153104.

Steenweg, R., M. Hebblewhite, R. Kays, J. Ahumada, J. T. Fisher, C. Burton, S. E. Townsend, C.

Carbone, J. M. Rowcliffe, J. Whittington, J. Brodie, J. A. Royle, A. Switalski, A. P. Clevenger, N. Heim, and L. N. Rich. 2017. Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors. Frontiers in Ecology and the Environment 15:26-34.

Steenweg, R., M. Hebblewhite, J. Whittington, P. Lukacs, and K. McKelvey. 2018. Sampling scales define occupancy and underlying occupancy-abundance relationships in animals. Ecology 99:172-183.

Stein, A., K. Gerstner, and H. Kreft. 2014. Environmental heterogeneity as a universal driver of species richness acrss taxa, biomes, and spatial scales. Ecology Letters 17:866-880.

Townsend, P. A., J. Clare, N. Liu, J. L. Stenglein, C. Anhalt-Depies, T. R. Van Deelen, N. A. Gilbert, A. Singh, K. J. Martin, and B. Zuckerberg. Integrating remote sensing within jurisdictional observation networks to improve the resolution of ecological management. doi: 10.1101/2020.06.08.140848.

Tucker, M. A., J. Bohning-Gaese, W. F. Fagan…et al. 2018. Moving in the Anthropocene: global reductions in terrestrial mammalian movements. Science 359:466-469.

Van Moorter, B., N. Bunnefeld, M. Panzacchi, C. M. Rolandsen, E. J. Solberg, and B. Saether. 2013. Understanding scales of movement: animals ride waves and ripples of environmental change. Journal of Animal Ecology 82:770-780.

Visser, M. E., A. J. van Noordwijk, J. M. Tinbergen, and C. M. Lessells. 1998. Warmer springs lead to mistimed reproduction in great tits (*Parus major*). Proceedings of the Royal Society, B 265:1867-1870.

Wolkovich, E. M., and E. E. Cleland. 2010. The phenology of plant invasions: a community ecology perspective. Frontiers in Ecology and the Environment 9:287-294.

Wood, S. N. 2017. Generalized additive models: an introduction with R. CRC Press, London, UK.

Zipkin, E. R., J. A. Royle, D. K. Dawson, and S. Bates. 2010. Multi-species occurrence models to evaluate the effects of conservation and management actions. Biological Conservation 143:479-484.

Zuckberberg, B., D. Fink, F. A. La Sorte, W. M. Hochachka, and S. Kelling. 2016. Novel seasonal land cover associations for eastern North American forest birds identified through dynamic species distribution modelling. Diversity and Distributions 22:717-730.

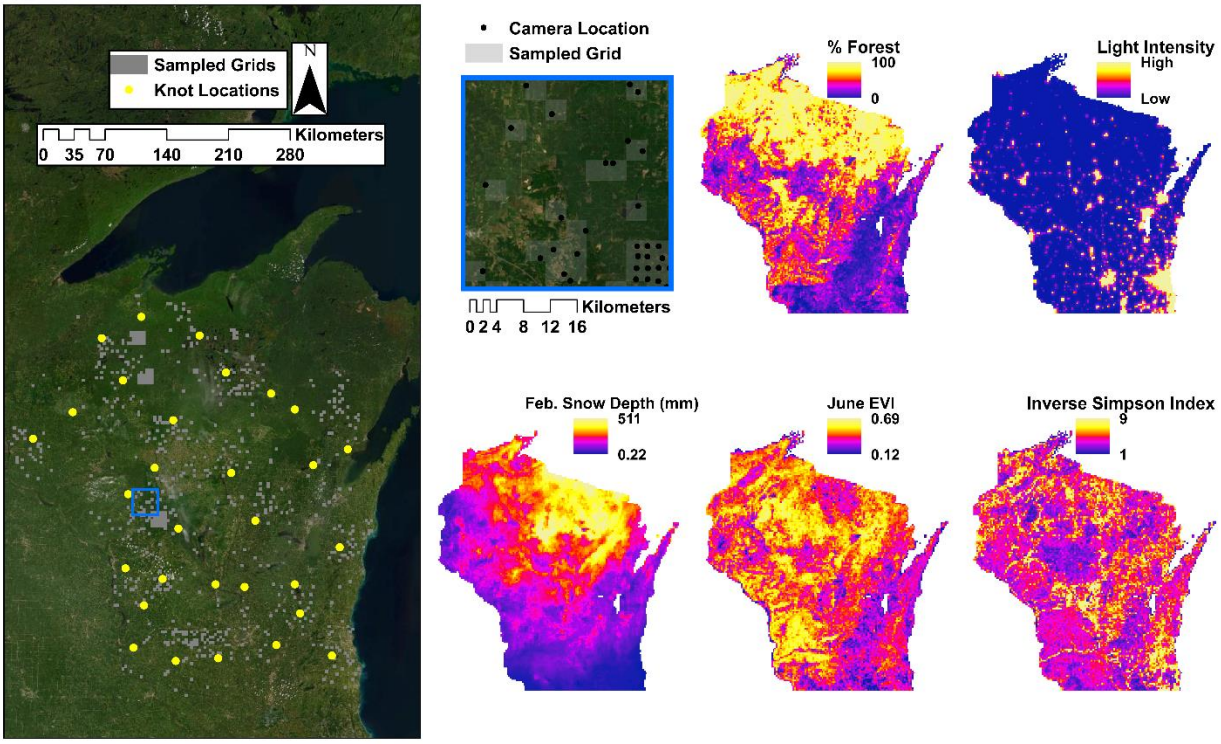

Figure 1. Sampled sites, spatial spline knot locations, and location of inset image in blue (left): the inset location depicts sampled areal units and sampling points within units. Other panels denote variation in spatial covariates of interest—note that snow depth and the enhanced vegetation index values depicted are static measures roughly corresponding to the values within primary period 2 and period 6, respectively.

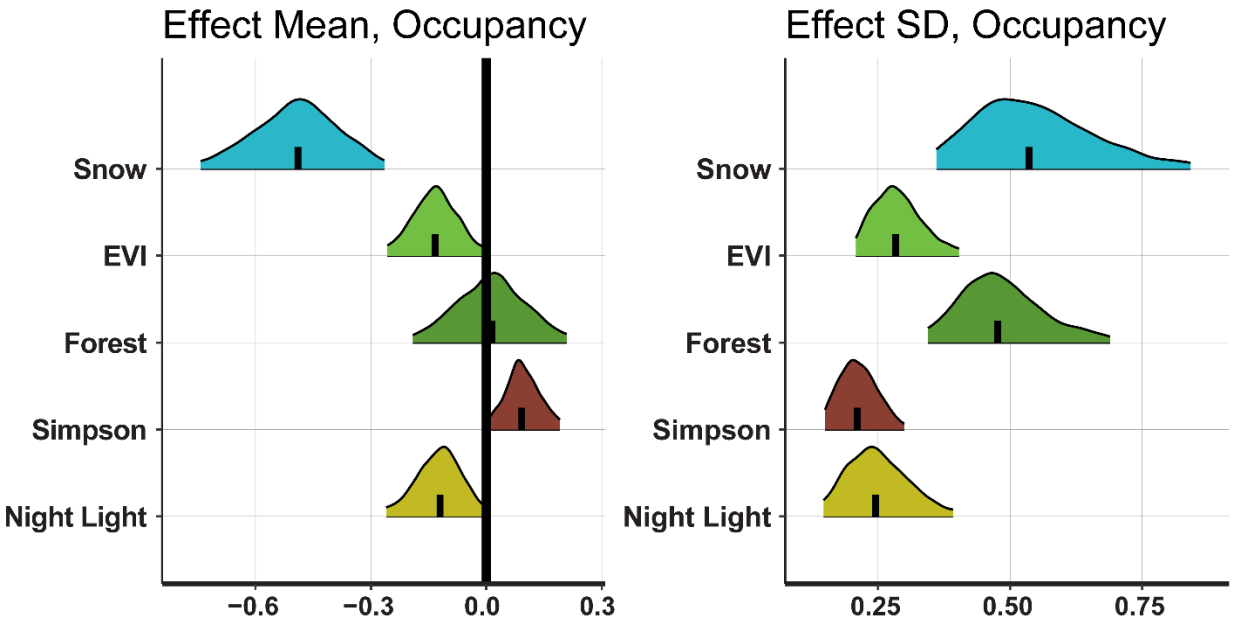

Figure 2. The posterior distributions of community-level hyperparameters (left: average effect across species; right: and inter-specific standard deviation in effect size). Snow depth exhibits a large negative expected effect with great inter-specific variance, forest cover has a near-zero expected effect that also varies sizably across species, and the inverse Simpson index (of land cover types) has a small positive expected effect with limited inter-specific variation.

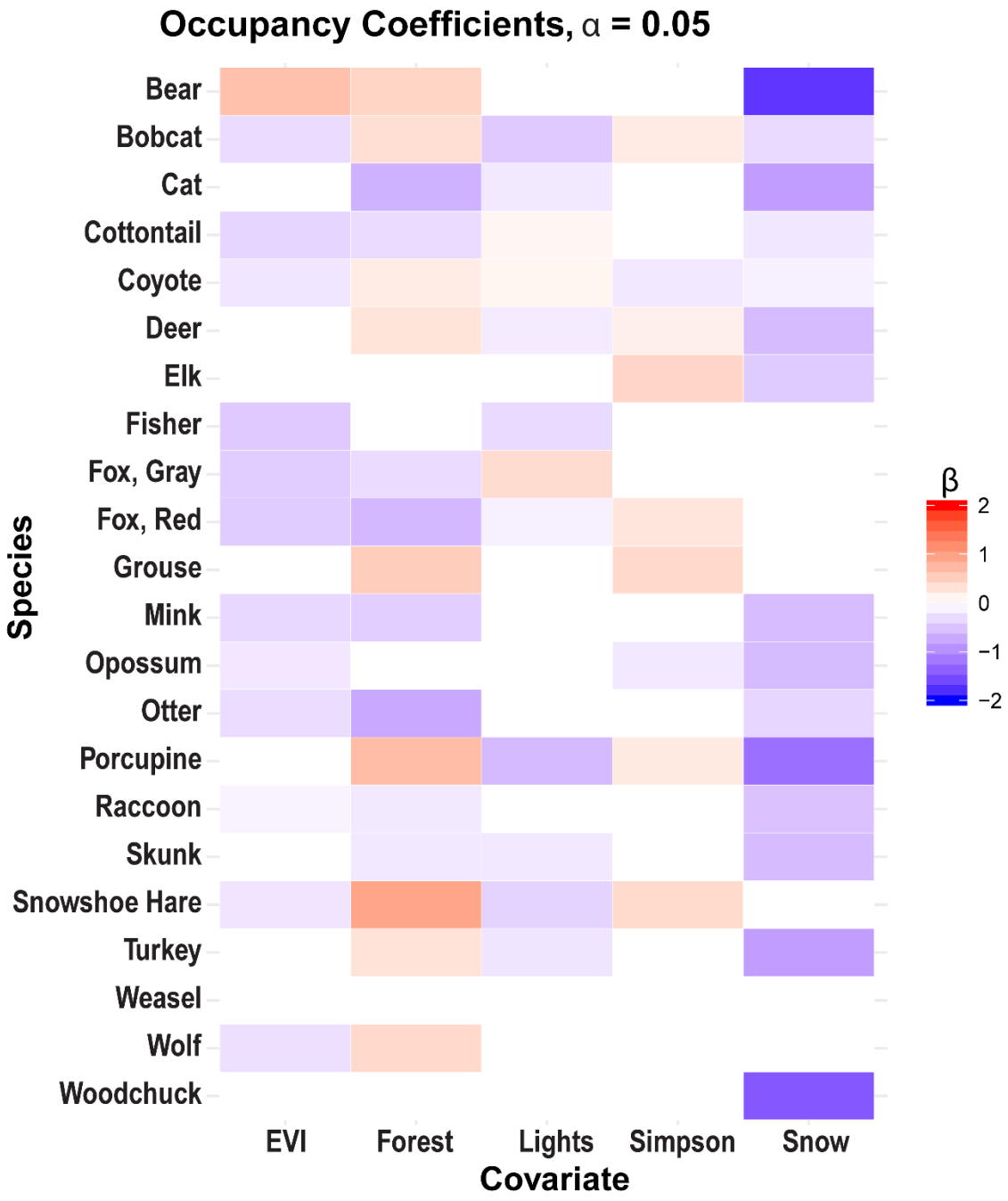

Figure 3. The effects of covariates of interest upon species-level occupancy probabilities. Note that effects with 95% credible intervals that overlapped zero are set to 0 and denoted in white.

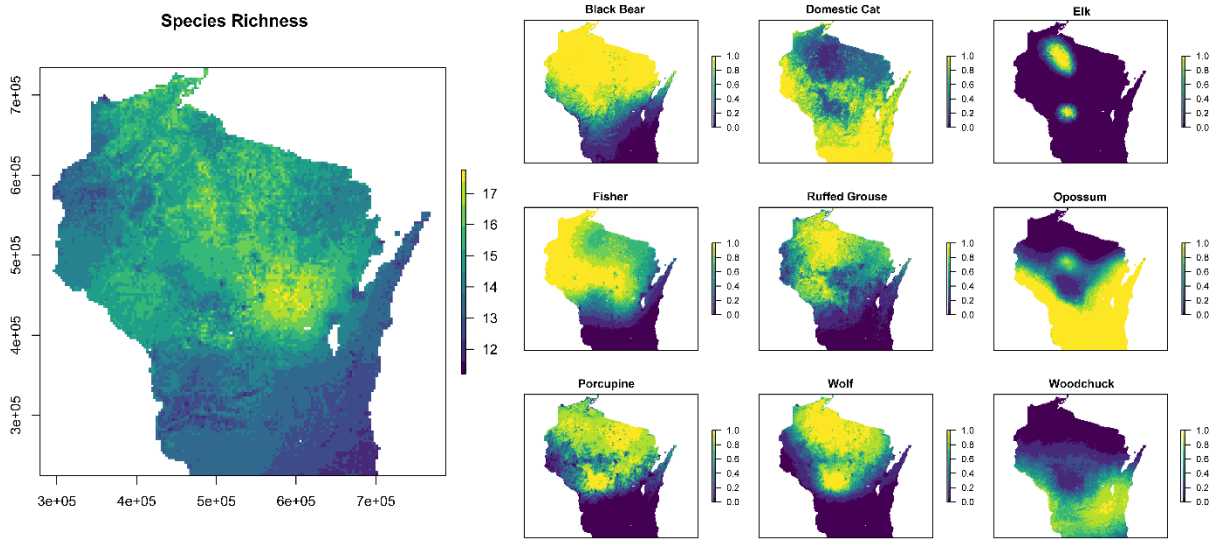

Figure 4. Annually-integrated predictions of species richness (given the species considered) across Wisconsin (left) show greatest expected richness in the central and northern parts of the state. Specific predictions of annually-integrated occupancy probability across a set of species depicted at right. Many species tended to exhibit either northerly (e.g., black bear) or southerly (e.g., opossum) distributions, with a smaller set of species that were either ubiquitously or more erratically distributed.

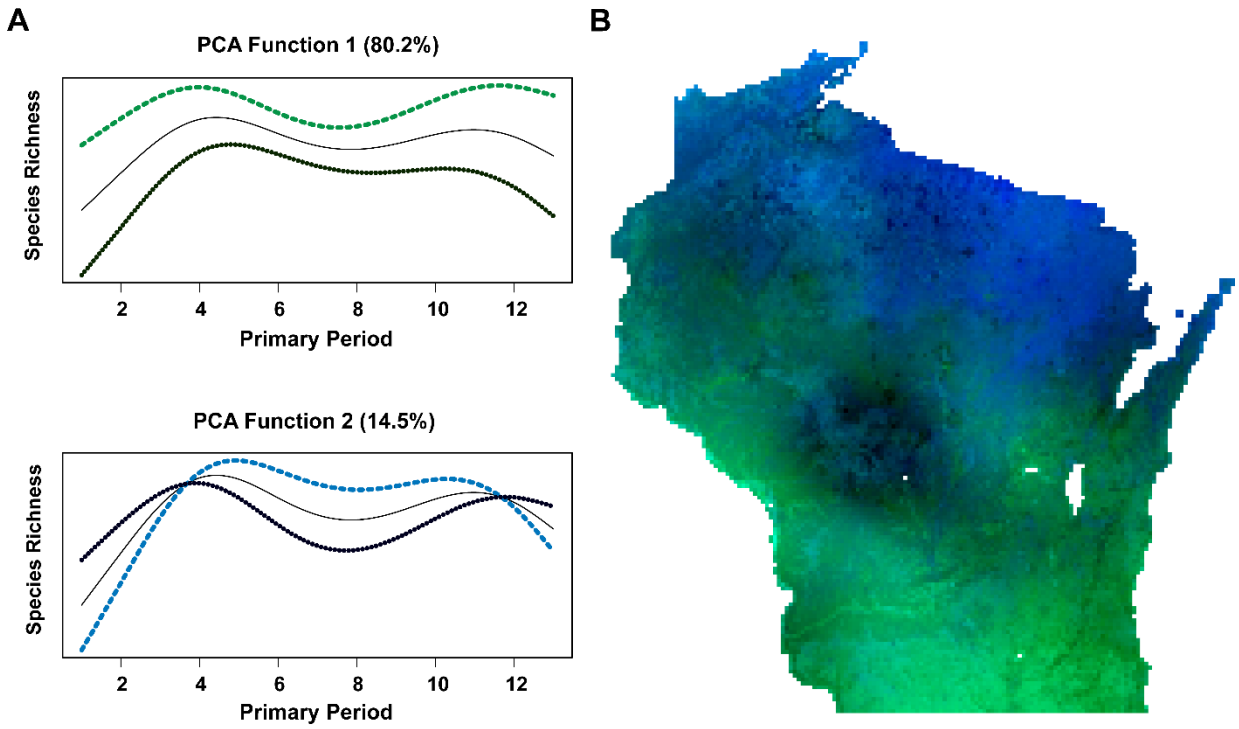

Figure 5. Two retained PCA functions (A) suggest that the primary axes for delineating spatial patterns in species richness time-series include overall richness per time period relative to expected values (top), and seasonal contrast between growing season richness and non-growing season richness (bottom). Spatial patterns in these functions (B, using green and blue bands to delineate functional values) suggest that Wisconsin's northern forests exhibit more pronounced seasonal variation in species richness, while southern Wisconsin tends to exhibit a greater and more consistent number of species.

Appendix S1. Supporting Tables and Figures.

Table S1. Species considered within this study

Species	Common Name
Neovison vison	American Mink
Ursus americanus	Black Bear
Lynx rufus	Bobcat
Canis latrans	Coyote
Felis catus	Domestic Cat
Sylvilagus floridanus	Eastern Cottontail
Cervus elaphus	Elk
Pekaniai pennanti	Fisher
Urocyon cinereoargenteus	Gray Fox
Canis lupus	Gray Wolf
Bonasa umbellus	Ruffed Grouse
Didelphis virginiana	Opossum
Erethizon dorsatum	Porcupine
Procyon lotor	Raccoon
Vulpes vulpes	Red Fox
Lutra Canadensis	River Otter
Lepus americanus	Snowshoe Hare
Mephitis mephitis	Striped Skunk
Meleagris gallopavo	Turkey
Multiple *Mustela spp.*	Weasel
Odocoileus virginianus	White-tailed Deer
Marmota monax	Woodchuck

Table S2. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with model hyper-parameters.

Parameter	Posterior Median	MC Error	LCI	UCI
μ_{α_0}	-1.72	0.01	-2.15	-1.27
σ_{α_0}	1.04	0.00	0.79	1.45
μ_{α_1}	0.48	0.00	0.29	0.66
σ_{α_1}	0.42	0.00	0.30	0.61
μ_{α_2}	0.18	0.00	-0.09	0.42
σ_{α_2}	0.65	0.00	0.47	0.91
μ_{β_1}	-0.49	0.00	-0.74	-0.27
σ_{β_1}	0.54	0.01	0.36	0.84
μ_{β_2}	-0.13	0.00	-0.26	-0.01
σ_{β_2}	0.28	0.00	0.21	0.40
μ_{β_3}	0.01	0.00	-0.19	0.21
σ_{β_3}	0.48	0.00	0.34	0.69
μ_{β_4}	0.09	0.00	-0.01	0.19
σ_{β_4}	0.21	0.00	0.15	0.30
μ_{β_5}	-0.12	0.00	-0.26	-0.01
σ_{β_5}	0.24	0.00	0.15	0.39
σ_v	1.06	0.01	0.82	1.47
σ_t	0.06	0.00	0.05	0.08

Table S3. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with variance parameters for species-specific spatial smoothing terms (σ_v^S).

Species	Posterior Median	MC Error	LCI	UCI
Bear	0.37	0.02	0.09	0.70
Bobcat	1.21	0.03	0.64	2.09
Cat	0.32	0.02	0.15	0.57
Cottontail	0.68	0.02	0.45	1.01
Coyote	0.74	0.03	0.44	1.15
Deer	0.23	0.03	0.09	0.54
Elk	3.11	0.05	1.74	6.23
Fisher	0.53	0.02	0.28	1.27
Fox, Gray	0.46	0.02	0.28	0.82
Fox, Red	1.04	0.03	0.48	1.67
Grouse	0.41	0.02	0.20	0.93
Mink	0.38	0.02	0.12	1.07
Opossum	1.52	0.03	0.95	2.61
Otter	0.59	0.02	0.26	1.08
Porcupine	0.78	0.02	0.32	1.67
Raccoon	0.32	0.01	0.17	0.55
Skunk	0.24	0.02	0.11	0.48
Snowshoe Hare	0.72	0.03	0.43	1.16
Turkey	0.42	0.03	0.23	0.76
Weasel	0.18	0.01	0.05	0.41
Wolf	0.67	0.03	0.36	1.10
Woodchuck	0.49	0.02	0.15	1.01

Table S4. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with variance parameters for species-specific temporal variance in spatial smoothing terms (σ_t^s).

Species	Posterior Median	MC Error	LCI	UCI
Bear	0.11	0.01	0.07	0.17
Bobcat	0.04	0.01	0.02	0.07
Cat	0.04	0.00	0.02	0.07
Cottontail	0.04	0.00	0.02	0.06
Coyote	0.03	0.00	0.02	0.06
Deer	0.05	0.01	0.01	0.09
Elk	0.06	0.01	0.02	0.10
Fisher	0.08	0.01	0.05	0.13
Fox, Gray	0.07	0.01	0.04	0.11
Fox, Red	0.04	0.00	0.02	0.06
Grouse	0.08	0.00	0.05	0.11
Mink	0.06	0.01	0.02	0.11
Opossum	0.07	0.00	0.04	0.11
Otter	0.06	0.01	0.02	0.11
Porcupine	0.07	0.01	0.03	0.12
Raccoon	0.06	0.00	0.03	0.08
Skunk	0.06	0.01	0.02	0.11
Snowshoe Hare	0.05	0.01	0.02	0.09
Turkey	0.07	0.00	0.04	0.10
Weasel	0.03	0.01	0.02	0.08
Wolf	0.04	0.01	0.02	0.09
Woodchuck	0.15	0.01	0.09	0.25

Table S5. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific detection intercepts.

Species	Posterior Median	MC Error	LCI	UCI
Bear	-1.82	0.00	-1.89	-1.75
Bobcat	-2.83	0.00	-2.98	-2.68
Cat	-1.67	0.00	-1.80	-1.53
Cottontail	-0.58	0.00	-0.63	-0.53
Coyote	-1.37	0.00	-1.41	-1.33
Deer	0.79	0.00	0.77	0.81
Elk	-2.43	0.00	-2.65	-2.23
Fisher	-2.67	0.00	-2.92	-2.45
Fox, Gray	-1.59	0.00	-1.79	-1.42
Fox, Red	-1.25	0.00	-1.33	-1.16
Grouse	-3.03	0.00	-3.30	-2.80
Mink	-2.77	0.00	-3.11	-2.44
Opossum	-0.65	0.00	-0.71	-0.59
Otter	-2.32	0.00	-2.66	-2.00
Porcupine	-2.31	0.00	-2.44	-2.19
Raccoon	-0.50	0.00	-0.52	-0.47
Skunk	-2.11	0.00	-2.25	-1.98
Snowshoe Hare	-1.29	0.00	-1.37	-1.21
Turkey	-1.15	0.00	-1.20	-1.11
Weasel	-2.96	0.01	-3.63	-2.45
Wolf	-3.19	0.00	-3.45	-2.94
Woodchuck	-2.85	0.01	-3.25	-2.47

Table S6. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific effects of maintained trails upon detection.

Species	Posterior Median	MC Error	LCI	UCI
Bear	0.42	0.00	0.28	0.56
Bobcat	1.04	0.00	0.85	1.23
Cat	0.81	0.00	0.60	1.01
Cottontail	0.39	0.00	0.29	0.49
Coyote	0.84	0.00	0.77	0.92
Deer	0.20	0.00	0.15	0.25
Elk	0.57	0.00	0.22	0.88
Fisher	-0.40	0.00	-0.79	-0.04
Fox, Gray	0.02	0.00	-0.37	0.40
Fox, Red	0.22	0.00	0.05	0.38
Grouse	0.62	0.00	0.27	0.94
Mink	-0.05	0.00	-0.48	0.34
Opossum	0.39	0.00	0.28	0.50
Otter	0.42	0.01	-0.07	0.88
Porcupine	0.18	0.00	-0.07	0.45
Raccoon	0.33	0.00	0.26	0.39
Skunk	0.89	0.00	0.68	1.07
Snowshoe Hare	0.75	0.00	0.61	0.90
Turkey	0.66	0.00	0.58	0.74
Weasel	0.37	0.01	-0.13	0.93
Wolf	1.24	0.00	0.93	1.53
Woodchuck	1.05	0.01	0.63	1.46

Table S7. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific effects of the presence of proximal water features on detection.

Species	Posterior Median	MC Error	LCI	UCI
Bear	-0.01	0.00	-0.13	0.12
Bobcat	0.24	0.00	0.04	0.44
Cat	-0.31	0.00	-0.50	-0.11
Cottontail	-0.10	0.00	-0.20	0.00
Coyote	-0.06	0.00	-0.13	0.02
Deer	0.08	0.00	0.04	0.12
Elk	-0.59	0.00	-0.94	-0.26
Fisher	0.06	0.00	-0.21	0.35
Fox, Gray	-0.09	0.00	-0.42	0.27
Fox, Red	-0.47	0.00	-0.63	-0.31
Grouse	-0.09	0.00	-0.38	0.21
Mink	1.96	0.01	1.42	2.50
Opossum	0.03	0.00	-0.08	0.13
Otter	1.59	0.01	1.03	2.16
Porcupine	0.44	0.00	0.24	0.64
Raccoon	0.13	0.00	0.07	0.18
Skunk	-0.25	0.00	-0.45	-0.06
Snowshoe Hare	-0.21	0.00	-0.36	-0.06
Turkey	-0.15	0.00	-0.22	-0.08
Weasel	-0.11	0.01	-0.62	0.44
Wolf	0.24	0.00	-0.04	0.51
Woodchuck	0.57	0.01	0.14	1.03

Table S8. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific occupancy intercepts.

Species	Posterior Median	MC Error	LCI	UCI
Bear	-3.34	0.01	-3.77	-2.99
Bobcat	-2.33	0.05	-3.02	-1.86
Cat	-2.65	0.00	-2.89	-2.44
Cottontail	-1.61	0.00	-1.75	-1.49
Coyote	0.31	0.01	0.19	0.44
Deer	2.51	0.00	2.38	2.65
Elk	-11.36	0.20	-15.38	-8.61
Fisher	-3.45	0.07	-4.44	-2.87
Fox, Gray	-3.75	0.01	-4.04	-3.47
Fox, Red	-1.94	0.01	-2.08	-1.77
Grouse	-3.60	0.05	-4.38	-3.00
Mink	-3.62	0.02	-4.07	-3.24
Opossum	-3.36	0.05	-4.09	-2.84
Otter	-4.42	0.03	-4.91	-3.97
Porcupine	-4.61	0.11	-6.48	-3.69
Raccoon	0.20	0.00	0.12	0.30
Skunk	-2.04	0.01	-2.24	-1.85
Snowshoe Hare	-5.85	0.05	-7.47	-4.98
Turkey	-0.50	0.00	-0.61	-0.39
Weasel	-3.68	0.01	-4.22	-3.03
Wolf	-4.35	0.06	-5.73	-3.40
Woodchuck	-5.22	0.06	-6.36	-4.41

Table S9. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific effects of snow depth on species occurrence.

Species	Posterior Median	MC Error	LCI	UCI
Bear	-1.70	0.01	-2.40	-1.12
Bobcat	-0.31	0.00	-0.49	-0.13
Cat	-0.84	0.01	-1.25	-0.46
Cottontail	-0.21	0.00	-0.36	-0.04
Coyote	-0.13	0.00	-0.23	-0.02
Deer	-0.57	0.00	-0.71	-0.44
Elk	-0.44	0.00	-0.80	-0.11
Fisher	-0.03	0.00	-0.26	0.22
Fox, Gray	0.04	0.00	-0.21	0.27
Fox, Red	0.03	0.00	-0.11	0.17
Grouse	-0.11	0.00	-0.37	0.14
Mink	-0.55	0.01	-1.06	-0.13
Opossum	-0.58	0.00	-0.89	-0.31
Otter	-0.35	0.00	-0.72	-0.01
Porcupine	-1.24	0.01	-1.68	-0.82
Raccoon	-0.52	0.00	-0.64	-0.40
Skunk	-0.57	0.00	-0.85	-0.32
Snowshoe Hare	-0.08	0.00	-0.25	0.09
Turkey	-0.85	0.00	-1.00	-0.70
Weasel	-0.02	0.01	-0.43	0.33
Wolf	-0.04	0.00	-0.26	0.17
Woodchuck	-1.41	0.02	-2.54	-0.58

Table S10. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific effects of EVI on species occurrence.

Species	Posterior Median	MC Error	LCI	UCI
Bear	0.65	0.00	0.45	0.86
Bobcat	-0.31	0.00	-0.46	-0.16
Cat	-0.06	0.00	-0.20	0.09
Cottontail	-0.36	0.00	-0.45	-0.26
Coyote	-0.21	0.00	-0.29	-0.12
Deer	0.13	0.00	-0.01	0.26
Elk	0.13	0.01	-0.15	0.38
Fisher	-0.46	0.00	-0.71	-0.21
Fox, Gray	-0.43	0.00	-0.67	-0.20
Fox, Red	-0.44	0.00	-0.55	-0.31
Grouse	-0.17	0.00	-0.41	0.08
Mink	-0.34	0.00	-0.56	-0.09
Opossum	-0.20	0.00	-0.35	-0.05
Otter	-0.29	0.00	-0.54	-0.04
Porcupine	-0.18	0.00	-0.39	0.03
Raccoon	-0.10	0.00	-0.18	-0.03
Skunk	0.01	0.00	-0.12	0.13
Snowshoe Hare	-0.22	0.00	-0.45	-0.03
Turkey	-0.05	0.00	-0.14	0.04
Weasel	-0.04	0.00	-0.32	0.23
Wolf	-0.26	0.00	-0.48	-0.04
Woodchuck	0.08	0.01	-0.29	0.44

Table S11. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific effects of Forest on species occurrence.

Species	Posterior Median	MC Error	LCI	UCI
Bear	0.43	0.00	0.22	0.66
Bobcat	0.34	0.01	0.10	0.57
Cat	-0.67	0.00	-0.87	-0.48
Cottontail	-0.30	0.00	-0.43	-0.17
Coyote	0.20	0.00	0.07	0.33
Deer	0.29	0.00	0.12	0.45
Elk	0.41	0.01	-0.16	1.05
Fisher	-0.22	0.01	-0.50	0.06
Fox, Gray	-0.30	0.01	-0.57	-0.01
Fox, Red	-0.61	0.01	-0.78	-0.43
Grouse	0.52	0.02	0.15	0.90
Mink	-0.41	0.01	-0.76	-0.06
Opossum	-0.14	0.00	-0.31	0.02
Otter	-0.74	0.01	-1.15	-0.37
Porcupine	0.70	0.01	0.41	1.01
Raccoon	-0.18	0.00	-0.29	-0.07
Skunk	-0.20	0.00	-0.37	-0.02
Snowshoe Hare	0.93	0.01	0.64	1.26
Turkey	0.29	0.00	0.17	0.42
Weasel	-0.09	0.01	-0.46	0.28
Wolf	0.40	0.01	0.06	0.74
Woodchuck	-0.35	0.01	-0.75	0.04

Table S12. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific effects of the inverse Simpson index (of land cover classes) on species occurrence.

Species	Posterior Median	MC Error	LCI	UCI
Bear	-0.10	0.00	-0.24	0.04
Bobcat	0.20	0.00	0.06	0.35
Cat	-0.10	0.00	-0.24	0.05
Cottontail	0.02	0.00	-0.07	0.12
Coyote	-0.19	0.00	-0.28	-0.11
Deer	0.16	0.00	0.05	0.29
Elk	0.43	0.00	0.17	0.71
Fisher	-0.02	0.00	-0.20	0.15
Fox, Gray	0.15	0.00	-0.02	0.32
Fox, Red	0.27	0.00	0.17	0.38
Grouse	0.39	0.00	0.20	0.58
Mink	0.19	0.00	-0.01	0.39
Opossum	-0.20	0.00	-0.32	-0.08
Otter	0.22	0.00	0.00	0.45
Porcupine	0.22	0.00	0.07	0.39
Raccoon	-0.04	0.00	-0.11	0.02
Skunk	-0.04	0.00	-0.17	0.08
Snowshoe Hare	0.38	0.00	0.24	0.54
Turkey	-0.06	0.00	-0.14	0.02
Weasel	0.03	0.00	-0.23	0.29
Wolf	-0.05	0.00	-0.24	0.12
Woodchuck	-0.03	0.00	-0.33	0.24

Table S13. Estimates, Monte Carlo error, and upper and lower 95% credible intervals associated with species-specific effects of night-time light intensity on species occurrence.

Species	Posterior Median	MC Error	LCI	UCI
Bear	-0.01	0.00	-0.31	0.23
Bobcat	-0.45	0.01	-0.85	-0.14
Cat	-0.19	0.00	-0.33	-0.05
Cottontail	0.10	0.00	0.01	0.18
Coyote	0.11	0.00	0.00	0.23
Deer	-0.17	0.00	-0.27	-0.05
Elk	-0.14	0.01	-0.67	0.33
Fisher	-0.31	0.00	-0.65	-0.02
Fox, Gray	0.37	0.00	0.18	0.54
Fox, Red	-0.12	0.00	-0.22	-0.01
Grouse	-0.24	0.01	-0.64	0.12
Mink	-0.05	0.00	-0.23	0.13
Opossum	0.04	0.00	-0.06	0.13
Otter	-0.07	0.00	-0.31	0.16
Porcupine	-0.57	0.01	-1.09	-0.21
Raccoon	-0.07	0.00	-0.16	0.01
Skunk	-0.19	0.00	-0.35	-0.04
Snowshoe Hare	-0.35	0.01	-0.81	-0.01
Turkey	-0.22	0.00	-0.32	-0.13
Weasel	-0.15	0.00	-0.52	0.14
Wolf	-0.22	0.01	-0.63	0.12
Woodchuck	-0.03	0.00	-0.26	0.19

Table S14. Posterior pair-wise correlation between species-specific responses to the environmental variables considered here.

Variable 1	Variable 2	Pearson *r*	LCI	UCI
Snow	EVI	-0.62	-0.87	-0.18
Snow	Forest	-0.09	-0.28	0.09
Snow	Simpson	0.24	0.02	0.45
Snow	Lights	0.10	-0.16	0.39
EVI	Forest	0.27	0.05	0.47
EVI	Simpson	-0.16	-0.39	0.07
EVI	Lights	0.00	-0.30	0.26
Forest	Simpson	0.27	0.04	0.48
Forest	Lights	-0.39	-0.62	-0.10
Simpson	Lights	-0.30	-0.54	0.01

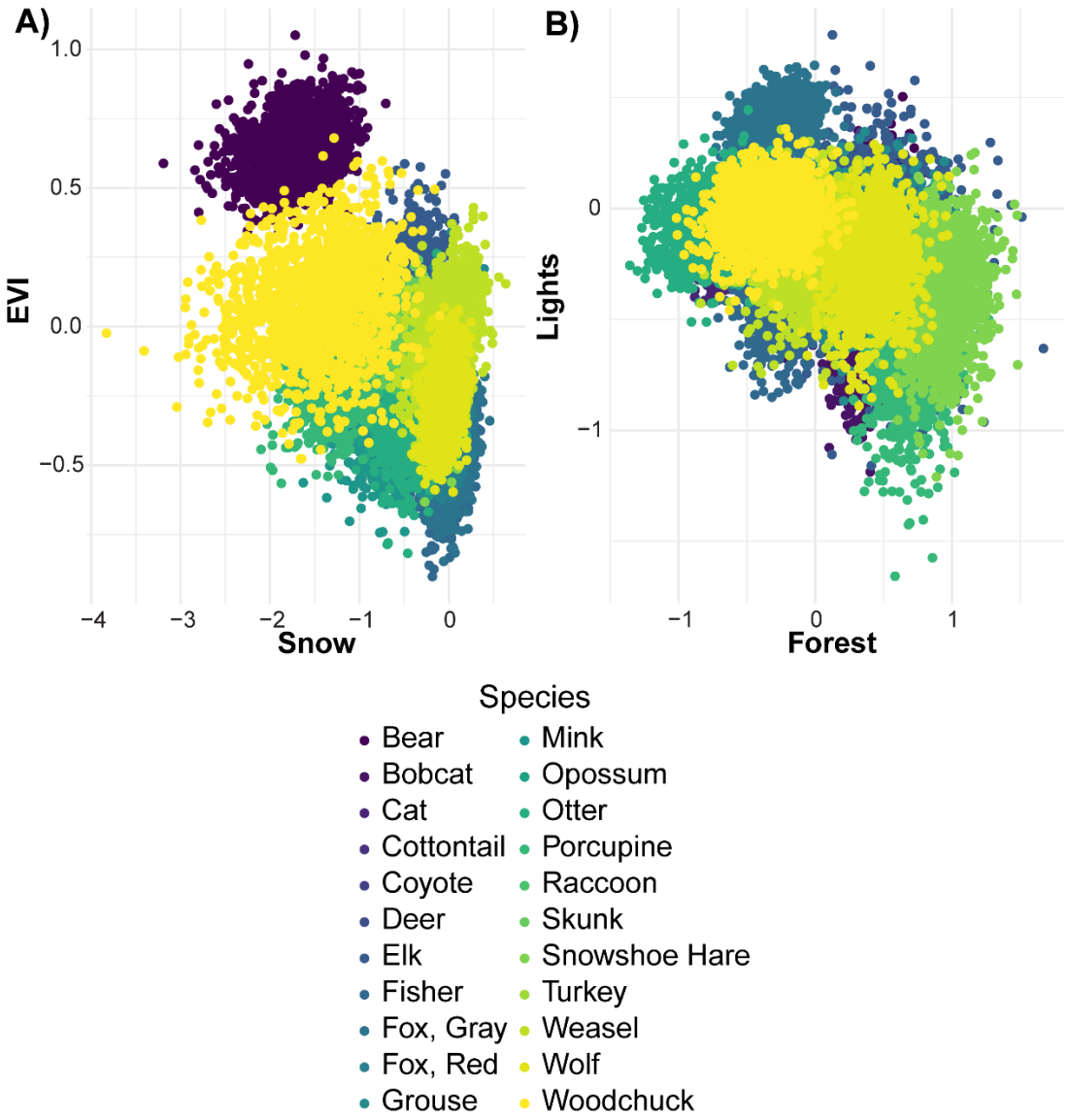

Figure S1. Posterior correlations between the occupancy effects of (A) EVI and snow-depth, and (B) night-time light intensity and the proportion of forest cover (right) indicates that, on average, species more positively associated with vegetation greenness were more negatively impacted by deeper snow, and that species positively associated with night-time light were generally less likely to occupy areas with greater forest cover.

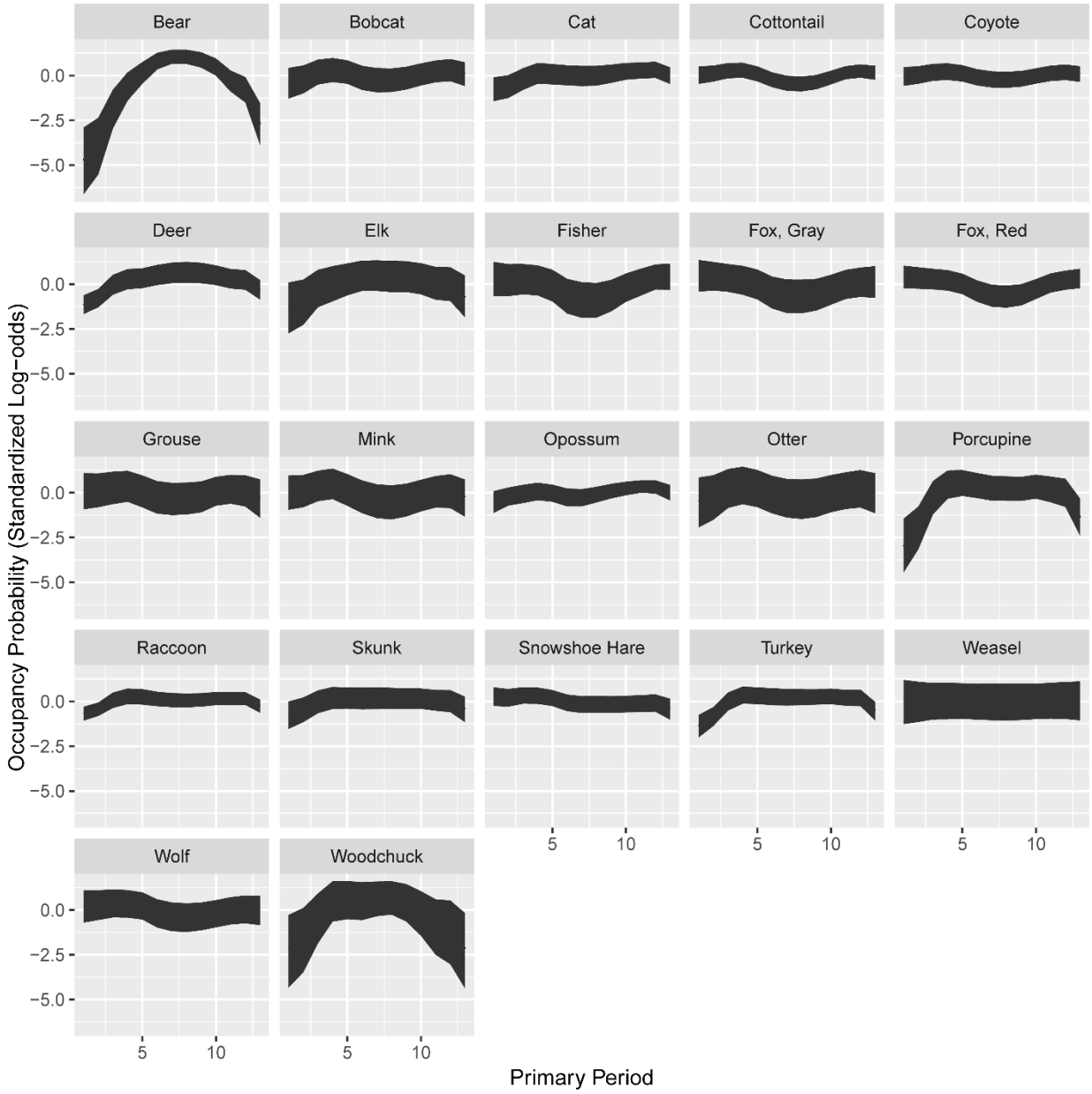

Figure S2. Temporal trends in the area of occupancy for the species considered over the annual cycle, standardized (using log-odds) relative to the annual mean for each species.