Ecological Applications, 0(0), 2019, e01849

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

John D. J. Clare,[1,5] Philip A. Townsend,[1] Christine Anhalt-Depies,[1] Christina Locke,[2] Jennifer L. Stenglein,[2] Susan Frett,[2] Karl J. Martin,[3] Aditya Singh,[1,4] Timothy R. Van Deelen,[1] and Benjamin Zuckerberg[1]

[1]*Department of Forest and Wildlife Ecology, University of Wisconsin–Madison, 1630 Linden Drive, Madison, Wisconsin 53706 USA*
[2]*Office of Applied Sciences, Wisconsin Department of Natural Resources, Madison, Wisconsin 53716 USA*
[3]*Division of Cooperative Extension, University of Wisconsin Extension, Madison, Wisconsin 53706 USA*

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 00(00):e01849. 10.1002/eap.1849

Abstract. Measurement or observation error is common in ecological data: as citizen scientists and automated algorithms play larger roles processing 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.

Key words: *automated classification; citizen science; crowdsourcing; false-positive error; misclassification; remote camera; species distribution model.*

Introduction

Applied ecologists increasingly study phenomena and tackle problems occurring at massive spatial scales (La Sorte et al. 2018). This shift has 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, Norouzzadeh 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

Manuscript received 1 August 2018; revised 25 October 2018; accepted 21 December 2018. Corresponding Editor: Viviana Ruiz-Gutierrez.

[4] Present address: Department of Agricultural and Biological Engineering, University of Florida, Gainesville, Florida 32611-0570 USA.

[5] E-mail: jclare2@wisc.edu

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 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 protocols, 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 (Norouzzadeh 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 data set 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 data set 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 (Fig. 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 data set, and (3) estimate a requisite level of accuracy or error within the raw data that allows study objectives

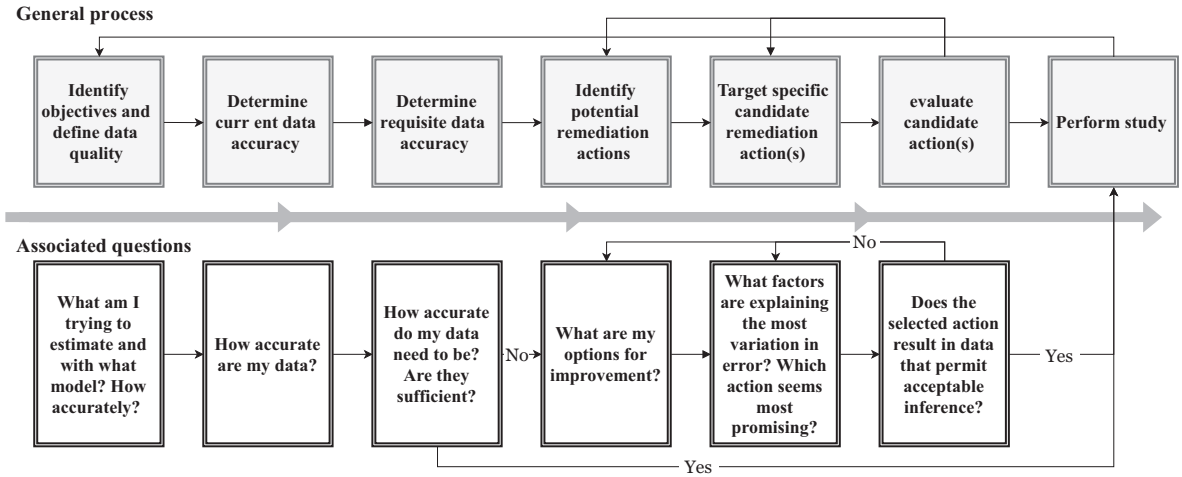

FIG. 1. Conceptual diagram of the sequential process described for evaluating data quality and data remediation actions introduced in Methods: *Framework*.

to be achieved. Remediation evaluation includes (4) identifying possible actions, (5) exploring important sources of variation in error within a data set 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. Although 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 <5% absolute expected bias and <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, 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-h 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 $(\psi_{i,\text{sim}}) = \beta_0 + \beta_1 X_{i,\text{sim},1} + \beta_2 X_{i,\text{sim},2}$, where β_0 = either -1 or 1, $\beta_1 = -2$, and $\beta_2 = 1$; logit $(p_{i,\text{sim}}) = \alpha_0 + \alpha_1 X_{i,\text{sim},1} + \alpha_2 X_{i,\text{sim},2}$, 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, 2016).

Observations were initially generated as $y_{i,\text{sim}} \sim$ Bernoulli $(z_{i,\text{sim}}, \times p_{i,sim})$, where $z_{i,\text{sim}}$ is the occupancy state for a site/simulation combination and was generated as $z_{i,\text{sim}} \sim$ Bernoulli $(\psi_{i,\text{sim}})$. We then induced additional false-negative and false-positive classification error within each simulated data set: 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 S3 and S4 for more discussion of this issue). We simulated 300 data sets for each combination of parameter values, and fit occupancy models to each data set using Markov chain Monte Carlo simulation (three chains each consisting of 2,000 adaptation steps and 3,000 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 a statistical data remediation action following the observation-confirmation protocol described by Chambert et al. (2015). Chambert et al.'s (2015) model assumes 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 $y_{i,j} \sim$ 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; Holmen, WI, USA), camera placement strategies (the seasons we evaluated variably focused upon sampling aquatic mammal monitoring or ungulates), how images were presented to online citizen classifiers (single photographs vs. sequences of three affiliated triggers), and minor changes to the user interface. If classification error varied strongly by season, it would suggest that 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 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 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 factor 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 whether the 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 the 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 four chains with 1,000 burn-in and 1,000 posterior samples, or if

necessary for convergence, 4,000 burn-in and 4,000 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?).

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

Predictor	Description
User proportion†	proportion of users voting for consensus species
jday†	day of year (1 January = 1)
dectime†	decimal hour photo was taken (military time)
TWSC†	time-weighted species change‡
Height	camera height above ground level (feet)
Distance	camera distance to target trail (feet)
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

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

‡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 $i_{x,b} \times (1/t_{x,b}) + i_{x,a} \times (1/t_{x,a})$. 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, and C are each separated by the minimum trigger interval of 15 s and record species A, B, A or B, A, B).

Results

Estimating existing data accuracy

Across the full data set, the accuracy of crowdsourced species classifications was 93.4%, but false-positive and false-negative error varied considerably across species (Fig. 2). More commonly encountered species were generally subject to less false-positive and false-negative error (Fig. 3). Exceptions include lagomorphs, as brown phenophase snowshoe hare (*Lepus americanus*) were commonly misclassified as cottontails (*Sylvilagus floridanus*, Appendix S1: Table S1 and Fig. S3), 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 (Fig. 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 four species appeared to be classified with sufficient baseline accuracy (Fig. 2). Occupancy models accounting for false-positive error provided unbiased inference across all error rates considered (Fig. 5). Estimator performance improved as more sampling intervals were validated (Fig. 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 >1,000, Appendix S1: Table S4), 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 $\omega_i = 1$), 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.

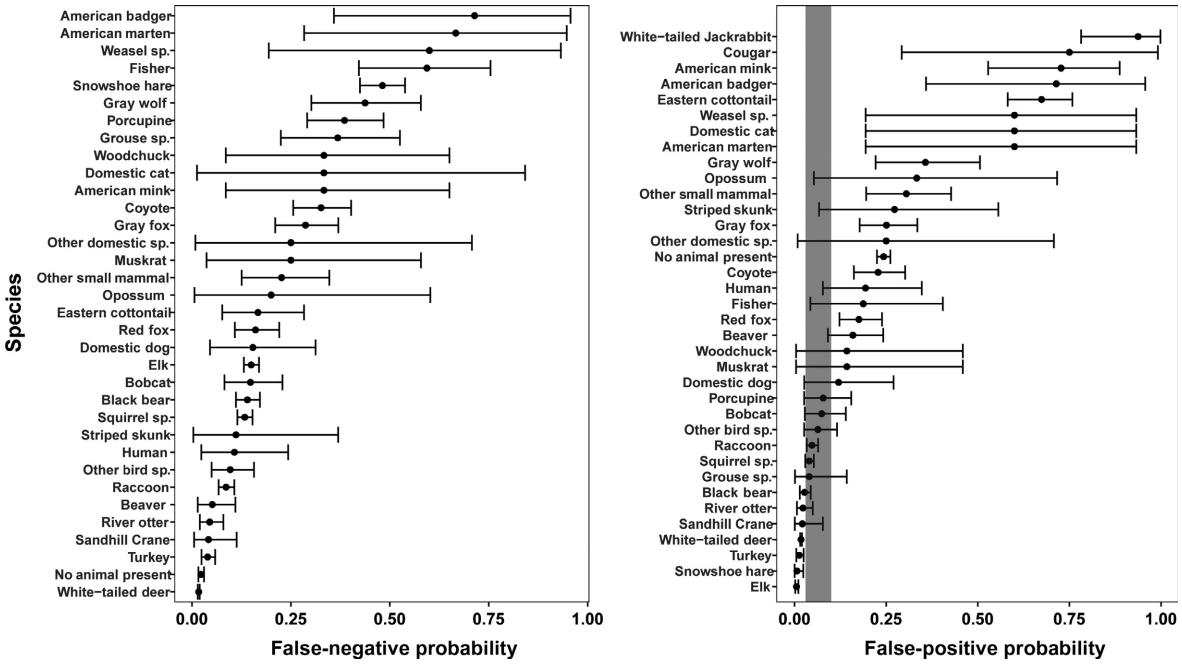

FIG. 2. False-negative (left) and false-positive (right) probabilities estimated with expert validation of crowdsourced trail camera image classification. Whiskers represent 95% credible intervals. The gray shaded area on the right panel contains a threshold for false-positive error that simulation suggested was requisite 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.

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, Appendix S1: Table S5), 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; Fig. 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 <3% within test-partitioned black bear (*Ursus americanus*) pictures required censuring <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. Empirical 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 much sampling effort is required and how to allocate available 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 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 (McClintock et al. 2010, Butt et al. 2013). Our specific results are most germane for the growing

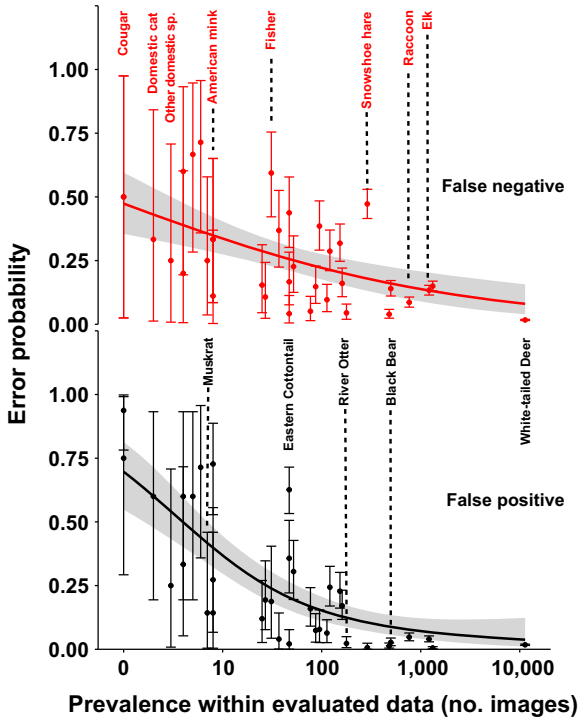

FIG. 3. Negative association between species prevalence in the data set (total number of images; log-transformed) and false-positive (black) and false-negative (red) classification error probabilities, which suggests that the distribution of rare or cryptic species was more likely to be estimated with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

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 into the state of data reliability associated with one of the most ubiquitous data processing tasks (species identification), one of the most common goals in ecology (estimating species distribution), and one of the most widely used models for estimating species distribution. The first data quality concern associated with estimating species distribution 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, Norouzzadeh 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, Norouzzadeh 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 data sets processed by humans with limited training (volunteer or not) or automated algorithms, there is a non-trivial risk of substantially overestimating the distribution of many species using many commonly used data types. Furthermore, motivation to further expedite data processing has led to the development of hybrid 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 second 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, 2015, Ruiz-Gutiérrez et al. 2016); false-positive parameters are often distinct from 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 (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% bias in animal abundance estimated using certain models (Clare et al. 2018). For other models, the overall number or rate of detections rather than their locations may be more important. Although classification error appears to have reshuffled species observations across

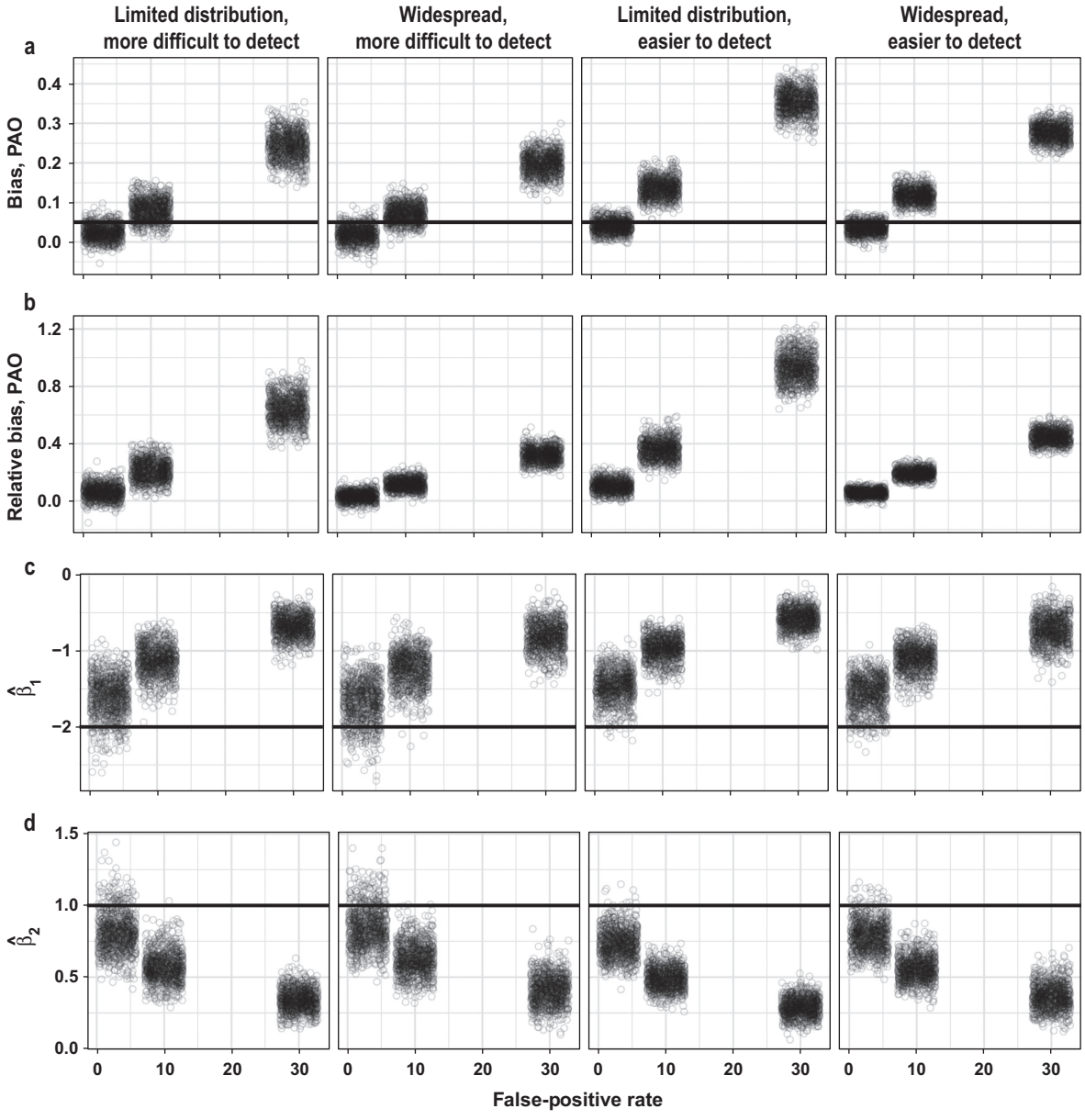

FIG. 4. False-positive classification error is the strongest determinant of (a) mean error or bias and (b) relative bias in finite sample estimates (PAO indicates proportion of area occupied). The solid line in panel a represents the predefined threshold described in Methods: *Defining data quality*. Occupancy coefficient estimates are displayed in panels c and d, and false-positive error shrinks coefficients toward 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.

locations, the overall prevalence of species within our data set was largely preserved (Appendix S1: Table S1 and Fig. S1). Had we considered, for example, 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 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

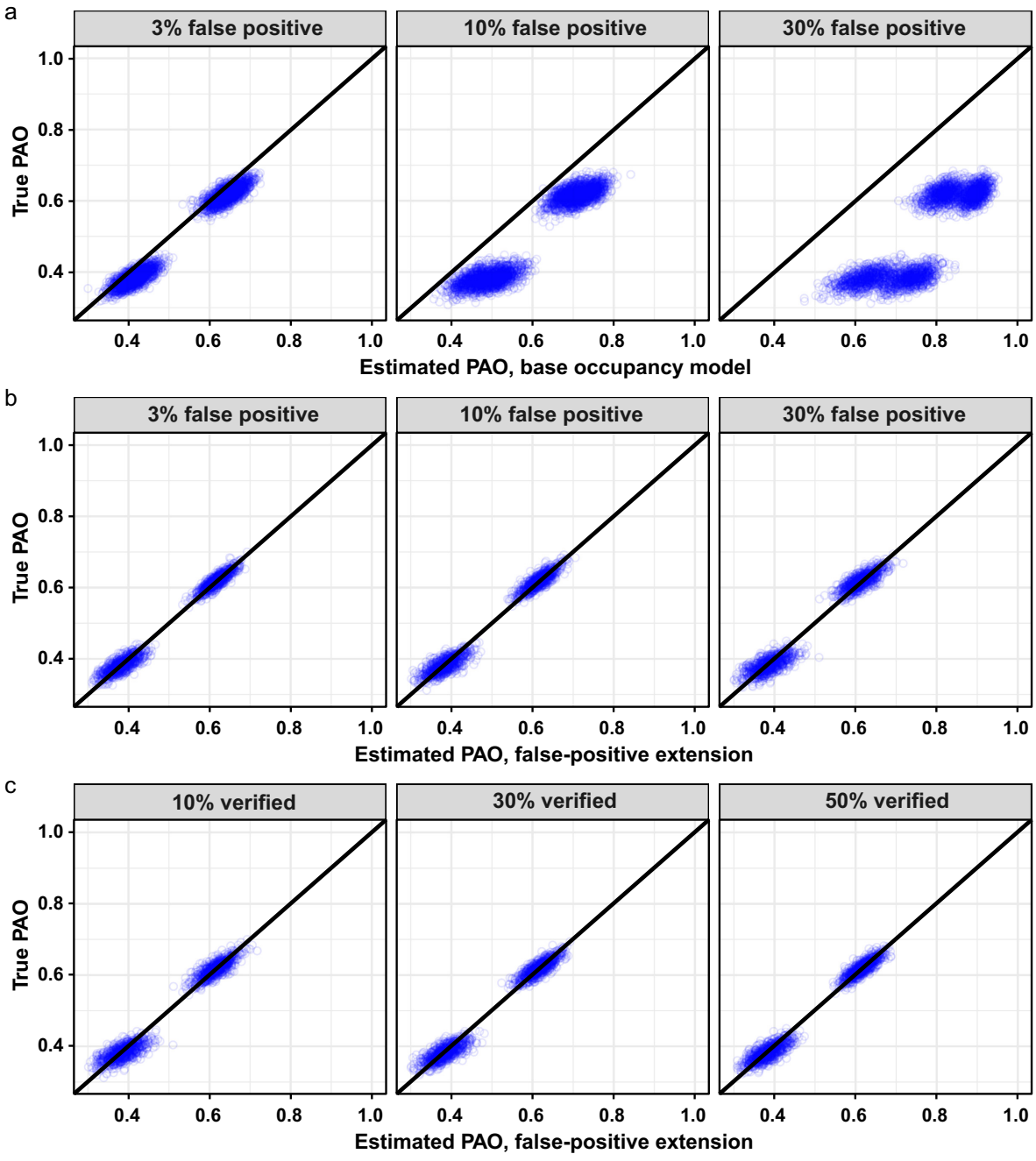

FIG. 5. (a) Standard occupancy models assuming only false-negative error are strongly biased as the proportion of false-positive observations within the sample increases. (b) Models that also incorporate false-positive error estimated using sample validation are generally accurate even when error rates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

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 measure or estimate 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

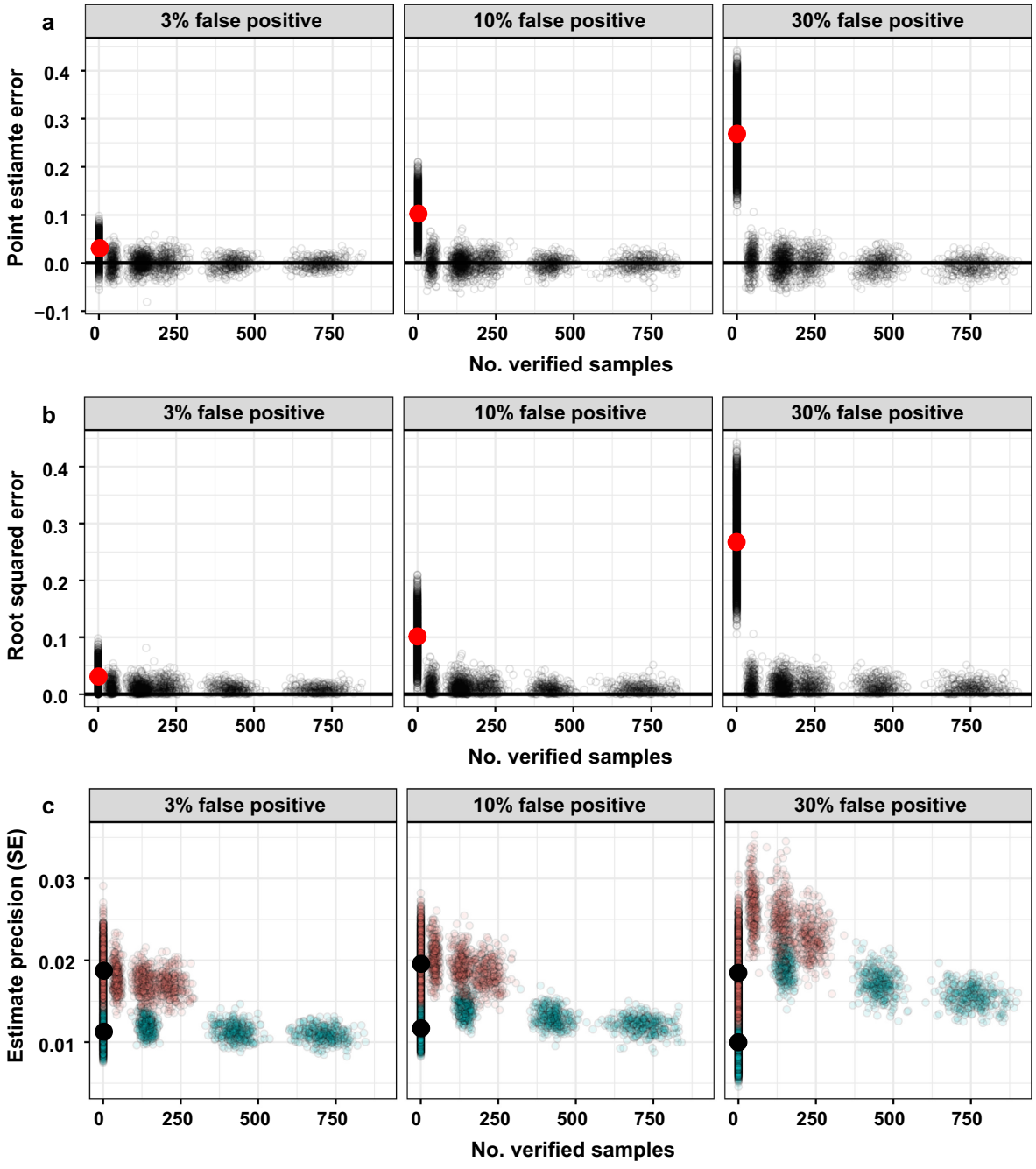

FIG. 6. (a) Point estimate bias and (b) root-squared error of finite-sample occupancy estimates decreased as more simulated samples were verified, but also that the expected decrease in either loss function decelerated. (c) Using a model incorporating false positives inflates estimate uncertainty (standard 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).

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

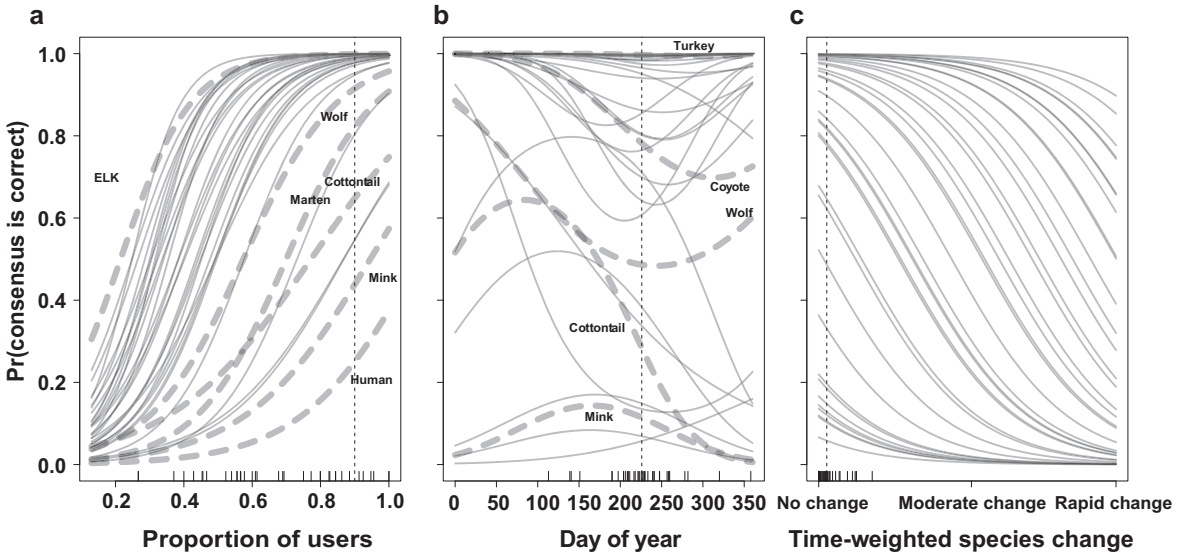

FIG. 7. Marginal modeled effects suggest that (a) the consensus crowdsourced classification was more likely to be correct as the proportion of users voting for the consensus species increased, (b) interspecifically variable depending upon the Julian day on which the image was taken, and (c) less likely to be accurate if the classifications immediately previous and/or subsequent at a given camera station reported different species in quick succession. 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 data set. In panel a, the divergent response associated with a human classification is likely a function of different retirement rules associated with human images.

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 components. 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 the classification interface, or as a more expensive example, switching to different camera models) were likely to be inefficient remediation actions. 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 implementing data censuring as a remediation action for our objective (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 carries several 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 aimed at reducing baseline rates of classification error. Quantifying interspecific variability in error and user agreement as useful indicators of accuracy directly informed protocol changes such as highlighting commonly confused species (Appendix S1: Fig. S3) 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 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 (30–50 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 evaluating data quality 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 or improve inference, and that any effort allocated toward 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 toward 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, while 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.

Literature Cited

Abra, F. D., M. P. Huijser, C. S. Pereira, and K. Ferraz. 2018. How reliable are your data? Verifying species identification of

road-killed 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., et al. 2014. Statistical solutions for error and bias in global citizen science datasets. Biological Conservation 173:144–154.

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

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, New York, USA.

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., 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

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. Goodrich. 2016. rstanarm: Bayesian applied regression modeling via Stan. https://cran.r-project.org/web/packages/rstanarm

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

Kissling, W. D., 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.

Mac Aohda, O., et al. 2018. Bat detective—deep learning tools for bat acoustic signal detection. PLoS Computational Biology 14:e1005995.

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.

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica 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. 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.

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., 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.

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

Plummer, M. 2003. JAGS: a program for analysis of Bayesian graphical models using GIBBS sampling. Pages 20–22 *in* Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vol. 124. Technische Universit at Wien, Wien, Austria. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf

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.

R CoreTeam. 2015. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. http://www.r-project.org. Accessed October 21, 2015.

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.

Steenweg, R., 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.

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.

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. 2018. Machine learning to classify animal species in camera trap images: applications in ecology. BioRxiv 346809. https://doi.org/10.1101/346809

Wiggins, A., and K. Crowston. 2015. Surveying the citizen science landscape. First Monday 20. https://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. 2018. Identifying animal species in camera trap images using deep learning and citizen science. Methods in Ecology and Evolution. https://doi.org/10.1111/2041-210x.13099

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.1849/full