Appendix S2: Statistical assessment on determining local presence of rare bat species, *Ecosphere*

Authors: Irvine, K.M.; Banner, K.M.; Stratton, C.; Ford, W.M., Reichert, B.E.

- Updated: 2022-03-02

Statistical software used

R statistical Software (R Core Team 2021) was used to write all code and for all statistical simulations. The computing language NIMBLE (de Valpine et al. 2017, 2021) was used for implementation of Markov chain Monte Carlo (MCMC) to fit the three Bayesian hierarchical models. The `coda` (Plummer et al. 2006) and `rstan` (Stan Development Team 2020) packages were used for MCMC processing and diagnostic summaries. The `tidyverse` (Wickham et al. 2019) was used extensively for post-processing simulations, and `ggplot2` (Wickham 2016) and `gridExtra` (Auguie 2017) were used for creating figures.

Simulation Scenarios and MCMC

There were redundancies in the results for many of the simulation scenarios we investigated and we chose to present a subset of scenarios in the manuscript that highlighted the biggest differences and practical takeaways. Here, we present all simulation results *not presented* in the main text *(R code available in DataS1)*. For reference, the parameter settings used for each scenario are provided in Table S1 (this is the same as Table 1 in the paper).

A total of 50 datasets were generated under each scenario/visit combination (Table S1). We first tuned the implementation of the Bayesian models by investigating trace plots and other Markov Chain Monte Carlo (MCMC) convergence criteria like the potential scale reduction factor (Rhat) and the number of effective sample sizes. The burn in, number of iterations,

Scenario	No. Nights	No. Sites	Species	ϕ_k	ψ_k	λ_k	p_k
1	8, 16	55	Spp 1	(0.90, 0.35)	0.25	.3	0.26
			Spp 2	(0.10, 0.65)	0.75	10	0.999
2	8, 16	55	Spp 1	(0.65, 0.10)	0.25	.3	0.26
			Spp 2	(0.35, 0.90)	0.75	10	0.999
3	8, 16	55	Spp 1	(0.65, 0.40)	0.25	.3	0.26
			Spp 2	(0.35,0.60)	0.75	10	0.999
4	8, 16	55	Spp 1	(0.90, 0.35)	0.50	.3	0.26
			Spp 2	(0.10, 0.65)	0.75	10	0.999
5	8, 16	55	Spp 1	(0.90, 0.35)	0.25	.7	0.50
			Spp 2	(0.10, 0.65)	0.75	10	0.999

Table S1: Parameter settings used to generate autoID datasets for 55 sites and 8 or 16 nightly surveys. Simulated data assumes the information about autoclassifier performance identifying species is correct and constant for all sites and revisits. Note ϕ_k has the columns as the true species identity and the rows refer to the autoID label. For example, under scenario 1, autoID labels to species 1 are correctly assigned 90% of the time and are contaminated by false positives or incorrectly labeled as species 1 35% of the time, on average. ψ_k is the probability species k occurs at a site, λ_k is the relative activity for species k during a nightly survey, and p_k is the probability at least one call file is recorded for species k during a nightly survey.

and thin settings in NIMBLE that were required to achieve reasonable mixing of the chains (visual inspection of traceplots), Rhat < 1.1., and an effective sample size of at least 500 differed for each scenario.

- Scenarios 1,2,3,4: 16 and 8 visits; niter = 10000, nburn = 5000, thin = 5, nchains = 3
- Scenario 5: required very long burnin and a more aggressive thin
  - 8 visits: niter = 80000, nburn = 70000, thin = 10, nchains = 3
  - 16 visits: niter = 110000, nburn = 100000, thin = 10, nchains = 3

Despite tuning efforts, some of the datasets fit with the Bayesian models we investigated showed evidence that the MCMC used by NIMBLE to obtain draws from the posterior distribution did not converge or had poor mixing of chains. We removed results that exhibited issues with convergence (issue flagged if potential scale reduction factor (Rhat) was greater than 1.1, $n_{eff} < 500$, or credible intervals were extremely wide for the λ parameters) prior to summarizing simulation results. A summary of the total number of data sets that

were flagged for issues with convergence is provided in Table S2. These numbers also reflect the total number of intervals that went into the average credible interval calculation for each combination of method, scenario, and visits (Figures S1 - S2).

Sections in this supplement follow the organization of sections in the manuscript. We present parameter estimation results first, and then a detailed investigation into the effects of different threshold values with respect to final site-level decisions.

Parameter estimation

Wright et al. (2020) established that the *2SppCt* model results in unbiased estimates of the probability of presence ψ_k and the visit-level relative activity λ_k for a species assemblage with k species, when the species (miss)classification probabilities $\phi_{kk'}$ are unknown and need to be informed by the data or an informative prior (see Stratton et al., n.d. for details). Here, we investigate the ability of the *2SppCt* model to result in unbiased estimates for its parameters when the species (miss)classification probabilities are assumed to be known. The assumption that the species (miss)classification probabilities are known affords a direct comparison to the *MLESite* decision results, which also assumes $\phi_{kk'}$ are known. We also evaluate parameter estimation for the Bayesian single-species occupancy models, *Naive* and *Remove* (MLE-metric applied at the visit level prior to fitting single-species occupancy model), to assess the appropriateness of their use for making site-level decisions about species presence. Figures S1 and S2 show the results not shown in the paper for rare species and common species respectively.

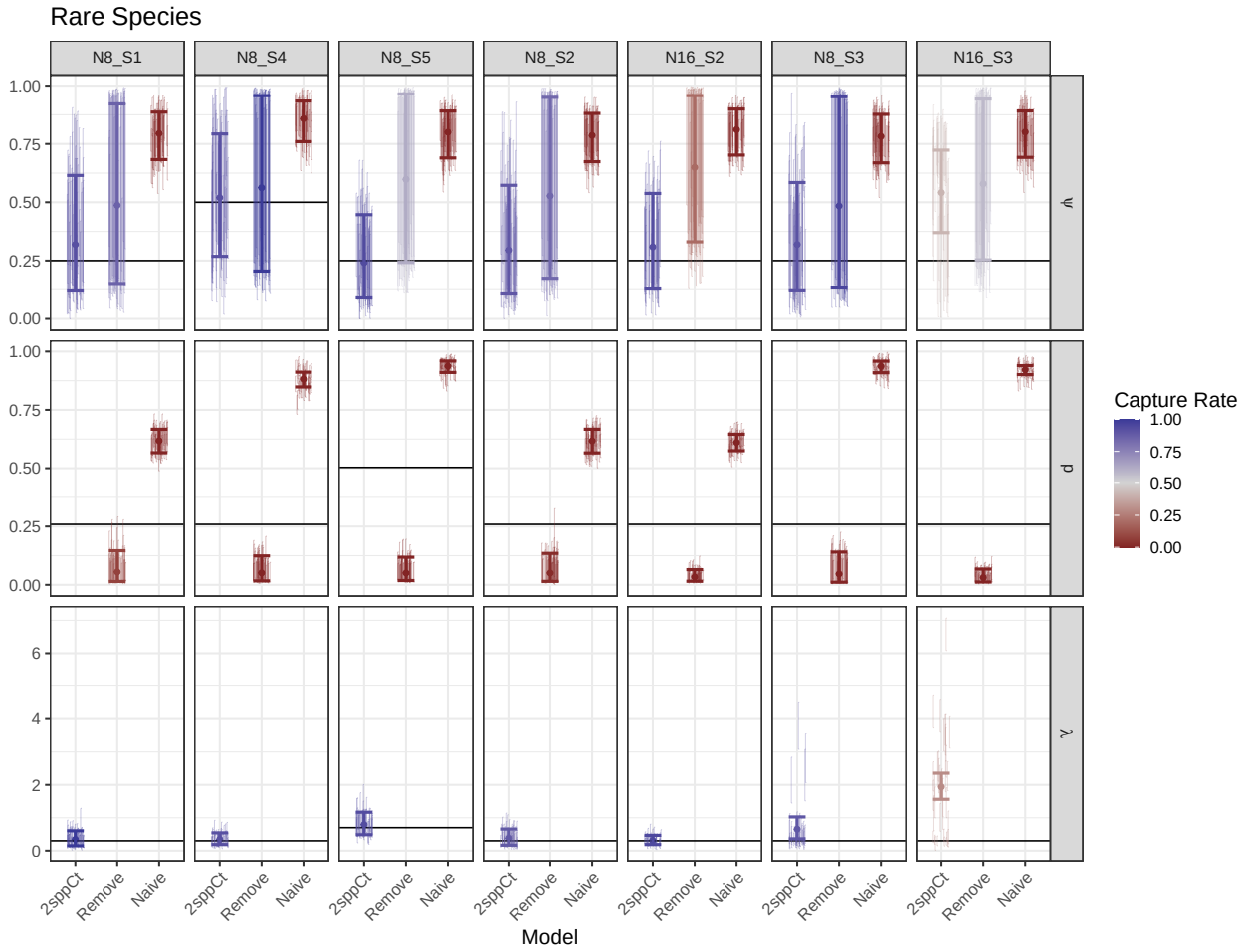

Figure S1: Average (thick line) and individual (thin lines) 95% posterior intervals for ψ (top row), p (middle row) and λ (bottom row) for the rare species for all scenarios not presented in the main text (columns). Coverage for each scenario is indicated by the color of the interval (red = poor, purple = decent). Average intervals and coverage are computed out of the total number of iterations of the simualtion that resulted in a model that converged (see Tables S2 and S3).

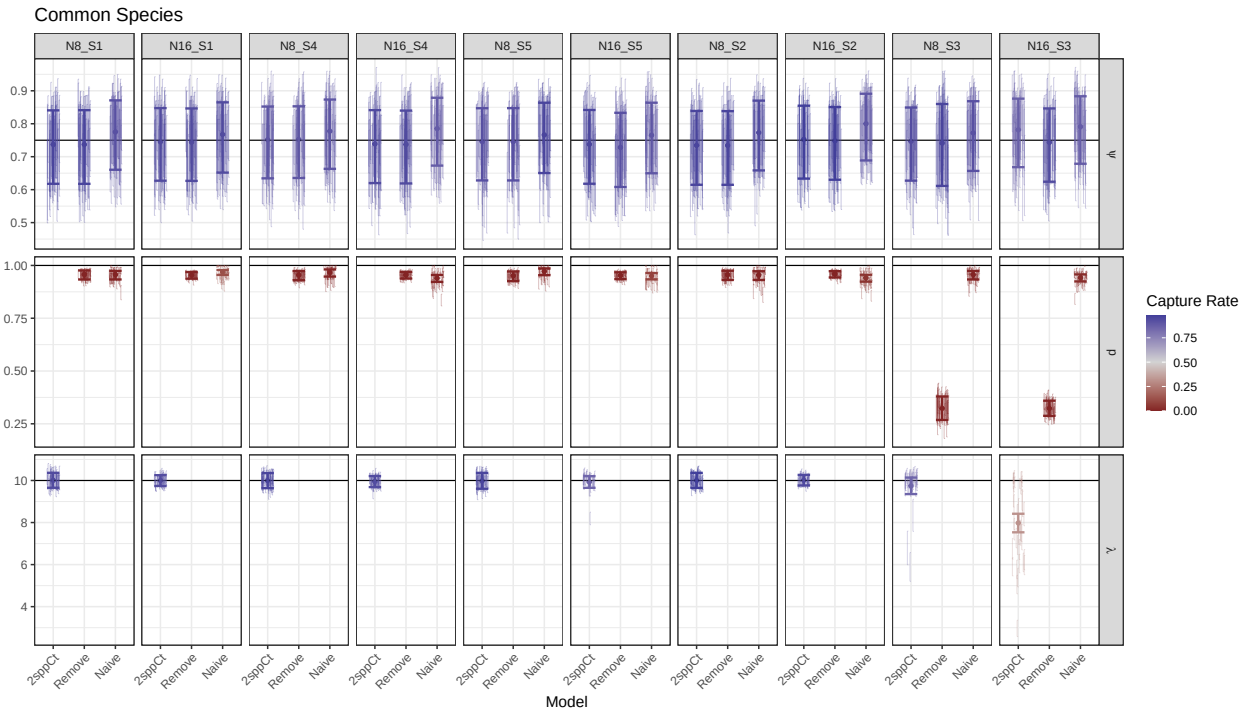

Figure S2: Average (thick line) and individual (thin lines) 95% posterior intervals for ψ (top row), p (middle row) and λ (bottom row) for the common species for all scenarios considered (columns). Coverage for each scenario is indicated by the color of the interval (red = poor, purple = decent). Average intervals and coverage are computed out of the total number of iterations of the simualtion that resulted in a model that converged (see Table S2 and S3).

group	param	model	num_unconverged	num_converged
N16_S1	lambda[1]	2SppCt	4	46
N16_S4	lambda[1]	2SppCt	3	47
N8_S5	lambda[1]	2SppCt	1	49
N16_S5	lambda[1]	2SppCt	11	39
N8_S2	lambda[1]	2SppCt	2	48
N16_S2	lambda[1]	2SppCt	12	38
N8_S3	lambda[1]	2SppCt	11	39
N16_S3	lambda[1]	2SppCt	12	38
N16_S1	lambda[2]	2SppCt	4	46
N16_S4	lambda[2]	2SppCt	3	47
N8_S5	lambda[2]	2SppCt	1	49
N16_S5	lambda[2]	2SppCt	17	33
N8_S2	lambda[2]	2SppCt	2	48
N16_S2	lambda[2]	2SppCt	13	37
N8_S3	lambda[2]	2SppCt	9	41
N16_S3	lambda[2]	2SppCt	12	38
N16_S1	psi[1]	2SppCt	2	48
N16_S4	psi[1]	2SppCt	1	49
N16_S5	psi[1]	2SppCt	3	47
N8_S2	psi[1]	2SppCt	2	48
N16_S2	psi[1]	2SppCt	8	42
N8_S3	psi[1]	2SppCt	5	45
N16_S3	psi[1]	2SppCt	9	41
N16_S4	psi[2]	2SppCt	1	49
N16_S5	psi[2]	2SppCt	3	47
N16_S2	psi[2]	2SppCt	2	48
N8_S3	psi[2]	2SppCt	4	46
N16_S3	psi[2]	2SppCt	4	46

Table S2: Summary of MCMC convergence for each scenario/visit/model and parameter combination. The total number data sets that resulted in a fitted model that converged is shown in the rightmost column. The `num_converged` also represents the total number of individual intervals plotted in Figures S1 and S2, as well as the number used in the denominator to create the overall average intervals for each scenario.

- There were no issues with MCMC with the *Naive* or *Remove* approaches and there are 50 intervals plotted for all scenario/visit/model combinations shown in Figures S1 and S2. The number of intervals plotted for parameters in the *2SppCt* model for each scenario are available in Table S2.

Threshold investigation

The parameter estimation investigation revealed that the *Naive* model should not be considered for site-level decisions, as the parameter estimation was severely biased. Here, we investigate the ramifications of choice of threshold (choice of α for *MLESite* and z_{cutoff}, $Pr(Z_{ik} = 1|y_{ijk}) \geq z_{cutoff} \implies Z_{ik} = 1$ for *2SppCt* and *Remove*) on each method's ability to result in correct site-level decisions. Given that parameter estimation results were similar for Scenario 2 and Scenario 1 and that there were issues with estimation for Scenario 3, we present results for both species for combinations of Scenarios 1, 4, and 5, each with 8 and 16 visits in Figures S3-S5. Investigating 8 vs 16 visits within a scenario allows us to get a sense for the impact the total number of recordings for each species has on site-level decisions regarding species presence (see Discussion in paper). Within a plot for a species-model combination (e.g., "MLESite: Rare Species" in Figure S3), correct decisions are shown on the diagonal. Correct determination of species presence is in the upper left corner and correct determination of species absence is in the bottom right corner. Because the decision rates are calculated conditional on the true Z state, each rate in the top row of a plot for a specific species type (i.e., rare or common) has a corresponding detection rate in the bottom row that is 1 minus the rate in the top row (*Note: 0s do not show up on the plots*).

Threshold investigation for *MLESite* method

An investigation into the influence of different *p-value* thresholds (α) on the (in)correct decision rates for determining species presence or absence using the *MLESite* decision rule is summarized by Figure S3. The top plot shows the more interesting results for the rare species and the bottom plots shows the ability of the *MLESite* method to make consistently correct decisions for the common species regardless of if its status at the site or scenario-visit combination. Some patterns emerge from investigating decisions for the rare species.

- As α increases, the median decision rate for correctly determining species presence also

increases across all scenario-visit combinations. This is a direct result of allowing a higher tolerance for erroneously concluding a species is present at a site when in fact it is not. The opposite can be seen in the medians of the decision rate for correctly determining species absence (as α increases, the median decreases).

- A slight increase in median correct decision rates for species presence when considering 8 vs. 16 visits is apparent.
- Variability is consistent among all scenario-visit combinations for all αs investigated.

We chose to present the $\alpha = 0.1$ results in the paper to reflect a larger, but commonly accepted tolerance for a erroneous conclusion of species presence than the USFWS guideline of $\alpha = 0.05$. This cutoff affords slightly higher correct decision rates than the $\alpha = 0.05$ cutoff, but this is a very important consideration for practitioners using the *MLESite* approach to consider in the context of their study.

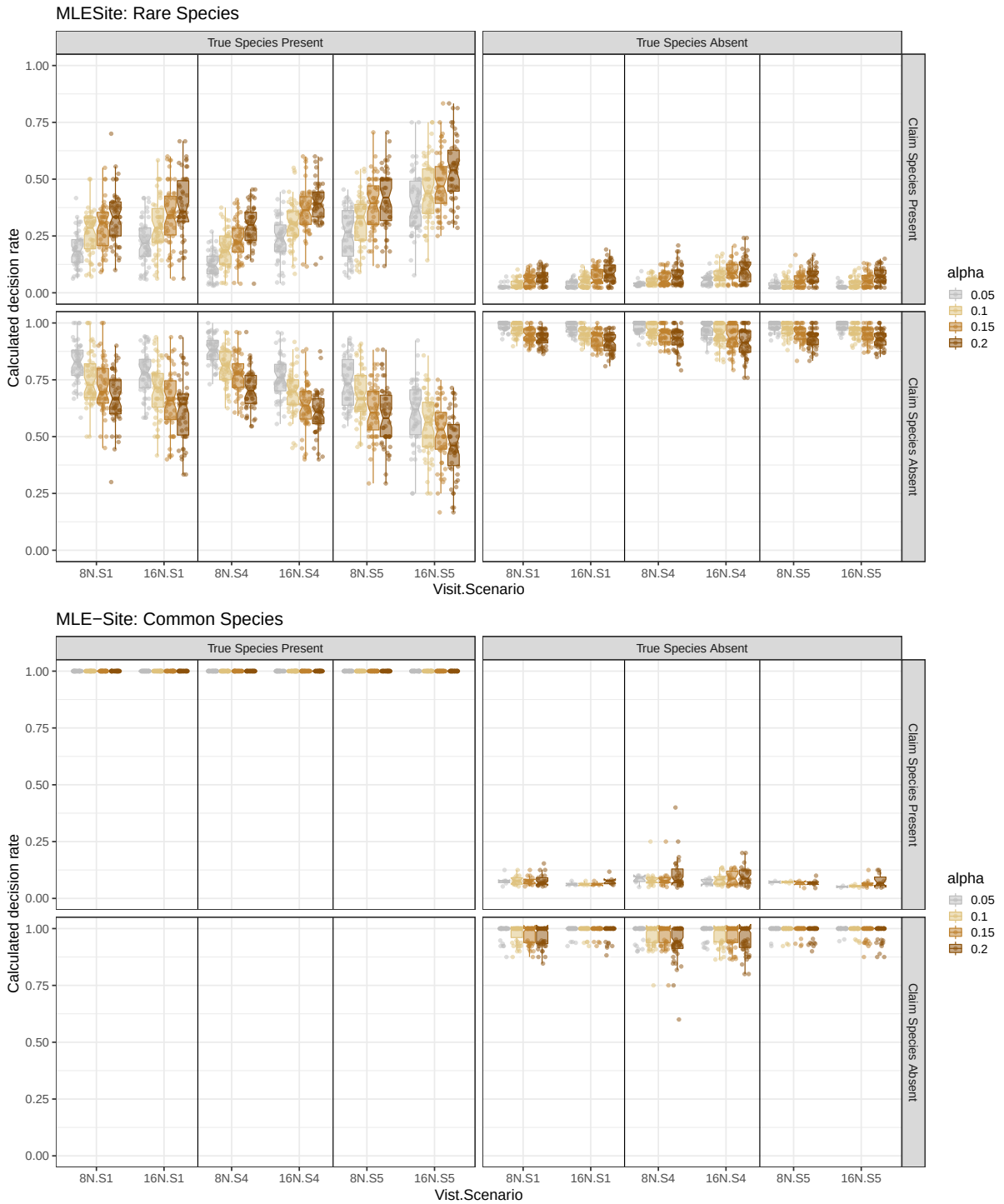

Figure S3: Calculated decision rates for MLE site approach (50 datasets with n = 55 sites) for deciding rare species presence or absence. Each dot represents a calculated conditional proportion based on the true Z state for each site. Each box represents the scenario for different p-value cutoffs related to the null hypothesis test. A species is claimed present if the p-value less than α. Each column is the true Z-state and each row is the species site-level decision (correct decision about species presence in top left and about species absence in bottom right. We present a comparison of results among Scenarios 1, 4, and 5 for 8 vs. 16 visits. These scneario-visit combinations are reprsented on the x-axis (e.g., '8N.S1' reflects the results assuming data generating values for scenario 1 and 8 visits (8 nights of recording).

Threshold investigation for Bayesian models

Investigations into the influence of different z_{cutoff} thresholds on the (in)correct decision rates for determining species presence or absence using the *2SppCt* and *Remove* models are summarized in Figure S4 (rare species) and Figure S5 (common Species).

For the common and easily detected species, altering the threshold value for the *Remove* approach and *2SppCt* model made no difference in the ability to discriminate between a site with the species present or absent (Figure S5). Alternatively, decisions about rare-species presence were very sensitive to the threshold used. For both *Remove* and *2sppCt*, we found that the preferred-cutoff choice for arriving at a binary decision for species presence based on the posterior probability a rare species occurred would be different than the preferred cutoff for determining the species was absent from a site. Smaller z_{cutoff}s are better if the rare species is truly present at the site (lighter colors in left column, Figure S4). Conversely, larger z_{cutoff}s appear better if the rare species is truly absent from the site (darker colors in right column, Figure S4).

The "risk" associated with incorrect decisions should be weighed in the context of the problem before choosing a cutoff. For example, if the greatest risk is claiming a species is absent when the species is truly present, then a lower posterior probability should be considered as evidence of species presence (Figure S4). A threshold choice for a decision regarding rare species presence should include a transparent consideration of the cost (or value) of making incorrect (or correct) decisions regarding species presence or absence. An improved approach that avoids relying solely on a threshold to create a binary decision would be to consider a Bayesian decision analysis (BDA) for determining local species presence or absence (Williams and Hooten 2016, see Discussion).

We also investigated receiver operating characteristic type curves (not shown here) to help us identify "optimal" cutoffs. Our investigation revealed that when a species is truly present, the z_{cutoff} of 0.25 strikes the best balance between correct decisions about species presence

and false conclusions of species presence when the species is truly absent (higher rates in top left panel and lower rates in top right panel, both plots in Figure S4). Conversely, the z_{cutoff} of 0.75 was found to strike the best balance between correct decisions about species absence and false conclusions of species absence when the species was truly present (higher rates in bottom right panel and lower rates in bottom left panel, both plots in Figure S4).

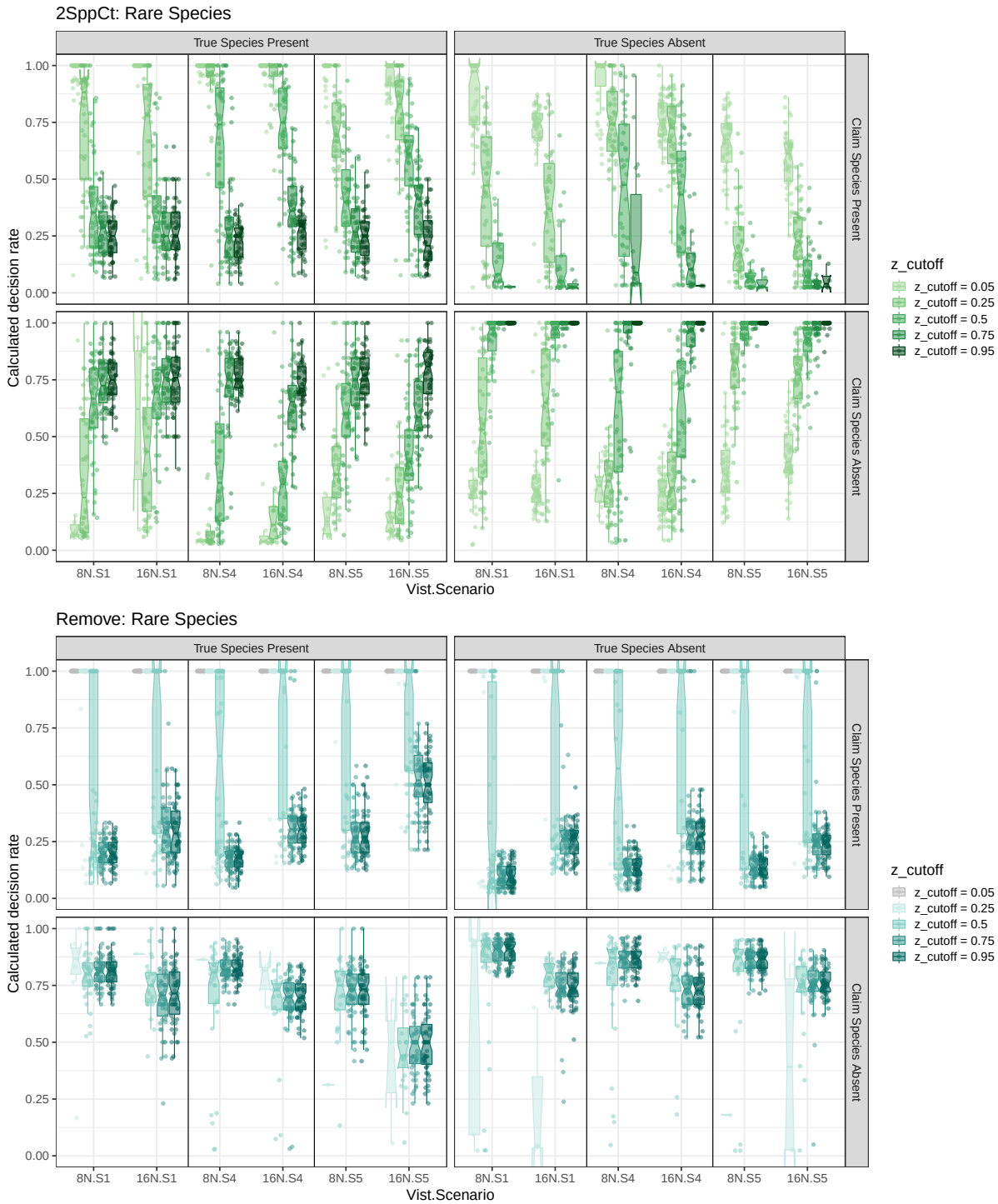

Figure S4: Calculated error rates for for deciding rare species presence or absence for the 2*SppCt* model (top plot) and *Remove* model (bottom plot). Each dot represents a calculated conditional proportion based on the true Z state for each site and simulatied dataset. A species is claimed present if $Pr(Z_{ik} = 1|y_{ijk}) \geq z_{cutoff}$. True Z-state (column) and site-level decision (row) show the rate of (in)correct decision rates from each simulated dataset (50 datasets with n = 55 sites). Different colors represent the different $z_{cutoffs}$ investigated and scneario-visit combinations are reprsented on the x-axis (e.g., '8N.S1' reflects the results assuming data generating values for scenario 1 and 8 visits (8 nights of recording). The top plot

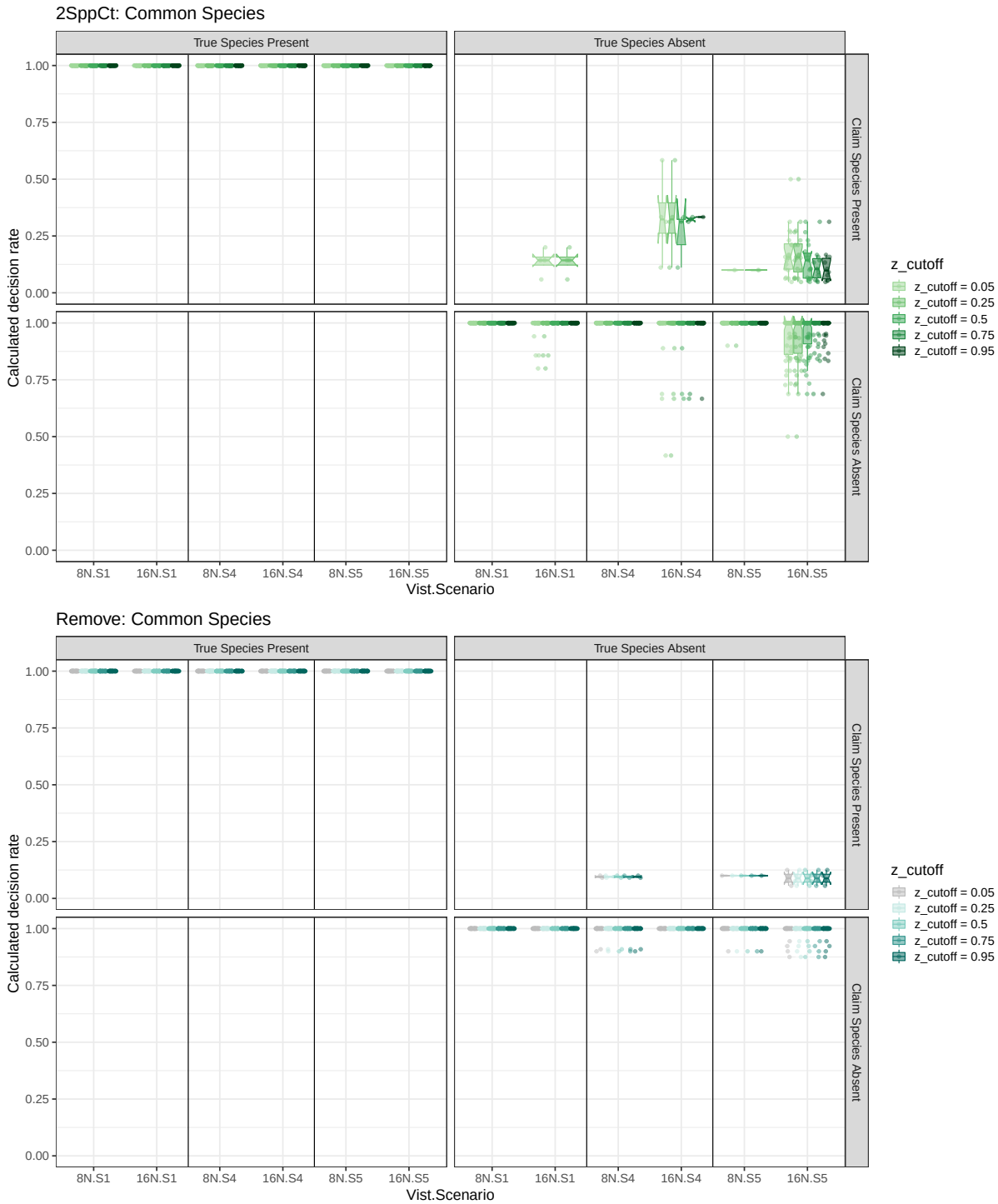

Figure S5: Calculated error rates for for deciding common species presence or absence for the 2*SppCt* model (top plot) and *Remove* model (bottom plot). Each dot represents a calculated conditional proportion based on the true Z state for each site and simulatied dataset. A species is claimed present if $Pr(Z_{ik} = 1|y_{ijk}) \geq z_{cutoff}$. True Z-state (column) and site-level decision (row) show the rate of (in)correct decision rates from each simulated dataset (50 datasets with n = 55 sites). Different colors represent the different $z_{cutoffs}$ investigated and scneario-visit combinations are reprsented on the x-axis (e.g., '8N.S1' reflects the results assuming data generating values for scenario 1 and 8 visits (8 nights of recording). The top plot

Table summary of comparsions presented in text

Summary statistics for the decision rates that make up the boxplots in Figure 3 in the main text are provided in Tables S3 and S4.

Approach	Decision.Truth	Species	Visit.Scenario	Median	IQR	Q1	Q3
2SppCt	1.1	1.00	16N.S5	0.83	0.26	0.67	0.93
2SppCt	1.1	1.00	16N.S4	1.00	0.05	0.95	1.00
2SppCt	1.1	1.00	16N.S1	0.78	0.49	0.43	0.92
Remove	1.1	1.00	16N.S5	1.00	0.00	1.00	1.00
Remove	1.1	1.00	16N.S4	1.00	0.00	1.00	1.00
Remove	1.1	1.00	16N.S1	1.00	0.00	1.00	1.00
MLESite	1.1	1.00	16N.S5	0.44	0.20	0.35	0.55
MLESite	1.1	1.00	16N.S4	0.30	0.11	0.25	0.36
MLESite	1.1	1.00	16N.S1	0.29	0.15	0.22	0.37
2SppCt	0.1	1.00	16N.S5	0.24	0.25	0.12	0.36
2SppCt	0.1	1.00	16N.S4	0.11	0.14	0.05	0.19
2SppCt	0.1	1.00	16N.S1	0.43	0.46	0.17	0.63
Remove	0.1	1.00	16N.S5	0.50	0.32	0.28	0.59
Remove	0.1	1.00	16N.S4	0.81	0.10	0.73	0.83
Remove	0.1	1.00	16N.S1	0.89	0.00	0.89	0.89
MLESite	0.1	1.00	16N.S5	0.56	0.20	0.45	0.65
MLESite	0.1	1.00	16N.S4	0.70	0.11	0.64	0.75
MLESite	0.1	1.00	16N.S1	0.71	0.15	0.63	0.78

Table S3: Summary statistics of calculated decision rates whe the true species is present shown in Figure 1 of the text. Calculated decision rate Medians, IQR (middle 50% of the decision rate distriution), Q1 (decision rate at which 25% of the calculated rates are below), and Q3 (decision rate at which 75% of the calculated rates fall below) within each visit.scenario combination.

Approach	Decision.Truth	Species	Visit.Scenario	Median	IQR	Q1	Q3
2SppCt	1.0	1.00	16N.S5	0.04	0.03	0.02	0.05
2SppCt	1.0	1.00	16N.S4	0.10	0.14	0.04	0.17
2SppCt	1.0	1.00	16N.S1	0.02	0.02	0.02	0.04
Remove	1.0	1.00	16N.S5	0.25	0.09	0.19	0.28
Remove	1.0	1.00	16N.S4	0.28	0.13	0.21	0.34
Remove	1.0	1.00	16N.S1	0.27	0.10	0.20	0.30
MLESite	1.0	1.00	16N.S5	0.03	0.03	0.03	0.05
MLESite	1.0	1.00	16N.S4	0.07	0.04	0.05	0.10
MLESite	1.0	1.00	16N.S1	0.05	0.05	0.02	0.07
2SppCt	0.0	1.00	16N.S5	1.00	0.03	0.97	1.00
2SppCt	0.0	1.00	16N.S4	0.97	0.10	0.90	1.00
2SppCt	0.0	1.00	16N.S1	1.00	0.00	1.00	1.00
Remove	0.0	1.00	16N.S5	0.76	0.09	0.72	0.81
Remove	0.0	1.00	16N.S4	0.73	0.13	0.67	0.79
Remove	0.0	1.00	16N.S1	0.74	0.10	0.70	0.81
MLESite	0.0	1.00	16N.S5	0.97	0.03	0.95	0.98
MLESite	0.0	1.00	16N.S4	0.95	0.08	0.92	1.00
MLESite	0.0	1.00	16N.S1	0.95	0.05	0.93	0.98

Table S4: Summary statistics of calculated decision rates whe the true species is absent shown in Figure 1 of the text. Calculated decision rate Medians, IQR (middle 50% of the decision rate distriution), Q1 (decision rate at which 25% of the calculated rates are below), and Q3 (decision rate at which 75% of the calculated rates fall below) within each visit.scenario combination.

- All tables generated by `xtable`, Dahl et al. (2019)

References

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Dahl, David B., David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. 2019. *Xtable: Export Tables to Latex or Html*. https://CRAN.R-project.org/package=xtable.

de Valpine, Perry, Christopher Paciorek, Daniel Turek, Nick Michaud, Cliff Anderson-Bergman, Fritz Obermeyer, Claudia Wehrhahn Cortes, Abel Rodrìguez, Duncan Temple Lang, and Sally Paganin. 2021. *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling* (version 0.11.0). https://doi.org/10.5281/zenodo.1211190.

de Valpine, Perry, Daniel Turek, Christopher Paciorek, Cliff Anderson-Bergman, Duncan Temple Lang, and Ras Bodik. 2017. "Programming with Models: Writing Statistical Algorithms for General Model Structures with NIMBLE." *Journal of Computational and Graphical Statistics* 26 (2): 403–13. https://doi.org/10.1080/10618600.2016.1172487.

Plummer, Martyn, Nicky Best, Kate Cowles, and Karen Vines. 2006. "CODA: Convergence Diagnosis and Output Analysis for Mcmc." *R News* 6 (1): 7–11. https://journal.r-project.org/archive/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stan Development Team. 2020. "RStan: The R Interface to Stan." http://mc-stan.org/.

Stratton, Christian, Kathryn M. Irvine, Katharine M. Banner, Wilson J. Wright, Cori Lausen, and Jason Rae. n.d. "Coupling Validation Effort with *in Situ* Bioacoustic Data Improves Estimating Relative Activity and Occupancy for Multiple Species with Cross-Species

Misclassifications." *Methods in Ecology and Evolution*.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Williams, Perry J., and Mevin B. Hooten. 2016. "Combining Statistical Inference and Decisions in Ecology." *Ecological Applications* 26 (6): 1930–42.

Wright, Wilson J., Kathryn M. Irvine, Emily S. Almberg, and Andrea R. Litt. 2020. "Modelling Misclassification in Multi-species Acoustic Data When Estimating Occupancy and Relative Activity." *Methods in Ecology and Evolution* 11 (1): 71–81.