Associate Editor:

I enjoyed reading this well-written manuscript, and believe that the modeling framework proposed is sufficiently versatile to be used in a wide range of applications. In addition to the points made by the referees, I have two comments:

1. *Reviewer 1 is not convinced that it is helpful to make explicit the link to survival analysis. I tend to disagree with this, since I think that this is what renders the approach widely applicable and also extendable. In fact, the general idea strongly reminded me of several papers published recently by David Borchers and colleagues (using concepts from survival analysis to model detection events within distance sampling and spatially explicit capture-recapture). I was wondering if there is a common denominator between his work and the present manuscript (but the authors need not necessarily explore this).*

   > We agree, this manuscript has strong parallels to the continuous-time models of David Borchers and colleagues. Their models are for removal-distance data, whereas ours is for removal-only data.

2. *In Section 3.4, models are built where the various TTDDs considered are linked to covariates via the rate parameter. It was not clear to me why the rate parameter was chosen to be modeled as a function of covariates, as opposed to the mean of the TTDD. For example, in gamma GLMs, the mean is modeled as a function of covariates, with the shape parameter fixed (the latter as done also in the present work). Basically I'm now wondering a) why the rate rather than the mean was modeled, and b) if that matters at all. In 3.1, a justification seems to be given, but I do not follow the reasoning here why would the inclusion of random and fixed effects be any different if not the rate but the mean would be used? Sorry if I'm just missing the obvious.*

   > We have added a statement clarifying that there is no mathematical difference. ADAM: In response to (a), while we fit our models with common distributions for $f_T(t)$, we could alternatively begin with a detection hazard rate $h(t)$. Given this context, the detection rate feels more like the focus of inference than the expected time to detection.

Reviewer #1:

I have included several items for the authors to consider below and provide this review as an Ecologist and potential model user. Several pieces of information are lacking from the manuscript that will be relevant to users of this method.

1. *What types of data are relevant to each examined distribution (e.g., gamma, Weibull)?*
2. *What sample sizes (i.e., number of detections) are needed to obtain reliable estimates with CIs that are not so wide as to be meaningless?*
3. *What are approximate computation times for models of different complexity (e.g., days? weeks?).*

Abstract

Line 14: Perhaps instead of saying its not justified, mention it can lead to bias

Done.

Line 29-31: Clarify that you conducted a simulation analysis and your results from that analysis suggest it can outperform non-mixture models.

Done.

Line 40-41. Also worth mentioning caveats - e.g., greater imprecision and computation times.

ADAM: Address this point

Introduction
Line 36-41: I find mention of survival analyses here and later in the manuscript confusing for most readers - especially Ecologists, presuming this is the intended audience. It may be due to the unorthodox use of time-to-detection instead of traditional time-to-event analyses often used to estimate survival. I suggest either omitting the reference to survival analysis here and elsewhere in the manuscript, or using the time-to-event terminology in this sentence and providing citations for the cdf and pdfs outlined here for reference. Then explain that within the manuscript you will subsequently refer to it as TTDD.

We have explicitly defined time-to-detection data as being time-to-event data.
ADAM: I am unclear what the reviewer means by 'providing citations for the cdf and pdfs outlined here for reference'.

3.1
Line 60: In the context of estimating abundance of wildlife, justifying the inclusion of the Weibull and lognormal distributions because they are often used in survival analysis is weak. Are there biologically plausible scenarios in which these distributions are appropriate given this type of data? Justify here similar to the Gamma distribution. Otherwise, omit.

We have added an example where detection rates vary across individuals, causing the marginal detection rate to decline over time.

Simulation Analyses
Line 57-59: This is the first mention of peaked and non-peaked TTDDs. Please define these

terms and how they are manifested during point-count surveys in the Introduction - including a biological justification that warrants their inclusion in simulation studies.

> We have made the meaning of 'peaked' and 'nonpeaked' more explicit in Section 4. We do not believe this distinction requires a separate explanation in the Introduction.

As a potential user of such a model presented in this manuscript, I find the simulation analyses somewhat lacking.

1. *Users will most certainly not bother using such complex modeling approach to model an intercept only - and thus, evaluating TTDDs based on intercept-only models seems overly simple and not very informative.*

   > We agree that an intercept-only model is simplistic and will not be of much practical use in applied settings. However, it can be informative. TTDDs that are biased (e.g. non-mixture lognormal) or not robust to misspecification (e.g. mixture exponential) may not be trustworthy in more complicated analyses.

2. *Simulations incorporating variable sample sizes to evaluate relative bias and precision would be very useful in terms of study design and applicability.*

   > ADAM: Investigate more/different simulations

3. *How might study design impact bias and precision? For example, length of survey relative to peaked TTDDs and pooling of time periods during data collection? These are also scenarios that would benefit readers. Alternatively, these issues could be addressed in the Discussion.*

   > ADAM: We have not investigated either of these issues. During my Masters work, we examined the effects of censoring on a mixture exponential model and found a 33% increase in credible interval widths for $p^{(det)}$ when switching from 9- to 3-interval data in a model with covariates. However, I would not dream of extending those findings to other distributions. We obviously would prefer longer durations, but those risk the violation of closure assumptions. Petit et al. (1995); Johnson (2008); Lee and Marsden (2008) and Reidy et al. (2011) all address the effects of duration. I am inclined not to add Discussion on topics about which I do not have well-informed opinions. We have added a mention of the trade-offs of a settling down period to reduce observer effects on behavior.

4. *Providing computation times for scenarios that vary by model complexity and sample size (including the computer used for analyses) may also provide a gauge for users as to whether this method is feasible for their needs.*

   > ADAM: Because of improvements in Stan, I want to run more simulation reps before quantifying this.

5. *I dont fully appreciate the need for both 50% and 90% coverage estimates. It seems*

one estimate would be sufficient, and typically 90% or 95% credible intervals are used with these types of analyses?

> We have retained just the 50% coverage. ADAM: Waiting until all new simulations are run.

Results

Page 18, Figure 2: Holy moly, the CIs of the mixture distributions are exceedingly wide given a sample size of 381 detections (which is a lot relative to many studies).

> Clarification: 947 detections at 381 sites (which is *really* a lot of detections)

For example, The Gamma distribution, touted as the most accurate based on the simulation study, suggests 95% CIs of p ranges from 0.3 to 1? And subsequently CIs around N are from 1.75 to 3.25 so we can say there are somewhere between 56 and 1,778 birds? Is that a useful estimate? There is a tradeoff between accuracy and precision happening here, and although precise, biased estimates are not what we want, any alternative needs to provide precision at levels that are still informative. This is a limitation not sufficiently addressed in the Discussion.

> We think our figure may have been unclear. The cited credible intervals are for *uncounted* birds not total abundance. We have now changed the figure to $\log_{10}$(Abundance) to avoid confusion. The gamma TTDD 95% credible interval for total abundance is (1007, 2836). This two- to three-fold scale of uncertainty is commensurate with other studies cited in this manuscript (Diefenbach et al., 2007; Reidy et al., 2011; Sólymos et al., 2013; Amundson et al., 2014).

Discussion

Page 21, Lines 12-16: I do not agree with this assessment. I think these models are appropriate for a particular subset of data - single species discrete-count surveys with a large sample size and relatively low availability (i.e., whether the animal provides a cue to an observer during the survey)

> ADAM: What is a 'discrete-count survey'? I have not found a clear reference.
> I accede the 'large sample' point. But based on our findings, I'd advocate against modeling abundance/perception from any small dataset, because analysis requires either a strong assumption about detection rates or a strong prior. Any comment on this point would be strengthened if we ran small-dataset simulations.
> I think the low-availability stipulation is redundant. If availability is really high, then a removal analysis is not necessary at all.

where non-constant detection through time is suspected based on the behavioral ecology of the species.

> ADAM: What I do not advocate is defaulting to the constant-rate assumption: by the precautionary principle, the burden of proof should rest with the analyst who wishes to assume constant detection rather than with the analyst who wishes to account for non-constant detection.

As the authors suggest, most point-count studies are designed to maximize availability, which is typically high for birds (e.g., 80-95%) with perceptibility (e.g., the probability an observer detects a cue that is given by an animal) being the much larger source of bias. Therefore, this model, without including a supplemental method to estimate perceptibility, is of limited use to many readers.

We agree. Our current research addresses this issue in more depth.

Further, although the method may be more accurate even for constant detection data, the loss in precision is such that users will likely not opt to use it unless they suspect non-constant detection through time. Further, given its presumably arduous computation times (albeit unknown), adding a perceptibility component to the model would likely greatly increase computational demands with unknown implications to bias and precision. Including a paragraph in the Discussion that outlines when the model would be most beneficial to users (e.g., what type of data, surveys, or sample sizes) and explicitly stating drawbacks (e.g., computation time, loss of precision) would greatly improve this section.

ADAM: This last recommendation seems feasible, though I think our final paragraph already addesses some of these issues.

Reviewer #2:

Although many statistical solutions to known issues in abundance estimation have been proposed to date, a plethora of relevant problems remain untreated. In this manuscript, the author(s?) put the finger on the wound, so to speak, and deal with a very relevant problem: allowing for heterogeneity in detection, or a non-constant detection rate during animal point-count surveys. This topic is relevant because, as the authors clearly state multiple times, the statistical properties of abundance estimates are sensitive to the statistical properties of the estimates of the detection probabilities. In that sense, I commend the manuscript's pertinence: the question treated is indeed relevant and needs careful examination. The advent of computer intensive approaches to estimate parameters in hierarchical models very quickly re-shaped the field of abundance estimation, and in less than a decade, we've moved from trying to explain a complicated natural signal with the simplest possible model to trying to explain a complicated natural signal with a 'realistic' but equally complicated sampling model. Then, the burden of the quality of the estimation is put into the process of the specification of the sampling model. However, there is little guarantee that the data, as considered in the paper, contains the necessary information to be able to reliably tease apart all the components of the statistical sampling model.

We appreciate the reviewers' insights and whole-heartedly agree.

In what follows, I not only expand on this topic but I outline a few major and minor comments that I hope the author(s) will regard as useful to improve the quality of the manuscript. I will certainly recommend this manuscript for publication after the author(s) include the modifications I request or she (he) successfully convinces me otherwise.

1. MAJOR COMMENTS

Diagnostics: Coverage. The idea of testing coverage is fantastic, but poorly implemented. First, the number of simulations (16 in Table 1) is exceedingly low to be able to reliably diagnose the patterns in coverage or the statistical explanation of these patterns. Second, the statistical properties of the estimator of p(det) are likely depending on the size of the true value of p(det). Therefore, I think Table 1 should have been repeated for a whole range of values of the true value of p(det)

ADAM: We should run more simulations. Reviewer 1 wants more variation in $n^{(obs)}$. Reviewer 2 wants a wide range or simulated $p^{(det)}$.

(Is a boundary like the one in Olkin et al 1981, for your case, a hard one such that below it estimation is bad, and above it estimation is "uniformly" good?).

It is not a hard boundary. It results from the relative sample mean and sample variance of large samples. However, similar thresholds continue to surface again and again in various abundance-estimating contexts:

- Veech et al. (2016): suggest $p > 0.5$ in the context of individual-level detection effects (no time-to-detection data).
- Field et al. (2016): plots indicate uncertainty in $\hat{N}$ increase rapidly as $p^{(det)}$ decreases below 0.40. Methods include removal, double-observer, distance,

and multiple-visit.

- Davis et al. (2016): report accurate $\hat{N}$ at $p^{(det)} > 0.40$ in a traditional removal-sampling context, though for small abundance ($N < 50$) they required $p^{(det)} > 0.7$.

ADAM: Jarad, I am unclear on the purpose of the Olkin-citing paragraph. I don't feel like the above arguments really merit inclusion in the text.

Without variation in "true" simulated scenarios, it is easy to inadvertendly "stack the deck" in favor or against a particular combination of settings (Particularly since your "true" value of p is high: 0.8). A wide range of simulated truths is also necessary because to be useful, these methods should be widely applicable in the tropics (low N's, low p's very often) as well as in temperate forests (large N's, large p's), for example.

Diagnostics: relation of the target parameter to other parameters in the models. Precisely because, as the authors mention, the statistical properties of the estimator of p(det) are tightly linked with the statistical properties of the abundance estimator, then the author(s) should detail explicitly how a bad coverage in p(det), for instance, affects the estimation of abundances. The author(s) mention that bias in p translates into a bias in abundance estimation, I would like to see diagnostics for the realized N values too. And to do that, two words come to mind: profile likelihoods. Unless you are explicitly adopting a subjective Bayesian approach, you basically declare a priori ignorance for your parameters. Furthermore, you state that typically you will be in a case where data sets aren't large, so the information in the data is not "swamping" the priors, so to speak.

We have posterior estimates of abundance from our model fits. We have added them to the manuscript.
ADAM: Thoughts on how to do this:

1. Generate Tables 1 & 2 for $\hat{N}$ as well as $p^{(det)}$, but this gets crowded.
2. Supplement Tables 1 & 2 with some manner of confidence intervals for $\hat{N}/N$, but averaging across replicates is not a trivial issue
3. Supplement Figure 1 with a row of representative abundance posteriors (no averaging required)
4. Following the profile likelihood comment, generate a plot of $\hat{N}$ vs. $p^{(det)}$ for a representative sample.

Lele et al 2010 (a frequent co-author of Solymos) propose a pretty neat diagnostic tool to assess estimability that is the by product of tricking a bayesian MCMC set up into Maximum Likelihood estimation for hierarchical models (see Lele et al 2007, 2010 and others. The keyword is "Data Cloning"). You are one step away from using Solymos and Lele's "Data Cloning" (DC) approach to get the full ML estimation working (See Lele et al 2010 and other Data Cloning papers). Now, this is relevant because using DC you can provide clear estimability diagnostics for the parameters of interest. Are there parameters that are technically not estimable? Given that you don't have informative priors, you are "driving in the dark" so to speak if you are not certain if some of your parameters are un-identifiable

(see Lele et al 2010 and discussion in Lele and Dennis 2009). Note that I am NOT asking you to re-do all the analysis using a ML approach, I am just suggesting the fact that using DC to get the ML estimates one can, as a very useful by product, get very neat diagnostics regarding the estimability of your parameters. Again, see Lele et al 2010. Implementing DC would only imply a simple modification to the programs you already have. And by the way, Solymos has a DC package easy to use.

ADAM:

- The Solymos package is built for BUGS and JAGS. Most of the tools are easily hand-coded, I think.
- To simplify our lives, we could choose to attempt this for only quick-running models on the actual data... there's no reason to suspect identifiability issues for the gamma TTDD that would not appear for the Weibull or lognormal.
- Technical note: they suggest using narrower priors as the number of clones increases. Likewise, we could input posterior means as initial values.

Generalized Pareto: In another area in biology, in evolutionary genetics, the "fitness" of individual variant strains has been modeled as coming from an exponential distribution. For various reasons having to do with the biology of the system, such model was quickly challenged and pretty soon papers questioning the validity of this or this other model for data akin to your waiting times abounded. There is one pretty interesting paper by Beisel et al, 2007 (Genetics, 10.1534/genetics.106.068585), written towards the epiloge of such discussions, that posits that a particular parameterization of the Generalized Pareto distribution, with a single stroke, encompassed many suitable probabilistic models in the Weibull, Gumbel and Freched domain of attraction. I wonder if the authors could write a general parameterization so that changing from one distribution to the other would simply amount to dialing a given parameter, much like Beisel et al do? This would be just a practical consideration that can at once deal with model selection and could speed up calculations

We appreciate the merit of the reviewer's suggestion and have added comments in the Discussion. We focus our comments on a possible generalized gamma TTDD, which is an umbrella distribution that includes exponential, gamma, lognormal, and Weibull distributions as special cases. The generalized Pareto analysis in Beisel et al. (2007) requires observations from the tail of the TTDD distribution, which we do not have. Additionally, its focus is on flexible domains of attraction, whereas the TTDDs in our analysis are all from the Gumbel domain of attraction. ADAM: After researching the generalized gamma, add comments in the paragraph beginning 'If the estimation of abundance is the primary interest...'.

Writing: The multiplicity of mathematical explicit meanings for the word "mixture" may result confusing. The distinction between the mixture component a la Farnsworth et al and the N-mixture models (e.g. binomial counts with N pois and p random) is clear to me, but it may not be for some of JABES' audience. I suggest re-shaping the introduction to that effect.

We have inserted clarifications: (i) when we first reference the N-Mixture frame-

work in the Introduction, and (ii) when we first use the term 'mixture model' in Section 3.3.

One thing that the authors could do is to take the reader by the hand by building equation 1 bit by bit: present first the simplest version of equation 1 and what is customary to date, and then add modifications/hierarchies plus text and build the model with explicit equations and little by little arrive to the final product, which is the current equation 1. Nothing better than clear simple math aided by short explanations to present a model unambiguously.

ADAM: This potentially requires a lot of rewiring. Two options are: (i) highlight each component of Equation (1) as we address it in the text, because we do address them all, or (ii) at the end of Section 3.2, provide a simpler version of the model without mixtures and without covariates.

Write self-contained paragraphs, with one main idea, not two or three (and each one half developed). As a reviewer, I enjoyed verifying that the cited papers in the introduction and in the discussion were indeed meaningful (by downloading them and reading those that I was not familiar with). However, when many papers are cited, it is useful to expand in the text why the different papers are being cited. Doing so not only clarifies your intentions, but does proper justice to the papers being referred to. It is also a useful exercise because it allows you to tell whether you have more than one clear ideas and hence, material for more than one paragraph. For instance, I think that the paragraph in the introduction, line 34, page 3, could be expanded and/or broken into two paragraphs. One introducing the time-to-detection as is done in survival analysis, and one presenting to the reader how data that seem not to conform to the constant-detection assumption is dealt with by using the idea of the increased detection probability via a mixture component. These are two different ideas, hence two different paragraphs. The same approach to construct paragraphs through the text should be taken.

We have reviewed the entire manuscript with these points in mind. We have split the cited paragraph and others. ADAM: The only times I feel I have cited many articles is when I've said "Many authors have done this.'.

The authors could have crafted a beautiful and very telling graph (Figure 1) representing the multinomial intervals, the TTDD, the right censoring due to the end of the observation period, and the increased detection probability mixture component. Such figure (which is what I did by hand as I was reading the manuscript) would be a very useful guide in the reading of the paper.

We have crafted a beautiful and very telling graph (Figure 1).

Process vs. observation error: Finally, here's a perspective outside the author's topic, but within statistical ecology that might result in a markedly improvement of these "N-mixture" techniques, along with all of its variants. I wonder if thinking a bit more in terms of modeling the biological process behind the data can result in a better estimation of the sampling variance and as a "by-product", some novel understanding of the biology behind the data. I suppose the author(s) are familiar (or at least have seen) the multiple papers where statistical inference is done for a stochastic population dynamics model using time series of

abundances while taking into account sampling error. In these settings, it is customary to deal with one observed abundance per time step. The sampling error and the ecologically-phrased variability are phrased using a hierarchical model. When statistical inference is done for this hieararchical model, the information to be able to tease apart the process from the observation variance lies within the structure of the temporal dependencies phrased in the model. Amazingly, with a single observation per time step, the sampling model is not ill posed as the time dependencies in the process brings about enough information in the data. As a result, thinking of the biological process (in that case, the population dynamics model) yields better, unbiased estimates of the observation error. In your case, I cannot help but wonder if having data varying along the axis of time or space result in a much better estimation of the sampling noise. Just as exponential waiting times are, as you surely are aware of, tightly linked to continuous time, discrete state Markov models, the other TTDD models could be linked to non-Markovian processes of moving/singing/behavior. In many instances, multiple observations in the same "point-count" are available. Anyways, this are just some thoughts rapidly put together, please comment on this if you think it's worthwhile.

Thank you for the opportunity to comment on your paper, I hope you find my views useful. I think yours will in the end, be a very important contribution.

References

Amundson, C. L., Royle, J. A., and Handel, C. M. (2014). A hierarchical model combining distance sampling and time removal to estimate detection probability during avian point counts. The Auk **131,** *476–494.*

Beisel, C. J., Rokyta, D. R., Wichman, H. A., and Joyce, P. (2007). Testing the extreme value domain of attraction for distributions of beneficial fitness effects. Genetics **176,** *2441–2449.*

Davis, A. J., Hooten, M. B., Miller, R. S., Farnsworth, M. L., Lewis, J., Moxcey, M., and Pepin, K. M. (2016). Inferring invasive species abundance using removal data from management actions. Ecological Applications **26,** *2339–2346.*

Diefenbach, D. R., Marshall, M. R., Mattice, J. A., Brauning, D. W., and Johnson, D. (2007). Incorporating availability for detection in estimates of bird abundance. The Auk **124,** *96–106.*

Field, C. R., Gjerdrum, C., and Elphick, C. S. (2016). How does choice of statistical method to adjust counts for imperfect detection affect inferences about animal abundance? Methods in Ecology and Evolution **7,** *1282–1290.*

Johnson, D. H. (2008). In defense of indices: the case of bird surveys. The Journal of Wildlife Management **72,** *857–868.*

Lee, D. C. and Marsden, S. J. (2008). Adjusting count period strategies to improve the accuracy of forest bird abundance estimates from point transect distance sampling surveys. Ibis **150,** *315–325.*

Petit, D. R., Petit, L. J., Saab, V. A., and Martin, T. E. (1995). Fixed-radius point counts in forests: factors influencing effectiveness and efficiency. Technical Report PSW-GTR-149, USDA Forest Service.

Reidy, J. L., Thompson, F. R., and Bailey, J. (2011). Comparison of methods for estimating density of forest songbirds from point counts. The Journal of Wildlife Management **75,** *558–568.*

Sólymos, P., Matsuoka, S. M., Bayne, E. M., Lele, S. R., Fontaine, P., Cumming, S. G., Stralberg, D., Schmiegelow, F. K., and Song, S. J. (2013). Calibrating indices of avian density from non-standardized survey data: making the most of a messy situation. Methods in Ecology and Evolution **4,** *1047–1058.*

Veech, J. A., Ott, J. R., and Troy, J. R. (2016). Intrinsic heterogeneity in detection probability and its effect on n-mixture models. Methods in Ecology and Evolution **7,** *1019–1028.*