Submitted to the Annals of Applied Statistics

A RANDOM-EFFECTS HURDLE MODEL FOR PREDICTING BYCATCH OF ENDANGERED MARINE SPECIES

By E. Cantoni* J. Mills Flemming† and A. H. Welsh‡

University of Geneva, Dalhousie University† and Australian National University‡*

Understanding and reducing the incidence of accidental bycatch, particularly for vulnerable species such as sharks, is a major challenge for contemporary fisheries management worldwide. Bycatch data, most often collected by at-sea observers during fishing trips, are clustered by trip and/or vessel and typically involve a large number of zero counts and very few positive counts. Though hurdle models are very popular for count data with excess zeros, models for clustered forms have received far less attention. Here we present a novel random-effects hurdle model for bycatch data that makes available accurate estimates of bycatch probabilities as well as other cluster-specific targets. These are essential for informing conservation and management decisions as well as for identifying bycatch hotspots, often considered the first step in attempting to protect endangered marine species. We validate our methodology through simulation and use it to analyse bycatch data on critically endangered hammerhead sharks from the U.S. National Marine Fisheries Service Pelagic Observer Program.

1. Introduction. The oceanic ecosystem is by far the largest on Earth, covering more than 70% of the planet. Human impacts on this ecosystem including overfishing, habitat destruction, pollution and climate change are causing serious conservation concern. In particular, industrial fishing has profoundly changed the biological state of the oceans and while the direct impacts of overfishing on target species are increasingly being addressed, accidental bycatch of nontarget species is a key challenge for contemporary fisheries management. Excess bycatch is particularly threatening for long-lived marine species like sharks (Lewison et al. 2004, Hall et al. 2000) so a core objective of the ecosystem approach to fisheries management is to reduce and eliminate bycatch (Pikitch et al. 2004).

Bycatch data are most often collected by at-sea observer programs and are composed of the presence (counts or mass) and absence (zeros) of non-

MSC 2010 subject classifications: Primary 62F99; secondary 62P10

Keywords and phrases: bycatch; clustered count data; excess of zeros; random-effects hurdle models; prediction.

target species along with information on vessel and gear specification, fishing effort, and environmental covariates. Specifically, we analyse bycatch data for a critically endangered marine species, the hammerhead shark obtained by Julia Baum (see Baum et al., 2003, Baum, 2007 and Myers et al., 2007 for further details) from the U.S. National Marine Fisheries Service Pelagic Observer Program (http://www.sefsc.noaa.gov/pop.jsp). In the spring of 2013 these sharks, which are commercially valuable and whose numbers have been declining dramatically in recent years, were given added protection by CITES (the Convention on International Trade in Endangered Species of Wild Fauna and Flora). We consider 1825 records of hammerhead shark bycatch from 292 fishing trips where 85% of these counts are zeros, indicating that no hammerhead sharks were caught as bycatch in many of the hauls. The few positive counts (obtained if one or more hammerhead sharks were caught as bycatch in a haul) range from 1 to 46. Counts are clustered because hauls are clustered within trips which may also be clustered within vessels, for example. The covariates considered are: year (YEAR, from 1 to 14, representing the period 1992-2005), average hook depth (AVGHKDEP, from 6.40 to 182.88 fathoms), area (4=South Atlantic Bight and 5=Mid Atlantic Bight), and season (SEASON, 464 observations in autumn, 543 in spring, 525 in summer and 293 in winter). The catch effort is measured using the logarithm of the number of hooks (TOTHOOK, ranging from 25 to 1548).

An excess of zeros is a feature of count data arising in many areas, particularly health research and ecology more broadly. For independent data, excess zeros reduce the usefulness of Poisson and negative binomial models (Welsh et al. 2000) because they underfit the probability of observing zeros. The simplest solution is a hurdle model (also two-part, zero-altered, separated or conditional model, see Mullahy 1986) or an overlapping model (or zero-inflated model, Lambert 1992). We describe these two alternative models in Section 2.

Further complications arise when we consider clustered counts with excess zeros like bycatch data. Incorporating the clusters into the analysis can be achieved via a marginal GEE approach as in Dobbie & Welsh (2001) or a conditional random-effects approach, the latter being a more natural way to account for within-cluster dependence when the interest is in within-cluster effects. Yau & Lee (2001) and Hur et al. (2002) extended overlapping models to include random effects to evaluate injury prevention strategies and model health care outcomes, respectively. Min & Agresti (2005) proposed a hurdle model with random effects for repeated measures count data with extra zeros and Liu et al. (2010) applied this type of model to correlated

medical cost data. Alfò & Maruotti (2010) used correlated random effects to analyse data on health care utilization and Neelon et al. (2013) recently presented a spatial Poisson hurdle model for emergency department visits. Finally, Molas & Lesaffre (2010) have suggested fitting random-effects hurdle models by h-likelihood. As highlighted in some of these papers, and fully demonstrated by our simulation study (Section 4), bias can be induced for fixed-effects regression coefficients when the two parts of these kinds of models are misspecified as independent.

We propose a new random-effects hurdle model framework for estimating the probability of bycatch and other management targets from bycatch data. It is applicable to any form of clustered count data with excess zeros and also readily extendable to small-area estimation problems where the variables of interest are small counts. The model has two parts: the first determines the presence or absence of bycatch in a haul and the second determines the size of the bycatch. To allow for dependence between the two parts of the model we introduce parameters which, if nonzero, indicate that the two parts are dependent: a simple classical test can be used here. For our bycatch data, we show that the two parts are dependent. We develop inferential procedures which, in contrast to all existing approaches, make available empirical best predictors of the random effects (Jiang & Lahiri 2001) and other cluster-specific targets (e.g. the probability of nonzero bycatch on a particular fishing trip). We are the first to provide a way of assessing the mean squared error of prediction of these quantities. For this we propose a new fast bootstrap procedure whose asymptotic distribution is the same as that of the maximum likelihood estimator. We apply these procedures to our data and show that bycatch of hammerhead sharks is declining through time.We also show the effect of the number of hooks, average hook depth and season on shark bycatch.

We show that our random-effects hurdle model is a powerful tool for dealing with bycatch data on endangered species. It generates reliable estimates and predictions that are essential for both understanding the processes underlying bycatch and those needed to help reduce and possibly eliminate its occurrence.

2. The Model. Here we describe our random-effects hurdle model for estimating probabilities of bycatch and other management targets from bycatch data. This model is applicable to any form of clustered count data with excess zeros and its full generalization is provided in the Supplementary Material, Section 1.

The hurdle model for independent counts of bycatch Y_i, $i = 1, \ldots, n$, can

be written as

$$P(Y_i = y_i) = \begin{cases} 1 - p(\boldsymbol{x}_i) & y_i = 0 \\ p(\boldsymbol{x}_i) f(y_i, \lambda(\boldsymbol{z}_i)) & y_i = 1, 2, 3, \ldots \end{cases}$$

where $p(\boldsymbol{x}_i)$ is the probability of crossing the hurdle, $f(y_i, \lambda(\boldsymbol{z}_i))$ is a discrete distribution on the positive integers (the truncated Poisson distribution, for example) and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are two sets of covariates, possibly overlapping. It is usual to model $p(\boldsymbol{x}_i)$ as $\text{logit}\{p(\boldsymbol{x}_i)\} = \boldsymbol{x}_i^T \boldsymbol{\alpha}$ and $\lambda(\boldsymbol{z}_i)$ as $\log\{\lambda(\boldsymbol{z}_i)\} = \boldsymbol{z}_i^T \boldsymbol{\beta}$. Alternatively, an overlapping model (often referred to as a zero-inflated Poisson or ZIP model) is a mixture model where

$$P(Y_i = y_i) = \begin{cases} \pi(\boldsymbol{x}_i) + (1 - \pi(\boldsymbol{x}_i)) \tilde{f}(0, \lambda(\boldsymbol{z}_i)) & y_i = 0 \\ (1 - \pi(\boldsymbol{x}_i)) \tilde{f}(y_i, \lambda(\boldsymbol{z}_i)) & y_i = 1, 2, \ldots \end{cases}$$

where $\tilde{f}(y_i, \lambda(\boldsymbol{z}_i))$ is a discrete distribution (the Poisson distribution, for example), $\text{logit}\{\pi(\boldsymbol{x}_i)\} = \boldsymbol{x}_i^T \widetilde{\boldsymbol{\alpha}}$ and $\log\{\lambda(\boldsymbol{z}_i)\} = \boldsymbol{z}_i^T \widetilde{\boldsymbol{\beta}}$. Min & Agresti (2002) give a good review of these models. The advantage of the hurdle model for both computation and interpretation is that it has two distinct parts that, for independent data, can be fitted and interpreted separately.

Now suppose that the data are clustered and each of the c clusters contains n_i $(i = 1, \ldots, c)$ units (e.g. hauls within trips). That is, on the jth haul of the ith trip, we observe a univariate count of bycatch y_{ij} and we consider two (possibly overlapping) sets of covariates, which can be written as $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, c$. For our bycatch data these covariates include information on gear specification (e.g. average fishing hook depth) and fishing effort (e.g. the number of fishing hooks utilized) as well as environmental information. We assume that the dependence structure in the data is described by unobserved independent random intercepts u_i and v_i. Building on the hurdle model for independent data our random-effects hurdle model specifies that, given u_i and v_i, the y_{ij} are independent with probability mass function

$$[y_{ij} \mid u_i, v_i] = \begin{cases} 1 - p(\boldsymbol{x}_{ij}, u_i) & y_{ij} = 0 \\ p(\boldsymbol{x}_{ij}, u_i) f\{y_{ij}, \lambda(\boldsymbol{z}_{ij}, u_i, v_i), \boldsymbol{\nu}\} & y_{ij} = 1, 2, 3, \ldots \end{cases} \tag{2.1}$$

where $[w|s]$ denotes the probability mass function of w given s, p is the probability of observing a positive count (i.e., "crossing the hurdle"), $f\{y_{ij}, \lambda(\boldsymbol{z}_{ij}, u_i, v_i), \boldsymbol{\nu}\}$ is the probability mass function of a discrete distribution defined on the positive integers with parameter λ which is a function of the covariates, the random effects u_i and v_i, and possibly additional nuisance parameters $\boldsymbol{\nu}$.

The hurdle model involves two random processes. For bycatch data, one process determines the presence or absence of bycatch in a haul, and in those hauls for which nonzero bycatch occurs, a second process determines the number sharks in the bycatch. Random intercepts in both random processes account for clustering in hauls during the same trip.

We model the probability of observing nonzero bycatch in the jth haul of the ith trip by

$$\text{logit}\{p(\boldsymbol{x}_{ij}, u_i)\} = \boldsymbol{x}_{ij}^T\boldsymbol{\alpha} + \sigma_u u_i. \tag{2.2}$$

We assume that the number of sharks in the nonzero bycatch event can be described using the truncated Poisson density

$$f(y_{ij}, \lambda(\boldsymbol{z}_{ij}, u_i, v_i)) = \frac{\exp(-\lambda(\boldsymbol{z}_{ij}, u_i, v_i))\lambda(\boldsymbol{z}_{ij}, u_i, v_i)^{y_{ij}}}{y_{ij}!(1 - \exp(-\lambda(\boldsymbol{z}_{ij}, u_i, v_i)))} \tag{2.3}$$

(which has no $\boldsymbol{\nu}$ parameter) and we model λ as

$$\log(\lambda(\boldsymbol{z}_{ij}, u_i, v_i)) = \boldsymbol{z}_{ij}^T\boldsymbol{\beta} + \gamma\sigma_u u_i + \sigma_v v_i. \tag{2.4}$$

We can extend what follows to incorporate other link functions in (2.2) and (2.4); the logit and log links are those most commonly used in hurdle models. In (2.2) and (2.4), $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are regression parameters, σ_u and σ_v are non-negative spread parameters and γ is a scalar parameter which controls the dependence between the random process determining the presence (or not) of bycatch $p(\boldsymbol{x}_{ij}, u_i)$ and that determining its amount $f(y_{ij}, \lambda(\boldsymbol{z}_{ij}, u_i, v_i))$. When $\gamma = 0$, p and λ are independent. Finally, we assume that the random intercepts u_i and v_i follow a $N(0, 1)$ distribution. This assumption corresponds to considering a random intercept $\tilde{u}_{i1} = \sigma_u u_i$ in the first part of the model and a random intercept $\tilde{u}_{i2} = \gamma\sigma_u u_i + \sigma_v v_i$ in the second part, with the distributional assumption $(\tilde{u}_{i1}, \tilde{u}_{i2})^T \sim N_2(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma_u^2 & \gamma\sigma_u^2 \\ \gamma\sigma_u^2 & \gamma^2\sigma_u^2 + \sigma_v^2 \end{pmatrix}.$$

This particular model as defined by (2.1)-(2.4), is equivalent to that proposed by Min & Agresti (2005), but our general formulation (see Appendix) encompasses a much larger class of models.

2.1. *Estimation.* Under the hurdle model defined by (2.1)-(2.4), for cluster i, the vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ has conditional density $[\boldsymbol{y}_i|u_i, v_i] = \prod_{j=1}^{n_i}[y_{ij}|u_i, v_i]$ and hence

$$[\boldsymbol{y}_i] = \int\int \prod_{j=1}^{n_i}[y_{ij}|u_i, v_i]\phi(u_i)\phi(v_i)du_i dv_i,$$

where ϕ denotes the density function of a $N(0,1)$ random variable. It follows that the likelihood for $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \sigma_u, \boldsymbol{\beta}, \sigma_v, \gamma)$, the vector of all the parameters, is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{c}[\boldsymbol{y}_i] = \prod_{i=1}^{c}\int\int\prod_{j=1}^{n_i}[y_{ij}|u_i,v_i]\phi(u_i)\phi(v_i)du_idv_i \\ &= \prod_{i=1}^{c}\int\int\exp\left(\sum_{j=1}^{n_i}\log\{1-p(\boldsymbol{x}_{ij},u_i)\}\right. \\ &\quad + \sum_{j=1}^{n_i}\iota(y_{ij}>0)\log[p(\boldsymbol{x}_{ij},u_i)/\{1-p(\boldsymbol{x}_{ij},u_i)\}] \\ &\quad \left. + \sum_{j=1}^{n_i}\iota(y_{ij}>0)\log f(y_{ij},\lambda(\boldsymbol{z}_{ij},u_i,v_i))\right)\phi(u_i)\phi(v_i)du_idv_i \end{aligned}$$

where $\iota()$ is an indicator function. The advantage of using the likelihood is that we can apply standard asymptotic theory to obtain the fixed-effects estimates for the covariates appearing in the two parts of the model. See, for example, Theorem 2.1 in the Supplementary Material, Section 2. In our case, the maximisation of the likelihood is complicated by the integrals. Many approaches exist for computing the likelihood, including analytical approximation techniques like the Laplace approximation (De Bruijn 1981, Huber et al. 2004) or adaptive Gaussian quadrature used in Rabe-Hesketh et al. (2002), data cloning (Lele et al. 2007) as well as Monte Carlo approaches such as simple Monte Carlo or importance sampling. We use simple Monte Carlo in this paper and approximate the likelihood by

$$\begin{aligned} L(\boldsymbol{\theta}) \simeq \tilde{L}(\boldsymbol{\theta}) &= \frac{1}{K^c}\prod_{i=1}^{c}\sum_{k=1}^{K}\exp\left(\sum_{j=1}^{n_i}\log\{1-p(\boldsymbol{x}_{ij},u_k^*)\}\right. \\ &\quad + \sum_{j=1}^{n_i}\iota(y_{ij}>0)\log[p(\boldsymbol{x}_{ij},u_k^*)/\{1-p(\boldsymbol{x}_{ij},u_k^*)\}] \\ &\quad \left. + \sum_{j=1}^{n_i}\iota(y_{ij}>0)\log f(y_{ij},\lambda(\boldsymbol{z}_{ij},u_k^*,v_k^*))\right), \end{aligned}$$

where K is the number of Monte Carlo replications and u_k^* and v_k^* are independent realizations of random $N(0,1)$ variables.

We maximise the approximated log-likelihood $\log(\tilde{L}(\boldsymbol{\theta}))$ numerically, using the function `optim` in R (R Development Core Team 2011) and use the inverse of the corresponding Hessian matrix to estimate the variances of

all of the parameter estimates in $\boldsymbol{\theta}$, namely the fixed-effect parameters, the variance components and γ as well as associated confidence intervals. See the help file for `optim` in R for additional computational details.

2.2. *Prediction.* With bycatch data many quantities need to be predicted at the cluster-specific level or estimated marginally with respect to the clusters. This requires predictions of the random effects u_i and v_i, hereafter denoted by u and v for ease of notation, along with expressions for the mean and variance of the response under our random-effects hurdle model. The mean and variance of the truncated Poisson distribution of the positive observations (see (2.3)) are

$$m\{\lambda(\boldsymbol{z}_{ij}, u, v)\} = \lambda(\boldsymbol{z}_{ij}, u, v)/[1 - \exp\{-\lambda(\boldsymbol{z}_{ij}, u, v)\}]$$

and

$$\begin{aligned}\operatorname{var}\{\lambda(\boldsymbol{z}_{ij}, u, v)\} =& \lambda(\boldsymbol{z}_{ij}, u, v)/[1 - \exp\{-\lambda(\boldsymbol{z}_{ij}, u, v)\}] \\ &- \lambda^2(\boldsymbol{z}_{ij}, u, v)\exp\{-\lambda(\boldsymbol{z}_{ij}, u, v)\}/[1 - \exp\{-\lambda(\boldsymbol{z}_{ij}, u, v)\}]^2\end{aligned}$$

respectively. The expected bycatch, and its variance during the jth haul of the ith trip are given by the conditional mean and variance of the count Y_{ij} given u, v. That is,

$$\mathrm{E}(Y_{ij}|u, v) = p(\boldsymbol{x}_{ij}, u)m\{\lambda(\boldsymbol{z}_{ij}, u, v)\}$$

and

$$\begin{aligned}\operatorname{var}(Y_{ij}|u, v) =& p(\boldsymbol{x}_{ij}, u)\operatorname{var}\{\lambda(\boldsymbol{z}_{ij}, u, v)\} \\ &+ p(\boldsymbol{x}_{ij}, u)\{1 - p(\boldsymbol{x}_{ij}, u)\}m\{\lambda(\boldsymbol{z}_{ij}, u, v)\}^2.\end{aligned}$$

We also need to predict the probability of nonzero bycatch for a particular haul of a trip

$$P(Y_{ij} > 0|u) = p(\boldsymbol{x}_{ij}, u),$$

and the expected number of sharks in the nonzero bycatch

$$\mathrm{E}(Y_{ij}|Y_{ij} > 0, u, v) = m\{\lambda(\boldsymbol{z}_{ij}, u, v)\}.$$

Analogous marginal estimates are also of interest. These are obtained by integrating the analogous cluster-specific quantities over u and v. Some examples are the probability of nonzero bycatch defined by

$$P(Y_{ij} > 0) = \int p(\boldsymbol{x}_{ij}, u)\phi(u)du,$$

the expected number of sharks in the nonzero bycatch

$$\mathrm{E}(Y_{ij}|Y_{ij}>0) = \int\int m\{\lambda(\boldsymbol{z}_{ij},u,v)\}\phi(u)\phi(v)dudv,$$

and finally the expected bycatch often used as a proxy for abundance

$$\mathrm{E}(Y_{ij}) = \int\int p(\boldsymbol{x}_{ij},u)m\{\lambda(\boldsymbol{z}_{ij},u,v)\}\phi(u)\phi(v)dudv.$$

A unified treatment is possible since the cluster-specific prediction targets of interest are all of the form $t(u,v,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})$, and the marginal estimation targets are

$$\int\int t(u,v,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})\phi(u)\phi(v)dudv. \tag{2.5}$$

The marginal estimation targets can be estimated by substituting the estimated parameters $\hat{\boldsymbol{\theta}}$ into expression (2.5) and evaluating the integral by Monte Carlo integration. To assess their precision, we need to compute at least an approximation to their standard errors. We treat the integral as being approximated to a high order, and obtain approximate standard errors as $(\hat{\boldsymbol{\delta}}^T\hat{V}\hat{\boldsymbol{\delta}})^{1/2}$, where $\hat{V}$ is the estimated variance of $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\delta}}$ is obtained by evaluating

$$\boldsymbol{\delta} \;=\; \int\int \partial_{\boldsymbol{\theta}}\{t(u,v,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})\}\phi(u)\phi(v)dudv$$

at $\hat{\boldsymbol{\theta}}$, where $\partial_{\boldsymbol{\theta}}$ means the derivative with respect to $\boldsymbol{\theta}$. The integrals in $\hat{\boldsymbol{\delta}}$ can be evaluated using the same methods as for estimating the targets.

Two main approaches exist for predicting functions of random effects, see for example Section 3.6.2 of Jiang (2007). The first approach uses the predictor $t(\hat{u},\hat{v},\boldsymbol{x},\boldsymbol{z},\hat{\boldsymbol{\theta}})$, where $\hat{u}$ and $\hat{v}$ are predictors of u and v. For example, $\hat{u}$ and $\hat{v}$ may be the values that maximize $[u,v|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_c]$, referred to as the conditional modes. This approach is used by Breslow & Clayton (1993), Lee & Nelder (1996), Jiang et al. (2001), and, to some extent, Booth & Hobert (1998) who also proposed using a conditional prediction mean squared error to measure variability. It is a straightforward approach for prediction in clusters from which we have observations (and hence estimates of their random effects) but it is not clear how to proceed for clusters for which $\boldsymbol{y}_i$ is not observed. A more satisfactory approach uses the minimum mean squared error predictor or "best predictor" of Jiang (2003) which, by Bayes'

Theorem, is

$$\begin{aligned} T_t(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{y}_i) &= \int\int t(u,v,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})[u,v|\boldsymbol{y}_i]dudv \\ &= \frac{\int\int t(u,v,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})[\boldsymbol{y}_i|u,v]\phi(u)\phi(v)dudv}{\int\int[\boldsymbol{y}_i|u,v]\phi(u)\phi(v)dudv}. \end{aligned}$$

The expression for $T_t(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{y}_i)$ shows that the best predictor for $t(u,v,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})$ is a ratio of integrals. These integrals can be estimated using the same methods as for the likelihood. By Monte Carlo approximation we obtain the empirical best predictor (EBP)

$$\hat{T}_t(\boldsymbol{x},\boldsymbol{z},\hat{\boldsymbol{\theta}};\boldsymbol{y}_i) = \frac{\sum_{k=1}^K t(u_k^*,v_k^*,\boldsymbol{x},\boldsymbol{z},\hat{\boldsymbol{\theta}})[\boldsymbol{y}_i|u_k^*,v_k^*]}{\sum_{k=1}^K[\boldsymbol{y}_i|u_k^*,v_k^*]},$$

where u_k^* and v_k^* are sampled from independent $N(0,1)$ distributions. We used the same u_k^* and v_k^* in the numerator and in the denominator to reduce computation; this also reduces the Monte Carlo variability of the predictor.

The mean squared error of prediction for the empirical best predictor $\hat{T}_t(\boldsymbol{x},\boldsymbol{z},\hat{\boldsymbol{\theta}};\boldsymbol{y}_i)$ is

$$\text{msep}_{ij}(\hat{T}_t,t) = E_{u,v,\boldsymbol{y}_i}\{\hat{T}_t(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij},\hat{\boldsymbol{\theta}};\boldsymbol{y}_i) - t(u,v,\boldsymbol{x}_{ij},\boldsymbol{z}_{ij},\boldsymbol{\theta})\}^2. \tag{2.6}$$

This quantity is not straightforward to estimate, arguably the main reason these predictors have not been much used in practice. We tried to follow Jiang (2007, p. 156) and linearize $\hat{T}_t(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij},\hat{\boldsymbol{\theta}};\boldsymbol{y}_i) - \hat{T}_t(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij},\boldsymbol{\theta};\boldsymbol{y}_i)$ around $\boldsymbol{\theta}$ and then use the additional approximations suggested by him to simplify the expressions, but this approach did not produce sensible results likely because of the series of approximations involved (in particular the "trick" producing Jiang's formula (3.57)). As an alternative here one could use the jackknife (Jiang et al. 2002). A second option to estimate the mean squared error of prediction (2.6) is the parametric bootstrap approach which is conceptually straightforward and can be used in the following way:

- Compute the estimate $\hat{\boldsymbol{\theta}}$ from the data.
- For $b = 1,\ldots,B$
  - use the parametric bootstrap to generate $\hat{\boldsymbol{\theta}}_b^*$ from (2.1)–(2.4)
  - generate each of $u_{b1}^*, u_{b0}^*, v_{b1}^*, v_{b0}^*$ independently from $N(0,1)$ and $\boldsymbol{y}_{bi}^*$ from $[\boldsymbol{y}_i|u_{b1}^*,v_{b1}^*]$.
- Compute the bootstrap estimate of $\text{msep}_{ij}(\hat{T}_t,t)$ as

$$\frac{1}{B}\sum_{b=1}^B\{\hat{T}_t(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij},\hat{\boldsymbol{\theta}}_b^*;\boldsymbol{y}_{bi}^*) - t(u_{b0}^*,v_{b0}^*,\boldsymbol{x}_{ij},\boldsymbol{z}_{ij},\hat{\boldsymbol{\theta}})\}^2. \tag{2.7}$$

By randomly generating u, v and $\boldsymbol{y}_i$ we are taking into account all of the sources of variability (they are all random variables) in the mean squared error of prediction.

The full parametric bootstrap option above is simple but the repeated estimation of $\boldsymbol{\theta}$ makes it very time-consuming to implement. We therefore developed a fast bootstrap scheme based on turning the estimating equation which defines our estimator $\widehat{\boldsymbol{\theta}}$ into a fixed-point equation and then using an adjusted one-step bootstrap estimator (Salibian-Barrera et al. 2008). This approach is described in detail in the Supplementary Material, Section 3 along with a theorem that states that the asymptotic distribution of the fast bootstrap estimator is the same as that of the maximum likelihood estimator. This holds when the model is correct and regular (so we can interchange the order of integration and differentiation).

3. Analysis of the Bycatch Data. We fitted our random-effects hurdle model (2.1)–(2.4) (hereafter referred to as our *dependent hurdle model*) to the hammerhead shark bycatch data. For the Monte Carlo approximation to the log-likelihood, we used $K = 5000$ random points to evaluate the integrand. (Generally, $K = 1000$ provides a sufficiently good approximation, but we chose to be conservative here.) We maximized the log-likelihood from 30 distinct starting points for the parameters generated using the package `fields` (see Furrer et al. 2012) and retained the solution with the largest log-likelihood. We then used our fast bootstrap procedure with $B = 1000$ to obtain estimates of the variability of the parameter estimates and the predicted functions of random effects.

We also fitted the two parts of the hurdle model separately, hereafter referred to as the *independent hurdle model*, using the R package `glmmADMB` (Skaug et al. 2012, Fournier et al. 2012).

Table 1 presents estimates of the fixed-effect regression parameters and the parameters describing the random structure for both the dependent and independent hurdle models. For our dependent hurdle model, two sets of standard errors and confidence intervals are provided: those based on the numerical Hessian matrix and those obtained from the fast bootstrap, these are in agreement. The dependence parameter γ of our dependent hurdle model is estimated at 1.116 and is significantly different from 0, indicating that the two parts of the model are indeed correlated. This information is important as it implies that it is (i) incorrect to use the independent model for these data and (ii) inappropriate to consider only the nonzero bycatch events, both of which are often done in practice without first testing for dependence. As we will see below, doing (i) or (ii) can translate into incorrect

TABLE 1

Estimated coefficients and standard errors with significant effects shown in bold. (SE-H, standard errors from the numerical Hessian; SE-b, standard errors from the bootstrap; CI-H, 95% confidence interval based on normal approximation with SE-H, CI-b, 95% confidence interval based on normal approximation with SE-b.)

	Dependent hurdle model					Independent hurdle model		
				Presence-absence				
Variable	Coeff.	SE-H	SE-b	CI-H	CI-b	Coeff.	SE	CI
Intercept	−2.123	1.453	1.636	(−4.970; 0.724)	(−5.330; 1.084)	−1.951	1.508	(−4.908; 1.006)
YEAR	−0.059	0.028	0.027	**(−0.115; −0.004)**	**(−0.113; −0.006)**	−0.044	0.030	(−0.103; 0.015)
AVGHKDEP	−0.011	0.011	0.016	(−0.031; 0.010)	(−0.042; 0.021)	0.007	0.010	(−0.011; 0.026)
AREA5	−0.241	0.242	0.317	(−0.716; 0.234)	(−0.862; 0.381)	−0.053	0.254	(−0.552; 0.446)
SEASONspring	1.609	0.341	0.351	**(0.940; 2.277)**	**(0.920; 2.297)**	1.630	0.358	**(0.928; 2.333)**
SEASONsummer	0.074	0.362	0.372	(−0.635; 0.784)	(−0.655; 0.803)	0.096	0.366	(−0.622; 0.814)
SEASONwinter	1.068	0.369	0.342	**(0.345; 1.792)**	**(0.397; 1.739)**	0.950	0.393	**(0.180; 1.720)**
log(TOTHOOK)	−0.008	0.212	0.206	(−0.424; 0.407)	(−0.412; 0.396)	−0.170	0.222	(−0.606; 0.267)
σ_u	1.413	0.145	0.134	**(1.129; 1.698)**	**(1.150; 1.676)**	1.387	n.a.	–
				Abundance				
Intercept	−4.871	0.661	0.678	**(−6.167; −3.576)**	**(−6.200; −3.542)**	−3.322	1.148	**(−5.571; −1.072)**
YEAR	−0.132	0.032	0.029	**(−0.195; −0.069)**	**(−0.187; −0.076)**	−0.105	0.042	**(−0.188; −0.023)**
AVGHKDEP	−0.067	0.008	0.013	**(−0.084; −0.051)**	**(−0.092; −0.043)**	−0.052	0.013	**(−0.077; −0.027)**
AREA5	−0.151	0.230	0.257	(−0.602; 0.300)	(−0.654; 0.352)	−0.182	0.249	(−0.669; 0.306)
SEASONspring	0.850	0.362	0.236	**(0.139; 1.560)**	**(0.387; 1.312)**	−0.121	0.519	(−1.138; 0.896)
SEASONsummer	−0.435	0.488	0.401	(−1.392; 0.522)	(−1.221; 0.350)	−0.843	0.590	(−1.998; 0.312)
SEASONwinter	1.278	0.342	0.226	**(0.608; 1.948)**	**(0.835; 1.721)**	0.288	0.574	(−0.836; 1.413)
log(TOTHOOK)	1.020	0.096	0.122	**(0.831; 1.208)**	**(0.780; 1.259)**	0.944	0.171	**(0.608; 1.280)**
σ_v	1.248	0.144	0.179	**(0.966; 1.530)**	**(0.897; 1.599)**	1.544	n.a.	–
γ	1.116	0.159	0.157	**(0.804; 1.429)**	**(0.808; 1.425)**	–	–	–

conservation and management decisions.

For both models, the coefficient of each covariate summarizes the effect of that covariate after adjusting for the contributions of the other covariates, and further, some that are significant (based on the confidence interval containing zero or not) in the abundance part are not significant in the presence-absence part. The latter is quite common in ecological problems; more factors tend to affect the abundance process than the presence-absence process. Table 1 also shows that fitting the independent hurdle model (rather than the dependent one) would lead us to mistakenly conclude that there is no effect of YEAR in the presence-absence part of the model and no effect of SEASON in the abundance part of the model. Further we would also underestimate the size of nonzero bycatch events.

In the dependent hurdle model, we find that YEAR is significant in both parts and has a negative sign. This means that hammerhead sharks are being (or at least reported as being) caught as bycatch less often through time, and additionally, when they are caught as bycatch there are fewer

of them. That is, with each additional year the adjusted odds of observing a nonzero bycatch event decrease by $1 - \exp(-0.059) = 5.7\%$ (but not at all according to the independent hurdle model), whereas λ is reduced by $1 - \exp(-0.132) = 12.4\%$ (but only by 10% according to the independent hurdle model) so the expected number of sharks in the nonzero bycatch $m(\lambda) = \lambda/\{1 - \exp(-\lambda)\}$ is reduced. Initially one might interpret positively this reduction in the number of nonzero bycatch events through time. However, the second part of the model indicates that the number of sharks in the bycatch is reducing even more quickly. Both are cause for alarm as they suggest a decrease in abundance of this critically endangered species (assuming fishing practices have remained stable). SEASON too plays a role in both parts, with spring and winter significantly different from the autumn reference. The catch effort (log(TOTHOOK)) does not impact the presence-absence part, but is significant in the abundance part. Its estimated coefficient is very close to 1. This value is consistent with using log(TOTHOOK) as an offset as is often done in statistical analyses of catch data. The hook depth (AVGHKDEP) impacts the number of sharks in the bycatch (given it is nonzero) significantly. Such information is useful to managers for determining time-area closures of fisheries for the prevention/reduction of bycatch.

For the independent hurdle model predictions of v_i are only possible for those clusters which have at least one positive outcome. Desirably, with our dependent hurdle model we can predict v_i for all the clusters even those trips that didn't report any bycatch. This is a nice feature given the reduction in bycatch events that we are seeing through time. Figure 1 displays the predicted values for the random components u_i and v_i, for all of the $i = 1, \ldots, 292$ trips (clusters). The predictions for v_i are in general smaller in magnitude than the predictions for u_i and sometimes very close to zero. These very small $\hat{v}_i$ correspond to negative $\hat{u}_i$ (as can be seen from the leftmost panel of Figure 1). The clusters for which this happens are clusters whose responses are all equal to zero (172 and all have $\hat{v}_i$ close to zero). The $\hat{u}_i$ for these clusters are negative, which reduces the estimated probability of crossing the hurdle.

The predicted values (black dots) are shown in the two rightmost panels of Figure 1 with their corresponding prediction intervals (constructed by subtracting and adding 1.96 times the square-root of the msep estimates). The results are presented with the $\hat{u}_i$ and $\hat{v}_i$ ordered separately. By examining the predicted u_i and v_i in the lower and upper tails we can look for structure related to covariates. We found the pattern to be mainly seasonal. That is, small values of $\hat{u}_i$ (which reduce the probability of crossing the

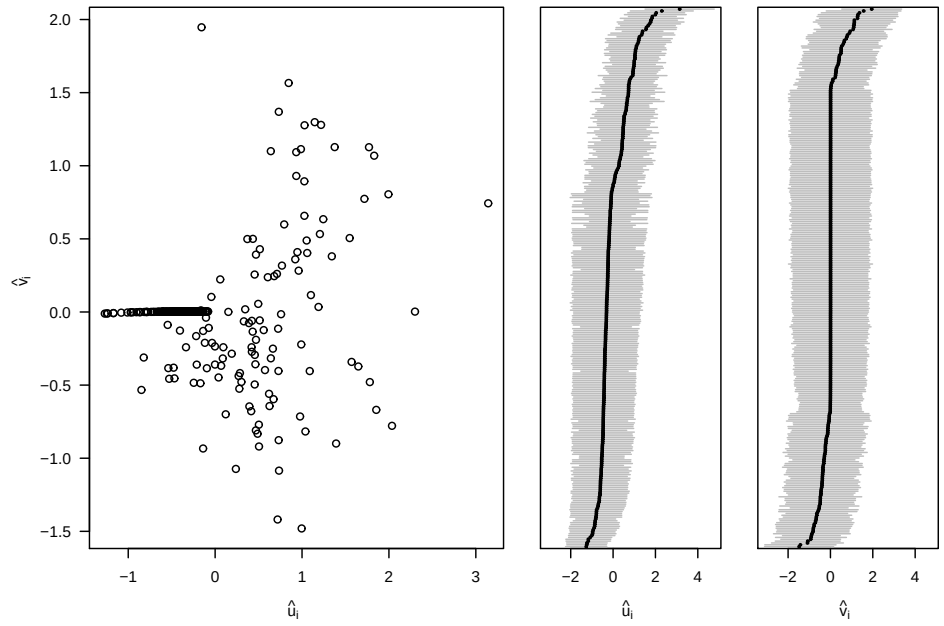

FIG 1. *Predictions (leftmost panel) of the random components u_i and v_i for the $i = 1, \ldots, 292$ trips. Prediction intervals for the ordered predictions of u_i (middle panel) and v_i (rightmost panel) for $i = 1, \ldots, 292$.*

hurdle) tended to be associated with spring and winter rather than summer and fall.

Figure 2 shows the predicted probability of nonzero bycatch $P(Y_{ij} > 0|u_i, v_i)$ and the conditional expectations $E(Y_{ij}|Y_{ij} > 0, u_i, v_i)$ with their corresponding prediction intervals ($\pm 1.96\sqrt{msep}$) for the first five fishing trips. In the left panel of Figure 2, trip 01A019 suggests two groups of predictions. The covariates of the observations for this trip only differ for average hook depth (AVGHKDEP) and catch effort (log(TOTHOOK)), with the coefficient of the latter being virtually zero in the presence-absence part of the model. For this trip AVGHKDEP has two values, one resulting in both higher predicted probabilities of bycatch and larger expected counts. This is useful information, for bycatch mitigation. In the right panel of Figure 2, the length of the prediction intervals can be quite variable. For trip 01A019, the two much larger prediction intervals correspond to a combination of AVGHKDEP and log(TOTHOOK), which is significant in the abundance part of the model and therefore has an impact on the estimation of

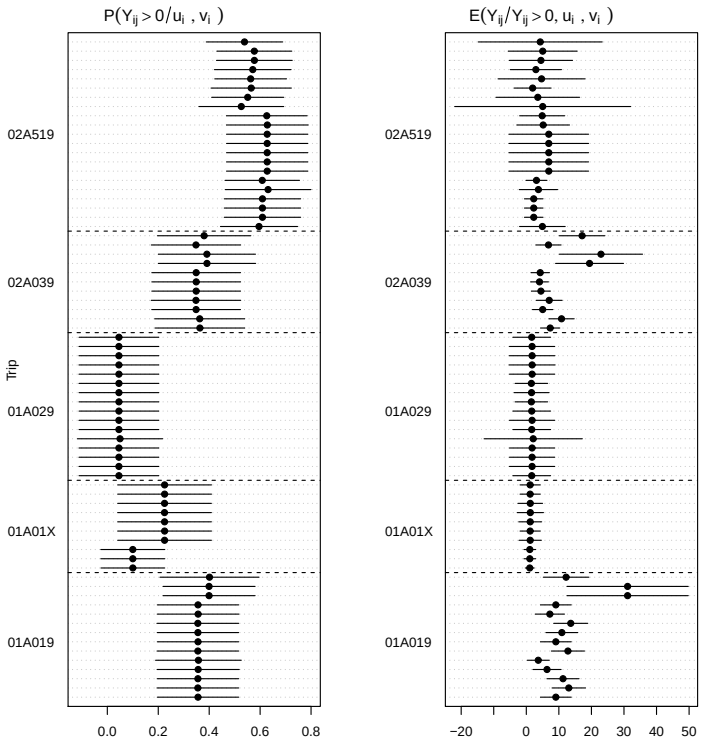

FIG 2. *Prediction intervals for the predictions of* $P(Y_{ij} > 0|u_i, v_i)$ *(left) and* $E(Y_{ij}|Y_{ij} > 0, u_i, v_i)$ *(right) for trips* $i = 1, \ldots, 5$.

$E(Y_{ij}|Y_{ij} > 0, u_i, v_i)$. Similarly, the longer prediction interval for trip 01A029 corresponds to its dissimilar value of AVGHKDEP. The smaller variations in length are associated with the values of log(TOTHOOK).

To better understand the coverage of the prediction intervals in Figure 2 we revisit the random-effects' distributional assumptions. In Table 1 we see that the estimates for σ_u and σ_v are quite large with σ_v overestimated in the independent hurdle model. McCulloch & Neuhaus (2011) have suggested that for the goal of predicting the random effects one can expect only modest impacts on the mean squared error of prediction due to misspecification. For confirmation we did a sensitivity analysis (as summarized in Figure 3) where we assume mixtures of normal distributions for the predicted random effects (as suggested by their empirical distributions) and found that the estimated coefficients and corresponding errors are generally robust to these misspecifications.

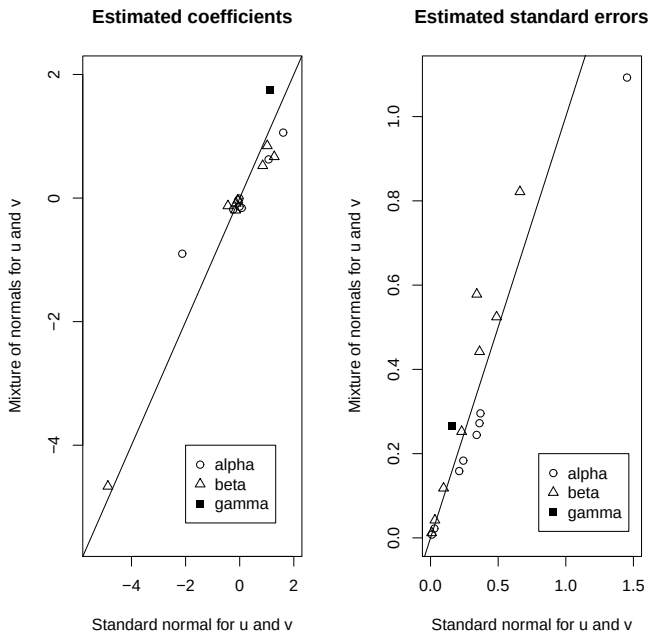

FIG 3. *Estimates (left panel) and standard errors (right panel) for the intercept and variables (YEAR, AVGHKDEP, AREA5, SEASON, and log(TOTHOOK)) corresponding to two different specifications of the random-effects distribution.* α *and* β *correspond to the presence-absence and abundance part of the model, respectively.*

4. Simulation Study. We carried out a simulation study to assess whether parameters are estimated accurately using our methodology as well as to understand the properties of our cluster-specific predictions. We simulated data from our hurdle model (2.1)–(2.4). Each simulated data set comprised $c = 100$ clusters, half with 5 measurements and half with 10 measurements per cluster, for a total of 750 observations. We included in $\boldsymbol{x}_{ij}$ an intercept, a $N(0,1)$ covariate and a Bernoulli(1/2) covariate (all independent of each other). The covariates $\boldsymbol{z}_{ij}$ included an intercept, the same $N(0,1)$ variable as in $\boldsymbol{x}_{ij}$, a Bernoulli(1/2) variable and another N(0,1) variable, so that $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ were partially overlapping. For each of 400 simulations, we used $K = 1000$ for the Monte Carlo approximation to the likelihood, 10 starting points for its numerical optimisation and took $B = 1000$ in the fast bootstrap.

For the parameters, we considered four settings:

- Setting I: $\boldsymbol{\alpha} = (0.5, 0.5, 1)^T$, $\boldsymbol{\beta} = (-0.5, 0.5, 1, 0.5)$, $\gamma = 1$, $\sigma_u = 0.75$ and $\sigma_v = 0.5$, which gives 30% zeros on average.
- Setting II: Same as Setting I but with $\gamma = 0$.
- Setting III: $\boldsymbol{\alpha} = (-2, 0.5, 1)^T$, $\boldsymbol{\beta} = (-0.5, 0.5, 1, 0.5)$, $\gamma = 1$, $\sigma_u = 0.75$ and $\sigma_v = 0.5$, which gives 75% zeros on average.
- Setting IV: Same as Setting III, but with $\gamma = 0$.

Setting III produces smaller, simplified versions of the bycatch data; the other settings are included to allow comparison with simpler situations.

4.1. *Results for parameter estimation.* In interpreting the results of fitting models for data with excess zeros, it is important to keep in mind that, although in general it is more difficult to fit models with random effects to binary data (i.e. the presence-absence part) than to count data (i.e. the abundance part), fewer observations contribute to estimation of the parameters in the abundance part of the model than the presence-absence part, so, in general, it tends to be more difficult to estimate and make inferences about the abundance parameters.

Figure 4 presents boxplots of the sampling distributions of the centered parameter estimates for Settings III and IV (analogous results for Settings I and II are given in the Supplementary Material, Figure 1). For the dependent hurdle model, all the regression parameters are estimated unbiasedly in all four settings. The dependence parameter γ is also estimated approximately unbiasedly, but has quite large variability. The spread parameters σ_u and σ_v are slightly underestimated on average when $\gamma \neq 0$ (Settings I and III), but this is expected given the negative bias associated with maximum likelihood estimation of variance components. The larger bias and variance

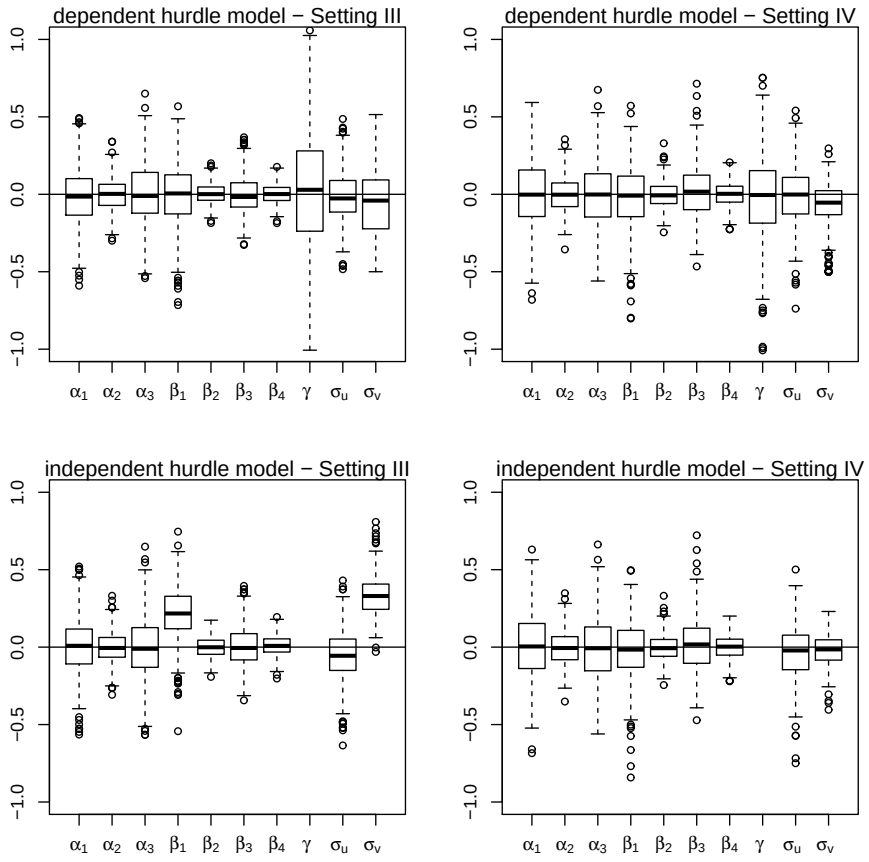

FIG 4. *Setting III and IV: boxplots of* $(\hat{\theta}_l - \theta_l)$ *for* $l = 1, \ldots, 10$.

in the estimates of σ_v relative to σ_u are due to the smaller contributing sample sizes.

For the independent hurdle model, some of the regression parameter estimates are biased when $\gamma \neq 0$ (Setting I) and particularly in Setting III when the proportion of zeros is larger. This finding agrees with observations by Su et al. (2009), who considered two-part models for semicontinuous data, as well as those of Fulton et al. (2015), who modelled multivariate binary responses. As noted, an incorrect assumption of independence between the random parts of the model produces biases in the parameter estimates, in particular, the intercept for the abundance part, because correlated random effects are informative about cluster size (since parameters in the binary part influence the number of observations in the abundance part of the model).

TABLE 2

The couple (l, r) represents the percentage of confidence intervals that miss the true value, on the left (l) and on the right side (r), for a nominal 95% confidence interval. Since 400 simulations were run each percentage must be a multiple of 0.25%. If the true value of l and r is 2.5% then the simulation standard error of their estimates is 0.78 percentage points. Similarly, if the true coverage is 95% the simulation standard error of an estimate of coverage or non-coverage is 1.09 percentage points. (SE-H, standard errors from the numerical Hessian; SE-b, standard errors from the bootstrap; boot., bootstrap percentile method.)

	Setting III				Setting IV			
	dependent hurdle model			indep. hurdle	dependent hurdle model			indep. hurdle
	SE-H	SE-b	boot.		SE-H	SE-b	boot.	
α_1	(3.75, 1.75)	(3.75, 1.50)	(2.00, 1.25)	(3.00, 2.75)	(4.00, 1.50)	(4.25, 2.00)	(1.75, 1.50)	(4.50, 1.50)
α_2	(1.75, 2.50)	(2.50, 2.75)	(0.50, 3.00)	(2.75, 3.25)	(2.75, 3.50)	(2.25, 4.25)	(1.25, 4.25)	(2.50, 3.75)
α_3	(2.50, 3.00)	(3.00, 3.25)	(0.50, 3.50)	(3.25, 4.25)	(2.25, 3.50)	(3.00, 3.00)	(0.75, 4.00)	(2.00, 3.25)
σ_u	(2.00, 3.75)	(3.00, 5.75)	(1.25, 5.75)	n.a.	(1.25, 3.25)	(1.50, 3.25)	(0.50, 3.25)	n.a.
β_1	(3.25, 3.00)	(3.75, 3.50)	(2.00, 3.25)	(2.50, 2.75)	(2.00, 3.00)	(2.25, 4.00)	(1.25, 3.25)	(2.75, 3.00)
β_2	(2.75, 2.50)	(4.00, 5.00)	(1.75, 5.50)	(2.75, 3.00)	(2.25, 3.00)	(3.50, 3.75)	(1.50, 3.25)	(2.00, 3.25)
β_3	(2.75, 1.50)	(4.25, 4.50)	(1.25, 4.50)	(3.50, 3.00)	(2.75, 2.50)	(4.25, 3.25)	(1.00, 3.50)	(2.50, 2.50)
β_4	(3.00, 1.25)	(5.00, 3.75)	(1.25, 4.25)	(2.25, 3.00)	(3.25, 4.00)	(3.50, 4.50)	(1.25, 4.00)	(3.00, 4.00)
σ_v	(13.50, 15.25)	(17.25, 30.00)	(12.50, 30.75)	n.a.	(0.50, 5.00)	(1.25, 10.00)	(0.50, 9.25)	n.a.
γ	(1.25, 14.50)	(5.50, 16.50)	(2.00, 17.00)	–	(2.00, 3.50)	(3.50, 6.25)	(0.75, 6.00)	–

Further, when $\gamma = 0$ (Settings II and IV), the independent hurdle model is correct, but our dependent hurdle model performs as well as the independent hurdle model.

Table 2 shows the complement of coverage of 95% confidence intervals for Settings III and IV (analogous results for Settings I and II are given in the Supplementary Material, Table 1). For the dependent model such intervals are constructed using (i) a normal approximation with standard error estimates obtained from either the numerical Hessian or the bootstrap, or (ii) the bootstrap percentile method. For the independent hurdle model, a normal approximation is used with the (numerical) standard errors from the `glmmADMB` output. For the dependent hurdle model, the three methods give similar results. For the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the confidence intervals have good coverage and are fairly symmetric. Results are less satisfactory for the parameters related to the random-effects structure where missing to the right is more probable (due to the underestimation of the variances) and the coverage is below the 95% nominal level. These parameters are more difficult to estimate; in particular σ_v and γ are more variable because they are estimated only from the nonzero observations which are a small proportion of the total number of observations. In Setting IV, where $\gamma = 0$, the actual coverage is better. For the independent hurdle model, only the standard errors for the regression parameters are available from

`glmmADMB`. The same comments apply for Settings I and II.

One clear advantage of our model is the ability to perform tests on γ. In fact, our simulation results support the use of a simple significance test (t-test). For a 5% nominal level, the actual level of such a test can be deduced from the confidence interval results. That is, for Setting II and IV (based on the standard errors from the numerical Hessian, for example) the actual levels are: 6.5% for Setting II and 5.5% for Setting IV. In cases where γ is found to be non-zero we should always favor our dependent hurdle model as failing to do so by using instead the independent hurdle model could lead to incorrect conclusions regarding the fixed-effects. We did explore both likelihood ratio testing and information criterion based procedures as alternatives here but difficulties in establishing their distributions (in particular the appropriate degrees of freedom) necessarily precluded their use.

4.2. *Results for prediction.* For a prediction target $t(u_i, v_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{\theta})$ we decompose the mean squared error of prediction as

$$\begin{aligned}
\text{msep}_i &= E\left\{\hat{T}_t(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \hat{\boldsymbol{\theta}}; \boldsymbol{y}_i) - t(u_i, v_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{\theta})\right\}^2 \\
&= \mathrm{E}\left[\hat{T}_t(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \hat{\boldsymbol{\theta}}; \boldsymbol{y}_i) - \mathrm{E}\{\hat{T}_t(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \hat{\boldsymbol{\theta}}; \boldsymbol{y}_i)\}\right]^2 \\
&\quad + \left[\mathrm{E}\{\hat{T}_t(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \hat{\boldsymbol{\theta}}; \boldsymbol{y}_i)\} - \mathrm{E}\{t(u_i, v_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{\theta})\}\right]^2 \\
&\quad + \mathrm{E}\left[\mathrm{E}\{t(u_i, v_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{\theta})\} - t(u_i, v_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{\theta})\right]^2 \\
&= \mathrm{se}\{\hat{T}_t(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \hat{\boldsymbol{\theta}}; \boldsymbol{y}_i)\}^2 + \mathrm{bias}^2 + \mathrm{sd}\{t(u_i, v_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{\theta})\}^2,
\end{aligned}$$

and we estimate se, bias and sd empirically via 5%-trimmed means of the 400 predictions obtained by simulation. Results for predictions in four distinct clusters (two with $n_i = 5$ and two with $n_i = 10$) are given in the Supplementary Material, Tables 2-9. The bias is generally quite small, but more often negative. This is due to the underestimation of the spread parameters and the built-in shrinkage effect in optimal prediction. There is reasonable agreement between $\sqrt{msep}$ (estimated using 5%-trimmed means) and $\sqrt{msep_t^*}$ (5%-trimmed mean of $\sqrt{msep^*}$), but with some exceptions.

Finally, in Table 3 we present the actual coverage of the normal prediction intervals (constructed by normal approximation using the bootstrap estimates of msep) for u_i and v_i for these clusters. There is good coverage for u_i, but not so good for v_i, when $\gamma \neq 0$ (Settings I and III) because v_i is estimated from a smaller sample.

5. Discussion. In this paper we propose a random-effects hurdle model for bycatch data and address all aspects of estimation, prediction and inference. In so doing we make available much anticipated tools for marine

TABLE 3
Actual coverages of nominal 95% prediction intervals for u_i and v_i.

	CI($\hat{u}_i, u_i$)				CI($\hat{v}_i, v_i$)			
Setting	I	II	III	IV	I	II	III	IV
Cluster 1 ($n_i = 5$)	0.955	0.970	0.978	0.978	0.895	0.948	0.900	0.922
Cluster 2 ($n_i = 5$)	0.928	0.945	0.968	0.962	0.910	0.955	0.932	0.950
Cluster 51 ($n_i = 10$)	0.952	0.962	0.952	0.958	0.908	0.975	0.922	0.948
Cluster 52 ($n_i = 10$)	0.958	0.948	0.950	0.958	0.918	0.950	0.912	0.910

conservation research, specifically to predict cluster-specific targets like the probablity of bycatch of endangered hammerhead sharks for particular fishing trips. Although we develop our estimation, prediction and inference procedures for a random-effects hurdle model, they are easily adapted to a broad variety of situations. In fact, they can be used to obtain predictions and their mean squared errors for the entire class of generalized linear mixed models and models with multiple mixed linear predictors, where no alternative methods are currently available. As well, our general model formulation (Supplementary Material) contains numerous special cases for the random structure, including, for example, a two-level nested structure.

The random effects, used to describe the dependence structure of bycatch data, are parametrized so as to be independent, which is convenient for non-Gaussian random effects and, additionally, allows dependence between the two parts of the model to be both optional and simply tested. For our bycatch data, the dependence parameter is found to be significantly different from zero, leading us to conclude that the random process determining the presence/absence of bycatch is not independent of that determining the size of the nonzero bycatch events. As a result it would be inappropriate to model the nonzero bycatch events separately as is often done in practice. Further, our data analysis and simulation results demonstrate that ignoring this dependence can lead to bias in the fixed-effects regression parameters as well as an inability to predict random effects in some cases. In fact we would underestimate both the extent to which the probablity of hammerhead shark bycatch events is decreasing with time and the size of these events.

We derive empirical best predictors and obtain estimates of the mean squared error of these predictions using a fast bootstrap approach. Valuable insight can be gained from these predictions and their variability. For example we can predict the probability of nonzero bycatch for particular trips as well as the expected number of hammerhead sharks in these events. A comprehensive simulation study demonstrates the effectiveness and reliability of our proposals, for both the fixed-effects and the predictions. In particular, we see that the asymptotic theory applies well for the fixed-effects regres-

sion parameters but that the parameters of the random structure are more difficult to estimate. For the random-target predictions, we observe across simulations almost no bias and mean squared error estimates that are very often in agreement with those computed by bootstrap. Prediction intervals constructed using normal approximations are found to be reliable.

A natural next step is to incorporate spatially structured random effects into our framework so that we can more fully describe the spatial dependence in bycatch data and more accurately identify bycatch hotspots.

6. Acknowledgements. We thank the referees, Editor and Associate Editor for helpful comments. This research has been supported by the Natural Sciences and Engineering Research Council (Canada) and the Australian Research Council (DP0559135).

SUPPLEMENTARY MATERIAL

Supplementary material for the paper "A random-effects hurdle model for predicting bycatch of endangered marine species": (doi: 10.1214/00-AOASXXXXSUPP; .pdf). The supplementary file contains four sections. In the first section we give a general formulation of the random effects hurdle model. The second section presents a result about maximum likelihood estimation of the model. The third section introduces a fast bootstrap estimator and establishes its asymptotic distribution. Finally the fourth section gives additional simulation results, as discussed in this paper.

References.

Alfò, M. & Maruotti, A. (2010). Two-part regression models for longitudinal zero-inflated count data. *Canadian Journal of Statistics* **38**, 197–216.

Baum, J. (2007). *Population- and community-level consequences of the exploitation of large predatory marine fishes*. Ph.D. thesis, Biology Department, Dalhousie University, Halifax, Canada.

Baum, J., Myers, R., Kehler, D., Worm, B., Harley, S. & Doherty, P. (2003). Collapse and conservation of shark populations in the northwest atlantic. *Science* **299**, 389–392.

Booth, J. G. & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 262–272.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

De Bruijn, N. G. (1981). *Asymptotic Methods in Analysis*. Dover, New York.

Dobbie, M. J. & Welsh, A. H. (2001). Modelling correlated zero-inflated count data. *Australian & New Zealand Journal of Statistics* **43**, 431–444.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M., Nielsen, A. & Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **27**, 233–249.

Fulton, K. A., Liu, D., Haynie, D. L., Albert, P. S. et al. (2015). Mixed model and estimating equation approaches for zero inflation in clustered binary response data with application to a dating violence study. *The Annals of Applied Statistics* **9**, 275–299.

Furrer, R., Nychka, D. & Sain, S. (2012). *fields: Tools for spatial data*. R package version 6.6.3.

Hall, M. A., Alverson, D. L. & Metuzals, K. I. (2000). By-catch: problems and solutions. *Marine Pollution Bulletin* **41**, 204–219.

Huber, P., Ronchetti, E. & Victoria-Feser, M. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 893–908.

Hur, K., Hedeker, D., Henderson, W., Khuri, S. & Daley, J. (2002). Modeling clustered count data with excess zeros in health care outcomes research. *Health Services and Outcomes Research Methodology* **3**, 5–20.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference* **111**, 117–127.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Verlag.

Jiang, J., Jia, H. & Chen, H. (2001). Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica* **11**, 97–120.

Jiang, J. & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics* **53**, 217–243.

Jiang, J., Lahiri, P., Wan, S.-M. et al. (2002). A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics* **30**, 1782–1810.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Lee, Y. & Nelder, J. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 619–678.

Lele, S. R., Dennis, B. & Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using bayesian markov chain monte carlo methods. *Ecology letters* **10**, 551–563.

Lewison, R. L., Crowder, L. B., Read, A. J. & Freeman, S. A. (2004). Understanding impacts of fisheries bycatch on marine megafauna. *Trends in Ecology and Evolution* **19**, 598–604.

Liu, L., Strawderman, R., Cowen, M. & Shih, Y. (2010). A flexible two-part random effects model for correlated medical costs. *Journal of Health Economics* **29**, 110–123.

McCulloch, C. E. & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science* **26**, 388–402.

Min, Y. & Agresti, A. (2002). Modeling nonnegative data with clumping at zero: A survey. *Journal of the Iranian Statistical Society* **1**, 7–33.

Min, Y. & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5**, 1.

Molas, M. & Lesaffre, E. (2010). Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in Medicine* **29**, 3294–3310.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–365.

Myers, R., Baum, J., Shepherd, T., Powers, S. & Peterson, C. (2007). Cascading effects of the loss of apex predatory sharks from a coastal ocean. *Science* **315**, 1846–1850.

Neelon, B., Ghosh, P. & Loebs, P. F. (2013). A spatial poisson hurdle model for

exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **176**, 389–413.

Pikitch, E., Santora, C., Babcock, E., Bakun, A., Bonfil, R., Conover, D., Dayton, P., Doukakis, P., Fluharty, D., Heneman, B., Houde, E., Link, J., Livingston, P., Mangel, M., McAllister, M., Pope, J. & Sainsbury, K. (2004). Ecosystem-based fishery management. *Science* **305**, 346–347.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* **2**, 1–21.

Salibian-Barrera, M., Van Aelst, S. & Willems, G. (2008). Fast and robust bootstrap. *Statistical Methods and Applications* **17**, 41–71.

Skaug, H., Fournier, D., Nielsen, A., Magnusson, A. & Bolker, B. (2012). *Generalized Linear Mixed Models using AD Model Builder*. R package version 0.7.4.

Su, L., Tom, B. D. M. & Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10**, 374–389.

Welsh, A., Cunningham, R. & Chambers, R. (2000). Methodology for estimating the abundance of rare animals: seabird nesting on north east herald cay. *Biometrics* **56**, 22–30.

Yau, K. & Lee, A. (2001). Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine* **20**, 2907–2920.

E. Cantoni
Research Center for Statistics and
Geneva School of Economics and Management
University of Geneva
Bd Pont d'Arve 40
CH-1211 Geneva 4, Switzerland
E-mail: Eva.Cantoni@unige.ch

J. Mills Flemming
Department of Mathematics and Statistics
Dalhousie University
Halifax N.S., Canada, B3H 3J5
E-mail: Joanna.Flemming@Dal.Ca

A. H. Welsh
Centre for Mathematics and its Applications
Australian National University
Canberra, ACT 0200, Australia
E-mail: Alan.Welsh@anu.edu.au