NRC Research Press

ARTICLE

# Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative

James T. Thorson

**Abstract:** Ecologists often analyse biomass sampling data that result in many zeros, where remaining samples can take any positive real number. Samples are often analysed using a "delta-model" that combines two separate generalized linear models, GLMs (for encounter probability and positive catch rates), or less often using a compound Poisson-gamma (CPG) distribution that is computationally expensive. I discuss three theoretical problems with the conventional delta-model: difficulty interpreting covariates for encounter probability, the assumed independence of the two GLMs, and the biologically implausible form when eliminating covariates for either GLM. I then derive an alternative "Poisson-link model" that solves these problems. To illustrate, I use biomass samples for 113 fish populations to show that the Poisson-link model improves fit (and decreases residual spatial variation) for >80% of populations relative to the conventional delta-model. A simulation experiment illustrates that CPG and Poisson-link models estimate covariate effects that are similar and biologically interpretable. I therefore recommend the Poisson-link model as a useful alternative to the conventional delta-model with similar properties to the CPG distribution.

**Résumé :** Les écologistes analysent souvent des données d'échantillonnage de la biomasse qui donnent de nombreux zéros, les échantillons restants pouvant prendre n'importe quel nombre réel positif. Les échantillons sont souvent analysés en utilisant un « modèle delta » qui combine deux modèles linéaires généralisés (MLG) différents (pour la probabilité de rencontre et les taux de prises positifs) ou, moins souvent, une distribution Poisson-gamma composite (PGC) plus onéreuse sur le plan computationnel. J'aborde trois problèmes théoriques associés au modèle delta classique, soit la difficulté d'interpréter les covariables en ce qui concerne la probabilité de rencontres, l'indépendance présumée des deux MLG et la forme non plausible du point de vue biologique quand les covariables sont éliminées pour l'un ou l'autre des MLG. Je développe ensuite un nouveau « modèle Poisson-lien » qui résout ces problèmes. À des fins d'illustration, j'utilise des échantillons de biomasse pour 113 populations de poissons pour démontrer que le modèle Poisson-lien améliore le calage (et réduit la variation spatiale résiduelle) pour >80 % des populations par rapport au modèle delta classique. Une expérience de simulation illustre le fait que les modèles PGC et Poisson-lien estiment les effets de covariables qui sont semblables et permettent une interprétation biologique. Je recommande donc le modèle de Poisson-lien comme solution de rechange utile au modèle delta classique avec des propriétés semblables à la distribution PGC. [Traduit par la Rédaction]

## Introduction

Ecologists often estimate unknown biological rates (e.g., survival, stage-transition probabilities, per-capita productivity) by fitting ecological models to available data. Ecologists frequently collect such data from biological surveys, where observers visit a predefined site and either record the species they encounter (occupancy data) or measure the quantity of each species (counts or biomass). Many common analyses (including species distribution models, climate envelope analysis, and habitat utilization models) then involve fitting a regression to available data, often using a generalized linear mixed model (GLMM).

Ecological surveys (particularly for marine fishes) will often measure the biomass of a given species at each site. For example, this is common in marine fish sampling where thousands of individual fish can be captured simultaneously by a trawl gear. In this case, it is easiest to sort the sample by species, weigh the biomass for each species, and potentially subsample to determine individual mass, sex, and age (which can then be used to estimate the number sampled, even though numbers are not directly counted). Other examples of sampling species biomass include insect traps and leaf-litter traps (e.g., Clark 2016). In each case, sampling yields some proportion of zeros (e.g., where no individuals of a given taxon were encountered) and also a continuous-valued measure of density (e.g., biomass for samples where at least one individual of a given taxon was encountered).

Biomass sampling data are often analysed using a "delta" (or "hurdle") model (Aitchison 1955; Lo et al. 1992; Stefansson 1996) that includes two components: the probability of encountering the species and the expected biomass given that the species is encountered. This "delta-model" remains one of the most common types of regression used by ecologists and fisheries scientists (Maunder and Punt 2004; Zuur et al. 2009). However, it has several theoretical and practical drawbacks (as discussed below). One increasingly popular alternative to the conventional delta-model is using a compound Poisson-gamma (CPG) model (Smyth 1996; Foster and Bravington 2013; Lecomte et al. 2013), which is derived by assuming that biomass samples capture a Poisson-distributed number of individuals, where the biomass of each individual follows an independent gamma distribution. This CPG model is a special case of the Tweedie distribution (Foster and Bravington 2013), but it remains computationally expensive to evaluate and

Received 29 June 2017. Accepted 11 October 2017.

**J.T. Thorson.*** Fisheries Resource Assessment and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98112, USA.

**Email for correspondence:** James.Thorson@noaa.gov.

*James Thorson currently serves as an Associate Editor; peer review and editorial decisions regarding this manuscript were handled by Verena Trenkel.

Can. J. Fish. Aquat. Sci. **75**: 1369–1382 (2018) dx.doi.org/10.1139/cjfas-2017-0266

Published at www.nrcresearchpress.com/cjfas on 13 October 2017.

**Table 1.** Names and symbols used in the main text, indicating whether each refers to data ("Data"), an index ("Index"), a fixed ("Fixed") or a random ("Random") effect, or a derived quantity ("DQ").

| Name | Symbol | Type |
|---|---|---|
| Observed biomass for a survey sample *i* | $c_i$ | Data |
| Area swept for sample *i* | $a_i$ | Data |
| Measured covariates for sample *i* | $\mathbf{x}_i$ | Data |
| Unmeasured variables (treated as random) for sample *i* | $\mathbf{z}_i$ | Data |
| Sample index | $i$ | Index |
| Site index | $s$ | |
| Time index | $t$ | |
| Dispersion for probability density function for positive catch rates in conventional or Poisson-link delta-models | $\sigma_M$ | Fixed |
| Shape parameter for variation in individual mass in compound Poisson-gamma (CPG) distribution | $k$ | Fixed |
| Fixed effects for delta-model | $\boldsymbol{\beta}_p$, $\boldsymbol{\beta}_r$ | Fixed |
| Fixed effects for Poisson-link model | $\boldsymbol{\beta}_n$, $\boldsymbol{\beta}_w$ | Fixed |
| Fixed effects for CPG model | $\boldsymbol{\beta}_\lambda$, $\boldsymbol{\beta}_\mu$ | Fixed |
| Variance of random effects affecting $p_i$ in conventional delta-model | $\sigma^2_{p\omega}$, $\sigma^2_{p\varepsilon}$ | Fixed |
| Variance of random effects affecting $r_i$ in conventional delta-model | $\sigma^2_{r\omega}$, $\sigma^2_{r\varepsilon}$ | Fixed |
| Variance of random effects affecting $n_i$ in Poisson-link delta-model | $\sigma^2_{n\omega}$, $\sigma^2_{n\varepsilon}$ | Fixed |
| Variance of random effects affecting $w_i$ in Poisson-link delta-model | $\sigma^2_{w\omega}$, $\sigma^2_{w\varepsilon}$ | Fixed |
| Random effects affecting $p_i$ in conventional delta-model | $\omega_p\ (s)$, $\varepsilon_p\ (s,t)$ | Random |
| Random effects affecting $r_i$ in conventional delta-model | $\omega_r\ (s)$, $\varepsilon_r\ (s,t)$ | Random |
| Random effects affecting $n_i$ in Poisson-link delta-model | $\omega_n\ (s)$, $\varepsilon_n\ (s,t)$ | Random |
| Random effects affecting $w_i$ in Poisson-link delta-model | $\omega_w\ (s)$, $\varepsilon_w\ (s,t)$ | Random |
| Taylor's power law parameter | $v$ | DQ |
| Predicted number of individuals in CPG distribution | $\lambda_i$ | DQ |
| Predicted density for sample *i* | $d_i$ | DQ |
| Predicted group density for sample *i* | $n_i$ | DQ |
| Predicted average mass for each individual or group for sample *i* | $w_i$ | DQ |
| Predicted encounter probability for sample *i* | $p_i$ | DQ |
| Predicted biomass when a taxon is encountered for sample *i* | $r_i$ | DQ |
| Predicted individual mass in CPG distribution | $\mu_i$ | DQ |
| Mean of Tweedie parameterization for CPG distribution | $\eta_i$ | DQ |
| Dispersion of Tweedie parameterization of CPG distribution | $\phi_i$ | DQ |

therefore is difficult to combine with other detailed model components (e.g., spatio-temporal variation (Cressie and Wikle 2011)).

In the following, I first describe the most widely used version of the delta-model in detail, which involves a logistic regression for encounter probability and a separate generalized linear model (GLM) for biomass when the taxon is encountered, and outline three theoretical problems with using this conventional delta-model. In response, I then define an alternative "Poisson-link" model for analysing biomass sampling data and describe how this Poisson-link model rectifies all three theoretical problems. I then discuss similarities between the Poisson-link and CPG models, i.e., that both estimate numbers density and average mass via log-linked linear predictors. Next, I compile biomass sampling data from 113 fishes from seven marine ecosystems in North America and Europe and show (1) that the Poisson-link model does not sacrifice model fit relative to the CPG distribution and (2) that the Poisson-link model often has better fit and reduces unexplained variation relative to the conventional delta-model. Finally, I use a simulation experiment to confirm that the Poisson-link and CPG distributions both provide a simple interpretation of covariates and estimate covariates similarly.

## Methods

### Defining the conventional delta-model

Fisheries scientists have analysed biomass sampling data using delta-models for nearly 30 years (Lo et al. 1992; Stefansson 1996). Historically, these delta-models have been fitted to data by estimating parameters for two separate and independent GLMs: (1) encounter probability: the probability of encountering the species at a given place and time and (2) positive catch rates: the probability density function for catch in biomass given that the species is encountered. Predictions from the two GLMs can then be multiplied together to predict local density, and this in turn is used to predict total abundance across a prespecified spatial domain.

The "encounter probability" component of the delta-model defines the probability $p_i$ that catch $C_i$ for the *i*th sample is nonzero:

$$(1a)\qquad \mathrm{I}(C_i > 0) \sim \mathrm{Bernoulli}(p_i)$$

where $\mathrm{I}(C_i > 0)$ is an indicator function equal to one if $C_i > 0$ and zero otherwise (where all symbols are summarized in Table 1). In a GLMM, $p_i$ is modelled via a link function $g$, where $g(p_i)$ is a linear function of fixed and random effects. The "encounter probability" GLM involves a Bernoulli distribution for each sample, and the canonical link-function for this distribution is a logit-link. Presumably for this reason, researchers have often specified a logit-link with little consideration of alternatives (e.g., Stefansson 1996; Maunder and Punt 2004; Thorson and Ward 2013):

$$(1b)\qquad \mathrm{logit}(p_i) = \boldsymbol{\beta}_p^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_p^{\mathrm{T}}\mathbf{z}_i$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are predictors for fixed-effects $\boldsymbol{\beta}_p$ and random effects $\boldsymbol{\gamma}_p$ associated with the *i*th observation affecting encounter probability *p*, although other potential link-functions include the probit and complementary log–log link-functions (Zuur et al. 2009 p. 248). In the following, we refer to the logit-link as the "conventional" delta-model due to its widespread use in fisheries science.

The "positive catch rate" component defines the probability density function for catch $C_i$ given that it is nonzero. In the following, we use a bias-corrected lognormal density function:

$$(2a)\qquad C_i\big|(C_i > 0) \sim \mathrm{Lognormal}\left(\log(r_i) - \frac{\sigma_M^2}{2}, \sigma_M^2\right)$$

where $\log(r_i)$ and $\sigma_M^2$ are the mean and variance of $\log(C_i)$ and we use $\log(r_i) - \frac{\sigma_M^2}{2}$ such that $r_i$ is interpreted as the mean (rather than median) catch $C_i$ given that sample $i$ encounters a given taxon. The model again specifies a linear predictor for $\log(r_i)$:

$$(2b)\qquad \log(r_i) = \boldsymbol{\beta}_r^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_r^{\mathrm{T}}\mathbf{z}_i + \log(a_i)$$

where $\boldsymbol{\beta}_r$ and $\boldsymbol{\gamma}_r$ are again fixed and random effects (now affecting positive catch rate $r$) and $a_i$ is the area swept during the $i$th sample (i.e., $a_i$ is a linear offset for $r_i$). Previous research has often explored the performance of the lognormal versus gamma distribution (or other alternatives), and comparisons using simulated data have generally supported to the use of the Akaike information criterion (AIC) (Akaike 1974) to properly identify the distribution used to generate simulated data (Dick 2004).

Given these two model components, population density $d(s, t)$ at location $s$ and time $t$ can be predicted, $d(s, t) = p(s, t) \times r(s,t)$. Total abundance, center of distribution, or effective area occupied can then be easily calculated from predicted population density and spatial information about the population or sampling domain (Thorson et al. 2016*b*).

## Three "theoretical" problems with conventional delta-models

I note three major drawbacks to using the conventional delta-model: (1) difficulties in interpreting covariates, (2) assumed independence between model components, and (3) biologically implausible form when removing covariates.

### *Drawback 1: Difficulties in interpreting coefficients*

Using the conventional delta-model, an ecologist can include covariates affecting the logit-encounter probability, logit($p$), and (or) the log-positive catch rate, log($r$). For predictors affecting logit($p$), fixed and random effects then affect the "odds ratio", defined as the log-ratio of encounter probability and nonencounter probability, i.e., $\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$. However, it is not easy to summarize the average effect of covariates for logit($p$) on population density $d$ because this effect depends upon the value of all covariates and samples. Furthermore, a random effect $\boldsymbol{\gamma}_p$ for logit($p$) may have a variance of $\sigma_\gamma^2$ but there is no closed-form equation for calculating the resulting variance in population density $\boldsymbol{d}$. I suspect that many ecologists would prefer to estimate the impact of a covariate affecting encounter probability (e.g., bottom temperature) on expected population densities rather than the "odds ratio". Although an ecologist could use predictive sampling to approximate the variance in population density for any link function, the lack of a closed-form solution still complicates interpretation for these models. This drawback would be solved by defining all covariates via the log-link function, in which case an estimated coefficient $\beta$ for covariate $\mathbf{x}$ indicates that a 0.01 increase in $x_i$ results in a $\beta$% increase in predicted population density. However, defining all covariates via log-link is inconvenient using a conventional delta-model because a log-link for encounter probability $p$ could exceed 1.0 (the upper bound for a probability).

### *Drawback 2: Assumed independence among components*

Using the conventional delta-model, the "encounter probability" and "positive catch rate" components are assumed to be statistically independent, i.e., knowledge about encounter probability $p$ gives no information about the likely distribution for positive catches $r$. This assumption is contrary to a large body of evidence suggesting (1) that abundant species have wide ranges, such that frequently encountered species also have higher density throughout their range (Gaston 1994), and also (2) that an increase in local density will decrease the probability of failing to detect a species that is present (Royle and Nichols 2003). Both phenomena suggest that a location with increased probability of encounter (higher $p$) will tend to have greater catch rates given an encounter (higher $r$), as has been argued previously for nonparametric zero-inflated models (Liu and Chan 2011). As one concrete example, the conventional delta-model specifies that positive catch rates $r_i$ increase linearly with increased area swept $a_i$ for a given sample but that increased area swept has no effect on encounter probability $p_i$. This specification is inconvenient because an increase in area swept for many species will increase the probability of sampling at least one occupied patch (Lecomte et al. 2013). In response, an ecologist could chose to also include area swept as a predictor for encounter probability $p$. However, there is no way to interpret the estimated coefficient as a "linear offset" when using the conventional logit-link for encounter probability (see Drawback 1 above), so this area swept covariate would then be estimated to have a nonlinear impact on expected catches.

### *Drawback 3: Biologically implausible form when removing covariates*

Ecologists often have little data with which to estimate a multitude of potential ecological processes. The presence of "tapering effects" (i.e., many ecological processes with gradually declining effect sizes for any given system) has driven interest in using model selection to identify "parsimonious" ecological models (Burnham and Anderson 2002). Parsimony in this case is defined as an appropriate number of parameters that minimizes total predictive error for a given data set (simultaneously low bias and imprecision). In many cases, parsimony is achieved by identifying a flexible family of models, where analysts can use model selection to identify the appropriate degree of model complexity. This approach is most effective, however, when the model that eliminates covariates remains biologically plausible (e.g., is likely to provide a good fit for species on average). As a corollary of Drawback 2, it will often be more statistically efficient to assume that a covariate associated with high encounter probability will also likely be associated with high positive catch rates (and vice versa). For example, if the density of rocky substrate is associated with increased encounter probability for a refuge-seeking fish, then it is also likely associated with increased positive catch rates because sampling will likely include a greater number of occupied habitat patches. By contrast, removing covariates in a delta-model generally involves specifying that a given covariate affects encounter probability but not positive catch rates (or vice versa).

## Solutions from using an alternative "Poisson-link" model

As an alternative to the conventional delta-model, I propose a "Poisson-link" model for biomass sampling data with many zeros. This Poisson-link model is derived by defining $n_i$ as the predicted density of individuals or groups at sample $i$, where the number of observed individuals is assumed to follow a Poisson process with expectation $n_i$. The encounter probability from this Poisson process is then

$$(3)\qquad p_i = 1 - \exp(-a_i \times n_i)$$

such that $p_i \to 1$ as $a_i n_i \to \infty$, i.e., an increased area swept increases the expected number of individuals observed (Foster and Bravington 2013; Lecomte et al. 2013). Predicted group density $n_i$ is then modelled via a log-linked linear predictor:

$$(4)\qquad \log(n_i) = \boldsymbol{\beta}_n^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_n^{\mathrm{T}}\mathbf{z}_i$$

where a 0.1 increase in the right-hand side of eq. 4 (due to fixed effects $\boldsymbol{\beta}_n^{\mathrm{T}}\mathbf{x}_i$ or random effects $\boldsymbol{\gamma}_n^{\mathrm{T}}\mathbf{z}_i$) results in an approximately 10% increase in predicted group density $n_i$.

I then combine two equations for biomass density $d$ to derive an expression for positive catch rates $r$, i.e., (1) predicted biomass density is the product of predicted group density and predicted biomass per group ($d_i = n_i \times w_i$, where $w$ is predicted biomass per group of individuals) and (2) biomass density is the product of encounter probability and positive catch rates ($d_i = p_i \times r_i$). After rearranging, these definitions imply that

$$(5) \qquad r_i = \frac{n_i}{p_i} \times w_i$$

When data are few, predicted biomass per group $w$ can be estimated as a single parameter. However, a more general treatment involves specifying $w$ via a log-linked linear predictor:

$$(6) \qquad \log(w_i) = \boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_w^{\mathrm{T}}\mathbf{z}_i$$

where this reduces to constant predicted biomass per group, $w_i = \exp(\beta_w)$ when $\mathbf{x} = \mathbf{1}$ and $\mathbf{z} = \varnothing$.

The probability distribution for biomass sample $C_i$ is then calculated for the Poisson-link model by converting predicted numbers density ($n_i$) to encounter probability ($p_i$) using eq. 3, converting predicted biomass per group ($w_i$) to positive catch rates ($r_i$) using eq. 5, and then applying the same likelihood function as the conventional delta-model (i.e., eqs. 1*a* and 2*a*). Consequently, this likelihood function requires essentially the same computational time as the conventional delta-model (eqs. 3, 5, 1*a*, and 2*a*). Similarly, the Poisson-link model can be interpreted as a reparameterization of a delta-model using a complementary log–log link for encounter probability (eq. 3) and a biologically interpretable linkage between encounter probability and positive catch rates (eq. 5). However, the Poisson-link model is not identical to a conventional delta-model using a complementary log–log link because group density $n$ affects both encounter probability $p$ and positive catch rates $r$. In the following, I specify a lognormal distribution for biomass given encounters for both conventional delta- and Poisson-link models, although future studies could use model selection to select among alternative distribution functions.

## The Poisson-link model responds to all three theoretical problems with the conventional delta-model

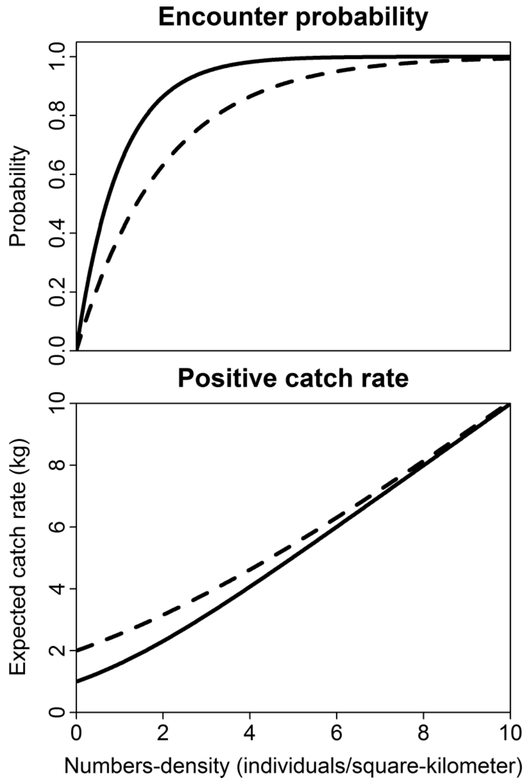
### *1. Difficulties in interpreting coefficients*

The Poisson-link model simplifies interpretation of covariates. In particular, covariates $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_w$ both predict changes in log-density, so, e.g., a 0.01 increase in $\boldsymbol{\beta}_n\mathbf{x}_i$ is associated with approximately a 1% increase in density. Similarly, a random effect $\boldsymbol{\gamma}_j$ with a standard deviation of 0.01 explains approximately a 1% coefficient of variation in density. Both fixed and random effects therefore have a similar interpretation in predicting variation in population density because both affect density via a log-link function.

### *2. Independence among components*

The Poisson-link model induces a correlation between predicted encounter probability $p$ and predicted positive catch rates $r$ that is interpretable biologically. When expected counts are low ($na \ll 1$), an increase in group density results in a proportional increase in encounter probability ($n \propto p$). In this case, increasing group density results in a greater proportion of encounters, where each encounter is likely to sample a single individual with mass $w$. As group-density becomes large ($na \gg 1$), encounter probability will plateau ($p \rightarrow 1$), and further increases in group density are accompanied by an increase in positive catch rates ($n \propto r$). In summary, encounter probability $p$ and positive catch rates $r$ are correlated via a joint dependence on group density $n$ (see Fig. 1).

**Fig. 1.** Conceptual diagram showing encounter probability $p$ (top panel $y$-axis) and positive catch rate $r$ (bottom panel $y$-axis) as a function of group density $n$ ($x$-axis) for the Poisson-link model given different values for average biomass per group $w$ (solid line: $w = 1$ kg; broken line: $w = 2$ kg) while holding area swept constant ($a = 1$ km²). An increased average mass results in a smaller increase in $p$ with increasing $n$ (with identical form to a complementary log–log link function) and also a slower convergence to the linear relationship between numbers density $n$ and positive catch rates $r$.



### *3. Biologically implausible form when removing covariates*

Finally, the Poisson-link model by default specifies that a covariate affecting group density (i.e., $\boldsymbol{\beta}_n$) influences predictions of both encounter probability $p$ and positive catch rates $r$. To continue our previous example, a covariate representing the local density of rocky substrate might be selected for group density but not for average mass, and this reduction in the number of estimated parameters still retains a biologically meaningful impact of substrate on both encounter probability and positive catch rates. This specification will allow a smaller number of estimated parameters to explain variation in both encounter probability and positive catch rates whenever $p_i$ and $r_i$ are positively correlated in the form predicted by the Poisson-link model.

Despite these improvements over the conventional delta-model, the Poisson-link and conventional delta-model are identical (e.g., have identical maximum likelihood and provide identical predictions of density) for several potential model configurations. For example, delta-model parameters ($\beta_p$ and $\beta_r$) can be converted to Poisson-link parameters ($\beta_n$ and $\beta_w$) and vice versa when using an intercept-only model (e.g., $\mathbf{x}$ is a design matrix and $\mathbf{z} = \varnothing$) and a constant area offset (i.e., $a_i = a$ for all observations $i$) via the relations

$$(7) \qquad \begin{aligned} \operatorname{logit}^{-1}(\beta_p) &= 1 - \exp(-a \times \exp(\beta_n)) \\ \exp(\beta_r) &= \frac{\exp(\beta_n)}{1 - \exp(-a \times \exp(\beta_n))}\exp(\beta_w) \end{aligned}$$

where these relations are derived from the definition of predicted encounter probability $p$ and positive catch rate $r$ (see Table 2) and the identical likelihood used in each model (eqs. 1*a* and 2*a*). The conventional delta- and Poisson-link models generally differ (i.e., result in different maximum likelihoods and density predictions) whenever they include either a covariate, variable area offset, or random effects.

### Comparison with the CPG distribution

The proposed Poisson-link model has many similarities to a CPG distribution, which is a special case of the Tweedie distribution (Smyth 1996; Lecomte et al. 2013). The CPG distribution is derived from the assumption that biomass samples arise from a Poisson distribution for the number of individuals captured:

$$(8) \qquad N_i \sim \text{Poisson}(\lambda_i \times a_i)$$

where $\lambda_i$ is the group density in the vicinity of sampling (I use different symbols for variables than in the Poisson-link model to indicate that estimated variables may differ between CPG and Poisson-link models). The CPG then specifies that the mass $W_{ij}$ of each individual follows a gamma distribution:

$$(9) \qquad W_{ij} \sim \text{Gamma}(k, \mu_i k^{-1})$$

where $k$ is the gamma shape parameter and $\mu_i k^{-1}$ is the scale parameter, such that total catch $C_i = \sum_{j=1}^{N_i} W_{ij}$. The parameterization used here involves estimating $\lambda_i$, $\mu_i$, and $k$, where $\lambda_i$ and $\mu_i$ can be specified via a link-function and linear predictors and $k$ is assumed constant for all samples $i$ (see Foster and Bravington (2013) for further discussion). The CPG distribution generates a "power-law" relationship between the expected value, $E(C_i) = \eta_i$, and variance, $Var(C_i) = \phi_i \eta_i^{\nu}$. By contrast, the typical Tweedie parameterization directly estimates the power parameter ($1 < \nu < 2$), uses a constant dispersion ($\phi_i = \phi$ for all $i$) and a linear predictor for log($\eta$), and has been used extensively elsewhere (Candy 2004; Shono 2008; Lecomte et al. 2013; Berg et al. 2014). The CPG is identical to the Tweedie parameterization given mean $\eta_i = \lambda_i a_i \mu_i$ and dispersion $\phi_i = \frac{1}{\lambda_i a_i}\frac{(\lambda_i a_i \mu_i)^{2-\nu}}{2-\nu}$ (based on Foster and Bravington (2013) for derivation).

Similar to the Poisson-link model, the CPG distribution (using the Foster and Bravington (2013) parameterization) specifies a log-link for both group density and average mass:

$$(10) \qquad \begin{aligned} \log(\lambda_i) &= \boldsymbol{\beta}_\lambda^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_\lambda^{\mathrm{T}}\mathbf{z}_i \\ \log(\mu_i) &= \boldsymbol{\beta}_\mu^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_\mu^{\mathrm{T}}\mathbf{z}_i \end{aligned}$$

and this results in an identical derivation for expected encounter probability $p$ and positive catch rates $r$ as the Poisson-link model (Table 2). The CPG distribution therefore responds to all three theoretical problems similarly to the Poisson-link model (Foster and Bravington 2013). The Foster–Bravington parameterization of the CPG distribution then involves estimating fixed effects $\boldsymbol{\beta}_\lambda$, $\boldsymbol{\beta}_\mu$, and $k$ by finding their values that maximize the likelihood function.

However, the CPG likelihood function (Smyth 1996, Dunn and Smyth 2005) is computationally expensive to evaluate because it involves approximating an integration constant $W$ as the sum of an infinite series (Appendix A) or approximating the CPG distribution using numerical sampling (e.g., Lauderdale 2012). Approximating the sum of an infinite series has computational cost determined by the number of terms in the summation, and numerical sampling requires introducing a large number of discrete-valued random effects. In the following, I evaluate the CPG likelihood using an upper limit of 1000 for calculating $W$ and confirm that the log-likelihood (given maximum likelihood estimates of all parameters) is identical to the value generated by the package *fishMod* (Foster et al. 2016) to a tolerance of $10^{-6}$. In practice, approximating this infinite series can be efficiently implemented by specialized numerical techniques (Dunn and Smyth 2008; Foster and Bravington 2013), e.g., by analytically determining an efficient lower and upper bound for the summation. I encourage further comparison of numerical techniques as a topic for future research but claim that the CPG likelihood is computationally expensive relative to the Poisson-link model because the former requires summing across as many as 50 terms (Foster and Bravington 2013 p. 539), while the Poisson-link model requires evaluating only a single term.

Unlike the CPG distribution, the Poisson-link model permits a fast, closed-form calculation of the model likelihood (using eqs. 1*a*, 2*a*, and 3–6). Both the CPG and Poisson-link model specify the density of groups ($\lambda$ and $n$, respectively for CPG and Poisson-link) and average mass per group ($\mu$ and $w$, respectively) via log-linked linear predictors. The main difference, however, is that the proposed Poisson-link specifies a different mean–variance relationship than the CPG model.

### Case study data: bottom trawl survey database

In the following, I first compare the fit of the conventional and alternative Poisson-link models with the CPG using real-world data and a simple model (estimating annual intercepts as fixed effects). I then compare the conventional and alternative Poisson-link models using a more complicated model (estimating fixed annual intercepts plus spatial and spatio-temporal variation), which is computationally infeasible using implementations of the CPG distribution available in Template Model Builder (Kristensen et al. 2016).

For each comparison, I use bottom trawl survey data from seven marine ecosystems: (1) Eastern Bering Sea: survey operated by the Alaska Fisheries Science Center (AFSC) obtained from a fixed-station design (Lauth and Conner 2016), (2) Gulf of Alaska: survey operated by the AFSC obtained from a randomized design (Von Szalay and Raring 2016), (3) Aleutian Islands: survey operated by the AFSC obtained from a randomized design (Raring et al. 2016), (4) US West Coast:the West Coast groundfish bottom trawl survey operated by the Northwest Fisheries Science Center (NWFSC) obtained from a stratified-random design (Keller et al. 2017), (5) North Sea: the North Sea international bottom trawl survey (NS-IBTS), restricting data to 1991–2015 obtained using a "Gov" gear in quarter 1 (winter) (ICES 2012), (6) Scottish West Coast: the Scottish West Coast international bottom trawl survey (SWC-IBTS), restricting data to 1991–2015 obtained using a "Gov" gear in quarter 1, and (7) Celtic Sea and Bay of Biscay: the French demersal survey (EVHOE) of the Celtic Sea and Bay of Biscay, operating by the French Research Institute for Exploitation of the Sea (IFREMER) from 1997 to 2015 in quarter 4 (fall) (Mahé and Poulard 2005).

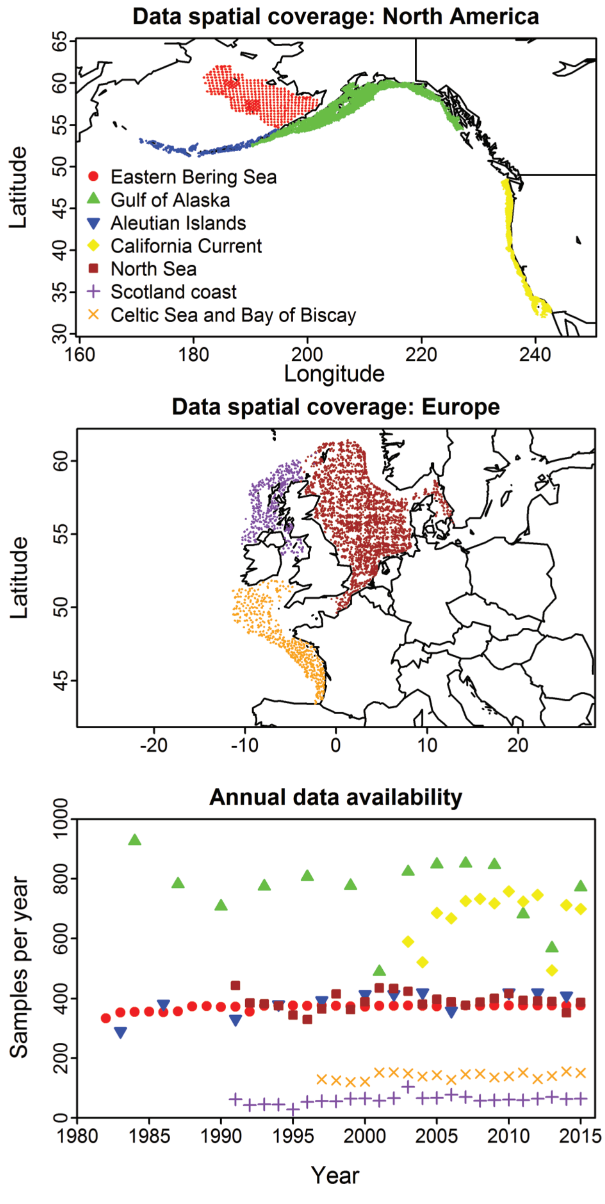
US survey protocols (1–4) contain biomass-per-unit-area data (i.e., samples are standardized to a constant area swept), so I assume that area swept is constant for these surveys. European survey protocols (5–7) are described in ICES (2012). Public databases for surveys 1–4 contain biomass-per-unit-area (i.e., samples are standardized to a constant area swept), while for those for surveys 5–7 contain raw biomass/numbers and a measure of fishing effort (the duration of tows in minutes). I therefore assume that "tow duration" is proportional to area swept $a_i$ for surveys 5–7 and that area swept is constant for surveys 1–4. US survey protocols (1–4) are described in Stauffer (2004), and publicly available databases for these surveys contain biomass-per-unit-area (i.e., samples are standardized to a constant area swept). I therefore analyse "biomass-per-area" as catch and fix area swept $a_i$ at a constant value for all samples in these surveys. European survey

**Table 2.** Comparison of variables for the conventional delta-model, an alternative Poisson-link model, and the compound Poisson-gamma model including how to calculate biomass density for each model.

| | Conventional delta-model | Poisson-link delta-model | Compound Poisson-gamma model |
|---|---|---|---|
| Component 1 | Encounter probability $p$:<br>$\text{logit}(p_i) = \boldsymbol{\beta}_n^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_n^{\mathrm{T}}\mathbf{z}_i$ | Group density $n$: $\log(n_i) = \boldsymbol{\beta}_n^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_n^{\mathrm{T}}\mathbf{z}_i$ | Group density $\lambda$: $\log(\lambda_i) = \boldsymbol{\beta}_\lambda^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_\lambda^{\mathrm{T}}\mathbf{z}_i$ |
| Component 2 | Positive catch rates $r$:<br>$\log(r_i) = \boldsymbol{\beta}_r^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_r^{\mathrm{T}}\mathbf{z}_i + \log(a_i)$ | Average biomass per group $w$:<br>$\log(w_i) = \boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_w^{\mathrm{T}}\mathbf{z}_i$<br>or equivalently positive catch rates:<br>$\log(r_i) = \log(n_i) - \log(p_i) + \boldsymbol{\beta}_r^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_r^{\mathrm{T}}\mathbf{z}_i$ | Average biomass per group $\mu$:<br>$\log(\mu_i) = \boldsymbol{\beta}_\mu^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\gamma}_\mu^{\mathrm{T}}\mathbf{z}_i$ |
| Density | Biomass density $d$: $d_i = p_i \times r_i$ | Biomass density $d$: $d_i = n_i \times w_i$ | Biomass density $d$: $d_i = \lambda_i \times \mu_i$ |
| Predicted encounter probability | $p_i = \dfrac{1}{1 + \exp(-\boldsymbol{\beta}_p^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\gamma}_p^{\mathrm{T}}\mathbf{z}_i)}$ | $p_i = 1 - \exp(-a_i \times n_i)$ | $p_i = 1 - \exp(-a_i \times \lambda_i)$ |
| Predicted positive catch rate | $r_i = a_i \times \exp(-\boldsymbol{\beta}_r^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{\gamma}_r^{\mathrm{T}}\boldsymbol{z}_i)$ | $r_i = \dfrac{n_i}{p_i} \times w_i$ | $r_i = \dfrac{\lambda_i}{p_i} \times \mu_i$ |
| Likelihood function | $\Pr(C = c_i) = \begin{cases} 1 - p_i & \text{if} \quad c_i = 0 \\ p_i \times f(C; r_i; \sigma_M^2) & \text{if} \quad c_i > 0 \end{cases}$ | $\Pr(C = c_i) = \begin{cases} 1 - p_i & \text{if} \quad c_i = 0 \\ p_i \times f(C; r_i; \sigma_M^2) & \text{if} \quad c_i > 0 \end{cases}$ | $\Pr(C = c_i) = \begin{cases} \exp(-\lambda_i) & \text{if} \quad c_i = 0 \\ W(c_i, \lambda_i, k, \mu_i) \times \exp\left(-\dfrac{c_i}{\mu_i} - \lambda_i - \log(c_i)\right) & \text{if} \quad c_i > 0 \end{cases}$<br>where $W(c_i, \lambda_i, k, \mu_i) = \sum_{j=1}^{\infty} \dfrac{\lambda_i^j \left(\dfrac{c_i}{\mu_i}\right)^{jk}}{j!\Gamma(jk)}$ |
| Process to simulate data | $P \sim \text{Bernoulli}(p_i)$ and<br>$c_i = \begin{cases} 0 & \text{if} \quad P = 0 \\ \text{LN}\left(\log(r_i) - \dfrac{\sigma_M^2}{2}, \sigma_M^2\right) & \text{if} \quad P = 1 \end{cases}$ | $P \sim \text{Bernoulli}(p_i)$ and<br>$c_i = \begin{cases} 0 & \text{if} \quad P = 0 \\ \text{LN}\left(\log(r_i) - \dfrac{\sigma_M^2}{2}, \sigma_M^2\right) & \text{if} \quad P = 1 \end{cases}$ | $N_i \sim \text{Poisson}(\lambda_i)\ W_{i,j} \sim \text{Gamma}(k^{-2}, \mu_i k^2)$ and<br>$c_i = \begin{cases} 0 & \text{if} \quad N_i = 0 \\ \sum_{j=1}^{N_i} W_{i,j} & \text{if} \quad N_i > 0 \end{cases}$ |

**Note:** We also include equations to convert from variables in the alternative Poisson-link model ($n$ and $w$) and compound Poisson-gamma model ($\lambda$ and $\mu$) to variables in the conventional delta-model model ($p$ and $r$), calculate the likelihood function, and simulate data given each model. The likelihood function is identical between conventional and Poisson-link delta-models, where we use a bias-corrected lognormal density function in the main text: $f(C; r_i; \sigma_M^2) = \dfrac{1}{c_i \sigma_M \sqrt{2\pi}} \exp\left(-\dfrac{\left(\log(c_i) - \log(r_i) - \dfrac{\sigma_M^2}{2}\right)^2}{2\sigma_M^2}\right)$. However, evaluating the likelihood function for the Poisson-link model requires converting predicted group density $n_i$ and biomass per group $w_i$ to encounter probability $p_i$ and positive catch rates $r_i$. The likelihood for the compound Poisson-gamma model is from Foster and Bravington (2013) (see their eq. 6) (after fixing a typo where they were missing a negative sign before their first summand on the right-hand side of the second row).

**Fig. 2.** Spatial location of sampling data for four surveys in North America (middle panel) and three surveys in Europe (top panel) and annual sample size (bottom panel) for all seven bottom trawl surveys with publicly available Application Programming Interfaces, used for comparing performance of conventional and "Poisson-link" delta-models (colors are defined in the legend in the top panel, identical between panels, and can be used to match spatial coverage to the annual sample size for each survey in the bottom panel). [Color online.]



protocols (5–7) are described in ICES (2012), and the public Datras database for these surveys contains numbers caught for multiple length bins and a measure of fishing effort (the duration of tows in minutes) for each sample as well as records of individual biomass and length. I calculate a length–mass key from records of individual biomass and length, use this key to convert numbers-at-length to biomass-at-length, and then calculate total biomass for each sample.

For each survey, I restrict data to the 20 most frequently encountered fishes (see Fig. 2 for annual sample sizes). Surveys 6 and 7 had sufficient mass-at-length records to calculate biomass data for fewer than 20 species, so I used biomass data for as many species as were available. All surveys are publicly available and can be accessed using R package *FishData* (https://github.com/james-thorson/FishData), which in turn uses R package *icesDatras* (https://github.com/ices-tools-prod/icesDatras) to download data for surveys 5–7.

## Comparison 1: Annual intercept models

I first compare the conventional delta-model and Poisson-link model against the CPG distribution using a simple model where each model component has a separate intercept by year. Parameters for all models are estimated via maximum likelihood using release number 1.5.0 (https://doi.org/10.5281/zenodo.834777) of package *VAST* (www.github.com/james-thorson/VAST) (Thorson and Barnett (2017)), which estimates parameters using Template Model Builder (Kristensen et al. 2016) within the R statistical platform (R Core Team 2015). Model selection is conducted using the marginal AIC (Akaike 1974) as is widely used in ecology and fisheries (Burnham and Anderson 2002) based on the marginal log-likelihood and the number of fixed effects. I do not attempt to calculate the conditional AIC (Vaida and Blanchard 2005), which measures complexity by the number of fixed effects plus the effective degrees of freedom for random effects. To my knowledge, conditional AIC has not been used in fisheries science, and I recommend its exploration as a topic for future research. The likelihood is identical for the conventional delta-model and Poisson-link model (see eq. 8 and associated text) and different from that for the CPG distribution (because the CPG has a different exponent for Taylor's power law; see section Comparison with the CPG distribution). I therefore present the difference in AIC between the Poisson-link model and the CPG model. I present this comparison to determine whether the Poisson-link model gains computational efficiency while maintaining a model fit comparable to that of the CPG model.

## Comparison 2: Spatio-temporal model

I next compare the conventional delta-model and Poisson-link model using a spatio-temporal modelling framework that includes both spatial and spatio-temporal variation among sites *s* and years *t* and also estimates a fixed effect for each year in each model component. I do not include the CPG distribution in this comparison because it is not computationally feasible to include the CPG within the spatio-temporal modelling framework in Template Model Builder (although see Arcuti et al. (2013) or Augustin et al. (2013) for a spatio-temporal implementations using the *mgcv* package (Wood et al. 2016) in R). For the conventional delta-model, I specify

$$\begin{aligned} \text{logit}^{-1}(p(s,t)) &= \beta_p(t) + \omega_p(s) + \varepsilon_p(s,t) \\ \log(r(s,t)) &= \beta_r(t) + \omega_r(s) + \varepsilon_r(s,t) \end{aligned} \tag{11}$$

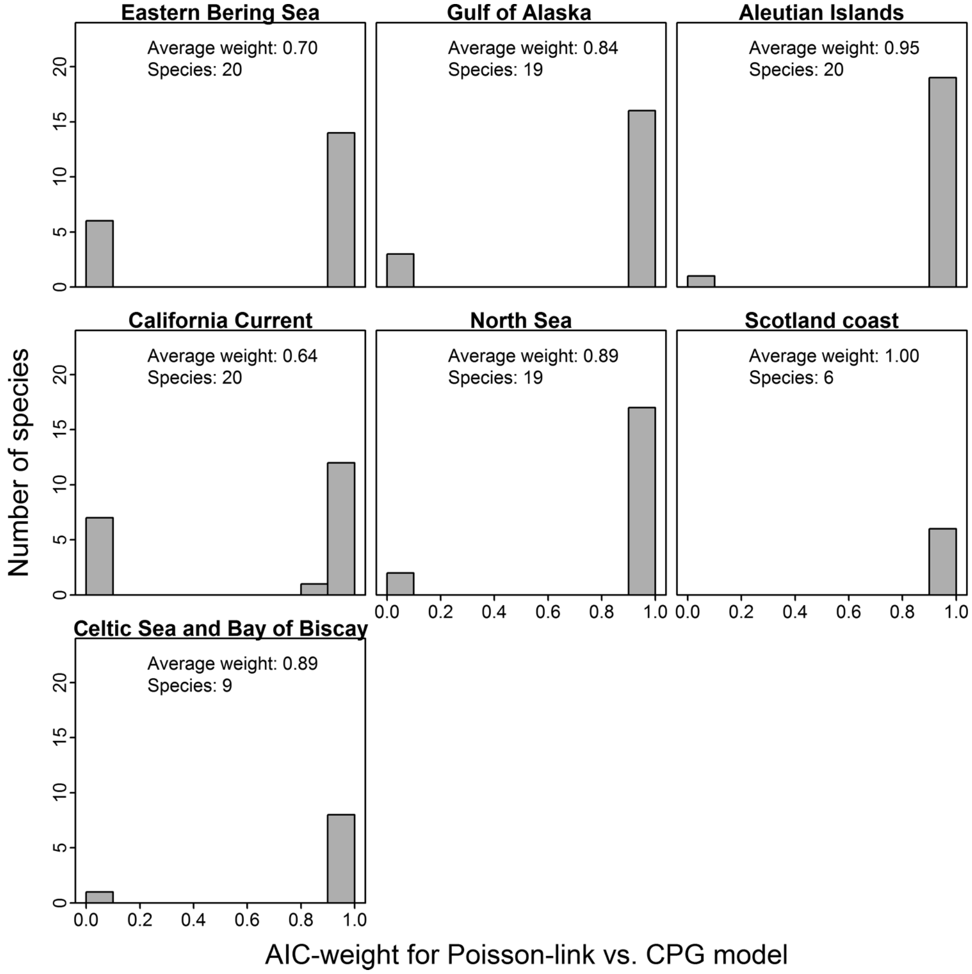
where intercepts $\beta_p(t)$ and $\beta_r(t)$ for each year are estimated as fixed effects and

$$\begin{aligned} \boldsymbol{\omega}_p &\sim \text{MVN}\left(\mathbf{0}, \sigma^2_{p\omega}\mathbf{R}\right) \\ \boldsymbol{\varepsilon}_p(t) &\sim \text{MVN}\left(\mathbf{0}, \sigma^2_{p\varepsilon}\mathbf{R}\right) \end{aligned} \tag{12}$$

where **R** is the spatial correlation given estimated decorrelation distance $\kappa$, $\sigma^2_{p\omega}$ is the estimated pointwise variance of spatial variation in $p$, $\sigma^2_{p\varepsilon}$ is the estimated pointwise variance of spatio-temporal variation in $p$, and $\boldsymbol{\omega}_r$ and $\boldsymbol{\varepsilon}_r(t)$ are defined identically but with separate estimates of spatial variance $\sigma^2_{r\omega}$ and spatio-temporal variance $\sigma^2_{r\varepsilon}$ (Thorson et al. 2015). For the alternative Poisson-link model, I specify

$$\begin{aligned} \log(n(s,t)) &= \beta_n(t) + \omega_n(s) + \varepsilon_n(s,t) \\ \log(w(s,t)) &= \beta_w(t) + \omega_w(s) + \varepsilon_w(s,t) \end{aligned} \tag{13}$$

**Fig. 3.** Akaike information criterion (AIC) weight for the "Poisson-link" model compared with a compound Poisson-gamma (CPG) model (where a species with a AIC weight of 1.0 means that AIC strongly favors the Poisson-link over the CPG model) for a simple (fixed intercept only) model applied to survey biomass sampling data in seven bottom trawl surveys (see Fig. 2 for spatial and temporal coverage of each survey), where each panel also lists the average AIC weight for the Poisson-link model and the number of species analysed in that region.



where spatial and spatio-temporal terms (e.g., eq. 12) are defined identically to the conventional delta-model (but using different subscripts to indicate the difference in variables). Parameters for both conventional and alternative models are estimated using maximum marginal likelihood, using the Laplace approximation to approximate the integral across the joint probability of fixed and random effects. Parameter estimation is again performed using package *VAST*, using a stochastic partial differential equation approximation to the multivariate normal distribution used in spatial and spatio-temporal processes (Lindgren et al. 2011), and model selection is conducted using AIC.

After estimating parameters, I then evaluate model performance by comparing the estimated standard deviation of spatial and spatio-temporal variation for positive catch rates $r$ in the conventional delta-model with these standard deviations for average mass $w$ in the Poisson-link model. I do not compare the variance for encounter probability $p$ because it is not easily interpretable in the conventional delta model (as explained in the previous section "Drawback 1: Difficulties in interpreting coefficients"). However, this comparison is appropriate for positive catch rates $r$ because the Poisson-link model decomposes variance in log($r$) into three additive components (see the conversion from $w$ and $p$ to $r$ in Table 2):

$$\text{(14)}\quad \begin{aligned}\mathrm{Var}[\log(r)] &= \mathrm{Var}\left[\log\left(\frac{n}{p}\right)\right] + \mathrm{Var}[\log(w)]\\ &= \mathrm{Var}\left[\log\left(\frac{n}{p}\right)\right] + \sigma^2_{w\omega} + \sigma^2_{w\varepsilon}\end{aligned}$$

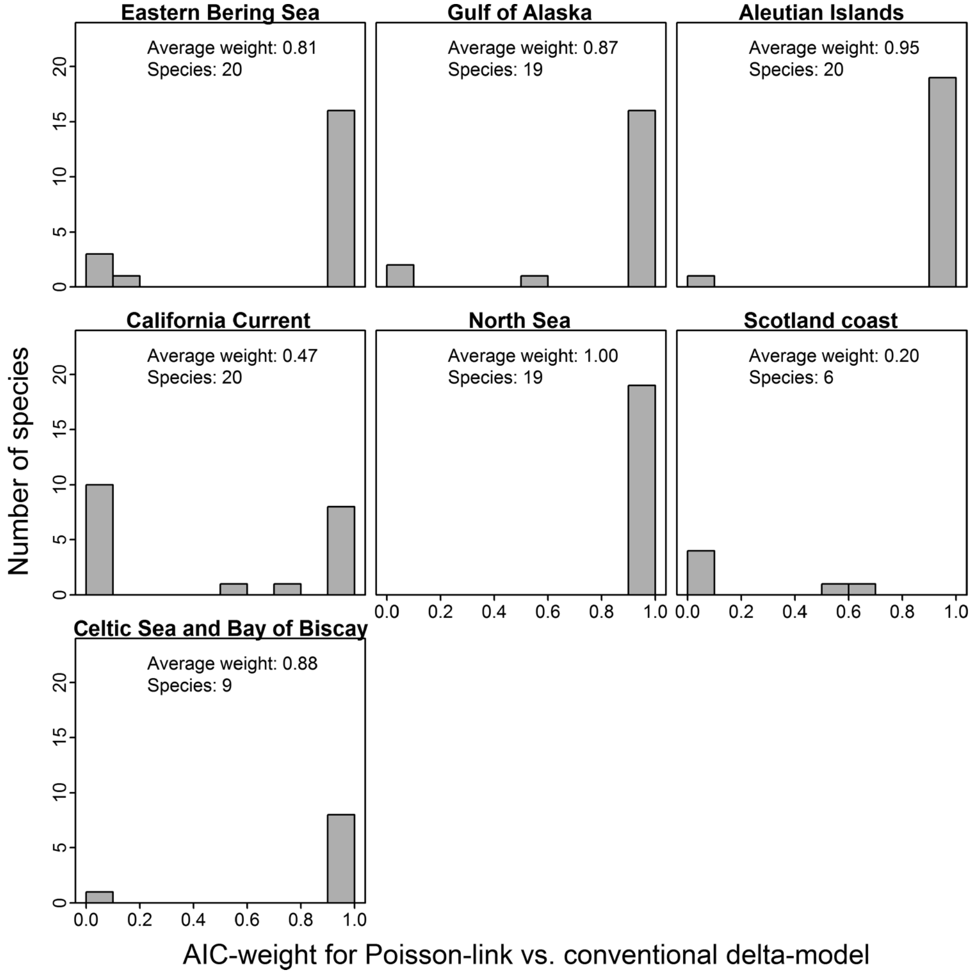
whereas the conventional model decomposes variance into two components (see eq. 11):

$$\text{(15)}\quad \mathrm{Var}[\log(r)] = \sigma^2_{r\omega} + \sigma^2_{r\varepsilon}$$

Therefore, if knowledge of encounter probability $p$ is informative about positive catch rates $r$, then this will cause average mass in the alternative model to have lower spatial and (or) spatio-temporal variance than for positive catch rates in the conventional model (i.e., if $\mathrm{Cov}\left(\log(r), \log\left(\frac{n}{p}\right)\right) > 0$, then $\sigma^2_{w\omega} + \sigma^2_{w\varepsilon} < \sigma^2_{r\omega} + \sigma^2_{r\varepsilon}$). Alternatively, if encounter probability $p$ is statistically independent or negatively associated about positive catch rates $r$, then the opposite will occur (i.e., if $\mathrm{Cov}\left(\log(r), \log\left(\frac{n}{p}\right)\right) \leq 0$, then $\sigma^2_{w\omega} + \sigma^2_{w\varepsilon} \geq \sigma^2_{r\omega} + \sigma^2_{r\varepsilon}$).

I therefore record (1) the proportion of species for each region where the conventional or alternative model was selected as parsimonious using the AIC, (2) the pointwise (marginal) standard deviation of spatial and spatio-temporal variance for both model components, and (3) the predictive standard deviation of an abun-

**Fig. 4.** Akaike information criterion (AIC) weight for the "Poisson-link" model compared with a conventional delta-model for a complicated (fixed intercept plus random spatial and spatio-temporal effects) model applied to survey biomass sampling data in seven bottom trawl surveys (see Fig. 3 caption for details).



dance index derived from each model (indices are area-weighted following Thorson et al. (2015)). I hypothesize that the Poisson-link model will be more parsimonious than the conventional delta-model for the majority of species. The pointwise variances $\sigma^2_{r\omega}$ and $\sigma^2_{r\varepsilon}$ from the conventional model and $\sigma^2_{w\omega}$ and $\sigma^2_{w\varepsilon}$ from the alternative model are directly comparable, and I hypothesize that spatial and spatio-temporal variances for the alternative model will be lower because the encounter probability $p$ (estimated from proportion of nearby samples that encounter the species) is informative about local positive catch rates $r$.

### Simulation experiment

Finally, I conduct a simulation experiment to evaluate relative performance of three alternative models (conventional delta-model: eqs. 1–2; Poisson-link model: eqs. 3–6; CPG model: eqs. 8–10) when estimating a covariate. This experiment involves the following steps:

(1) I obtain data for a single species (arrowtooth flounder, *Atheresthes stomias*, in the EBSBTS data; survey 1 above) including the depth for each sampling location.

(2) I fit each of three models to these data while including as fixed effect both year (as an annual intercept) and depth (standardized to have a mean of 0 and a standard deviation of 1) and while not including any random effects.

(3) For each model in step 2, I generate 100 simulated data sets, using the estimated depth effect, variance parameters, and simulating new annual intercepts that have the same mean and stan-

**Fig. 5.** Distribution of standard deviation estimates of residual variation in positive catch rates for a complicated (fixed intercept plus random spatial and spatio-temporal effects) model for each of 113 stocks (in total across seven surveys), using the conventional delta-model (solid line) or alternative Poisson-link delta-model (broken line). I display the average standard deviation for each model in the top right corner ("delta": conventional delta-model; "Poisson": Poisson-link delta-model).

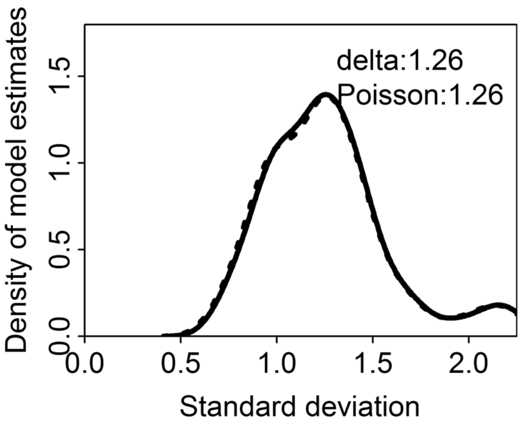
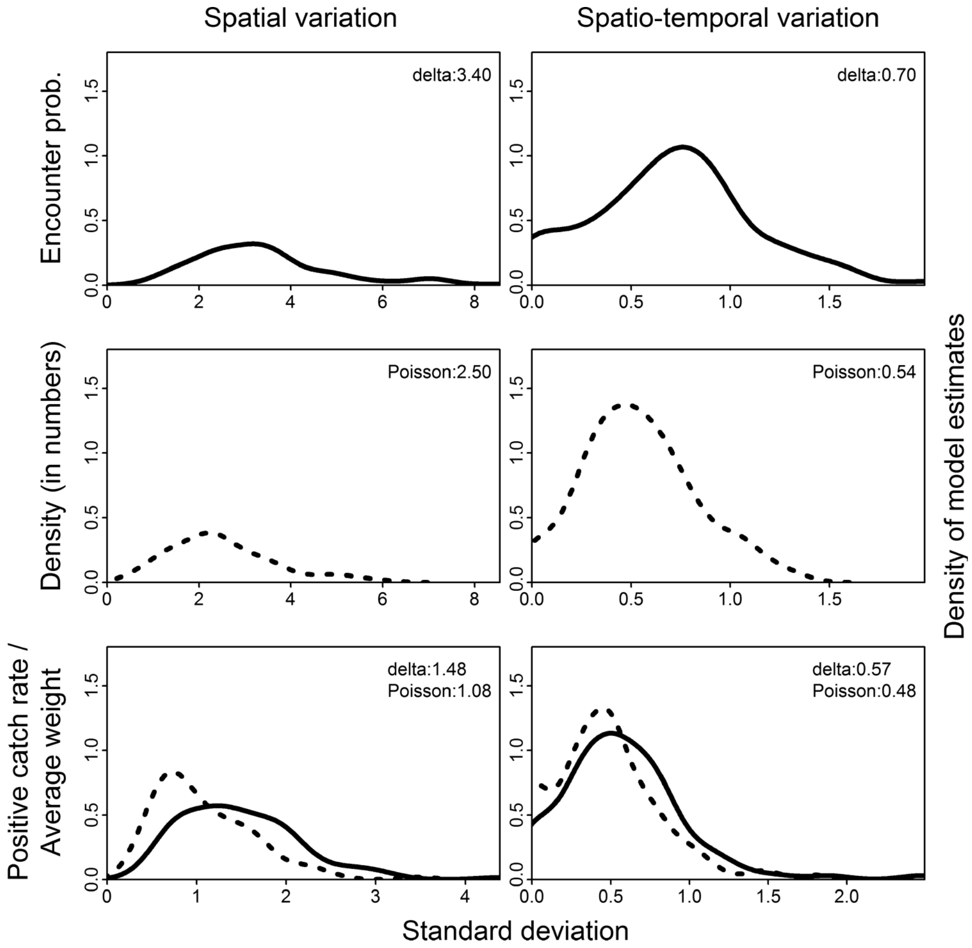


**Fig. 6.** Standard deviation estimates for a complicated (fixed intercept plus random spatial and spatio-temporal effects) model (see Fig. 5 caption for plot details) showing spatial variation (left column) and spatio-temporal variation (right column). Standard deviations for encounter probability *p* (top row) and density in numbers *n* (middle row) are not directly comparable (because encounter probability uses a logit-link, while density uses a log-link), but standard deviations for positive catch rates in the conventional delta-model and average mass in the Poisson-link delta model (bottom row) are directly comparable (because both use a log-link).



dard deviation as the sample mean and standard deviation of estimated intercepts (from step 2). Each simulated data set has same the annual sample size and sampling locations as the original data set (in step 1).

(4) For each of these 300 simulated data sets, I fit each of the three models (i.e., 900 model fits total). For each model fit, I record the estimated depth effect for both model components.

I assess model performance in two ways. First, I compare the estimated depth effect when fitted to real data (in step 2) among models to explore how interpretable these estimates are. Second, I compare the estimated and true depth effect from each combination of simulation model (in step 2) and estimation model (in step 4). Based on previous arguments, I hypothesize that the Poisson-link and CPG models will have similar performance when fitted to data generated by either model (i.e., because both specify depth effects via a log-link for both model components).

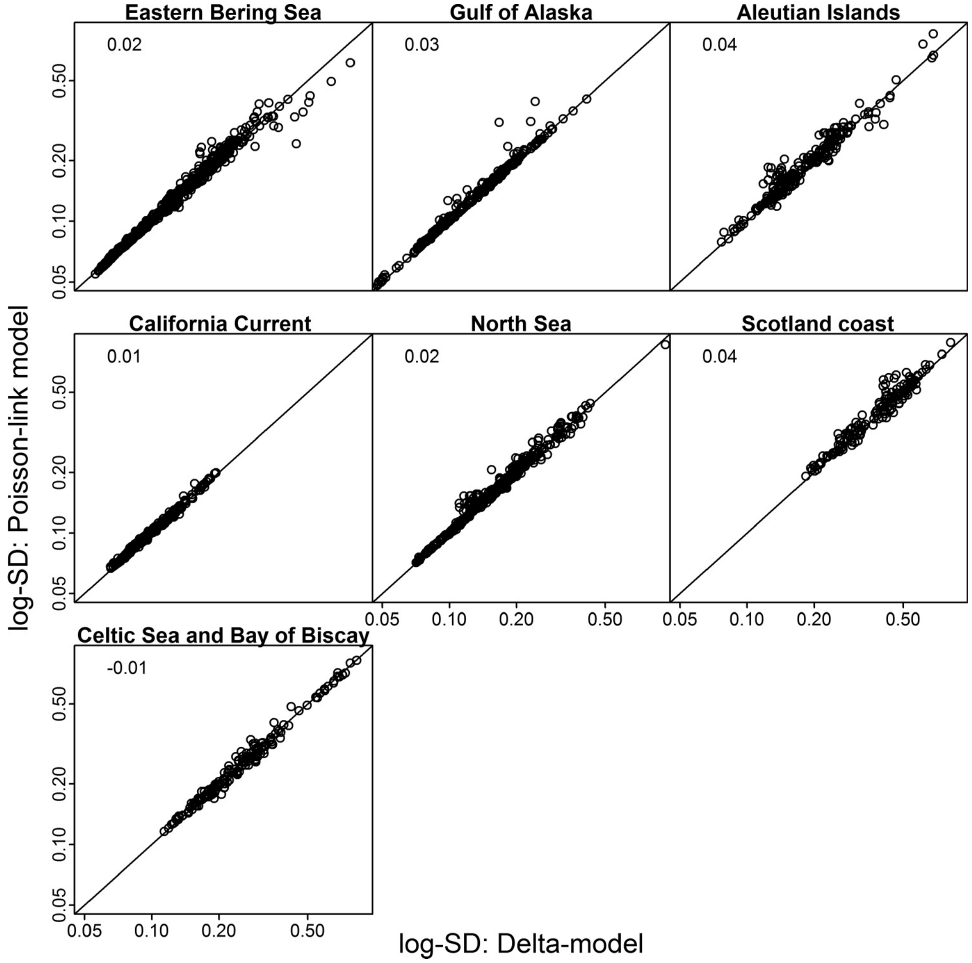
## Results

Available data show that the Poisson-link model results in better fit (a higher log-likelihood of available data) relative to the CPG model using the simple "annual intercept" structure (Fig. 3). Both models have the same number of parameters and differ in the relationship between mean and variance for positive catch rates in each year. This suggests that the Poisson-link model improved computational efficiency without sacrificing fit relative to the CPG for fish biomass sampling data distribution given this simple intercept-only model structure.

The complicated spatio-temporal model applied to these same data shows that using a Poisson-link model also results in better fit that the conventional delta-model for the vast majority of populations in five out of seven regions (Fig. 4). Models again have an identical number of estimated parameters, so a higher log-likelihood also indicates greater parsimony (e.g., using the AIC). The average AIC weight for the Poisson-link model is >80% for the same six regions. The exception is for the California Current, where each model is each selected for 10 of 20 species. In this region, the implied correlation between encounter probability and positive catch rates apparently does not improve fit relative to assuming independence between detection probability and positive catch rates. However, the implied correlation does improve fit for the majority of populations in other regions.

The conventional and alternative models have essentially identical estimates of residual variation in positive catch rates ($\sigma_M$ = 1.25 or 1.26), indicating that both models attribute a roughly identical portion of sampling variance to the combination of spatial and spatio-temporal variation (Fig. 5). As hypothesized, however, the Poisson-link model results in a lower standard deviation for spatial and spatio-temporal variation (Fig. 6). The standard deviation is not directly comparable for the first-model component between models because $\sigma_{r\omega}$ and $\sigma_{r\varepsilon}$ (from the conventional model;

**Fig. 7.** Comparison of estimated log-standard deviation of total population-wide abundance for the conventional delta-model and alternative Poisson-link model for a complicated (fixed intercept plus random spatial and spatio-temporal effects) model applied to biomass sampling data in seven bottom trawl surveys (the solid line shows a 1:1 relationship, indicating equal precision between models, and circles below the line indicate greater precision for the Poisson-link model for a given population and year; the number in the upper left corner indicates the average log-ratio between models).



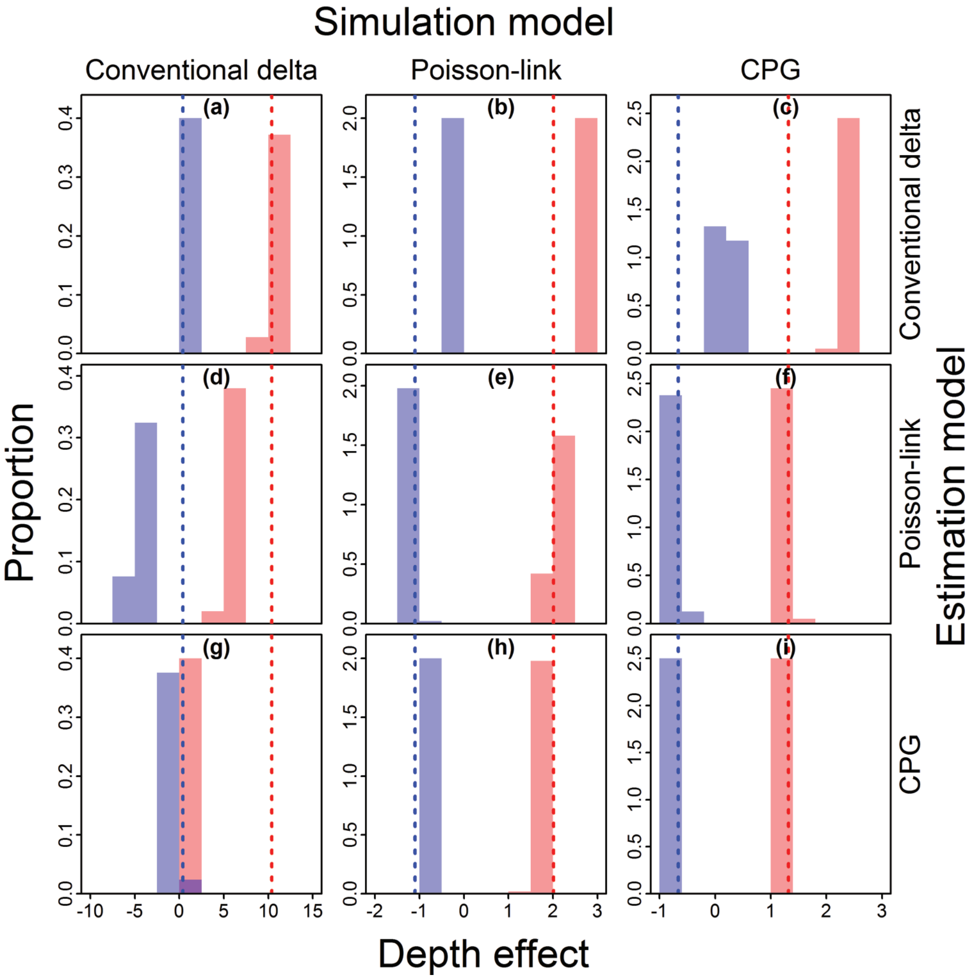
top row of Fig. 6) affect $r$ via a logit-link function, while $\sigma_{n\omega}$ and $\sigma_{n\varepsilon}$ (from the alternative model; middle row of Fig. 6) affect $r$ via a complementary log–log link function. However, the standard deviations for the second component are comparable (both $\sigma_{r\omega}$, $\sigma_{r\varepsilon}$ and $\sigma_{w\omega}$, $\sigma_{w\varepsilon}$ affect positive catch rates $r$ via the log-link function). For this second component (Fig. 6, bottom row), the delta-model has a pointwise standard deviation of 1.47, whereas the Poisson-link model has 1.05 for spatial variation. Therefore, including local densities and encounter probabilities ($n$ and $p$) as a predictor of $r$ shrinks the magnitude of unexplained spatial variation by $1 - \frac{1.08^2}{1.48^2} = 47\%$. Similarly, the Poisson-link model shrinks the magnitude of unexplained spatio-temporal variation by $1 - \frac{0.47^2}{0.56^2} = 29\%$ on average across populations.

Despite resulting in better fit and also shrinking the magnitude of explained variation in positive catch rates, the Poisson-link model does not consistently decrease the log-standard deviation of confidence intervals for estimated abundance indices relative to the conventional delta-model (Fig. 7). Across all seven regions, the Poisson-link model has similar or slightly wider confidence intervals on average (0%–4% wider) for all seven regions and for almost every stock within each region. Inspection of residual diagnostics (supplementary materials[1]) shows little difference in fit between models to two species selected for illustration purposes (arrowtooth flounder in the Eastern Bering Sea and shortraker rockfish, *Sebastes borealis*, in the Aleutian Islands).

Finally, the simulation experiment (Fig. 8) shows that the Poisson-link model estimates a 2.0% increase in group density and a −0.9% decrease in biomass per group when depth increases by 1% of its standard deviation for arrowtooth flounder in the Eastern Bering Sea (Fig. 8, vertical broken lines in middle column). The CPG estimates qualitatively similar depth effects (a 1.3% increase in group density and a −0.7% decrease in biomass per group), while the depth effect for encounter probability in the conventional delta model ($\beta_p$) is highly different. This difference arises because the conventional delta-model uses a logit-link function, and therefore the estimated depth coefficient for the delta-model cannot be used to calculate a single value for the increase in encounter probability per change in depth (instead the predicted increase in encounter probability with depth changes each year depending on the intercept for that year). The simulation experiment also confirms (1) that all three models generate unbiased

[1] Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/cjfas-2017-0266.

**Fig. 8.** Comparison of estimated depth effects (histogram) versus true value (broken vertical lines) when estimating parameters using the delta-model (top row, red: $\beta_p$; blue: $\beta_r$), Poisson-link model (middle row, red: $\beta_n$; blue: $\beta_w$), and compound Poisson-gamma (CPG) model (bottom row, red: $\beta_\lambda$; blue: $\beta_\mu$) applied to data generated using each model (left column: delta-model; middle column: Poisson-link model; right column: CPG) fitted to data for arrowtooth flounder, *Atheresthes stomias*, in the Eastern Bering Sea. Note that the true depth effect (broken vertical line) is identical for each panel in a given column (because these all use the same operating model to generate data) and that panels along the diagonal involve a correctly specified estimation model, while other panels involve a misspecified estimation model.



estimates of depth effects when the simulation and estimation models match (Figs. 8*a*, 8*e*, and 8*i*) and (2) that the Poisson-link and CPG estimation models have similar performance to one another, regardless of whether data are generated by using the Poisson-link or CPG simulation models (Figs. 8*e*, 8*f*, 8*h*, and 8*i*). Finally, a comparison of model selection results from the simulation experiment (results not shown) confirms that AIC identifies the data-generating model as the most parsimonious estimation model in nearly 100% of simulation replicates. This result confirms that AIC is a useful metric to evaluate model performance using real-world data (i.e., in Figs. 3 and 4).

## Discussion

Delta-models using a logit-link for encounter probabilities and a log-link for positive catch rates have a long history in fisheries science (Stefansson 1996; Maunder and Punt 2004), and I have presented three theoretical arguments for why this conventional delta-model is unsatisfactory, namely (1) difficulties in interpreting how covariates for encounter probability affect population density, (2) the lack of dependence between encounter probability and positive catch rates, and (3) the biologically implausible form when removing covariates for one or the other model component. I have then shown how these three difficulties are addressed using a new "Poisson-link" model for biomass sampling data that can be interpreted as a computationally efficient alternative to the CPG distribution. Application to 113 populations in seven marine regions shows that the Poisson-link model substantially improves fit by using knowledge of encounter probabilities to decrease otherwise unexplained variation in positive catch rates. However, this Poisson-link model decreases average confidence interval width for abundance indices in only one of seven regions. I therefore conclude that the Poisson-link model is not likely to substantially increase the information available to stock assessments when used to estimate abundance indices. However, improvements in fit, interpretability, and parsimony relative to a conventional delta-model are still likely to be useful when estimating habitat maps, estimating habitat associations, and fitting ecological models to samples of fish biomass.

I envision several useful avenues for future research. Most obviously, the Poisson-link model could be compared more exhaustively with the CPG distribution as well as other alternatives (e.g., the Law-of-Leaks "LoL" model: Ancelet et al. (2010)), including more detailed comparison of the different mean–variance relationships implied by these potential models. This comparison could then identify taxa and model structures where the CPG, LoL, and Poisson-link models are more or less statistically efficient.

Given the many potential numerical techniques to implement the CPG (Dunn and Smyth 2005, 2008; Foster and Bravington 2013), one of these will hopefully prove to be computationally feasible for the spatio-temporal models as explored here. I note that the CPG distribution automatically follows Taylor's rule (i.e., a power law mean–variance relationship) and therefore has stronger theoretical support for ecological processes. I also recommend future research exploring the potential consequences of ignoring variation among samples when predicting biomass per group (i.e., fixing $w_i = w$). This restriction is particularly appealing when introducing additional model complexity (i.e., modelling multiple species simultaneously: Thorson et al. 2016*a*). The current application to 113 populations worldwide shows that there is substantial variation in *w* even after accounting for the effect of encounter probabilities, but determining the impact of restricting $w_i = w$ on model performance will require further simulation testing. This simulation experiment could presumably be conditioned on the range of spatial and spatio-temporal variances estimated in this study.

Finally, the past decade has seen rapid growth in a variety of useful approximations for otherwise slow or intractable processes that arise in ecology. Examples include approximating individual birth–death demographics using Markov chains (Hubbell 2011), estimating the likelihood of ecological rates given unobserved (latent) variables via the Laplace approximation (Skaug and Fournier 2006; Kristensen et al. 2016), or approximating spatial variation and individual movement using finite-element analysis methods (Lindgren et al. 2011; Thorson et al. 2017). Collectively, these approximations are useful when they permit the development of models with increased realism regarding otherwise neglected processes in ecological systems (e.g., a "zero-sum" linkage between regional and local species pools for describing community richness: Hubbell 2011). In this light, the Poisson-link model can be viewed as a computationally efficient approximation to a common sampling design, where biomass samples arise from a weighing of individuals that vary in individual biomass. I recommend ongoing development and testing of efficient approximations to sampling processes and hope that these approximations will collectively allow biological rates (births, deaths, and movement) to be simultaneously estimated for entire communities occurring on heterogenous landscapes using available data worldwide. Hopefully, this will then allow us to "fill in the missing spaces" where messy or opportunistic data exist but ecologist have no previously conducted comparative analyses (e.g., in the white spaces in Fig. 1, top and middle panels).

## Acknowledgements

I thank E. Dick, J. Hastie, M. McClure, and four anonymous reviewers for comments on a previous draft. I also thank A. Magnusson, S. Large, C. Millar, and C. Berg for contributions to the *icesDatras* R package that have allowed me to include biomass data from waters near Europe. Finally, I thank the many participants who conducted sampling in these seven bottom trawl surveys: these are an invaluable resource for current and future marine ecologists.

## References

Aitchison, J. 1955. On the distribution of a positive random variable having a discrete probability mass at the origin. J. Am. Stat. Assoc. **50**(271): 901. doi:10.2307/2281175.

Akaike, H. 1974. A new look at the statistical-model identification. IEEE Trans. Autom. Control, **19**(6): 716–723. doi:10.1109/TAC.1974.1100705.

Ancelet, S., Etienne, M.-P., Benoît, H., and Parent, E. 2010. Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process. Environ. Ecol. Stat. **17**(3): 347–376. doi:10.1007/s10651-009-0111-6.

Arcuti, S., Calculli, C., Pollice, A., D'Onghia, G., Maiorano, P., and Tursi, A. 2013. Spatio-temporal modelling of zero-inflated deep-sea shrimp data by Tweedie generalized additive. Statistica, **73**(1): 87–101. doi:10.6092/issn.1973-2201/3987.

Augustin, N.H., Trenkel, V.M., Wood, S.N., and Lorance, P. 2013. Space-time modelling of blue ling for fisheries stock management. Environmetrics, **24**(2): 109–119. doi:10.1002/env.2196.

Berg, C.W., Nielsen, A., and Kristensen, K. 2014. Evaluation of alternative age-based methods for estimating relative abundance from survey data in relation to assessment models. Fish. Res. **151**: 91–99. doi:10.1016/j.fishres.2013.10.005.

Burnham, K.P., and Anderson, D. 2002. Model selection and multi-model inference. 2nd ed. Springer, New York.

Candy, S.G. 2004. Modelling catch and effort data using generalised linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. CCAMLR Sci. **11**: 59–80.

Clark, J.S. 2016. Why species tell more about traits than traits about species: predictive analysis. Ecology, **97**(8): 1979–1993. doi:10.1002/ecy.1453.

Cressie, N., and Wikle, C.K. 2011. Statistics for spatio-temporal data. John Wiley & Sons, Hoboken, N.J.

Dick, E.J. 2004. Beyond "lognormal versus gamma": discrimination among error distributions for generalized linear models. Fish. Res. **70**(2–3): 351–366. doi:10.1016/j.fishres.2004.08.013.

Dunn, P.K., and Smyth, G.K. 2005. Series evaluation of Tweedie exponential dispersion model densities. Stat. Comput. **15**(4): 267–280. doi:10.1007/s11222-005-4070-y.

Dunn, P.K., and Smyth, G.K. 2008. Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. Stat. Comput. **18**(1): 73–86. doi:10.1007/s11222-007-9039-6.

Foster, S.D., and Bravington, M.V. 2013. A Poisson–Gamma model for analysis of ecological non-negative continuous data. Environ. Ecol. Stat. **20**(4): 533–552. doi:10.1007/s10651-012-0233-0.

Foster, S.D., Foster, M.S., and SystemRequirements, C. 2016. Package "fishMod." Available from http://cran.itam.mx/web/packages/fishMod/fishMod.pdf.

Gaston, K.J. 1994. Measuring geographic range sizes. Ecography, **17**(2): 198–205. doi:10.1111/j.1600-0587.1994.tb00094.x.

Hubbell, S.P. 2011. The unified neutral theory of biodiversity and biogeography (MPB-32). Princeton University Press, Princeton, N.J.

ICES. 2012. Manual of the International Bottom Trawl Surveys. Series of the ICES Survey Protocols, International Council for the Exploration of the Sea (ICES), Copenhagen, Denmark. Available from https://www.ices.dk/sites/pub/Publication%20Reports/ICES%20Survey%20Protocols%20(SISP)/SISP1-IBTSVIII.pdf.

Keller, A.A., Wallace, J.R., and Methot, R.D. 2017. The Northwest Fisheries Science Center's West Coast Groundfish Bottom Trawl Survey: history, design, and description. NOAA Technical Memorandum, Northwest Fisheries Science Center, Seattle, Wash.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B.M. 2016. TMB: automatic differentiation and Laplace approximation. J. Stat. Softw. **70**(5): 1–21. doi:10.18637/jss. v070.i05.

Lauderdale, B.E. 2012. Compound Poisson–Gamma regression models for dollar outcomes that are sometimes zero. Polit. Anal. **20**(3): 387–399. doi:10.1093/pan/mps018.

Lauth, R.R., and Conner, J. 2016. Results of the 2013 eastern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate resources. NOAA Technical Memorandum, Seattle, Wash.

Lecomte, J.B., Benoît, H.P., Etienne, M.P., Bel, L., and Parent, E. 2013. Modeling the habitat associations and spatial distribution of benthic macroinvertebrates: a hierarchical Bayesian model for zero-inflated biomass data. Ecol. Model. **265**: 74–84. doi:10.1016/j.ecolmodel.2013.06.017.

Lindgren, F., Rue, H., and Lindström, J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. **73**(4): 423–498. doi:10.1111/j.1467-9868.2011.00777.x.

Liu, H., and Chan, K.-S. 2011. Generalized additive models for zero-inflated data with partial constraints. Scand. J. Stat. **38**(4): 650–665. doi:10.1111/j.1467-9469.2011.00748.x.

Lo, N.C., Jacobson, L.D., and Squire, J.L. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Can. J. Fish. Aquat. Sci. **49**(12): 2515–2526. doi:10.1139/f92-278.

Mahé, J.C., and Poulard, J.C. 2005. Manuel des protocoles de campagne halieutique. IFREMER. Available from http://archimer.ifremer.fr/doc/00036/14707/12013.pdf.

Maunder, M.N., and Punt, A.E. 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. **70**(2-3): 141–159. doi:10.1016/j.fishres.2004.08.002.

R Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/.

Raring, N.W., Laman, E.A., Von Szalay, P.G., Rooper, C.N., and Martin, M.H. 2016. Data report: 2012 Aleutian Islands bottom trawl survey. NOAA Technical Memorandum, US Department of Commerce, National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Alaska Fisheries Science Center, Seattle, Wash. Available from https://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-332.pdf.

Royle, J.A., and Nichols, J.D. 2003. Estimating abundance from repeated presence–absence data or point counts. Ecology, **84**(3): 777–790. doi:10.1890/0012-9658(2003)084[0777:EAFRPA]2.0.CO;2.

Shono, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. Fish. Res. **93**(1–2): 154–162. doi:10.1016/j.fishres.2008.03.006.

Skaug, H., and Fournier, D. 2006. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Comput. Stat. Data Anal. **51**(2): 699–709. doi:10.1016/j.csda.2006.03.005.

Smyth, G.K. 1996. Regression analysis of quantity data with exact zeros. *In* Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management, Gold Coast, Australia. pp. 17–19.

Stauffer, G. 2004. NOAA protocols for groundfish bottom trawl surveys of the nation's fishery resources. NOAA Technical Memorandum, National Oceanic and Atmospheric Administration (NOAA), Seattle, Wash. Available from http://www.mafmc.org/s/NOAA-protocols-for-bottom-trawl-surveys.pdf.

Stefansson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J. Mar. Sci. **53**(3): 577–588. doi:10.1006/jmsc.1996.0079.

Thorson, J.T., and Barnett, L.A.K. 2017. Comparing estimates of abundance trends and distribution shifts using single- and multispecies models of fishes and biogenic habitat. ICES J. Mar. Sci. **74**(5): 1311–1321. doi:10.1093/icesjms/fsw193.

Thorson, J.T., and Ward, E. 2013. Accounting for space–time interactions in index standardization models. Fish. Res. **147**: 426–433. doi:10.1016/j.fishres.2013.03.012.

Thorson, J.T., Shelton, A.O., Ward, E.J., and Skaug, H.J. 2015. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. ICES J. Mar. Sci. J. Conserv. **72**(5): 1297–1310. doi:10.1093/icesjms/fsu243.

Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C., and Zipkin, E.F. 2016*a*. Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. Glob. Ecol. Biogeogr. **25**(9): 1144–1158. doi:10.1111/geb.12464.

Thorson, J.T., Pinsky, M.L., and Ward, E.J. 2016*b*. Model-based inference for estimating shifts in species distribution, area occupied and centre of gravity. Methods Ecol. Evol. **7**(8): 990–1002. doi:10.1111/2041-210X.12567.

Thorson, J.T., Jannot, J., and Somers, K. 2017. Using spatio-temporal models of population growth and movement to monitor overlap between human impacts and fish populations. J. Appl. Ecol. **54**(2): 577–587. doi:10.1111/1365-2664.12664.

Vaida, F., and Blanchard, S. 2005. Conditional Akaike information for mixed-effects models. Biometrika, **92**(2): 351–370. doi:10.1093/biomet/92.2.351.

Von Szalay, P.G., and Raring, N.W. 2016. Data report: 2015 Gulf of Alaska bottom trawl survey. NOAA Technical Memorandum, US Department of Commerce, National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Alaska Fisheries Science Center, Seattle, Wash. Available from https://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-325.pdf.

Wood, S.N., Pya, N., and Säfken, B. 2016. Smoothing parameter and model selection for general smooth models. J. Am. Stat. Assoc. **111**(516): 1548–1563. doi:10.1080/01621459.2016.1180986.

Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., and Smith, G.M. 2009. Mixed effects models and extensions in ecology with R. 1st ed. Springer, New York.

## Appendix A: Comparing the Poisson-link delta model with the compound Poisson-gamma distribution

The Tweedie distribution is sometimes used to analyse biomass sampling data for marine fishes (Foster and Bravington 2013; Lecomte et al. 2013). This distribution specifies expected catch rates $D$ such that catch rate $C$ follows a stochastic process with expectation and variance:

$$\mathbb{E}(C) = D$$

$$\mathbb{V}(C) \propto D^{\nu}$$

where this formula for the variance is Taylor's power law and $v$ is the power parameter (Foster and Bravington 2013). When $1 < v < 2$, the Tweedie distribution can be derived from a compound Poisson-gamma (CPG) distribution, where the number of "individuals" captured is

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

and where the weight of each individual $j$ follows a gamma distribution:

$$W_j \sim \text{Gamma}(k, \theta)$$

such that total catch $C = \sum_{j=1}^{N} W_j$. In the main text, I present a reparameterization in terms of numbers density $\lambda_i$ and expected individual mass $\mu_i$ for each sample $i$, where gamma shape parameter $k$ is constant among samples but expected individual mass $\mu_i$ differs among samples (where $\mu_i = k\theta_i$). Following Foster and Bravington (2013), I specify that both numbers density $\lambda_i$ and expected individual mass $\mu_i$ are predicted using a log-linked linear predictor, and where the offset $a_i$ affects expected catch in numbers. I also show that this parameterization generates a similar functional form for expected encounter probability $r$ and positive catch rates $r$ as our alternative Poisson-link model.

Unfortunately, the CPG likelihood function (Smyth 1996) is expensive to evaluate:

$$\Pr(c_i = C) = \begin{cases} \exp(-\lambda_i) & \text{if} \quad c_i = 0 \\ W(c_i, \lambda_i, k, \mu_i) \times \exp\left(-\dfrac{c_i}{\mu_i} - \lambda_i - \log(c_i)\right) & \text{if} \quad c_i > 0 \end{cases}$$

where $W(c_i, \lambda_i, k, \mu_i)$ is a integration constant that requires calculating the sum of an infinite series:

$$W(c_i, \lambda_i, k, \mu_i) = \sum_{j=1}^{\infty} \frac{\lambda_i^j \left(\dfrac{c_i}{\mu_i}\right)^{jk}}{j!\,\Gamma(jk)}$$

where this likelihood can instead be approximated using Markov-chain sampling of $N_i$ (e.g., Lauderdale 2012). However, numerical techniques to approximate this likelihood function are a topic of ongoing research (Dunn and Smyth 2005, 2008).

## References

Dunn, P.K., and Smyth, G.K. 2005. Series evaluation of Tweedie exponential dispersion model densities. Stat. Comput. **15**(4): 267–280. doi:10.1007/s11222-005-4070-y.

Dunn, P.K., and Smyth, G.K. 2008. Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. Stat. Comput. **18**(1): 73–86. doi:10.1007/s11222-007-9039-6.

Foster, S.D., and Bravington, M.V. 2013. A Poisson–Gamma model for analysis of ecological non-negative continuous data. Environ. Ecol. Stat. **20**(4): 533–552. doi:10.1007/s10651-012-0233-0.

Lauderdale, B.E. 2012. Compound Poisson–gamma regression models for dollar outcomes that are sometimes zero. Polit. Anal. **20**(3): 387–399. doi:10.1093/pan/mps018.

Lecomte, J.-B., Benoît, H.P., Ancelet, S., Etienne, M.-P., Bel, L., and Parent, E. 2013. Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume. Methods Ecol. Evol. **4**(12): 1159–1166. doi:10.1111/2041-210X.12122.

Smyth, G.K. 1996. Regression analysis of quantity data with exact zeros. *In* Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management, Gold Coast, Australia. pp. 17–19.