

 Open access • Journal Article • DOI:10.1139/CJFAS-2017-0266

Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative — [Source link](#)

James T. Thorson

Institutions: National Marine Fisheries Service

Published on: 01 Sep 2018 - Canadian Journal of Fisheries and Aquatic Sciences (NRC Research Press)

Topics: Sampling (statistics)

Related papers:

- [Guidance for decisions using the Vector Autoregressive Spatio-Temporal \(VAST\) package in stock, ecosystem, habitat and climate assessments](#)
- [Comparing estimates of abundance trends and distribution shifts using single- and multispecies models of fishes and biogenic habitat](#)
- [TMB: Automatic Differentiation and Laplace Approximation](#)
- [Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes](#)
- [An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/three-problems-with-the-conventional-delta-model-for-biomass-11r3ncad3m>



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

Journal:	<i>Canadian Journal of Fisheries and Aquatic Sciences</i>
Manuscript ID	cjfas-2017-0266.R2
Manuscript Type:	Article
Date Submitted by the Author:	10-Oct-2017
Complete List of Authors:	Thorson, James; Northwest Fisheries Science Center,
Is the invited manuscript for consideration in a Special Issue? :	N/A
Keyword:	delta-generalized linear model, Tweedie distribution, compound Poisson-gamma distribution, biomass sampling data, spatio-temporal model

SCHOLARONE™
Manuscripts

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

3

4 Running header: Efficient model for biomass sampling

5

6 James Thorson

7

8 Fisheries Resource Assessment and Monitoring Division, Northwest Fisheries Science

9 Center, National Marine Fisheries Service, NOAA, Seattle, WA, USA

10 James.Thorson@noaa.gov

11

12 Keywords: delta-generalized linear model; Tweedie distribution; compound Poisson-gamma
13 distribution; spatio-temporal model; encounter probability; bottom trawl sampling;
14 biomass sampling data

15

16

17 Abstract

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

32

33 **Introduction**

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

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

51 Biomass-sampling data are often analysed using a “delta” (a.k.a. hurdle) model
52 (Aitchison 1955, Lo et al. 1992, Stefansson 1996) that includes two components: the
53 probability of encountering the species, and the expected biomass given that the species is
54 encountered. This “delta-model” remains one of the most common types of regression used
55 by ecologists and fisheries scientists (Maunder and Punt 2004, Zuur et al. 2009). However, it
56 has several theoretical and practical draw-backs (as discussed below). One increasingly
57 popular alternative to the conventional delta-model is using a compound Poisson-gamma

58 (CPG) model (Smyth 1996, Foster and Bravington 2013, Lecomte et al. 2013), which is
59 derived by assuming that biomass samples capture a Poisson-distributed number of
60 individuals, where the biomass of each individual follows an independent gamma
61 distribution. This CPG model is a special case of the Tweedie distribution (Foster and
62 Bravington 2013), but it remains computationally expensive to evaluate and therefore is
63 difficult to combine with other detailed model components (e.g., spatio-temporal variation;
64 (Cressie and Wikle 2011)).

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

79 **Methods**

80 **Defining the conventional delta-model**

81 Fisheries scientists have analysed biomass-sampling data using delta-models for nearly thirty
82 years (Lo et al. 1992, Stefansson 1996). Historically, these delta-models have been fitted to

83 data by estimating parameters for two separate and independent generalized linear models
 84 (GLMs):

- 85 1. *Encounter probability*: the probability of encountering the species at a given place and
 86 time;
- 87 2. *Positive catch rates*: the probability density function for catch in biomass given that the
 88 species is encountered.

89 Predictions from the two GLMs can then be multiplied together to predict local density, and
 90 this in turn is used to predict total abundance across a pre-specified spatial domain.

91 The “encounter probability” component of the delta-model defines the probability p_i that
 92 catch C_i for the i th sample is non-zero:

$$I(C_i > 0) \sim \text{Bernoulli}(p_i) \quad (1a)$$

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

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

100 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
 101 i th observation affecting encounter probability p , although other potential link-functions
 102 include the probit and complementary log-log link-functions (Zuur et al. 2009 pg 248). In the
 103 following, we refer to the logit-link as the “conventional” delta-model due to its widespread
 104 use in fisheries science.

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

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

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

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

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

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

120 **Three “theoretical” problems with conventional delta-models**

121 I note three major draw-backs to using the conventional delta-model:

- 122 1. Difficulties in interpreting covariates;
- 123 2. Assumed independence between model components;
- 124 3. Biologically implausible form when removing covariates.

125 *Drawback #1: Difficulties in interpreting coefficients*

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

144 *Drawback #2: Assumed independence among components*

145 Using the conventional delta-model, the “encounter probability” and “positive catch rate”
146 components are assumed to be statistically independent, i.e., knowledge about encounter
147 probability p gives no information about the likely distribution for positive catches r . This
148 assumption is contrary to a large body of evidence suggesting (1) that abundant species have
149 wide ranges, such that frequently encountered species also have higher density throughout
150 their range (Gaston 1994), and also (2) that an increase in local density will decrease the

151 probability of failing to detect a species that is present (Royle and Nichols 2003). Both
152 phenomena suggest that a location with increased probability of encounter (higher p) will
153 tend to have greater catch rates given an encounter (higher r), as has been argued previously
154 for nonparametric zero-inflated models (Liu and Chan 2011). As one concrete example, the
155 conventional delta-model specifies that positive catch rates r_i increases linearly with
156 increased area-swept a_i for a given sample, but that increased area-swept has no effect on
157 encounter probability p_i . This specification is inconvenient, because an increase in area-
158 swept for many species will increase the probability of sampling at least one occupied patch
159 (Lecomte et al. 2013). In response, an ecologist could chose to also include area-swept as a
160 predictor for encounter probability p . However, there is no way to interpret the estimated
161 coefficient as a “linear offset” when using the conventional logit-link for encounter
162 probability (see *Drawback #1* above), so this area-swept covariate would then be estimated to
163 have a nonlinear impact on expected catches.

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

165 Ecologists often have little data with which to estimate a multitude of potential ecological
166 processes. The presence of “tapering effects” (i.e., many ecological processes with gradually
167 declining effect-sizes for any given system) has driven interest in using model selection to
168 identify “parsimonious” ecological models (Burnham and Anderson 2002). Parsimony in this
169 case is defined as an appropriate number of parameters that minimizes total predictive error
170 for a given data set (simultaneously low bias and imprecision). In many cases, parsimony is
171 achieved by identifying a flexible family of models, where analysts can use model selection
172 to identify the appropriate degree of model complexity. This approach is most effective,
173 however, when the model that eliminates covariates remains biologically plausible (e.g., is
174 likely to provide a good fit for species on average). As corollary of *Drawback #2*, it will
175 often be more statistically efficient to assume that a covariate associated with high encounter

176 probability will also likely be associated with high positive catch rates (and vice versa). For
 177 example, if the density of rocky substrate is associated with increased encounter probability
 178 for a refuge-seeking fish, then it is also likely associated with increased positive catch rates
 179 because sampling will likely include a greater number of occupied habitat patches. By
 180 contrast, removing covariates in a delta-model generally involves specifying that a given
 181 covariate affects encounter probability but not positive catch rates (or vice versa).

182 **Solutions from using an alternative “Poisson-link” model**

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

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

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

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

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

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

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

198 When data are few, predicted biomass per group w can be estimated as a single parameter.

199 However, a more general treatment involves specifying w via a log-linked linear predictor:

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

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

201 $\mathbf{z} = \emptyset$.

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

216 The Poisson-link model responds to all three theoretical problems with the conventional
 217 delta-model:

218 *1. Difficulties in interpreting coefficients:* The Poisson-link model simplifies interpretation
 219 of covariates. In particular, covariates $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_w$ both predict changes in log-density, so
 220 e.g., a 0.01 increase in $\boldsymbol{\beta}_n \mathbf{x}_i$ is associated with approximately a 1% increase in density.

221 Similarly, a random effect γ_j with a standard deviation of 0.01 explains approximately a
222 1% coefficient of variation in density. Both fixed- and random-effects therefore have a
223 similar interpretation in predicting variation in population density, because both affect
224 density via a log link function.

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

235 3. *Biologically implausible form when removing covariates*: Finally, the Poisson-link
236 model by default specifies that a covariate affecting group-density (i.e., β_n) influences
237 predictions of both encounter probability p and positive catch rates r . To continue our
238 previous example, a covariate representing the local density of rocky substrate might be
239 selected for group-density but not for average weight, and this reduction in the number of
240 estimated parameters still retains a biologically meaningful impact of substrate on both
241 encounter probability and positive catch rates. This specification will allow a smaller
242 number of estimated parameters to explain variation in both encounter probability and
243 positive catch rates whenever p_i and r_i are positively correlated in the form predicted by
244 the Poisson-link model.

245 Despite these improvements over the conventional delta-model, the Poisson-link and
 246 conventional delta-model are identical (e.g., have identical maximum likelihood and provide
 247 identical predictions of density) for several potential model configurations. For example,
 248 delta-model parameters (β_p and β_r) can be converted to Poisson-link parameters (β_n and β_w)
 249 and vice-versa when using an intercept-only model (e.g., \mathbf{x} is a design matrix and $\mathbf{z} = \emptyset$) and
 250 a constant area-offset (i.e., $a_i = a$ for all observations i) via the relations:

$$\text{logit}^{-1}(\beta_p) = 1 - \exp(-a \times \exp(\beta_n)) \quad (7)$$

$$\exp(\beta_r) = \frac{\exp(\beta_n)}{1 - \exp(-a \times \exp(\beta_n))} \exp(\beta_w)$$

251 where these relations are derived from the definition of predicted encounter probability p and
 252 positive catch rate r (see Table 2) and the identical likelihood used in each model (Eq. 1a and
 253 2a). The conventional delta and Poisson-link models generally differ (i.e., result in different
 254 maximum likelihoods and density predictions) whenever they include either a covariate,
 255 variable area-offset, or random effects.

256 **Comparison with compound Poisson-gamma distribution**

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

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

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

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

265 where k is the gamma shape parameter and $\mu_i k^{-1}$ is the scale parameter, such that total catch
 266 $C_i = \sum_{j=1}^{N_i} W_{ij}$. The parameterization used here involves estimating λ_i , μ_i , and k , where λ_i
 267 and μ_i can be specified via a link-function and linear predictors and k is assumed constant for
 268 all samples i (see Foster and Bravington (2013) for further discussion). The CPG distribution
 269 generates a “power-law” relationship between the expected value, $E(C_i) = \eta_i$, and variance,
 270 $\text{Var}(C_i) = \phi_i \eta_i^\nu$. By contrast, the typical Tweedie parameterization directly estimates the
 271 power parameter ($1 < \nu < 2$), uses a constant dispersion ($\phi_i = \phi$ for all i) and a linear
 272 predictor for $\log(\eta)$, and has been used extensively elsewhere (Candy 2004, Shono 2008,
 273 Lecomte et al. 2013, Berg et al. 2014). The CPG is identical to the Tweedie parameterization
 274 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
 275 (2012) for derivation).

276 Similar to the Poisson-link model, the CPG distribution (using the Foster and Bravington
 277 (2013) parameterization) specifies a log-link for both group-density and average weight:

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

278 and this results in an identical derivation for expected encounter probability p and positive
 279 catch rates r as the Poisson-link model (Table 2). The CPG distribution therefore responds to
 280 all three theoretical problems similarly to the Poisson-link model (Foster and Bravington
 281 2013). The Foster-Bravington parameterization of the CPG distribution then involves
 282 estimating fixed effects $\boldsymbol{\beta}_\lambda$, $\boldsymbol{\beta}_\mu$, and k by finding their values that maximize the likelihood
 283 function.

284 However, the CPG likelihood function (Smyth 1996, Dunn and Smyth 2005) is
 285 computationally expensive to evaluate because it involves approximating an integration

286 constant W as the sum of an infinite series (Appendix S1), or approximating the CPG
287 distribution using numerical sampling (e.g., Lauderdale 2012). Approximating the sum of an
288 infinite series has computational cost determined by the number of terms in the summation,
289 and numerical sampling requires introducing a large number of discrete-valued random
290 effects. In the following, I evaluate the CPG likelihood using an upper limit of 1000 for
291 calculating W , and confirm that the log-likelihood (given maximum-likelihood estimates of
292 all parameters) is identical to the value generated by package *fishMod* (Foster et al. 2016) to
293 tolerance of 10^{-6} . In practice, approximating this infinite series can be efficiently
294 implemented by specialized numerical techniques (Dunn and Smyth 2008, Foster and
295 Bravington 2013), e.g., by analytically determining an efficient lower and upper bound for
296 the summation. I encourage further comparison of numerical techniques as a topic for future
297 research, but claim that the CPG likelihood is computationally expensive relative to the
298 Poisson-link model because the former requires summing across as many as 50 terms (Foster
299 and Bravington 2013 pg. 539), while the Poisson-link model requires evaluating only a single
300 term.

301 Unlike the CPG distribution, the “Poisson-link” model permits a fast, closed-form
302 calculation of the model likelihood (using Eq. 1a, 2a, and 3-6). Both the CPG and Poisson-
303 link model specify the density of groups (λ and n , respectively for CPG and Poisson-link)
304 and average weight per group (μ and w , respectively) via log-linked linear predictors. The
305 main difference, however, is that the proposed Poisson-link specifies a different mean-
306 variance relationship than the CPG model.

307 **Case study data: Bottom trawl survey database**

308 In the following, I first compare the fit of the conventional and alternative Poisson-link
309 models with the CPG using real-world data and a simple model (estimating annual intercepts
310 as fixed effects). I then compare the conventional and alternative Poisson-link models using

311 a more complicated model (estimating fixed annual intercepts, plus spatial and spatio-
312 temporal variation), which is computationally infeasible using implementations of the CPG
313 distribution available in Template Model Builder (Kristensen et al. 2016).

314 For each comparison I use bottom-trawl survey data from seven marine ecosystems:

- 315 1. *Eastern Bering Sea* – Survey operated by the Alaska Fisheries Science Center (AFSC)
316 obtained from a fixed-station design (Lauth and Conner 2016);
- 317 2. *Gulf of Alaska* – Survey operated by the AFSC obtained from a randomized design (Von
318 Szalay and Raring 2016);
- 319 3. *Aleutian Islands* – Survey operated by the AFSC obtained from a randomized design
320 (Raring et al. 2016);
- 321 4. *US West Coast* –The West Coast groundfish bottom trawl survey operated by the
322 Northwest Fisheries Science Center (NWFSC), obtained from a stratified-random design
323 (Keller et al. 2017);
- 324 5. *North Sea* – The North Sea international bottom trawl survey (NS-IBTS), restricting data
325 to 1991-2015 obtained using a “Gov” gear in quarter 1 (winter) (ICES 2012);
- 326 6. *Scottish West Coast* – The Scottish West Coast international bottom trawl survey (SWC-
327 IBTS), restricting data to 1991-2015 obtained using a “Gov” gear in quarter 1;
- 328 7. *Celtic Sea and Bay of Biscay* – The French demersal survey (EVHOE) of the Celtic Sea
329 and Bay of Biscay, operating by the French Research Institute for Exploitation of the Sea
330 (IFREMER) from 1997-2015 in quarter 4 (fall) (Mahé and Poulard 2005).
- 331 8. US survey protocols (#1-4) contain biomass-per-unit-area data (i.e., samples are
332 standardized to a constant area swept), so I assume that area-swept is constant for these
333 surveys. European survey protocols (#5-7) are described in ICES (2012), Public
334 databases for surveys #1-4 contain biomass-per-unit-area (i.e., samples are standardized
335 to a constant area swept), while for those for surveys #5-7 contain raw biomass/numbers

336 and a measure of fishing effort (the duration of tows in minutes). I therefore assume that
337 “tow duration” is proportional to area swept a_i for surveys #5-7, and that area-swept is
338 constant for surveys #1-4. Surveys
339 US survey protocols (#1-4) are described in Stauffer (2004), and publicly available databases
340 for these surveys contain biomass-per-unit-area (i.e., samples are standardized to a constant
341 area swept). I therefore analyse “biomass-per-area” as catch, and fix area swept a_i at a
342 constant value for all samples in these surveys. European survey protocols (#5-7) are
343 described in ICES (2012), and the public Datras database for these surveys contains numbers
344 caught for multiple length bins and a measure of fishing effort (the duration of tows in
345 minutes) for each sample, as well as records of individual biomass and length. I calculate a
346 length-weight key from records of individual biomass and length, use this key to convert
347 numbers-at-length to biomass-at-length, and then calculate total biomass for each sample.

348 For each survey, I restrict data to the twenty most frequently encountered fishes (see Fig.
349 2 for annual sample sizes). Surveys #6-7 had sufficient weight-at-length records to calculate
350 biomass data for fewer than twenty species, so I used biomass data for as many species as
351 were available. All surveys are publicly available and can be accessed using R package
352 *FishData* (<https://github.com/james-thorson/FishData>), which in turn uses R package
353 *icesDatras* (<https://github.com/ices-tools-prod/icesDatras>) to download data for surveys #5-7.

354 **Comparison #1: Annual-intercept models**

355 I first compare the conventional delta-model and Poisson-link model against the compound
356 Poisson-gamma distribution using a simple model where each model component has a
357 separate intercept by year. Parameters for all models are estimated via maximum likelihood
358 using release number 1.5.0 (<https://doi.org/10.5281/zenodo.834777>) of package *VAST*
359 (www.github.com/james-thorson/VAST; Thorson and Barnett (2017)), which estimates
360 parameters using Template Model Builder (Kristensen et al. 2016) within the R statistical

361 platform (R Core Team 2015). Model selection is conducted using the marginal Akaike
 362 information criterion, AIC (Akaike 1974) as is widely used in ecology and fisheries
 363 (Burnham and Anderson 2002), based on the marginal log-likelihood and the number of fixed
 364 effects. I do not attempt to calculate the conditional AIC (Vaida and Blanchard 2005), which
 365 measures complexity by the number of fixed effects plus the effective degrees of freedom for
 366 random effects. To my knowledge, conditional AIC has not been used in fisheries science,
 367 and I recommend its exploration as a topic for future research. The likelihood is identical for
 368 conventional delta-model and Poisson-link models (see Eq. 8 and associated text), and
 369 different than that for the compound Poisson-gamma distribution (because the CPG has a
 370 different exponent for Taylor's power law, see section *Comparison with compound Poisson-*
 371 *gamma distribution*). I therefore present the difference in AIC between the Poisson-link
 372 model and the compound Poisson-gamma model. I present this comparison to determine
 373 whether the Poisson-link model gains computational efficiency while maintaining
 374 comparable model fit to the CPG model.

375 **Comparison #2: Spatio-temporal model**

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

$$\text{logit}(p(s, t)) = \beta_p(t) + \omega_p(s) + \varepsilon_p(s, t) \quad (11)$$

$$\log(r(s, t)) = \beta_r(t) + \omega_r(s) + \varepsilon_r(s, t)$$

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

$$\boldsymbol{\omega}_p \sim \text{MVN}(\mathbf{0}, \sigma_{p\omega}^2 \mathbf{R}) \quad (12)$$

$$\boldsymbol{\varepsilon}_p(t) \sim \text{MVN}(\mathbf{0}, \sigma_{p\varepsilon}^2 \mathbf{R})$$

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

$$\log(n(s, t)) = \beta_n(t) + \omega_n(s) + \varepsilon_n(s, t) \quad (13)$$

$$\log(w(s, t)) = \beta_w(t) + \omega_w(s) + \varepsilon_w(s, t)$$

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

398 After estimating parameters, I then evaluate model performance by comparing the
 399 estimated standard deviation of spatial and spatio-temporal variation for “positive catch
 400 rates” r in the conventional delta-model with these standard deviations for “average weight”
 401 w in the Poisson-link model. I do not compare the variance for encounter probability p
 402 because it is not easily interpretable in the conventional delta model (as explained in the
 403 previous section “*Drawback #1: Difficulties in interpreting coefficients*”). However, this
 404 comparison is appropriate for positive catch rates r because the Poisson-link model

405 decomposes variance in $\log(r)$ into three additive components (see the conversion from w
 406 and p to r in Table 2):

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

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

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

408 Therefore, if knowledge of encounter probability p is informative about positive catch rates

409 r , then this will cause average weight in the alternative model to have lower spatial and/or

410 spatio-temporal variance than for positive catch rates in the conventional model (i.e., if

411 $\text{Cov}\left(\log(r), \log\left(\frac{n}{p}\right)\right) > 0$ then $\sigma_{w\omega}^2 + \sigma_{w\varepsilon}^2 < \sigma_{r\omega}^2 + \sigma_{r\varepsilon}^2$). Alternatively, if encounter

412 probability p is statistical independent or negatively associated about positive catch rates r ,

413 then the opposite will occur (i.e., if $\text{Cov}\left(\log(r), \log\left(\frac{n}{p}\right)\right) \leq 0$ then $\sigma_{w\omega}^2 + \sigma_{w\varepsilon}^2 \geq \sigma_{r\omega}^2 + \sigma_{r\varepsilon}^2$).

414 I therefore record (1) the proportion of species for each region where the conventional or
 415 alternative model was selected as parsimonious using the Akaike Information Criterion
 416 (AIC); (2) the pointwise (a.k.a. marginal) standard deviation of spatial and spatio-temporal
 417 variance for both model components; (3) the predictive standard deviation of an abundance
 418 index derived from each model (indices are area-weighted following Thorson et al. (2015)). I
 419 hypothesize that the Poisson-link model will be more parsimonious than the conventional
 420 delta-model for the majority of species. The pointwise variances $\sigma_{r\omega}^2$ and $\sigma_{r\varepsilon}^2$ from the
 421 conventional model and $\sigma_{w\omega}^2$ and $\sigma_{w\varepsilon}^2$ from the alternative model are directly comparable, and
 422 I hypothesize that spatial and spatio-temporal variances for the alternative model will be
 423 lower because the encounter probability p (estimated from proportion of nearby samples that
 424 encounter the species) is informative about local positive catch rates r .

425 Simulation experiment

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

- 430 1. I obtain data for a single species (arrowtooth flounder in the EBSBTS data; Survey #1
431 above) including the depth for each sampling location.
- 432 2. I fit each of three models to these data, while including as fixed effect both year (as an
433 annual intercept) and depth (standardized to have a mean of zero and a standard deviation
434 of one), and while not including any random effects.
- 435 3. For each model in Step #2, I generate 100 simulated data sets, using the estimated depth
436 effect, variance parameters, and simulating new annual intercepts that have the same
437 mean and standard deviation as the sample mean and standard deviation of estimated
438 intercepts (from Step #2). Each simulated data set has same the annual sample size and
439 sampling locations as the original data set (in Step #1).
- 440 4. For each of these 300 simulated data sets, I fit each of the three models (i.e., 900 model
441 fits total). For each model fit, I record the estimated depth effect for both model
442 components.

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

450 **Results**

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

458 The complicated “spatio-temporal” model applied to these same data shows that using a
459 Poisson-link model also results in better fit than the conventional delta-model for the vast
460 majority of populations in 6 out of 7 regions (Fig. 4). Models again have an identical number
461 of estimated parameters, so a higher log-likelihood also indicates greater parsimony (e.g.,
462 using the Akaike Information Criterion). The average AIC weight for the Poisson-link model
463 is >80% for the same 6 regions. The exception is for the California Current, where each
464 model is each selected for 10 of 20 species. In this region, the implied correlation between
465 encounter probability and positive catch rates apparently does not improve fit relative to
466 assuming independence between detection probability and positive catch rates. However, the
467 implied correlation does improve fit for the majority of populations in other regions.

468 The conventional and alternative models have essentially identical estimates of residual
469 variation in positive catch rates ($\sigma_M = 1.25$ or 1.26), indicating that both models attribute a
470 roughly identical portion of sampling variance to the combination of spatial and spatio-
471 temporal variation (Fig. 5). As hypothesized, however, the Poisson-link model results in a
472 lower standard deviation for spatial and spatio-temporal variation (Fig. 6). The standard
473 deviation is not directly comparable for the first-model component between models, because
474 $\sigma_{r\omega}$ and $\sigma_{r\epsilon}$ (from the conventional model; top row of Fig. 6) affect r via a logit-link function

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

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

493 Finally, the simulation experiment (Fig. 8) shows that the Poisson-link model estimates a
494 2.0% increase in group-density and a -0.9% decrease in biomass-per-group when depth
495 increases by 1% of its standard deviation for arrowtooth flounder in the Eastern Bering Sea
496 (Fig. 8, vertical dotted lines in middle columns). The CPG estimates qualitatively similar
497 depth effects (a 1.3% increase in group-density and a -0.7% decrease in biomass-per-group),
498 while the depth effect for encounter probability in the conventional delta model (β_p) is highly
499 different. This difference arises because the conventional delta-model uses a logit-link

500 function, and therefore the estimated depth coefficient for the delta-model cannot be used
501 calculate a single value for the increase in encounter probability per change in depth (instead
502 the predicted increase in encounter probability with depth changes each year depending on
503 the intercept for that year). The simulation experiment also confirms (1) that all three models
504 generate unbiased estimates of depth effects when the simulation and estimation models
505 match (Fig. 8 panels a/e/i), and (2) that the Poisson-link and CPG estimation models have
506 similar performance to one-another, regardless of whether data are generated by using the
507 Poisson-link or CPG simulation models (Fig. 8 panels e/f/h/i). Finally, a comparison of
508 model-selection results from the simulation experiment (results not shown) confirms that AIC
509 identifies the data-generating model as the most parsimonious estimation model in nearly
510 100% of simulation replicates. This result confirms that AIC is a useful metric to evaluate
511 model performance using real-world data (i.e., in Fig. 3-4).

512 **Discussion**

513 Delta-models using a logit-link for encounter probabilities and a log-link for positive catch
514 rates have a long history in fisheries science (Stefansson 1996, Maunder and Punt 2004), and
515 I have presented three theoretical arguments for why this conventional delta-model is
516 unsatisfactory, namely (1) difficulties in interpreting how covariates for encounter probability
517 affect population density, (2) the lack of dependence between encounter probability and
518 positive catch rates, and (3) the biologically implausible form when removing covariates for
519 one or the other model component. I have then shown how these three difficulties are
520 addressed using a new “Poisson-link” model for biomass sampling data that can be
521 interpreted as a computationally efficient alternative to the compound Poisson-gamma
522 distribution. Application to 113 populations in seven marine regions shows that the Poisson-
523 link model substantially improves fit by using knowledge of encounter probabilities to
524 decrease otherwise-unexplained variation in positive catch rates. However, this Poisson-link

525 model decreases average confidence-interval width for abundance indices in only two of
526 seven regions. I therefore conclude that the Poisson-link model is not likely to substantially
527 increase the information available to stock assessments when used to estimate abundance
528 indices. However, improvements in fit, interpretability, and parsimony relative to a
529 conventional delta-model are still likely to be useful when estimating habitat maps,
530 estimating habitat associations, and fitting ecological models to samples of fish biomass.

531 I envision several useful avenues for future research. Most obviously, the Poisson-link
532 model could be compared more exhaustively with the compound Poisson-gamma
533 distribution, as well as other alternatives (e.g., the Law-of-Leaks “LoL” model, Ancelet
534 (2010)), including more detailed comparison of the different mean-variance relationships
535 implied by these potential models. This comparison could then identify taxa and model-
536 structures where the CPG, LoL, and Poisson-link models are more or less statistically
537 efficient. Given the many potential numerical techniques to implement the CPG (Dunn and
538 Smyth 2005, 2008, Foster and Bravington 2013), one of these will hopefully prove to be
539 computationally feasible for the spatio-temporal models as explored here. I note that the
540 CPG distribution automatically follows Taylor’s rule (i.e., a power law mean-variance
541 relationship), and therefore has stronger theoretical support for ecological processes. I also
542 recommend future research exploring the potential consequences of the ignoring variation
543 among samples when predicting biomass per group (i.e., fixing $w_i = w$). This restriction is
544 particularly appealing when introducing additional model complexity (i.e., modelling
545 multiple species simultaneously; Thorson et al. (2016a)). The current application to 113
546 populations worldwide shows that there is substantial variation in w even after accounting for
547 the effect of encounter probabilities, but determining the impact of restricting $w_i = w$ on
548 model performance will require further simulation testing. This simulation experiment could

549 presumably be conditioned on the range of spatial and spatio-temporal variances estimated in
550 this study.

551 Finally, the past decade has seen rapid growth in a variety of useful approximations for
552 otherwise slow or intractable processes that arise in ecology. Examples include
553 approximating individual birth-death demographics using Markov chains (Hubbell 2011),
554 estimating the likelihood of ecological rates given unobserved (latent) variables via the
555 Laplace approximation (Skaug and Fournier 2006, Kristensen et al. 2016), or approximating
556 spatial variation and individual movement using finite-element analysis methods (Lindgren et
557 al. 2011, Thorson et al. 2017). Collectively, these approximations are useful when they
558 permit the development of models with increased realism regarding otherwise-neglected
559 processes in ecological systems (e.g., a “zero-sum” linkage between regional and local
560 species pools for describing community richness; Hubbell (2011)). In this light, the Poisson-
561 link model can be viewed as a computationally-efficient approximation to a common
562 sampling design, where biomass samples arise from a weighing individuals that vary in
563 individual biomass. I recommend ongoing development and testing of efficient
564 approximations to sampling processes, and hope that these approximations will collectively
565 allow biological rates (births, deaths, and movement) to be simultaneously estimated for
566 entire communities occurring on heterogeneous landscapes using available data worldwide.
567 Hopefully this will then allow us to “fill in the missing spaces” where messy or opportunistic
568 data exist but ecologists have not previously conducted comparative analyses (e.g., in the white
569 spaces in Fig. 1).

570 **Acknowledgements**

571 I thank E Dick, J Hastie, M McClure, and four anonymous reviewers for comments on a
572 previous draft. I also thank A Magnusson, S Large, C Millar, and C Berg for contributions to
573 the *icesDatras* R package, which have allowed me to include biomass data from waters near

574 Europe. Finally, I thank the many participants who conducted sampling in these seven
575 bottom-trawl surveys: these are an invaluable resource for current and future marine
576 ecologists.

577

Draft

578 **Bibliography**

- 579 Aitchison, J. 1955. On the Distribution of a Positive Random Variable Having a Discrete
580 Probability Mass at the Origin. *J. Am. Stat. Assoc.* **50**(271): 901.
581 doi:10.2307/2281175.
- 582 Akaike, H. 1974. New look at statistical-model identification. *IEEE Trans. Autom. Control*
583 **AC19**(6): 716–723.
- 584 Ancelet, S., Etienne, M.-P., Benoît, H., and Parent, E. 2010. Modelling spatial zero-inflated
585 continuous data with an exponentially compound Poisson process. *Environ. Ecol.*
586 *Stat.* **17**(3): 347–376. doi:10.1007/s10651-009-0111-6.
- 587 Arcuti, S., Calculli, C., Pollice, A., D’Onghia, G., Maiorano, P., and Tursi, A. 2013. Spatio-
588 temporal modelling of zero-inflated deep-sea shrimp data by Tweedie generalized
589 additive. *Statistica* **73**(1): 87–101. doi:10.6092/issn.1973-2201/3987.
- 590 Augustin, N.H., Trenkel, V.M., Wood, S.N., and Lorange, P. 2013. Space-time modelling of
591 blue ling for fisheries stock management. *Environmetrics* **24**(2): 109–119.
592 doi:10.1002/env.2196.
- 593 Berg, C.W., Nielsen, A., and Kristensen, K. 2014. Evaluation of alternative age-based
594 methods for estimating relative abundance from survey data in relation to assessment
595 models. *Fish. Res.* **151**: 91–99. doi:10.1016/j.fishres.2013.10.005.
- 596 Burnham, K.P., and Anderson, D. 2002. Model Selection and Multi-Model Inference. *In* 2nd
597 edition. Springer, New York.
- 598 Candy, S.G. 2004. Modelling catch and effort data using generalised linear models, the
599 Tweedie distribution, random vessel effects and random stratum-by-year effects.
600 *CCAMLR Sci.* **11**: 59–80.
- 601 Clark, J.S. 2016. Why species tell more about traits than traits about species: predictive
602 analysis. *Ecology* **97**(8): 1979–1993. doi:10.1002/ecy.1453.
- 603 Cressie, N., and Wikle, C.K. 2011. Statistics for spatio-temporal data. John Wiley & Sons,
604 Hoboken, New Jersey.
- 605 Dick, E.J. 2004. Beyond “lognormal versus gamma”: discrimination among error
606 distributions for generalized linear models. *Fish. Res.* **70**(2-3): 351–366.
607 doi:10.1016/j.fishres.2004.08.013.
- 608 Dunn, P.K., and Smyth, G.K. 2005. Series evaluation of Tweedie exponential dispersion
609 model densities. *Stat. Comput.* **15**(4): 267–280.
- 610 Dunn, P.K., and Smyth, G.K. 2008. Evaluation of Tweedie exponential dispersion model
611 densities by Fourier inversion. *Stat. Comput.* **18**(1): 73–86. doi:10.1007/s11222-007-
612 9039-6.
- 613 Foster, S.D., and Bravington, M.V. 2013. A Poisson–Gamma model for analysis of
614 ecological non-negative continuous data. *Environ. Ecol. Stat.* **20**(4): 533–552.
615 doi:10.1007/s10651-012-0233-0.
- 616 Foster, S.D., Foster, M.S., and SystemRequirements, C. 2016. Package “fishMod.” Available
617 from <http://cran.itam.mx/web/packages/fishMod/fishMod.pdf>.
- 618 Gaston, K.J. 1994. Measuring geographic range sizes. *Ecography* **17**(2): 198–205.
619 doi:10.1111/j.1600-0587.1994.tb00094.x.
- 620 Hubbell, S.P. 2011. *The Unified Neutral Theory of Biodiversity and Biogeography* (MPB-
621 32). Princeton University Press.
- 622 ICES. 2012. Manual of the International Bottom Trawl Surveys. Series of the ICES Survey
623 Protocols, International Council for the Exploration of the Sea (ICES), Copenhagen,
624 Denmark. Available from
625 [https://www.ices.dk/sites/pub/Publication%20Reports/ICES%20Survey%20Protocols](https://www.ices.dk/sites/pub/Publication%20Reports/ICES%20Survey%20Protocols%20(SISP)/SISP1-IBTSVIII.pdf)
626 [%20\(SISP\)/SISP1-IBTSVIII.pdf](https://www.ices.dk/sites/pub/Publication%20Reports/ICES%20Survey%20Protocols%20(SISP)/SISP1-IBTSVIII.pdf).

- 627 Keller, A.A., Wallace, J.R., and Methot, R.D. 2017. The Northwest Fisheries Science
628 Center's West Coast Groundfish Bottom Trawl Survey: History, Design, and
629 Description. NOAA Technical Memorandum, Northwest Fisheries Science Center,
630 Seattle, WA.
- 631 Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B.M. 2016. TMB: Automatic
632 Differentiation and Laplace Approximation. *J. Stat. Softw.* **70**(5): 1–21.
633 doi:10.18637/jss.v070.i05.
- 634 Lauderdale, B.E. 2012. Compound Poisson–Gamma Regression Models for Dollar Outcomes
635 That Are Sometimes Zero. *Polit. Anal.* **20**(3): 387–399. doi:10.1093/pan/mps018.
- 636 Lauth, R.R., and Conner, J. 2016. Results of the 2013 eastern Bering Sea continental shelf
637 bottom trawl survey of groundfish and invertebrate resources. NOAA Technical
638 Memorandum, Seattle, WA.
- 639 Lecomte, J.B., Benoît, H.P., Etienne, M.P., Bel, L., and Parent, E. 2013. Modeling the habitat
640 associations and spatial distribution of benthic macroinvertebrates: A hierarchical
641 Bayesian model for zero-inflated biomass data. *Ecol. Model.* **265**: 74–84.
642 doi:10.1016/j.ecolmodel.2013.06.017.
- 643 Lindgren, F., Rue, H., and Lindström, J. 2011. An explicit link between Gaussian fields and
644 Gaussian Markov random fields: the stochastic partial differential equation approach.
645 *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**(4): 423–498. doi:10.1111/j.1467-
646 9868.2011.00777.x.
- 647 Liu, H., and Chan, K.-S. 2011. Generalized Additive Models for Zero-Inflated Data with
648 Partial Constraints. *Scand. J. Stat.* **38**(4): 650–665. doi:10.1111/j.1467-
649 9469.2011.00748.x.
- 650 Lo, N.C., Jacobson, L.D., and Squire, J.L. 1992. Indices of Relative Abundance from Fish
651 Spotter Data based on Delta-Lognormal Models. *Can. J. Fish. Aquat. Sci.* **49**(12):
652 2515–2526.
- 653 Mahé, J.C., and Poulard, J.C. 2005. Manuel des protocoles de campagne halieutique.
654 IFREMER. Available from <http://archimer.ifremer.fr/doc/00036/14707/12013.pdf>.
- 655 Maunder, M.N., and Punt, A.E. 2004. Standardizing catch and effort data: a review of recent
656 approaches. *Fish. Res.* **70**(2-3): 141–159. doi:10.1016/j.fishres.2004.08.002.
- 657 Raring, N.W., Laman, E.A., Von Szalay, P.G., Rooper, C.N., and Martin, M.H. 2016. Data
658 report: 2012 Aleutian Islands bottom trawl survey. NOAA Technical Memorandum,
659 US Department of Commerce, National Oceanic and Atmospheric Administration,
660 National Marine Fisheries Service, Alaska Fisheries Science Center, Seattle, WA.
661 Available from [http://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-
662 AFSC-189/Main%20Body%20of%20Report.pdf](http://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-189/Main%20Body%20of%20Report.pdf).
- 663 R Core Team. 2015. R: A Language and Environment for Statistical Computing. R
664 Foundation for Statistical Computing, Vienna, Austria. Available from
665 <https://www.R-project.org/>.
- 666 Royle, J.A., and Nichols, J.D. 2003. Estimating abundance from repeated presence-absence
667 data or point counts. *Ecology* **84**(3): 777–790.
- 668 Shono, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis.
669 *Fish. Res.* **93**(1-2): 154–162.
- 670 Skaug, H., and Fournier, D. 2006. Automatic approximation of the marginal likelihood in
671 non-Gaussian hierarchical models. *Comput. Stat. Data Anal.* **51**(2): 699–709.
- 672 Smyth, G.K. 1996. Regression analysis of quantity data with exact zeros. *In* Proceedings of
673 the Second Australia-Japan Workshop on Stochastic Models in Engineering,
674 Technology and Management, Gold Coast, Australia. pp. 17–19.
- 675 Stauffer, G. 2004. NOAA protocols for groundfish bottom trawl surveys of the nation's
676 fishery resources. NOAA Technical Memorandum, National Oceanic and

- 677 Atmospheric Administration (NOAA), Seattle, WA. Available from
678 <http://www.mafmc.org/s/NOAA-protocols-for-bottom-trawl-surveys.pdf>.
- 679 Stefansson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and
680 delta approaches. *ICES J Mar Sci* **53**(3): 577–588.
- 681 Thorson, J.T., and Barnett, L.A.K. 2017. Comparing estimates of abundance trends and
682 distribution shifts using single- and multispecies models of fishes and biogenic
683 habitat. *ICES J. Mar. Sci.* **74**(5): 1311–1321. doi:10.1093/icesjms/fsw193.
- 684 Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C., and
685 Zipkin, E.F. 2016a. Joint dynamic species distribution models: a tool for community
686 ordination and spatio-temporal monitoring. *Glob. Ecol. Biogeogr.* **25**(9): 1144–1158.
687 doi:10.1111/geb.12464.
- 688 Thorson, J.T., Jannot, J., and Somers, K. 2017. Using spatio-temporal models of population
689 growth and movement to monitor overlap between human impacts and fish
690 populations. *J. Appl. Ecol.* **54**(2): 577–587. doi:10.1111/1365-2664.12664.
- 691 Thorson, J.T., Pinsky, M.L., and Ward, E.J. 2016b. Model-based inference for estimating
692 shifts in species distribution, area occupied and centre of gravity. *Methods Ecol. Evol.*
693 **7**(8): 990–1002. doi:10.1111/2041-210X.12567.
- 694 Thorson, J.T., Shelton, A.O., Ward, E.J., and Skaug, H.J. 2015. Geostatistical delta-
695 generalized linear mixed models improve precision for estimated abundance indices
696 for West Coast groundfishes. *ICES J. Mar. Sci. J. Cons.* **72**(5): 1297–1310.
697 doi:10.1093/icesjms/fsu243.
- 698 Thorson, J.T., and Ward, E. 2013. Accounting for space-time interactions in index
699 standardization models. *Fish. Res.* **147**: 426–433. doi:10.1016/j.fishres.2013.03.012.
- 700 Vaida, F., and Blanchard, S. 2005. Conditional Akaike information for mixed-effects models.
701 *Biometrika* **92**(2): 351–370.
- 702 Von Szalay, P.G., and Raring, N.W. 2016. Data report: 2015 Gulf of Alaska bottom trawl
703 survey. NOAA Technical Memorandum, US Department of Commerce, National
704 Oceanic and Atmospheric Administration, National Marine Fisheries Service, Alaska
705 Fisheries Science Center, Seattle, WA. Available from
706 [http://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-](http://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-189/Main%20Body%20of%20Report.pdf)
707 [189/Main%20Body%20of%20Report.pdf](http://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-189/Main%20Body%20of%20Report.pdf).
- 708 Wood, S.N., Pya, N., and Säfken, B. 2016. Smoothing Parameter and Model Selection for
709 General Smooth Models. *J. Am. Stat. Assoc.* **111**(516): 1548–1563.
710 doi:10.1080/01621459.2016.1180986.
- 711 Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., and Smith, G.M. 2009. *Mixed Effects*
712 *Models and Extensions in Ecology with R*, 1st edition. Springer, New York.
713
- 714

715 Table 1 – Names and symbols used in the main text, indicating whether each refers to data
 716 (“Data”), an index (“Index”), a fixed (“Fixed”) or random (“Random”) effect, or a derived
 717 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	σ_M	Fixed
Shape parameter for variation in individual weight in CPG distribution	k	Fixed
Fixed effects for delta-model	β_p, β_r	Fixed
Fixed effects for Poisson-link model	β_n, β_w	Fixed
Fixed effects for Compound Poisson-Gamma model	β_λ, β_μ	Fixed
Variance of random effects affecting p_i in conventional delta-model	$\sigma_{p\omega}^2, \sigma_{p\varepsilon}^2$	Fixed
Variance of random effects affecting r_i in conventional delta-model	$\sigma_{r\omega}^2, \sigma_{r\varepsilon}^2$	Fixed
Variance of random effects affecting n_i in Poisson-link delta-model	$\sigma_{n\omega}^2, \sigma_{n\varepsilon}^2$	Fixed
Variance of random effects affecting w_i in Poisson-link delta-model	$\sigma_{w\omega}^2, \sigma_{w\varepsilon}^2$	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	ν	DQ
Predicted number of individuals in compound Poisson-gamma (CPG) distribution	λ_i	DQ
Predicted density for sample i	d_i	DQ
Predicted group-density for sample i	n_i	DQ
Predicted average-weight for each individuals 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 weight in CPG distribution	μ_i	DQ
Mean of Tweedie parameterization for CPG distribution	η_i	DQ
Dispersion of Tweedie parameterization of CPG distribution	ϕ_i	DQ

718

719

720 Table 2 – Comparison of variables for the conventional delta-model, an alternative Poisson-link model, and the compound Poisson-gamma
 721 (CPG) model including how to calculate biomass density for each model. We also include equations to convert from variables in the alternative
 722 Poisson-link model (n and w) and CPG model (λ and μ) to variables in the conventional delta-model model (p and r), calculate the likelihood
 723 function, and simulate data given each model. The likelihood function is identical between conventional and Poisson-link delta-models, where

724 we use a bias-corrected lognormal density function in the main text, $f(C; r_i, \sigma_M^2) = \frac{1}{c_i \sigma_M \sqrt{2\pi}} \exp\left(-\frac{\left(\log(c_i) - \log(r_i) - \frac{\sigma_M^2}{2}\right)^2}{2\sigma_M^2}\right)$. However, evaluating

725 the likelihood function for the Poisson-link model requires converting predicted group density n_i and biomass-per-group w_i to encounter
 726 probability p_i and positive catch rates r_i . The likelihood for the compound Poisson-gamma model is from Foster and Bravington (2013), see
 727 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 2nd row).

	Conventional delta-model	Poisson-link delta-model	Compound Poisson-gamma model
<i>Component #1</i>	Encounter probability p : $\logit(p_i) = \boldsymbol{\beta}_p^T \mathbf{x}_i + \boldsymbol{\gamma}_p^T \mathbf{z}_i$	Group density n : $\log(n_i) = \boldsymbol{\beta}_n^T \mathbf{x}_i + \boldsymbol{\gamma}_n^T \mathbf{z}_i$	Group density λ : $\log(\lambda_i) = \boldsymbol{\beta}_\lambda^T \mathbf{x}_i + \boldsymbol{\gamma}_\lambda^T \mathbf{z}_i$
<i>Component #2</i>	Positive catch rates r : $\log(r_i) = \boldsymbol{\beta}_r^T \mathbf{x}_i + \boldsymbol{\gamma}_r^T \mathbf{z}_i + \log(a_i)$	Average biomass per group w : $\log(w_i) = \boldsymbol{\beta}_w^T \mathbf{x}_i + \boldsymbol{\gamma}_w^T \mathbf{z}_i$ or equivalently, positive catch rates: $\log(r_i) = \log(n_i) - \log(p_i) + \boldsymbol{\beta}_r^T \mathbf{x}_i + \boldsymbol{\gamma}_r^T \mathbf{z}_i$	Average biomass per group μ : $\log(\mu_i) = \boldsymbol{\beta}_\mu^T \mathbf{x}_i + \boldsymbol{\gamma}_\mu^T \mathbf{z}_i$
<i>Density</i>	Biomass density d : $d_i = p_i \times r_i$	Biomass density d : $d_i = n_i \times w_i$	Biomass density d : $d_i = \lambda_i \times \mu_i$
<i>Predicted encounter probability</i>	$p_i = \frac{1}{1 + \exp(-\boldsymbol{\beta}_p^T \mathbf{x}_i - \boldsymbol{\gamma}_p^T \mathbf{z}_i)}$	$p_i = 1 - \exp(-a_i \times n_i)$	$p_i = 1 - \exp(-a_i \times n_i)$
<i>Predicted positive</i>	$r_i = a_i \times \exp(\boldsymbol{\beta}_r^T \mathbf{x}_i + \boldsymbol{\gamma}_r^T \mathbf{z}_i)$	$r_i = \frac{n_i}{p_i} \times w_i$	$r_i = \frac{\lambda_i}{p_i} \times \mu_i$

catch rate

Likelihood
function

$$\Pr(C = c_i) = \begin{cases} 1 - p_i & \text{if } c_i = 0 \\ p_i \times f(C; r_i, \sigma_M^2) & \text{if } c_i > 0 \end{cases}$$

$$\Pr(C = c_i) = \begin{cases} 1 - p_i & \text{if } c_i = 0 \\ p_i \times f(C; r_i, \sigma_M^2) & \text{if } c_i > 0 \end{cases}$$

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

where

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

Process to
simulate
data

$$c_i = \begin{cases} 0 & \text{if } P = 0 \\ LN\left(\log(r_i) - \frac{\sigma_M^2}{2}, \sigma_M^2\right) & \text{if } P = 1 \end{cases}$$

$P \sim \text{Bernoulli}(p_i)$
and

$P \sim \text{Bernoulli}(p_i)$
and

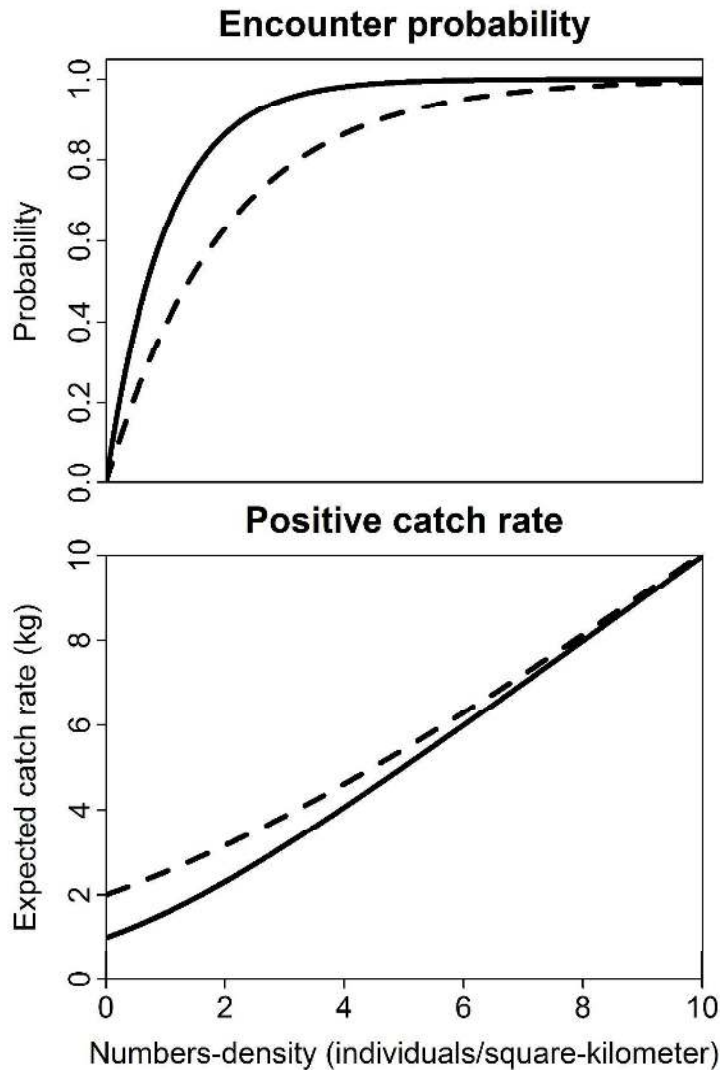
$N_i \sim \text{Poisson}(\lambda_i)$
 $W_{i,j} \sim \text{Gamma}(k^{-2}, \mu_i k^2)$

$$c_i = \begin{cases} 0 & \text{if } N_i = 0 \\ \sum_{j=1}^{N_i} W_{i,j} & \text{if } N_i > 0 \end{cases}$$

728

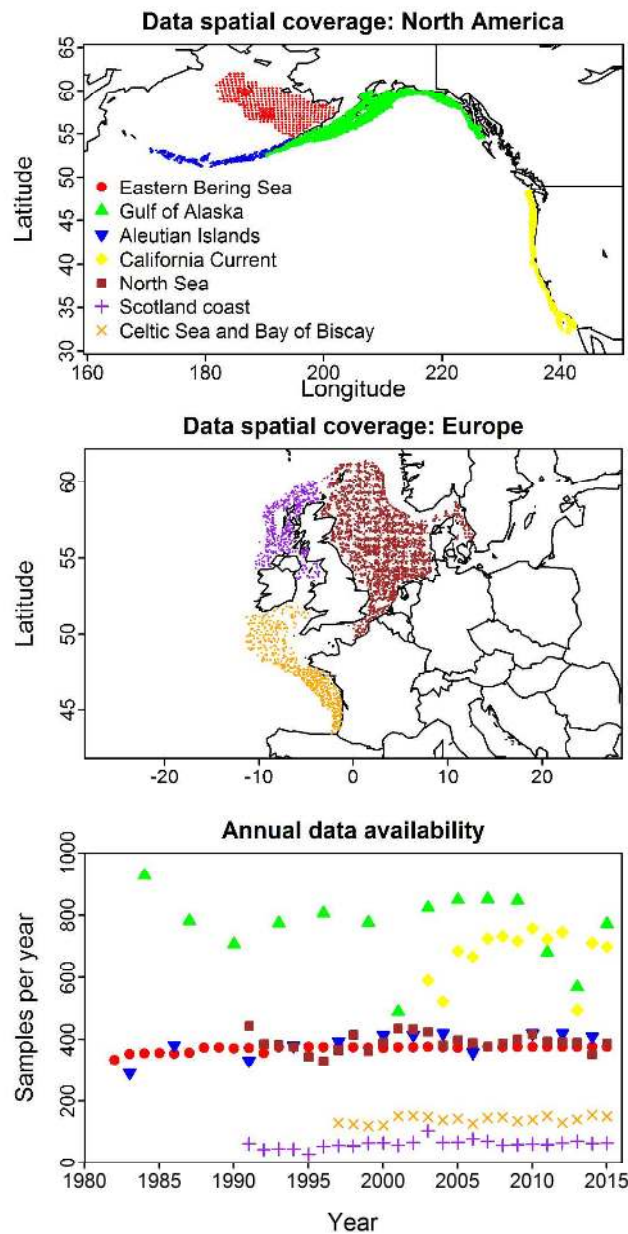
729

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



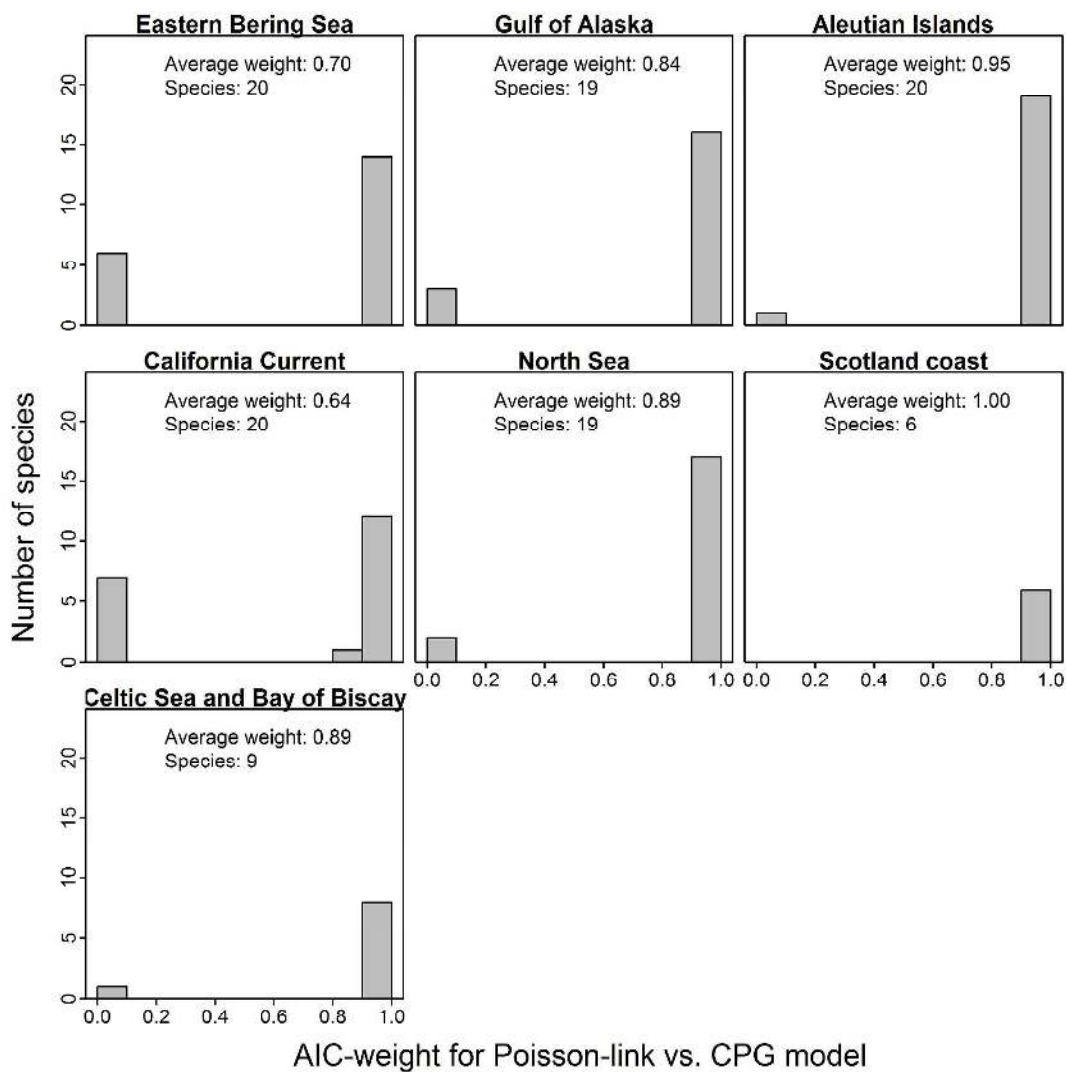
737

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



744

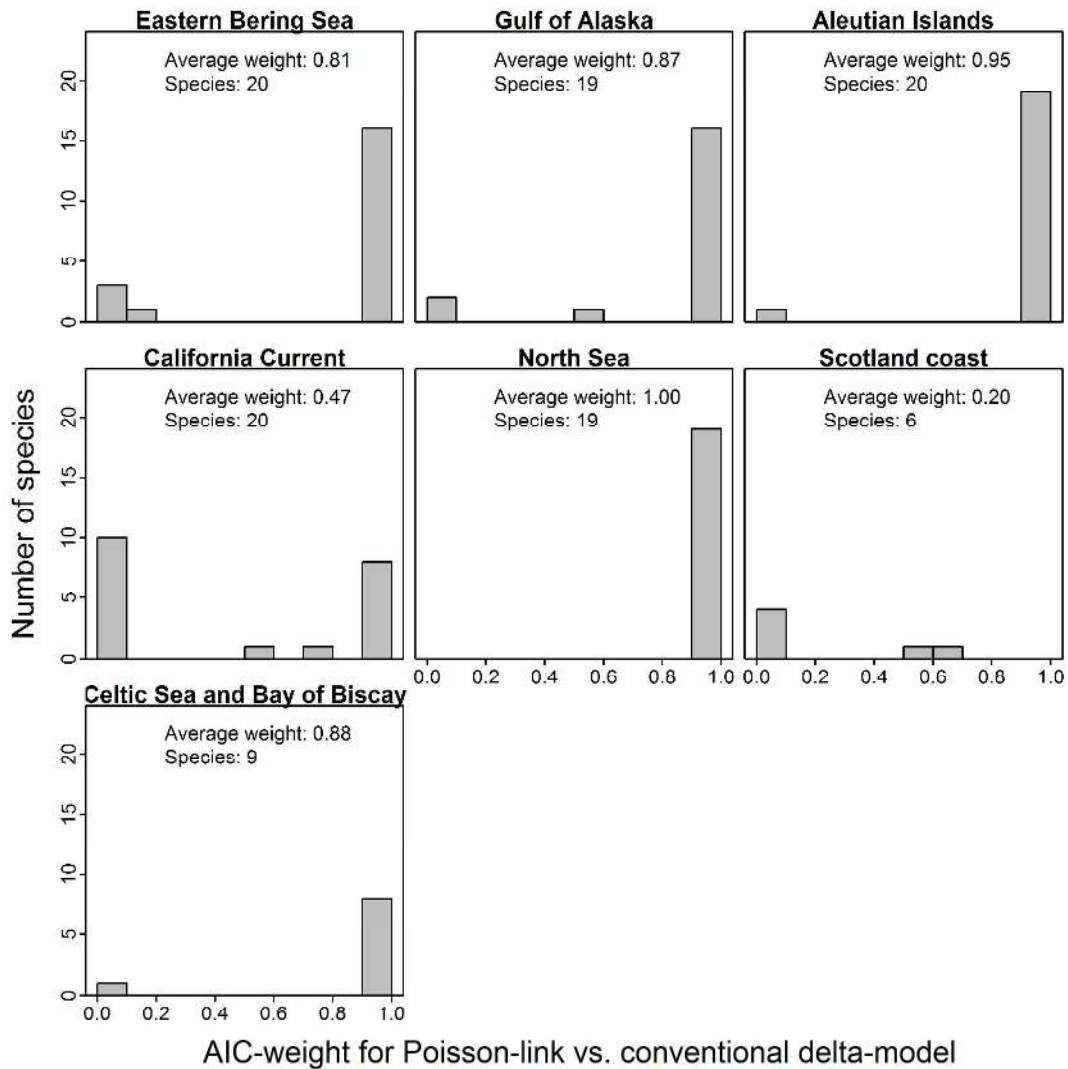
745 Fig. 3 – AIC weight for the “Poisson-link” model compared with a compound Poisson-
 746 gamma (CPG) model (where a species with a AIC-weight of 1.0 means that AIC strongly
 747 favors the Poisson-link over the CPG model) for a simple (fixed intercept-only) model
 748 applied to survey biomass-sampling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



751

752

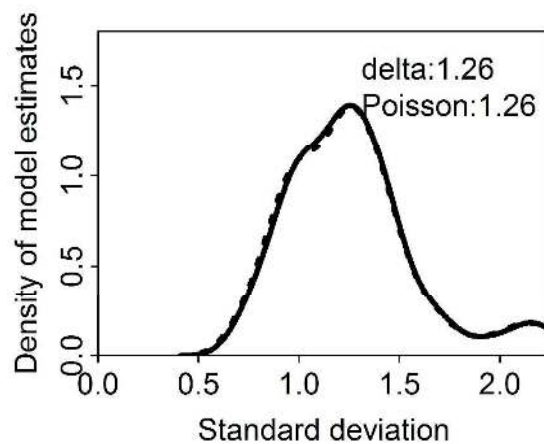
753 Fig. 4 – AIC weight for the “Poisson-link” model compared with a conventional delta-model
 754 for a complicated (fixed intercept, plus random spatial and spatio-temporal effects) model
 755 applied to survey biomass-sampling data in seven bottom trawl surveys (see Fig. 3 caption
 756 for details)



757

758

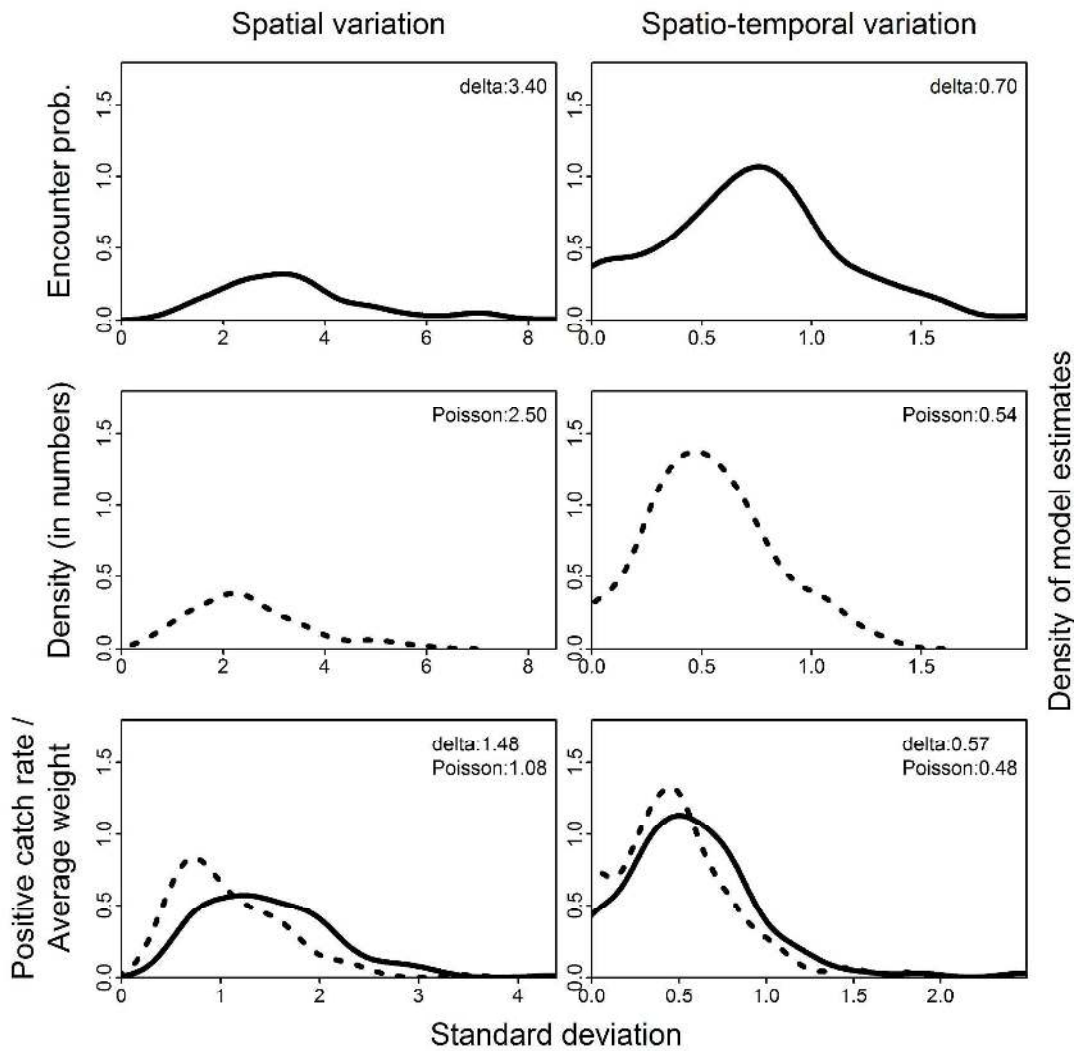
759 Fig. 5 – Distribution of standard deviation estimates of residual variation in positive catch
760 rates for a complicated (fixed intercept, plus random spatial and spatio-temporal effects)
761 model for each of 113 stocks (in total across seven surveys), using the conventional delta-
762 model (solid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



765

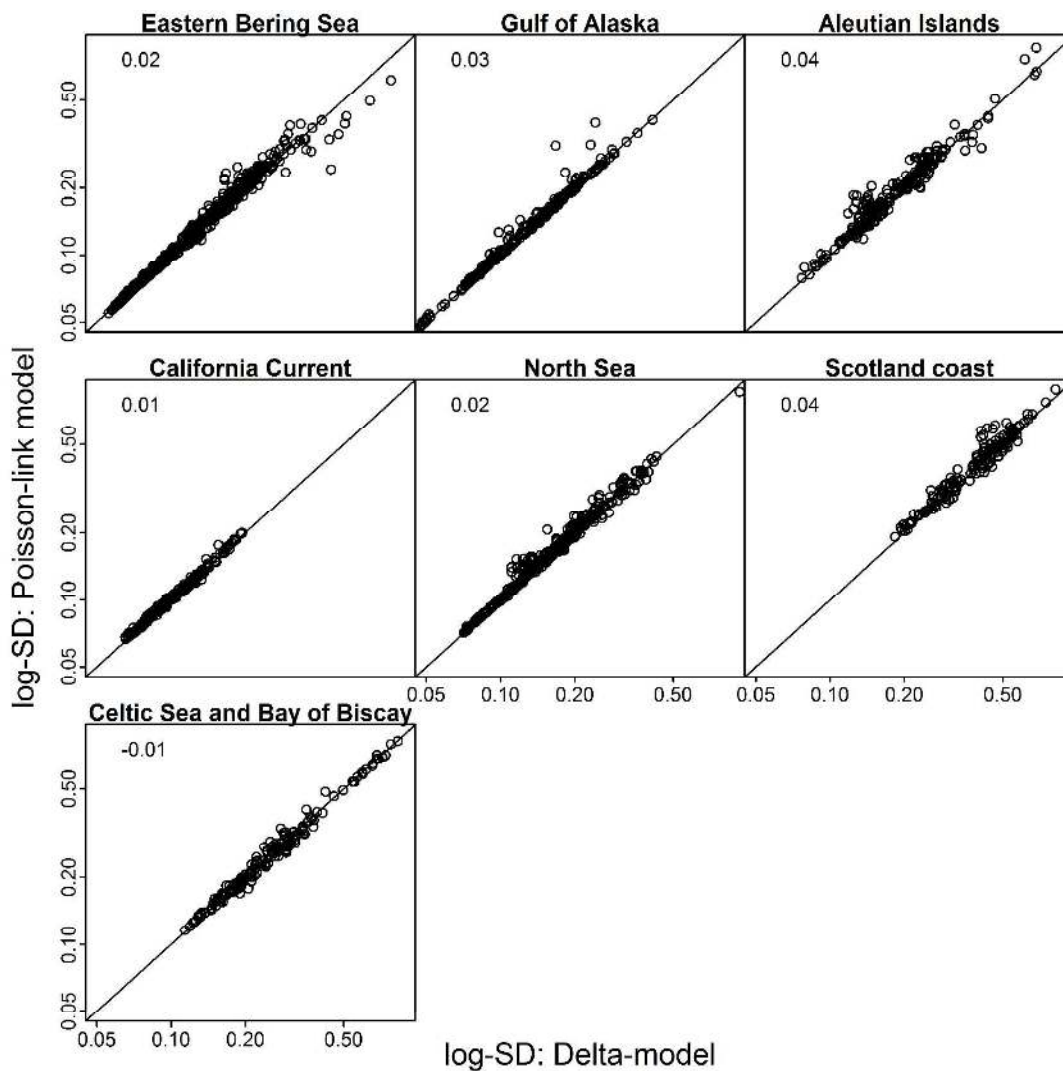
766

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



775

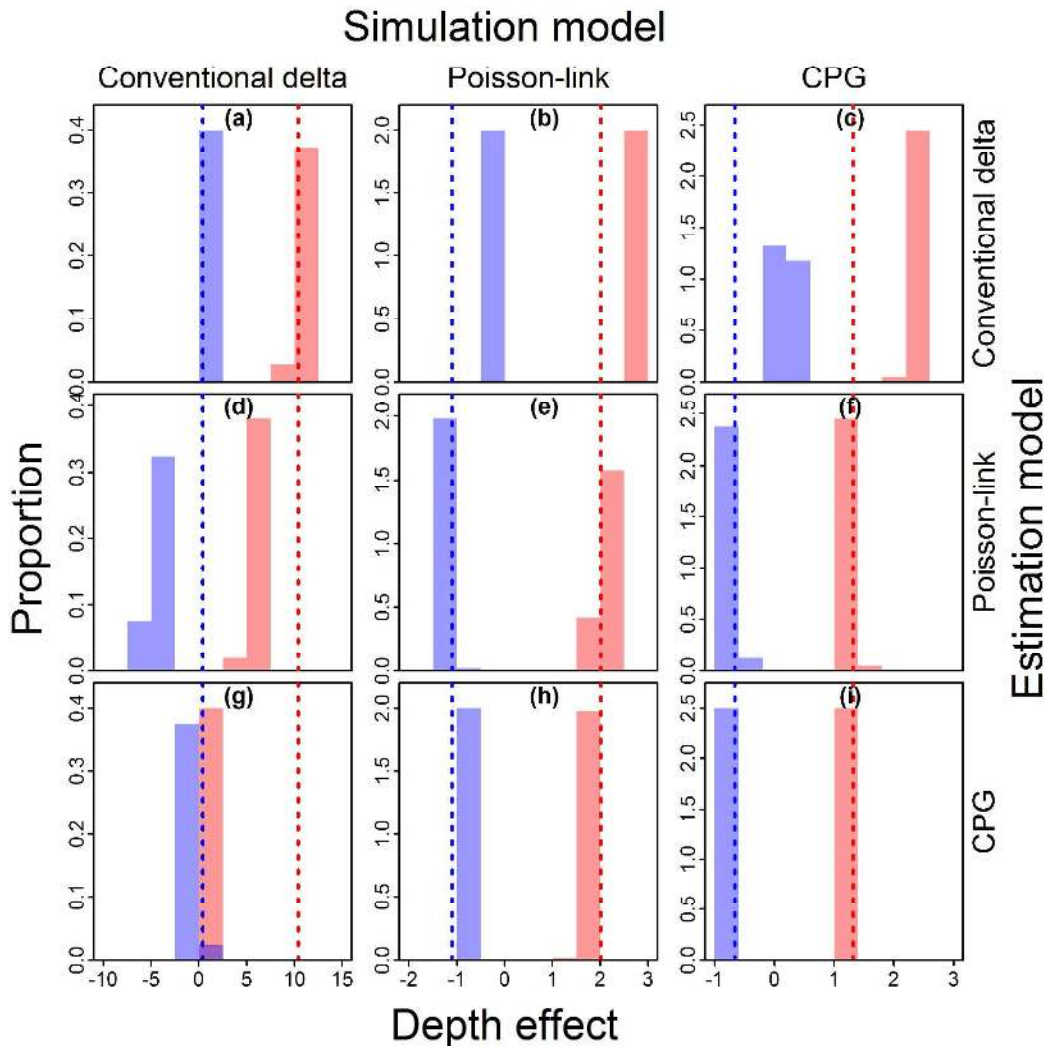
776 Fig. 7 – Comparison of estimated log-standard deviation of total population-wide abundance
 777 for the conventional delta-model and alternative Poisson-link model for a complicated (fixed
 778 intercept, plus random spatial and spatio-temporal effects) model applied to biomass-
 779 sampling data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



783

784

785 Fig. 8 – Comparison of estimated depth-effects (histogram) vs. true value (dotted vertical
 786 lines) when estimating parameters using the delta model (top row, red: β_p ; blue: β_r),
 787 Poisson-link model (middle row, red: β_n ; blue: β_w), and compound Poisson-gamma (CPG)
 788 model (bottom row, red: β_λ ; blue: β_μ), applied to data generated using each model (left
 789 column: delta-model; middle column: Poisson-link model; right column: CPG) fitted to data
 790 for arrowtooth flounder in the Eastern Bering Sea. Note that the true depth effect (dotted
 791 vertical line) is identical for each panel in a given column (because that these all use the same
 792 operating model to generate data), and that panels along the diagonal involve a correctly
 793 specified estimation model while other panels involve a mis-specified estimation model.



794

1 **Appendix A: Comparing the Poisson-link delta model with the compound**
 2 **Poisson-gamma distribution**

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

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

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

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

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

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

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

12 such that total catch $C = \sum_{j=1}^N W_j$. In the main text, I present a reparameterization in terms of
 13 numbers-density λ_i and expected individual weight μ_i for each sample i , where gamma shape
 14 parameter k is constant among samples but expected individual weight μ_i differs among
 15 samples (where $\mu_i = k\theta_i$). Following Foster and Bravington (2013), I specify that where
 16 both numbers-density λ_i and expected individual weight μ_i are predicted using a log-linked
 17 linear predictor, and where the offset a_i affects expected catch in numbers. I also show that

18 this parameterization generates a similar functional form for expected encounter probability r
 19 and positive catch rates r as our alternative Poisson-link model.

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

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

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

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

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

26 Works cited

- 27 Dunn, P.K., and Smyth, G.K. 2005. Series evaluation of Tweedie exponential dispersion
 28 model densities. *Stat. Comput.* **15**(4): 267–280.
- 29 Dunn, P.K., and Smyth, G.K. 2008. Evaluation of Tweedie exponential dispersion model
 30 densities by Fourier inversion. *Stat. Comput.* **18**(1): 73–86. doi:10.1007/s11222-007-
 31 9039-6.
- 32 Foster, S.D., and Bravington, M.V. 2013. A Poisson–Gamma model for analysis of
 33 ecological non-negative continuous data. *Environ. Ecol. Stat.* **20**(4): 533–552.
 34 doi:10.1007/s10651-012-0233-0.
- 35 Lauderdale, B.E. 2012. Compound Poisson–Gamma Regression Models for Dollar Outcomes
 36 That Are Sometimes Zero. *Polit. Anal.* **20**(3): 387–399. doi:10.1093/pan/mps018.
- 37 Lecomte, J.-B., Benoît, H.P., Ancelet, S., Etienne, M.-P., Bel, L., and Parent, E. 2013.
 38 Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data
 39 under a variable sampling volume. *Methods Ecol. Evol.* **4**(12): 1159–1166.
 40 doi:10.1111/2041-210X.12122.
- 41 Smyth, G.K. 1996. Regression analysis of quantity data with exact zeros. *In Proceedings of*
 42 *the Second Australia-Japan Workshop on Stochastic Models in Engineering,*
 43 *Technology and Management, Gold Coast, Australia.* pp. 17–19.

45