The surprising sensitivity of index scale to delta-model assumptions: recommendations for model-based index standardization

James T. Thorson[1], Curry Cunningham[2], Elaina Jorgensen[3], Andrea Havron[4], Peter-John F. Hulson[5], Cole C. Monnahan[6], Paul von Szalay[3]

1 Habitat and Ecological Processes Research Program, Resource Ecology and Fisheries Management, Alaska Fisheries Science Center, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, WA 98115, USA.

2 College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, 17101 Point Lena Loop Road, Juneau, AK 99801.

3 Groundfish Assessment Program, Resource Assessment and Conservation Engineering Division, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, 7600 Sand Point Way N.E., Seattle, WA 98115, USA.

4 School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA, 98195, USA

5 Marine Ecology and Stock Assessment, Auke Bay Laboratories, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, 17109 Point Lena Loop Road, Juneau, AK 99801, USA

6 Status of Stocks and Multispecies Assessments program, Resource Ecology and Fisheries Management, Alaska Fisheries Science Center, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, WA 98115, USA.

Abstract

Delta-models (a.k.a. hurdle models) are widely used to fit biomass samples that include zeros and a skewed response for positive catches, and spatio-temporal extensions of these models are increasingly used to quantify trends in abundance (i.e., estimate abundance indices). Previous research has shown estimated indices are proportional to changes in abundance. However, little research has tested the performance of delta-models for estimating “scale”; that is, whether abundance indices are not just proportional to population changes but also have the correct absolute value. We use data for twenty species in the eastern Bering Sea and Gulf of Alaska as well as a factorial experiment conditioned on data for Gulf of Alaska Pacific cod to support five conclusions related to scale in spatio-temporal delta-models. First, we show that conventional (nonspatial) delta-models are surprisingly sensitive to the *a priori* choice of probability distribution for positive catches, where gamma and Tweedie models give similar scale estimates but other distributions generally differ. Second, these same distributions also estimate widely different scales when using spatio-temporal delta-models, and the delta-gamma and Tweedie models provide similar scale to design-based indices. Third, model selection using marginal AIC often identifies the lognormal distribution as most parsimonious, despite it resulting in systematically higher abundance than design-based indices for many species. Fourth, scale is sensitive to the spatial resolution (i.e., number of knots) used in fitting the spatio-temporal model when using a naïve “empirical Bayes” estimator, but less sensitive when applying an epsilon bias-correction estimator. Fifth, the factorial simulation experiment suggests that the Tweedie and delta-gamma distributions perform well even when applied to data simulated from an inverse-Gaussian or lognormal distribution, whereas the opposite is not true. We conclude that index scale is sensitive to delta-model specification, and we make five recommendations when using spatio-temporal delta-models for index standardization: (1) apply the epsilon or other bias-

correction methods to reduce sensitivity of index scale on spatio-temporal model resolution; either (2) compare the scale of delta-model indices with that of design-based indices when design-based indices are available or (3) use the delta-gamma or Tweedie distribution by default when design-based indices are not available; (4) do not assume that AIC will identify the model specification that results in the most appropriate scale; and (5) consider apparent mismatches in index scale depending upon whether an assessment model specifies or estimates the associated catchability coefficient and whether the design-based index is believed to measure total abundance for a fully-selected age or length-class.

Keywords: Vector autoregressive spatio-temporal model; VAST; delta model; Tweedie distribution; stock assessment; abundance index; catchability coefficient

1. Introduction

Fisheries scientists worldwide support fisheries management by estimating stock status and sustainable fishing levels. They typically do this by fitting population-dynamics models to fishery catches, measurements of age and length composition, and indices of population abundance (Methot, 2009). Many common stock-assessment models are fitted to abundance-indices that are measures (or proxies) of biomass. Fisheries scientists have therefore developed a wide range of methods to sample local biomass and subsequently estimate total biomass over a pre-defined spatial domain. These methods include design-based indices, which are constructed from field-samples of biomass following a probabilistic design (wherein every sampling unit is sampled with a pre-specified probability) and an associated statistical estimator (Cochran, 1977; Smith, 1990; Petitgas, 2001). However, design-based indices are not appropriate for fishery-dependent data that are not collected under a probabilistic design, or for surveys where the design has changed substantially over time (i.e., adding northern stations in the eastern Bering Sea, or changing the southern extent in the West Coast triennial bottom trawl survey). The inability to apply design-based estimators in these instances has led to interest in model-based biomass estimators, including the widely used delta-model (Pennington, 1983; Lo et al., 1992; Stefansson, 1996).

The delta-model has been widely used for over 35 years, and separately models the probability that each sample encounters a given species (termed “encounter probability” here) and the probability distribution for sample biomass given that the species is encountered (termed “positive catch rate” here). Aitchison (1955) originally described the delta-model as a mixture distribution that contained a point mass at zero and a conditional distribution describing positive (non-zero) values. The delta-lognormal distribution was proposed in the follow-up paper by Aitchison and Brown (1957) and was first applied in fisheries by Pennington (1983) to describe

Atlantic mackerel egg production. Lambert (1992) first used a logit link function to approximate the probability of encountering zero as a linear function of covariates in the context of zero-inflated Poisson distributions. This approach was first applied in fisheries by Lo et al. (1992) to calculate an index of relative abundance for anchovy. Delta-models were then popularized in subsequent publications (Stefansson, 1996; Maunder and Punt, 2004).

Ongoing research has also developed spatio-temporal models that account for the correlation among survey observations resulting from their proximity in space and/or time (Banerjee et al., 2003; Cressie and Wikle, 2011), and these methods have recently been adapted to a delta-modelling framework (Shelton et al., 2014; Thorson et al., 2015). The benefit of a spatio-temporal delta-model can be seen by comparison with a design-based estimator. Specifically, spatially-correlated variability in habitat quality will result in residual variance among samples within each spatial stratum in a stratified-random design; this residual variance will result in increasing variance for the resulting index when using a stratified-random design-based estimator. In these cases, accounting for the randomized location of samples can control for this spatially-correlated variability, and therefore can substantially reduce standard errors for spatio-temporal indices (Shelton et al., 2014; Cao et al., 2017). In addition to increased index precision, spatio-temporal delta-models have been shown to reduce biologically-implausible variation in indices for long-lived species (Gertseva and Thorson 2013). When a spatio-temporal delta-model was fit to U.S. West Coast trawl survey data for 28 groundfish species, confidence intervals from the conventional design-based approach were 60% larger on average than those derived from the spatio-temporal estimator (Thorson et al. 2015).

Spatio-temporal delta-models have previously and continue to be implemented for a wide range of purposes. They have been used extensively for standardization of US West Coast groundfish trawl survey data and are seeing increased application to Alaska groundfish survey data (see list in Thorson, 2019a). In other US fisheries, spatio-temporal delta-models have been implemented to estimate indices using data form multiple trawl surveys (Perretti and Thorson 2019), or from a mix of trawl and fixed-gear survey observations (Gruss and Thorson 2019). Bayesian spatio-temporal delta-models have also been developed for standardization of crustacean indices from trawl survey data from the Mediterranean Sea (Arcuti et al. 2016) and shark bycatch in Canadian waters (Cosandey-Godin et al., 2014). For conservation planning, spatio-temporal models have been used to integrate data from seven fisheries-independent surveys, with the goal of quantifying spatial separation among target and non-target species in highly-mixed Celtic Sea fisheries (Dolder et al. 2018), and to quantify spatial bycatch risk in the Pacific Ocean (Stock et al. 2020). Finally, spatio-temporal delta-models have been utilized for ecological inference to describe changes in species distribution, concentration, and habitat association (Thorson et al. 2016a, Thorson et al. 2016b).

Design-based biomass indices derived from fishery-independent bottom trawl surveys are fitted within many age-structured stock assessments for fish stocks in the North Pacific (NPFMC, 2019a, 2019b). Age-structured models have the capacity to estimate the catchability coefficient representing the ratio of predicted and index biomass (Arreguín-Sánchez, 1996). Catchability coefficients are extremely influential with respect to the scale of biomass estimated by a stock assessment model and are typically either estimated as a parameter or fixed at some predetermined value (Wilberg et al., 2010). The estimated value for the catchability coefficient is affected by spatial overlap between the stock and the spatial extent of the survey (“horizontal availability”), the stocks’ vertical availability in the water column, and the stocks’ vulnerability to the gear used to capture the fish (Cordue, 2007). Given the potential sensitivity of survey index scale to standardization methods, and the interaction between index scale and the catchability coefficient on stock assessment results, it is useful to summarize the many ways catchability is specified within assessments currently.

We explore stock assessments at the Alaska Fisheries Science Center (AFSC) as an example of stock-assessment practices for specifying the catchability coefficient throughout the US and worldwide. Stock assessments at the AFSC treat the catchability coefficient using a variety of approaches (see Table 1 for summary) ranging from fixing it at a value a priori (e.g., Bryan, 2017) to estimated freely (Thompson and Thorson, 2019). When the catchability coefficient is fixed *a priori*, the survey biomass is treated as an absolute index and any change in the scale of survey biomass would have direct influence on parameters that determine the scale of the population (such as average recruitment and natural mortality rate). When the catchability coefficient is estimated freely, the survey biomass is treated as a relative index and any multiplicative change in index scale will be offset by a corresponding change in the estimated catchability coefficient. Between these two extremes, some stock assessments estimate the catchability coefficient using a prior distribution (either in a Bayesian or penalized likelihood framework) with an associated level of uncertainty; this specified uncertainty determines the degree to which the estimated catchability coefficient is able to deviate from the mean of this prior distribution. When specifying a prior distribution, an infinitesimally small uncertainty is equivalent to specifying a fixed value for the catchability coefficient, and an infinite level of uncertainty (using a normal prior distribution with arbitrarily large variance) is equivalent to freely estimating the catchability coefficient. As a consequence, the impact of changing the scale of the survey index on modeled quantities from an assessment, such as spawning biomass or management reference points, will be determined by the degree of precision ascribed to the assumed prior on catchability coefficients: changes in index scale will be more influential on modeled quantities in cases of a precise (low variance) prior and less influential in cases of imprecise (high variance) prior on catchability.

The probability distribution for positive catches specified in a delta-model can directly affect the absolute scale of the estimated index, and this is particularly important in stock assessments where the catchability coefficient is fixed *a priori* or has an informative prior distribution. For instance, the delta-model can result in a biased estimate of average biomass when the probability distribution is mis-specified with respect to the distribution of residuals (Hvingel et al., 2012; Myers and Pepin, 1990). Furthermore, delta-models can be highly sensitive to deviations from model assumptions that are otherwise difficult to detect using standard statistical diagnostics (Syrjala, 2000). In response, many approaches have been proposed and/or applied for selecting the most appropriate distribution. Graphical tests such as Taylor's power rule may help narrow the proposed set of distributions (Dick, 2004). Diagnostic tests like simple Pearson correlation and normality tests on residuals, but also the lesser-known Pregibon, modified Hosmer-Lemeshow, Kolmogorov-Smirnov, and Anderson-Darling tests have also been explored but without consensus about their performance (Hvingel et al., 2012; Ng and Cribbie, 2017). Researchers have also selected among alternative distributions using information criteria like the Akaike and Bayesian Information Criteria (Akaike, 1974; Schwarz, 1978; Burnham and Anderson, 2002), which appear reliable in simulations under ideal conditions and sufficient sample sizes (Dick, 2004; Mitchell et al., 2015). However, sometimes AIC will select models that fail diagnostic tests or can be unreliable with small sample sizes (Dick, 2004; Ng and Cribbie, 2017). Furthermore, these previous simulations used GLMs without spatial effects such that conclusions may not apply to spatio-temporal GLMMs. Consequently, the best statistical approach for selecting the distribution for positive catch rates in spatio-temporal delta-models remains unknown.

In this analysis, we first illustrate that the scale of an abundance-index estimated using a conventional (nonspatial) delta-model is highly dependent upon the assumed distribution for positive catch rates. We then compare index estimates from four spatio-temporal models (using delta-gamma, delta-lognormal, delta-inverse-Gaussian, and Tweedie distributions) with design-based estimates for twenty stocks in the eastern Bering Sea and Gulf of Alaska. Previous research has developed an epsilon bias-correction estimator (Thorson and Kristensen, 2016) that corrects for "retransformation bias" arising when random effects are transformed when calculating a quantity of interest (Thorson, 2019b), but no previous study has used a simulation experiment to demonstrate its importance when estimating abundance using a spatio-temporal model. Similarly, we are not aware of any previous simulation study exploring how alternative choices about spatial scale can affect the performance of a spatio-temporal index standardization model. We therefore compare performance within a factorial design of twenty species, four distributions, three spatial resolutions, and two estimators (either naïve or using the epsilon bias-correction estimator). We then identify which distribution(s) provide an approximately equal number of years where the abundance index is greater or less than the design-based index (i.e. equivalent scale of design and model-based indices), as well as which distribution(s) estimate a similar ratio between the modeled and design-based index. Finally, we use a factorial simulation design conditioned upon data for Pacific cod (*Gadus macrocephalus*) in the Gulf of Alaska, where we simulate data using each of the four models and fit each data set with these same four estimation models. Using this simulation design, we again determine the ratio of index-scale with the true population scale, as well as root-mean-squared error, to identify whether any model performs best on average. Based on these findings we provide generic advice for configuring delta-models for estimating abundance indices for use in stock assessments.

2. Methods

2.1 Overview

We seek to determine what specification for a spatio-temporal index standardization model results in an index scale that matches estimates from a design-based estimator. We specifically explore two alternative types of index standardization models: a delta-model involving two linear predictors, or a compound Poisson-gamma (a.k.a. Tweedie) distribution involving a single linear predictor. For the delta-model, we specifically explore three alternative distributions for positive catch rates: a lognormal, gamma, or inverse-Gaussian distribution. This then results in four model-specifications in total. All models are implemented using the Vector Autoregressive Spatio-Temporal (VAST) model (Thorson and Barnett, 2017; Thorson, 2019a), as implemented in package VAST release number 3.5.0 available online (https://github.com/James-Thorson-NOAA/VAST) for the R statistical environment (R Core Team, 2017). We do not explore the potential role of covariates in the following, although future research could continue to explore tradeoffs associated with their inclusion (e.g., Johnson et al., 2019).

We apply these four model specifications in two separate explorations:

1. Case study: The first is a case-study demonstration, where we fit these four model-specifications to data for twenty selected species in the Gulf of Alaska and eastern Bering Sea. We conduct two separate experiments using these case-study species. In the first, we fit nonspatial models that estimate a separate intercept for each linear predictor in each year to each species. This experiment is useful to show that differences in model scale arise between

alternative model specifications even in the simplest possible specification of an index-standardization model. In the second, we fit a spatio-temporal model to data for each species. In this experiment, we then compare results with a design-based estimator for each species, to see which model specification results in a similar index scale to the design-based estimator.

2. Factorial simulation experiment: The second is a factorial simulation experiment, where we fit each model specification to data for a single species (Pacific cod in the Gulf of Alaska). Given the estimated fixed and random effects for that species, we then simulate multiple replicate data sets. For each data set, we then fit all four estimation models. This then results in a 4×4 factorial cross of 4 operating models and 4 estimation models per simulation replicate. We refer to scenarios where the estimation model matches the operating model as a "self-test", while other scenarios explore the implications of model mis-specification on estimation model performance.

We describe each of these explorations in more detail below.

2.2 Model structure

In the following, we fit to observed biomass b_i for each sample i using either a Poisson-link delta-model (Thorson, 2018) or a compound Poisson-gamma model (Foster and Bravington, 2013). Delta-models have conventionally involved a logit-linked linear predictor for encounter probability, and a separate log-linked linear predictor for catch rates given an encounter (Stefansson, 1996). However, we instead use a Poisson-link delta model that previous research has shown to fit better while yielding a model structure that is more similar to the compound Poisson-gamma distribution.

Poisson-link delta-models involve two log-linked linear predictors:

$$\log\big(n(s_i,t_i)\big) = \beta_n(t_i) + \omega_n^*(s_i) + \varepsilon_n^*(s_i,t_i) \tag{1}$$

$$\log\big(w(s_i,t_i)\big) = \beta_w(t_i) + \omega_w^*(s_i) + \varepsilon_w^*(s_i,t_i),$$

where $\beta_n(t)$ is an annually varying intercept for each modeled year $t \in \{t_{min}, \dots, t_{max}\}$, $\omega_n^*(s)$ is spatial variation that is constant over time (termed "spatial variation") for location $s \in \Omega$ within a fixed spatial domain Ω, and ε_n^* is spatial variation that varies among years (termed "spatio-temporal variation") in the 1st log-linked linear predictor $n(s,t)$ and similar notation is used for the second log-linked linear predictor $w(s,t)$. The product of these linear predictors $d(s,t) = n(s,t)w(s,t)$ is then population density $d(s,t)$ at each location s and time t. By contrast, the compound Poisson-gamma model involves a single log-linked linear predictor for density:

$$\log\big(d(s_i,t_i)\big) = \beta_d(t_i) + \omega_d^*(s_i) + \varepsilon_d^*(s_i,t_i) \tag{2}$$

which again includes an annual intercept, spatial, and spatio-temporal variation.

These models then involve specifying a probability distribution B for each sample of biomass b_i. The Poisson-linked delta-models convert $n(s_i,t_i)$ and $w(s_i,t_i)$ to encounter probability p_i and positive catch rate r_i, which varies among samples i occurring at a given location s_i and time t_i due to differences in area-swept a_i. The Poisson-linked delta-model assumes that individuals are randomly distributed in the vicinity of sampling:

$$p_i = 1 - \exp\big(-a_i n(s_i,t_i)\big) \tag{3}$$

$$r_i = \frac{a_i n(s_i,t_i) w(s_i,t_i)}{p_i}$$

and all delta-models assume the same probability for encounter probability:

$$\Pr(B=0) = 1 - p_i \tag{4}$$

while alternative delta-models differ in the distribution for positive catches. Specifically we use a bias-corrected lognormal where dispersion parameter θ is the standard deviation in log-space:

$$\Pr(B = b_i|B > 0) = Lognormal\left(B; \log(r_i) - \frac{\theta^2}{2}, \theta^2\right) \quad (5A)$$

or use a shape-scale parameterization of the Gamma distribution where dispersion θ is the coefficient of variation:

$$\Pr(B = b_i|B > 0) = Gamma(B; \theta^{-2}, r_i\theta^2) \quad (5B)$$

or finally we use the mean-lambda parameterization of the inverse-Gaussian distribution, where dispersion θ is again the coefficient of variation

$$\Pr(B = b_i|B > 0) = Inv.Gaussian(B; r_i, \theta^{-2}). \quad (5C)$$

By contrast, the compound Poisson-gamma distribution replaces Eq. 4-5 with a single distribution for biomass B

$$\Pr(B = b_i) = Tweedie(B; a_i d_i, \theta, \phi). \quad (6)$$

While estimating dispersion θ and power parameter $1 < \phi < 2$. Lognormal, gamma, and inverse-Gaussian distributions are all parameterized such that r_i represents the mean of positive-catch rates, such that d_i is the mean of expected catches for all distributions. However, these distributions differ somewhat in how variance is assumed to vary as a function of the mean ("mean-variance relationship"). Similarly, these distributions assign a greater or lesser probability to "extreme catches" (i.e., catches greater than ten times the expected value), and these "extreme catch events" are a well-known property of demersal fish surveys (Thorson et al., 2011). For example, the lognormal has skewness of $CV^3 + 3CV$ (where CV is the measurement error coefficient of variation) while the gamma has skewness of $2CV$. Given that the estimated CV is typically above 1.0, these distributions can have substantially different skewness. As a consequence, extremely high (or low) catches will have a greater "leverage" on predicted density for some distributions than others.

All models adopt a predictive-process framework for predicting spatial and spatio-temporal variation at the location s_i of each sample i, or location s_g of each extrapolation-grid cell g, given the value at n_s knots (Banerjee et al., 2008). Specifically, we specify that the value of spatial and spatio-temporal variables at each knot follows a Gaussian Markov random field:

$$\boldsymbol{\omega}_n \sim MVN(\mathbf{0}, \sigma_\omega^2 \mathbf{Q}_n^{-1}) \tag{7}$$

$$\boldsymbol{\varepsilon}_n(t) \sim MVN(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{Q}_n^{-1}),$$

where $\mathbf{Q}$ is a sparse precision matrix that approximates a Matern correlation function with decorrelation rate κ_n that varies among linear predictors and a transformation matrix $\mathbf{H}$ that approximates geometric anisotropy and is shared among linear predictors. These spatial variables are then pre-multiplied by a matrix that represents bilinear interpolation (Lindgren and Rue, 2015):

$$\boldsymbol{\omega}_n^* = \mathbf{A}\boldsymbol{\omega}_n \tag{8}$$

$$\boldsymbol{\varepsilon}_n^*(t_i) = \mathbf{A}\boldsymbol{\varepsilon}_n(t_i)$$

and where spatial and spatio-temporal variables are treated similarly for other linear predictors $w(s,t)$ and $d(s,t)$. Specifically, interpolation matrix $\mathbf{A}$ has a row for each extrapolation-grid cell and a column for each knot. It is nonzero for only three elements of each row (hence a "sparse" matrix), with nonzero values corresponding to the weight assigned to three vertices surrounding a given location when interpolating from three neighboring knots within a triangulated mesh.

Parameters are estimated by identifying the value of fixed effects that maximizes the marginal likelihood when integrated across random effects. We approximate this multidimensional integral using the Laplace approximation, as implemented using Template Model Builder (Kristensen et al., 2016). After identifying fixed effects, we then apply an "empirical Bayes" estimator, which fixes random effects to their value that maximizes the joint likelihood conditional on estimated fixed effects. Derived quantities can then be calculated from the maximum likelihood estimate of fixed effects and empirical Bayes estimate of random effects. However, derived quantities that are calculated from a nonlinear transformation of random effects will be subject to "retransformation bias" when applying this naïve estimator. We therefore also apply the "epsilon bias-correction estimator" that corrects for the degree of nonlinearity and variance of random effects when calculating derived quantities, including biomass indices (Thorson and Kristensen, 2016).

To estimate parameters for these models the user must:

1. Choose which probability distribution to use for the positive catches (lognormal, gamma, etc.);
2. Choose the spatial resolution by specifying the number of interior knots n_x to use, which are then augmented with boundary knots to determine the size of spatial and spatio-temporal random effects n_s;
3. Choose whether to use the naïve or epsilon bias-correction estimator for derived quantities.

We seek to provide generic guidance for these three decisions while using the "predictive process" and exploring outcomes with modeled spatial resolution ranging from 100, 250, and 500 knots, $n_x = \{100, 250, 500\}$.

2.3 Case study design

Reviews for recent stock assessments at the Alaska Fisheries Science Center (AFSC) have recommended further exploration of VAST regarding model specification. We therefore conduct a case-study comparison of VAST models with design-based indices for twenty selected species in the Gulf of Alaska and eastern Bering Sea (see Table 1 for list). The eastern Bering Sea has followed a fixed-station design for bottom-trawl samples using the 83-112 gear from

1982-2019, where the number of samples has increased over time from approximately 350 to 375 per year (Lauth and Conner, 2016). The Gulf of Alaska has followed a random stratified design for bottom trawl samples from 1984-2019, using the Poly Nor'eastern gear from 1990-2019 and an earlier gear previously, sampling every third year from 1984-1999 and every second year from 1999-2019. The number of samples per year varies from 500-850, and the sampling intensity for each strata varies among years following a Neyman design based on strata-specific catch rates in previous years for all species. The stratified design followed an approximately consistent footprint for most years except for 2001 when the eastern Gulf of Alaska was not sampled, and also in other years when deep-water strata were dropped due to funding limitations (von Szalay and Raring, 2016).

For each of these stocks, we first fit model-based estimators that include only the annual intercept in each year (β) and exclude the spatial and spatio-temporal terms (ω and ε), resulting in a simple unstratified delta-model. We do not expect this specification of model-based indices to accurately measure population biomass because this specification ignores spatial stratification and other concerns about sampling design. However, we compare model-based indices for alternative models to demonstrate the extent to which index scale can differ even when fitting a simple index model.

For each stock, we next extract a design-based estimator using standard protocols and software for these two regions (Wakabayashi et al., 1985). We compare these with spatio-temporal model-based estimators that extrapolate density to the "standard" footprint of these surveys. The spatio-temporal estimator specifically predicts density at the centroid of grid cells within a 2km by 2km square extrapolation-grid that serves as "quadrature points" for integrating across density. This includes 36,140 grid cells for the "Eastern Bering Sea" extrapolation-grid and 23,339 grid cells for the "Gulf of Alaska" extrapolation-grid; each is included in package VAST and was developed previously by Angie Grieg (personal communication; retired from Alaska Fisheries Science Center). We expect that these spatio-temporal model-based estimators will appropriately account for spatial variation in inclusion probability (i.e., due to stratified sampling) given that this probability-sampling design is constructed based on results for a wide variety of species and is likely to be independent of density for any single species (Conn et al., 2017).

The design-based estimator will be an unbiased estimator for the portion of population biomass that is available to the survey in each year. We acknowledge that the design-based estimator will in many cases not be an accurate representation of fully-selected abundance or biomass, e.g., in cases when the stock moves into and out of the spatial footprint of a single survey (Ianelli et al., 2019), moves vertically out of the area accessible to bottom trawls (Kotwicki et al., 2015), or moves into areas where gear performs poorly (Thorson et al., 2013). Previous studies have evaluated performance for spatio-temporal models via comparison to stock-assessment model output (e.g., Cao et al., 2017; Thorson and Haltuch, 2018), but have not used a simulation experiment to compare performance against the scale of design-based indices. We therefore evaluate the model performance for estimating population scale relative to design-based indices by we calculating the average across n_t years for both the design-based index $\bar{B} = \frac{1}{n_t}\sum_{t=1}^{n_t} B_t$ and each model-based index $\bar{I} = \frac{1}{n_t}\sum_{t=1}^{n_t} I_t$. We then calculate the ratio of these two averages $R = \bar{I}/\bar{B}$ and record this ratio for each species and model specification, for each model resolution and when using either the naïve or epsilon-bias correction estimator. We seek to determine what model specification results in a similar scale to design-based indices and therefore identify a well performing model as one with a ratio R that is evenly distributed around one, indicating that the scale is similar on average to the scale of the design-based index.

For each model and spatial resolution, we also calculate the Akaike Information Criterion (AIC), calculated using the Laplace approximation to the marginal likelihood and the number of fixed effects. We specifically seek to determine whether AIC consistently favors any model specification, and if the model specification selected using this criteria varies with changes in spatial resolution.

2.4 Design for factorial simulation experiment

We also explore model performance by conducting a 4×4 factorial design of all four model specifications as both operating model and estimation model, when fixed and random effects for each operating model are determined by fitting them to the bottom trawl survey data for Pacific cod in the Gulf of Alaska. We note that the epsilon bias-correction estimator is computationally expensive using the predictive-process model formulation, and therefore facilitate parameter estimation within the replicated design by decreasing the number of extrapolation-grid cells. We specifically use a k-means algorithm to identify 2000 locations, and calculate their area as the sum of areas for those extrapolation-grid cells that are nearest to each. This procedure therefore integrates across density using 2000 “quadrature points” rather than the original 36,140 extrapolation-grid cells. This decreases the spatial resolution used when integrating density, and substantially reduces computation time in particular during the epsilon bias-correction estimator. By using this new technique for both the estimation and operating model, we decrease the time required for each simulation replicate by approximately 75%, and exploratory testing confirms that it does not introduce any bias when applied to both estimation and operating models.

We evaluate model performance by recording the true biomass $\tilde{B}_{mrt}$ in each operating model m, simulation replicate r, and year t, and comparing this true biomass with the estimated biomass I_{mrtd} for each each replicate, operating model, year, and estimation model d. We specifically calculate relative error $E_{m,r,t,d} = \left(I_{m,r,t,d} - \tilde{B}_{m,r,t}\right)/\tilde{B}_{m,r,t}$ and then visualize the average relative error across all years and replicates for a given operating and estimation model. A well-performing model will have a relative error centered on zero and a low root-mean-squared relative error. In particular a minimax estimator suggests that the best model is that which minimizes the maximum error across all model scenarios (Lehmann and Casella, 1998 pg. 309), in this case constituted by the four operating models.

We also evaluate model performance by calculating the correlation between the natural logarithm of true density (from the operating model) and predicted density (from the estimation model). In particular, we calculate the correlation separately for each year, and then average across years for a given simulation replicate; this calculation emphasizes model performance in identifying areas with high or low density. This comparison specifically addresses whether a particular estimation model performs better or worse at identifying spatial variation in density; we speculate that a different estimation model might be appropriate for accurately estimating spatial variation vs. estimating the scale when integrating across space for calculating an abundance index.

3. Results

Applying a nonspatial delta-model to biomass samples for twenty species in the Gulf of Alaska and eastern Bering Sea shows many cases where model specification has large effects on resulting index variability and scale (Fig. 1). For example, *Sebastes polyspinus* in the Gulf of Alaska shows an approximately stable index using the lognormal delta-model and an increasing trend for the inverse-Gaussian. By contrast, both gamma and Tweedie models show large spikes in estimated abundance in 2001 and 2013, and agree with the lower abundance in 2015-2019 estimated by the lognormal distribution rather than the elevated estimates of the inverse-Gaussian. Similarly, *S. alutus* in the Gulf of Alaska and both *Lepidopsetta polyxystra* and *Limanda aspera* in the eastern Bering Sea show similar indices for gamma and Tweedie models, but differ from indices arising from either lognormal or inverse-Gaussian distributions. These and other examples show that sensitivity to the assumed distribution of positive catch rates is a general characteristic of delta-models, rather than an issue specifically with spatio-temporal delta-models.

Next we compare spatio-temporal indices using three resolutions (100, 250, or 500 knots) with design-based indices. Illustrating results for three selected species shows that gamma and Tweedie models generate similar indices, which are also similar in terms of both variability and trend to the design-based indices (Fig. 2). However, models with lower resolutions (100 knots) tend to estimate a higher scale than increased resolutions (250 or 500 knots) or the design-based indices. For these species, the inverse-Gaussian and lognormal models produce indices that show similar index trends and variability to other models and design-based indices, but differ greatly in terms of scale as a function of the specified spatial resolution.

Notably, AIC selects the lognormal and inverse-Gaussian for 8-11 of the twenty species for these three resolutions (Fig. 3), and often selects the lognormal even for species where the Tweedie and gamma result in indices that have an index scale more similar to design-based indices (e.g., *Sebastes alutus* in the Gulf of Alaska in Fig. 2). Specifically, the ratio of average biomass for model-and design-based indices is 0.98 and 1.01 when using bias-correction and high resolution for the gamma and Tweedie models, while this ratio is 1.23 and 1.60 for the lognormal and inverse-Gaussian models (Fig. 4, black numbers in right column). The difference between design- and model-based scale increases for the gamma and Tweedie models either without epsilon bias-correction (e.g., red values in Fig. 4), or with decreasing resolution (e.g., left and middle columns in Fig. 4).

Finally, the factorial simulation design confirms that models generally have good performance (i.e., small bias and low root-mean-squared error) when the simulation and estimation model have matching specification (i.e., diagonal panels in Fig. 5). However, the estimation models (Fig. 5 columns) differ greatly in terms of average performance when applied to data from a mis-specified simulation model. For example, the inverse-Gaussian estimation model has poor performance (e.g., large positive bias) when applied to data simulated using a gamma or Tweedie distribution, and the lognormal distribution also shows a smaller but still substantial positive bias for these operating models. By contrast, the gamma and Tweedie estimation models have a bias between -4 to +1% when applied to data for any of the operating models. We therefore conclude that both gamma and Tweedie estimation models are identified by a “minimax” estimator as the estimation models that minimizes the maximum error across alternative operating models. By contrast, the lognormal estimation model performs somewhat better than the gamma and Tweedie models with respect to the correlation between true and estimated density, particularly when fitted to data generated by an inverse-Gaussian distribution (Fig. 6). However, we note that all three distributions all do well in general as estimation models (correlation > 0.84 for each operating model). We therefore conclude that the optimal distribution for estimating spatial variation in density will in some cases be different than the optimal distribution for estimating the scale of an abundance index that is in agreement with a design-based estimator.

4. Discussion

In this study, we have shown that delta-gamma and Tweedie distributions result in a similar scale for model-based abundance indices as design-based indices for twenty stocks in the North Pacific. Results also highlight that index scale is sensitive to the number of knots used to approximate spatial variation within a spatio-temporal model when using a naïve estimator, but this sensitivity is mitigated when using the epsilon bias-correction estimator that accounts for retransformation bias. Using the highest resolution and bias-correction estimator, the delta-gamma and Tweedie models have an average ratio of 0.98 and 1.01 relative to design-based indices, indicating that they have a similar scale on average to a design-based estimator. When averaging design and model-based indices across years, the root-mean-squared log-ratio between these averages is 0.16 and 0.24, respectively. This suggests that the difference in scale (i.e., difference in average value for design- and model-based indices for a given species) is approximately 20% between these alternative approaches. Similarly, a factorial simulation design suggests that delta-gamma and Tweedie models have minimal error even for data simulated using other distributions, and therefore minimize the maximum error arising from these candidate forms of model mis-specification. This result is similar to classical statistical studies aimed at comparing lognormal and gamma distributions within generalized linear models in general (Firth, 1988; Wiens, 1999). Finally, the lognormal distribution performs best (followed closely by gamma and Tweedie models) at estimating spatial variation in density, indicating that difficulties in estimating index scale are largely separate from model ability to accurately identify spatial variation in density.

Spatio-temporal models fitted to biomass samples are already seeing widespread use in stock, ecosystem, habitat, and climate-vulnerability assessments (Thorson, 2019a). In particular, model-based indices can be generated using data that do not strictly follow a probabilistic design (Ye and Dennis, 2009), or can account for failures to consistently implement a planned design. However, there is more to learn regarding the expected performance of delta-models when the estimation model is mis-specified with respect to the data-generating process. In particular, we are surprised by the strong dependence of abundance-index scale upon the choice of probability distribution for positive catch rates. Previous simulation studies have not highlighted this model sensitivity because they: (1) focused on the proportionality of index estimates and true abundance and thereby ignored scale (Dick, 2004; Thorson et al., 2015); (2) eliminated model mis-specification by using the same distribution for generating and estimation (Johnson et al., 2019); (3) explored bias for a single class of delta-model without comparing performance across distributions (Myers and Pepin, 1990; Smith, 1990); (4) focused simulation testing on features other than the process used to generate data used in index standardization (Berg et al., 2014; Lo et al., 1992); or (5) did not document this mismatch in scale even when the estimation and simulation models were mismatched (Ono et al., 2015). We recommend further testing of delta-models using a variety of operating models, including individual- and agent-based models whose properties will not exactly match any simple estimation model. Using a variety of operating models will allow a more complete picture of the magnitude of errors arising from mis-specifying the distribution for positive catch rates. We also recommend further exploration of optimal ways of generating the SPDE mesh used in INLA and VAST; we have not explored this in detail here, but it could be one line of research to explore the sensitivity of index scale to the specified resolution.

The appropriate use of information criteria such as AIC in hierarchical (e.g., spatio-temporal) models is an unresolved topic in statistics due to the difficulty in estimating the effective degrees of freedom associated with random-effects that are shrunk towards zero (Hodges and Sargent, 2001; Wikle et al., 2019 Chapter 6). Marginal AIC is defined as the AIC score when counting only fixed effects, while conditional AIC is defined as AIC while partially counting random effects based on their estimated variance (Vaida and Blanchard, 2005). Both marginal and conditional AIC have known types of poor behavior for mixed-effects models (Greven and Kneib, 2010), and our results confirm poor behavior for marginal AIC, which tended to select the lognormal distribution even in cases when its scale differed greatly from a design-based estimate. Multiple methods have also been proposed to improve performance for marginal and conditional AIC (Müller et al., 2013; Watanabe, 2013). For example, Shang and Cavanaugh (2008) developed a bootstrap method to calculate a more appropriate penalty term, Sakamoto (2019) developed a computationally efficient approach to correct for issues in the marginal AIC, and Grevin and Kneib (2010) developed an analytic correction to the conditional AIC.

In addition to model selection, new GLMM methods can allow for more rigorous model validation though hypothesis testing. The DHARMa R package (Hartig, 2017) offers a suite of tests and validation diagnostics to evaluate uniform residuals calculated from the empirical distribution function of simulated values for an observation evaluated at the observation value. One-step-ahead residuals are calculated iteratively by evaluating marginal likelihoods of observation subsets against predicted values (Thygesen et al., 2017). Residuals can be compared with specified distributions using tests such as the Shapiro-Wilk, Komogoroc-Smirnov or Anderson-Darling hypothesis tests. However, we recommend further research regarding quantitative tools for model selection and validation, to automate the process of identifying an appropriate distribution for positive catch rates in spatio-temporal delta-models.

We also recommend continued research to identify delta-model specifications that are less sensitive to likelihood choice. One idea is to develop and implement new generalized distributions in VAST that contain common distributions as nested submodels, thereby replacing a (categorical) model selection with (continuous) parameter estimation. Hvingel et al. (2012) used the generalized gamma distribution, which adds a third parameter to the gamma and contains the lognormal, gamma, Weibull, and exponential distributions as special cases (Stacy, 1962). This distribution is difficult to fit because its parameters are highly correlated (Stacy and Mihram, 1965), although there has also been some success with reparameterizations (Prentice, 1974). An alternative approach would be to use robust estimators that are designed to be insensitive to data drawn from a range of distributions (Maronna et al., 2019). Conceptually, a robust delta-lognormal estimator would minimize sensitivity to outliers, thereby serving as a reliable default. Some theoretical and simulation work has shown promise for models without covariates or other effects like space (Rosales, 2009), but research is needed to extend robust estimators to mixed-effects models like VAST. We encourage future studies to investigate these ideas as potential solutions to make estimation of absolute indices more stable and reliable.

Whether to use a model- or design-based survey index in a given stock assessment depends in part upon how the resulting index is subsequently treated within the assessment model. In particular, it depends upon whether the index is viewed as absolute (i.e., the catchability coefficient is fixed a priori), or if the survey index is treated as relative and the parameter(s) describing survey catchability are estimated. Differences in index trend between model and design-based indices would be important regardless of how the catchability coefficient is treated, but large differences in trend were not observed among estimation spatio-temporal delta-model specifications explored (e.g., Fig. 2). Differences in index scale between model and design-based indices are important if the assessment treats the index as absolute, but have limited impact on model results if the catchability coefficient is freely estimated. In practice, bottom trawl survey biomass indices at the AFSC typically fall somewhere on a continuum between absolute (q fixed at 1) and relative (q freely estimated) indices, with several assessments residing somewhere in between by specifying informative priors or likelihood penalties for q (Table 1). Delta-models using a gamma or Tweedie distribution generally differ from the design-based index scale by 10%, and this is usually within the standard deviation of the prior distribution assumed in Alaskan groundfish assessment models implementing an informative prior for q (Table 1).

Based upon our results and in light of issues noted above, we recommend the following practices when using spatio-temporal delta-models to generate abundance indices for use in stock assessments:

1. *Compare model-based index scale with design-based indices when possible*: Most importantly, our simulation and case-study examples highlight that the choice of distribution for positive catch rates can have large effect on estimated scale. In most cases, we envision that analysts will trust the scale from a design-based estimator, and that similarity in scale could be one criterion (among others) for selecting among potential distributions.
2. *Use the gamma or Tweedie distributions by default when it is not possible to compare with design-based scale*: In other cases, a design-based estimator may not be feasible, either because the data are opportunistic (i.e., fishery-dependent catches), the survey substantially departed from the planned design (i.e., a vessel broke down), or the design is not sufficient for inference about a given stock (i.e., data from multiple designs must be combined). In

these cases, our simulation experiment suggests that the gamma or Tweedie distribution have reasonable performance across a range of data-generating mechanisms.

3. *Correct for retransformation bias using the epsilon estimator*: Our case-study results suggest that the epsilon bias-correction estimator (Thorson and Kristensen, 2016) results in a much better match between model- and design-based index scale than the naïve empirical Bayes estimator, and decreases sensitivity to model resolution.
4. *Do not assume that AIC is the only criterion for model performance*: Our results also suggest that AIC will select the lognormal distribution even in cases where it has poor match to the scale of the design-based index. We therefore recommend multiple considerations (including index scale and diagnostics) when selecting a model. We also recommend future research to develop automated approaches to calculate conditional AIC for models implemented in Template Model Builder, including the VAST model used here. This development would then allow for a detailed performance comparison between marginal and AIC for index-standardization models.
5. *Consider assessment-model structure when deciding between model- and design-based indices*: Finally, we note a variety of practices for treating the catchability coefficient for stock assessments in the North Pacific, and suspect that this same variation arises in other management regions. Eight of the twenty case-study species use a catchability coefficient that is fixed a priori, and these assessments are likely to be highly sensitive to differences in index scale. In cases where a design-based index is available and believed to measure total abundance/biomass for a fully-selected age/length class (i.e., not missing entire spatial strata due to operational problems or gear restrictions), we encourage analysts to compare the scale of model-based indices with that of design-based indices and use this information to inform

their choice of which method to use. Six assessments estimate the catchability coefficient freely, and index scale will have no effect for these assessments; in these cases, comparison of scale between model- and design-based indices could be used as a diagnostic of the spatio-temporal model, but will have direct impact on assessment-model results. Finally, six are estimated with a prior or penalty, and prior/penalty standard deviation is typically larger than the expected difference in scale between model- and design-based indices for gamma and Tweedie distributions. In summary, we recommend that the index scale be compared between model- and design-based indices in all three cases. However, the match in scale is most important for assessments that assume a fixed catchability coefficient, and is relevant to consider in cases where the design-based index is believed to measure total abundance/biomass for a fully-selected age or length-class. We recognize that this recommendation requires contextual information to interpret, and recommend further research regarding situations when a model-based index is likely to provide a more useful estimate of scale (whether due to improved precision, accounting for densities in areas that are not measured within a design-based estimator, or other reasons).

Finally, we continue to recommend that regional authorities for scientific review establish regional "Terms of Reference" (Thorson, 2019a) such that criteria for model specification are clear, transparent, and easily replicated for any stock assessment within a given region.

5. Acknowledgements

We thank Kasper Kristensen, Hans Skaug, and the TMB development team, without which VAST would not be computationally feasible. We also thank the many scientists and volunteers who have contributed to the bottom trawl surveys in the Gulf of Alaska and eastern Bering Sea shelf. Finally, we thank Lewis Barnett, C. O'Leary, and two anonymous reviewers for helpful comments on an earlier draft.

6. 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, 901. https://doi.org/10.2307/2281175

Aitchison, J., Brown, J.A., 1957. The lognormal distribution with special reference to its uses in economics. Cambridge University Press, Cambridge, MA.

Akaike, H., 1974. New look at statistical-model identification. IEEE Trans. Autom. Control AC19, 716–723.

Arreguín-Sánchez, F., 1996. Catchability: a key parameter for fish stock assessment. Rev. Fish Biol. Fish. 6, 221–242. https://doi.org/10.1007/BF00182344

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2003. Hierarchical modeling and analysis for spatial data, 1st ed. Chapman & Hall/CRC, Boca Raton, FL.

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 825–848. https://doi.org/10.1111/j.1467-9868.2008.00663.x

Berg, C.W., Nielsen, A., 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. https://doi.org/10.1016/j.fishres.2013.10.005

Bryan, M.D., 2017. Assessment of the northern and southern rock sole (*Lepidopsetta polyxstra* and *bilineata*) stocks in the Gulf of Alaska (NPFMC Bering Sea and Aleutian Islands SAFE). North Pacific Fishery Management Council, Anchorage, AK.

Burnham, K.P., Anderson, D., 2002. Model Selection and Multi-Model Inference, 2nd ed. Springer, New York.

Cao, J., Thorson, J.T., Richards, R.A., Chen, Y., 2017. Spatiotemporal index standardization improves the stock assessment of northern shrimp in the Gulf of Maine. Can. J. Fish. Aquat. Sci. 74, 1781–1793. https://doi.org/10.1139/cjfas-2016-0137

Cochran, W.G., 1977. Sampling Techniques, 3rd Edition, 3rd ed. John Wiley & Sons.

Conn, P.B., Thorson, J.T., Johnson, D.S., 2017. Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. Methods Ecol. Evol. 8, 1535–1546. https://doi.org/10.1111/2041-210X.12803

Cordue, P.L., 2007. A note on non-random error structure in trawl survey abundance indices. ICES J. Mar. Sci. 64, 1333–1337. https://doi.org/10.1093/icesjms/fsm134

Cosandey-Godin, A., Krainski, E.T., Worm, B., Flemming, J.M., 2014. Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. Can. J. Fish. Aquat. Sci. 72, 186–197. https://doi.org/10.1139/cjfas-2014-0159

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

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

Firth, D., 1988. Multiplicative Errors: Log-Normal or Gamma? J. R. Stat. Soc. Ser. B Methodol. 50, 266–268. https://doi.org/10.1111/j.2517-6161.1988.tb01725.x

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

Greven, S., Kneib, T., 2010. On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika 97, 773–789. https://doi.org/10.1093/biomet/asq042
Hartig, F., 2017. DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. R Package Version 01 5.
Hodges, J.S., Sargent, D.J., 2001. Counting degrees of freedom in hierarchical and other richly-parameterised models. Biometrika 88, 367–379. https://doi.org/10.1093/biomet/88.2.367
Hvingel, C., Kingsley, M.C.S., Sundet, J.H., 2012. Survey estimates of king crab (*Paralithodes camtschaticus*) abundance off northern Norway using GLMs within a mixed generalized gamma-binomial model and Bayesian inference. ICES J. Mar. Sci. 69, 1416–1426. https://doi.org/10.1093/icesjms/fss116
Ianelli, J.N., Fissel, B., Holsman, K., Honkalehto, T., Kotwicki, S., Monnahan, C., Siddon, E., Stienessen, S., Thorson, J.T., 2019. Assessment of the walleye pollock stock in the Eastern Bering Sea (NPFMC Bering Sea and Aleutian Islands SAFE). North Pacific Fishery Management Council, Anchorage, AK.
Johnson, K.F., Thorson, J.T., Punt, A.E., 2019. Investigating the value of including depth during spatiotemporal index standardization. Fish. Res. 216, 126–137. https://doi.org/10.1016/j.fishres.2019.04.004
Kotwicki, S., Horne, J.K., Punt, A.E., Ianelli, J.N., 2015. Factors affecting the availability of walleye pollock to acoustic and bottom trawl survey gear. ICES J. Mar. Sci. J. Cons. 72, 1425–1439.
Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., Bell, B.M., 2016. TMB: Automatic differentiation and Laplace approximation. J. Stat. Softw. 70, 1–21. https://doi.org/10.18637/jss.v070.i05
Lambert, D., 1992. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. Technometrics 34, 1–14. https://doi.org/10.1080/00401706.1992.10485228
Lauth, R.R., Conner, J., 2016. Results of the 2013 eastern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate resources (NOAA Technical Memorandum No. NMFS-AFSC-331). Alaska Fisheries Science Center, Seattle, WA.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, 2nd edition. ed. Springer, New York.
Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with r-inla. J. Stat. Softw. 63, 1–25. https://doi.org/10.18637/jss.v063.i19
Lo, N.C., Jacobson, L.D., Squire, J.L., 1992. Indices of Relative Abundance from Fish Spotter Data based on Delta-Lognormal Models. Can. J. Fish. Aquat. Sci. 49, 2515–2526.
Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M., 2019. Robust Statistics: Theory and Methods, 2 edition. ed. Wiley, Hoboken, NJ.
Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. 70, 141–159. https://doi.org/10.1016/j.fishres.2004.08.002
Methot, R.D., 2009. Stock Assessment: Operational Models in Support of Fisheries Management, in: Beamish, R.J., Rothschild, B.J. (Eds.), The Future of Fisheries Science in North America. Springer Netherlands, Dordrecht, pp. 137–165.
Mitchell, E.M., Lyles, R.H., Schisterman, E.F., 2015. Positing, fitting, and selecting regression models for pooled biomarker data. Stat. Med. 34, 2544–2558. https://doi.org/10.1002/sim.6496
Müller, S., Scealy, J.L., Welsh, A.H., 2013. Model Selection in Linear Mixed Models. Stat. Sci. 28, 135–167. https://doi.org/10.1214/12-STS410
Myers, R.A., Pepin, P., 1990. The robustness of lognormal-based estimators of abundance. Biometrics 46, 1185–1192.
Ng, V.K.Y., Cribbie, R.A., 2017. Using the Gamma Generalized Linear Model for Modeling Continuous, Skewed and Heteroscedastic Outcomes in Psychology. Curr. Psychol. 36, 225–235. https://doi.org/10.1007/s12144-015-9404-0

NPFMC, 2019a. Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Gulf of Alaska. North Pacific Fishery Management Council, Anchorage, AK.

NPFMC, 2019b. Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea and Aleutian Islands Region. North Pacific Fishery Management Council, Anchorage, AK.

Ono, K., Punt, A.E., Hilborn, R., 2015. Think outside the grids: An objective approach to define spatial strata for catch and effort analysis. Fish. Res. 170, 89–101. https://doi.org/10.1016/j.fishres.2015.05.021

Pennington, M., 1983. Efficient Estimators of Abundance, for Fish and Plankton Surveys. Biometrics 39, 281–286.

Petitgas, P., 2001. Geostatistics in fisheries survey design and stock assessment: models, variances and applications. Fish Fish. 2, 231–249.

Prentice, R.L., 1974. A Log Gamma Model and Its Maximum Likelihood Estimation. Biometrika 61, 539–544. https://doi.org/10.2307/2334737

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosales, M.A.C., 2009. The Robustness of Confidence Intervals for the Mean of Delta Distribution. Western Michigan University.

Sakamoto, W., 2019. Bias-reduced marginal Akaike information criteria based on a Monte Carlo method for linear mixed-effects models. Scand. J. Stat. 46, 87–115.

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Stat. 6, 461–464. https://doi.org/10.1214/aos/1176344136

Shang, J., Cavanaugh, J.E., 2008. Bootstrap variants of the Akaike information criterion for mixed model selection. Comput. Stat. Data Anal. 52, 2004–2021. https://doi.org/10.1016/j.csda.2007.06.019

Shelton, A.O., Thorson, J.T., Ward, E.J., Feist, B.E., 2014. Spatial semiparametric models improve estimates of species abundance and distribution. Can. J. Fish. Aquat. Sci. 71, 1655–1666. https://doi.org/10.1139/cjfas-2013-0508

Smith, S.J., 1990. Use of statistical models for the estimation of abundance from groundfish trawl survey data. Can. J. Fish. Aquat. Sci. 47, 894–903.

Stacy, E.W., 1962. A Generalization of the Gamma Distribution. Ann. Math. Stat. 33, 1187–1192. https://doi.org/10.1214/aoms/1177704481

Stacy, E.W., Mihram, G.A., 1965. Parameter Estimation for a Generalized Gamma Distribution. Technometrics 7, 349–358. https://doi.org/10.1080/00401706.1965.10490268

Stefansson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J Mar Sci 53, 577–588.

Syrjala, S.E., 2000. Critique on the use of the delta distribution for the analysis of trawl survey data. ICES J. Mar. Sci. 57, 831–842. https://doi.org/10.1006/jmsc.2000.0571

Thompson, G., Thorson, J.T., 2019. Assessment of the Pacific cod stock in the Eastern Bering Sea. In Stock assessment and fishery evaluation report for the groundfish resources of the Bering Sea and Aleutian Islands (NPFMC Bering Sea and Aleutian Islands SAFE). North Pacific Fishery Management Council, Anchorage, AK.

Thorson, J.T., 2019a. Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. Fish. Res. 210, 143–161. https://doi.org/10.1016/j.fishres.2018.10.013

Thorson, J.T., 2019b. Perspective: Let's simplify stock assessment by replacing tuning algorithms with statistics. Fish. Res., Recruitment: Theory, Estimation, and Application in Fishery Stock Assessment Models 217, 133–139. https://doi.org/10.1016/j.fishres.2018.02.005

Thorson, J.T., 2018. Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative. Can. J. Fish. Aquat. Sci. 75, 1369–1382. https://doi.org/10.1139/cjfas-2017-0266

Thorson, J.T., 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, 1311–1321. https://doi.org/10.1093/icesjms/fsw193

Thorson, J.T., Haltuch, M.A., 2018. Spatiotemporal analysis of compositional data: increased precision and improved workflow using model-based inputs to stock assessment. Can. J. Fish. Aquat. Sci. 1–14. https://doi.org/10.1139/cjfas-2018-0015

Thorson, J.T., Kristensen, K., 2016. Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. Fish. Res. 175, 66–74. https://doi.org/10.1016/j.fishres.2015.11.016

Thorson, J.T., M. Elizabeth, C., Stewart, I.J., Punt, A.E., 2013. The implications of spatially varying catchability on bottom trawl surveys of fish abundance: a proposed solution involving underwater vehicles. Can. J. Fish. Aquat. Sci. 70, 294–306.

Thorson, J.T., Shelton, A.O., Ward, E.J., 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. Cons. 72, 1297–1310. https://doi.org/10.1093/icesjms/fsu243

Thorson, J.T., Stewart, I.J., Punt, A.E., 2011. Accounting for fish shoals in single-and multi-species survey data using mixture distribution models. Can. J. Fish. Aquat. Sci. 68, 1681–1693.

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K., Nielsen, A., 2017. Validation of ecological state space models using the Laplace approximation. Environ. Ecol. Stat. 24, 317–339. https://doi.org/10.1007/s10651-017-0372-4

Vaida, F., Blanchard, S., 2005. Conditional Akaike information for mixed-effects models. Biometrika 92, 351–370.

von Szalay, P.G., Raring, N.W., 2016. Data report: 2015 Gulf of Alaska bottom trawl survey (NOAA Technical Memorandum No. NMFS-AFSC-325). US Department of Commerce, National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Alaska Fisheries Science Center, Seattle, WA.

Wakabayashi, K., Bakkala, R.G., Alton, M.S., 1985. Methods of the U.S.-Japan demersal trawl surveys, in: Bakkala, R.G., Wakabayashi, K. (Eds.), Results of Cooperative US-Japan Groundfish Investigations in the Bering Sea during May-August 1979.

Watanabe, S., 2013. A widely applicable Bayesian information criterion. J. Mach. Learn. Res. 14, 867–897.

Wiens, B.L., 1999. When Log-Normal and Gamma Models Give Different Results: A Case Study. Am. Stat. 53, 89–93. https://doi.org/10.1080/00031305.1999.10474437

Wikle, C.K., Zammit-Mangion, A., Cressie, N., 2019. Spatio-Temporal Statistics with R, 1 edition. ed. Chapman and Hall/CRC, Boca Raton.

Wilberg, M.J., Thorson, J.T., Linton, B.C., Berkson, J., 2010. Incorporating time-varying catchability into population dynamic stock assessment models. Rev. Fish. Sci. 18, 7–24.

Ye, Y., Dennis, D., 2009. How reliable are the abundance indices derived from commercial catch-effort standardization? Can. J. Fish. Aquat. Sci. 66, 1169–1178.

Figures and Tables

Table 1: All stocks included in analysis, including the scientific and common name of the assessed species, the region for each stock (GOA=Gulf of Alaska, EBS=Eastern Bering Sea), and a reference for the stock assessment. We also list how the catchability coefficient for the bottom trawl survey is treated (either fixed at a value *a priori*, estimated with a prior distribution, or estimated freely without a prior distribution), the coefficient of variation for the associated prior when estimated using one, and whether catchability is varying over time either through a time-dependent parameterization or implicit variation due to estimated time-varying selectivity.

Scientific name	**Common name**	**Region**	**Assessment reference**	**Treatment of catchability coefficient**	**CV of prior on catchability coefficient**	**Time-varying catchability**
Atheresthes stomias	Arrowtooth Flounder	GOA	Spies et al., 2019a	Fixed	--	Not time-dependent
Microstomus pacificus	Dover Sole	GOA	McGilliard et al., 2019	Fixed and estimated with prior	85%	Time-blocks (fixed one block, estimated one block)
Hippoglossoides elassodon	Flathead Sole	GOA	Turnock et al., 2017	Fixed	--	Not time-dependent
Sebastes polyspinis	Northern Rockfish	GOA	Cunningham et al., 2018	Estimated with prior	45%	Not time-dependent
Gadus macrocephalus	Pacific Cod	GOA	Barbeaux et al., 2019	Estimated freely	--	Time-dependent through selectivity
Sebastes alutus	Pacific Ocean Perch	GOA	Hulson et al., 2019	Estimated with prior	45%	Not time-dependent

Lepidopsetta polyxystra and *L. bilineata*	Northern and Southern Rock Sole	GOA	Bryan, 2017	Fixed	--	Not time-dependent
Gadus chalcogrammus	Walleye Pollock	GOA	Dorn et al., 2019	Estimated with prior	10%	Not time-dependent
Pleuronectes quadrituberculatus	Alaska Plaice	EBS	Wilderbuer and Nichol, 2019	Fixed	--	Not time-dependent
Beringraja binoculata	Alaska Skate	EBS	Ormseth, 2018	Fixed	--	Not time-dependent
Atheresthes stomias	Arrowtooth Flounder	EBS	Spies et al., 2019a	Estimated freely	--	Time-dependent through annual deviations related to bottom water temperature
Reinhardtius hippoglossoides	Greenland Turbot	EBS	Bryan et al., 2018a	Fixed	--	Not time-dependent
Atheresthes evermanni	Kamchatka Flounder	EBS	Bryan et al., 2018b	Estimated freely	--	Time-dependent through annual deviations related to bottom water temperature
Lepidopsetta polyxystra	Northern Rock Sole	EBS	Wilderbuer et al., 2018	Fixed	--	Not time-dependent
Gadus macrocephalus	Pacific Cod	EBS	Thompson and Thorson, 2019	Estimated freely	--	Time-dependent through selectivity
Hippoglossus stenolepis	Pacific Halibut	EBS	--	Estimated freely in areas-as-fleets model	--	Not time-dependent
Gadus chalcogrammus	Walleye Pollock	EBS	Ianelli et al., 2019	Estimated freely	--	Time-dependent through selectivity
Limanda aspera	Yellowfin Sole	EBS	Spies et al., 2019b	Estimated with prior	90%	Time-dependent through annual deviations related to bottom water temperature

Anoplopoma fimbria	Sablefish	GOA and EBS	Hanselman et al., 2019	Estimated with prior	30%	Not time-dependent

Figure captions

Figure 1: Model-based abundance indices (y-axis) in each year (x-axis) for each of twenty species (panels), showing estimates from four nonspatial models: three Poisson-link delta-models using lognormal (red), gamma (green), and inverse-Gaussian (blue) distributions for positive catches, and a Tweedie distribution for modeling both encounter rate and positive catch rate (grey).

Figure 2: Visualizing model-based abundance indices (y-axis, shown on log-scale) in each year (x-axis) for each of three species (columns) using four alternative distributions (rows), where each panel shows the abundance index (line) and 95% confidence interval (shaded area) for three different spatial resolutions (see color legend in bottom-right panel indicating the number of knots) as well as the design-based estimators (black dots), and each panel also includes the percent AIC weight for each distribution and resolution across models (e.g., where percentages for a given color sum to 100% for each column)

Figure 3: Marginal AIC weights (y-axis) for each distribution (x-axis) using a given model resolution (rows). Each bar includes multiple colored segments, showing the AIC weight for each individual stock.

Figure 4: Histogram showing number of species (y-axis) with a given ratio between model- and design-based indices when each is averaged across years (x-axis, shown on log-scale) for three model resolutions (columns) and distributions (rows). A well-performing model will have an

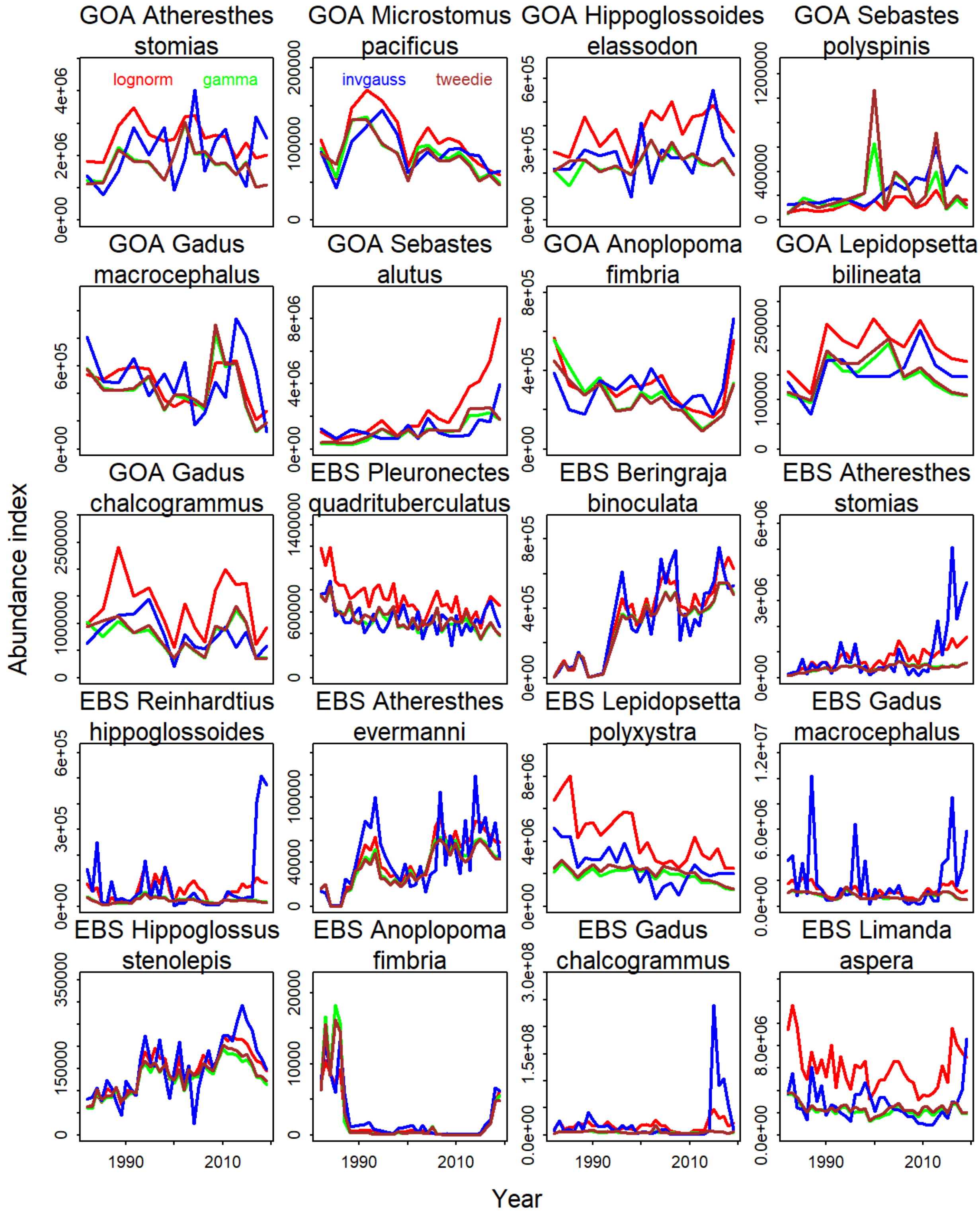
average ratio near 0 on the log scale or 1.0 on the linear scale. Each panel also has a set of numbers showing the average ratio (top-left, where 1.0 corresponds to a similar scale) and the root-mean-squared error (top-right, where 0.0 corresponds to a scale that is identical between model- and design-based approaches) when using epsilon bias-correction (black) or not using bias-correction (red).

Figure 5: Distributions of relative errors when comparing estimated and true abundance indices (x-axis) within a factorial simulation experiment conditioned on survey data for Pacific cod in the Gulf of Alaska, where the four distributions are used as operating models (rows, such that they are fitted to available data where fixed and random effects are then held constant when simulating new sampling data following the same sampling design), as well as estimation models (columns, i.e., fitted to simulated data from a given operating model). Panels on the diagonal involve the same estimation and operating model and are expected to have low error, while each column shows the performance of a given estimation model across different forms of model mis-specification. A generally well-performing estimation model will have a relative error near 0 (dashed vertical line) for all panels in a given column; each panel also lists the bias and root-mean-square-error (in parentheses) calculated for all replicates for a given operating and estimation model.

Figure 6: Distribution of Pearson correlation coefficients between estimated and true density, calculated for each year individually and then averaged across years for a given simulation replicate (x-axis), where the four distributions are used as operating models (rows) as well as estimation models (columns). See Fig. 5 caption for more details. A well-performing estimation

model will have a correlation near 1.0 for each panels in a given column; each panel also lists the average correlation calculated for all replicates for a given operating and estimation model.

856 Fig. 1



857

858

Fig. 2

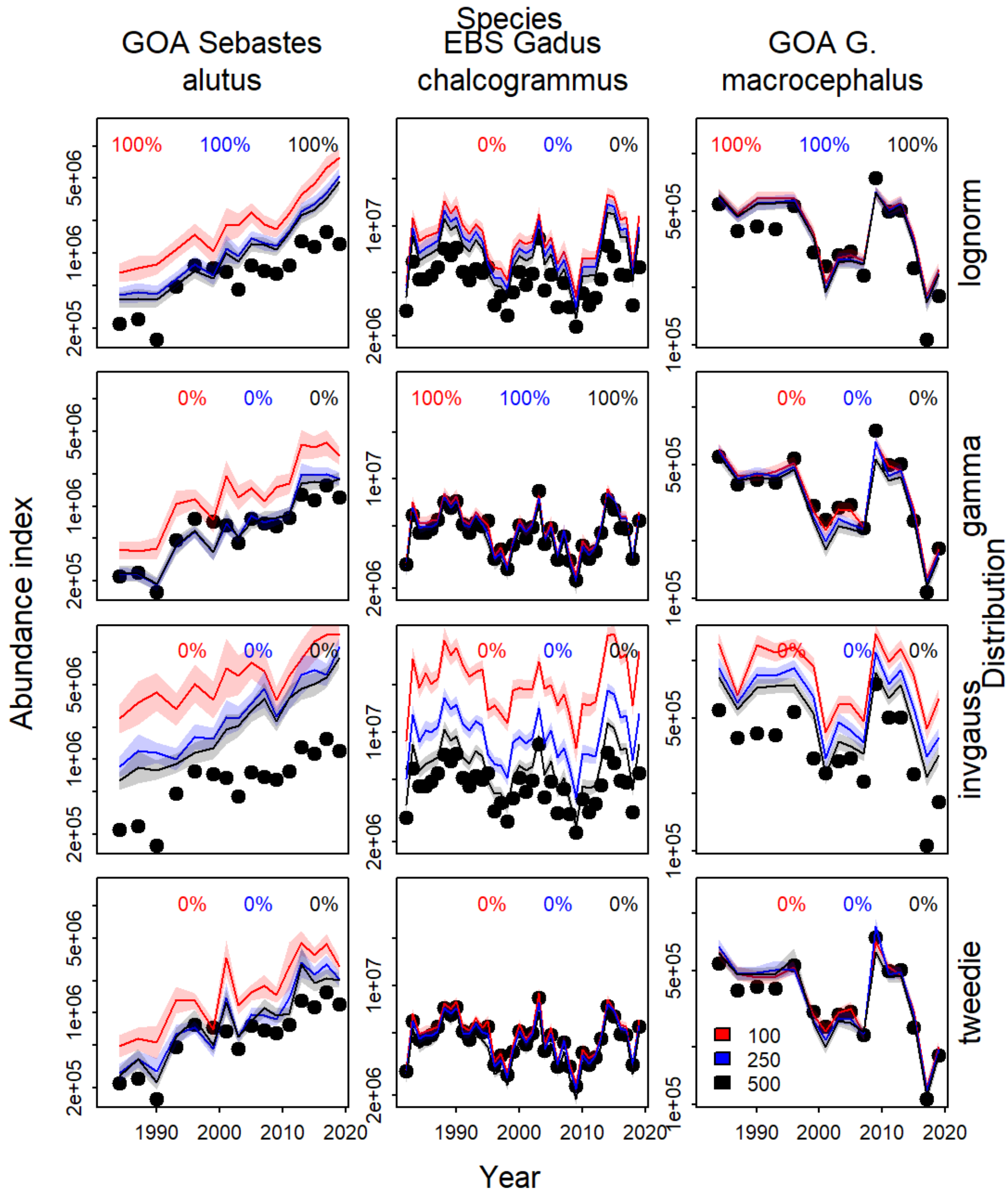

Species
GOA Sebastes alutus
EBS Gadus chalcogrammus
GOA G. macrocephalus
Distribution
lognorm
gamma
invgauss
tweedie
Abundance index
Year
100
250
500
1990
2000
2010
2020
2e+05
1e+06
5e+06
2e+06
1e+07
1e+05
5e+05
100%
0%

Fig. 3

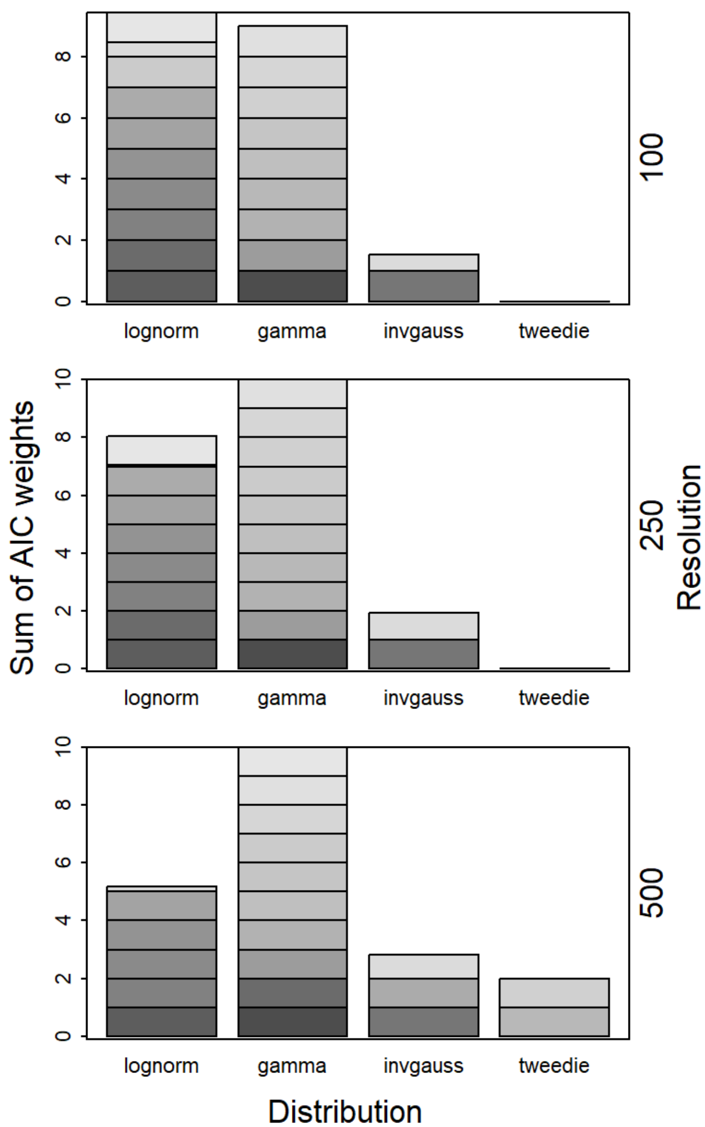
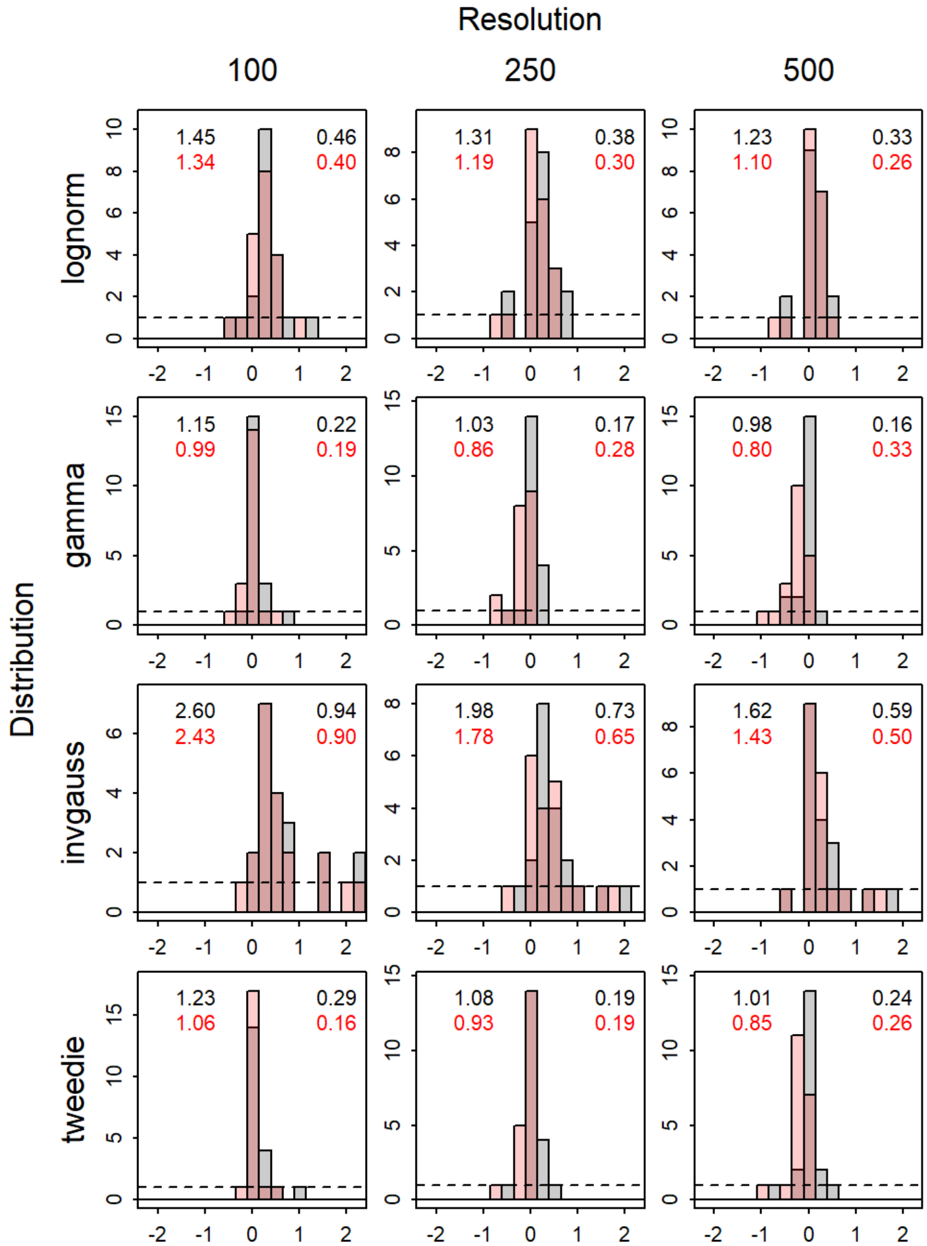

Sum of AIC weights
Distribution
Resolution
100
250
500
lognorm
gamma
invgauss
tweedie
0
2
4
6
8
10

Fig. 4



Resolution
100
250
500
Distribution
lognorm
gamma
invgauss
tweedie
1.45
1.34
0.46
0.40
1.31
1.19
0.38
0.30
1.23
1.10
0.33
0.26
1.15
0.99
0.22
0.19
1.03
0.86
0.17
0.28
0.98
0.80
0.16
0.33
2.60
2.43
0.94
0.90
1.98
1.78
0.73
0.65
1.62
1.43
0.59
0.50
1.23
1.06
0.29
0.16
1.08
0.93
0.19
0.19
1.01
0.85
0.24
0.26

Fig. 5

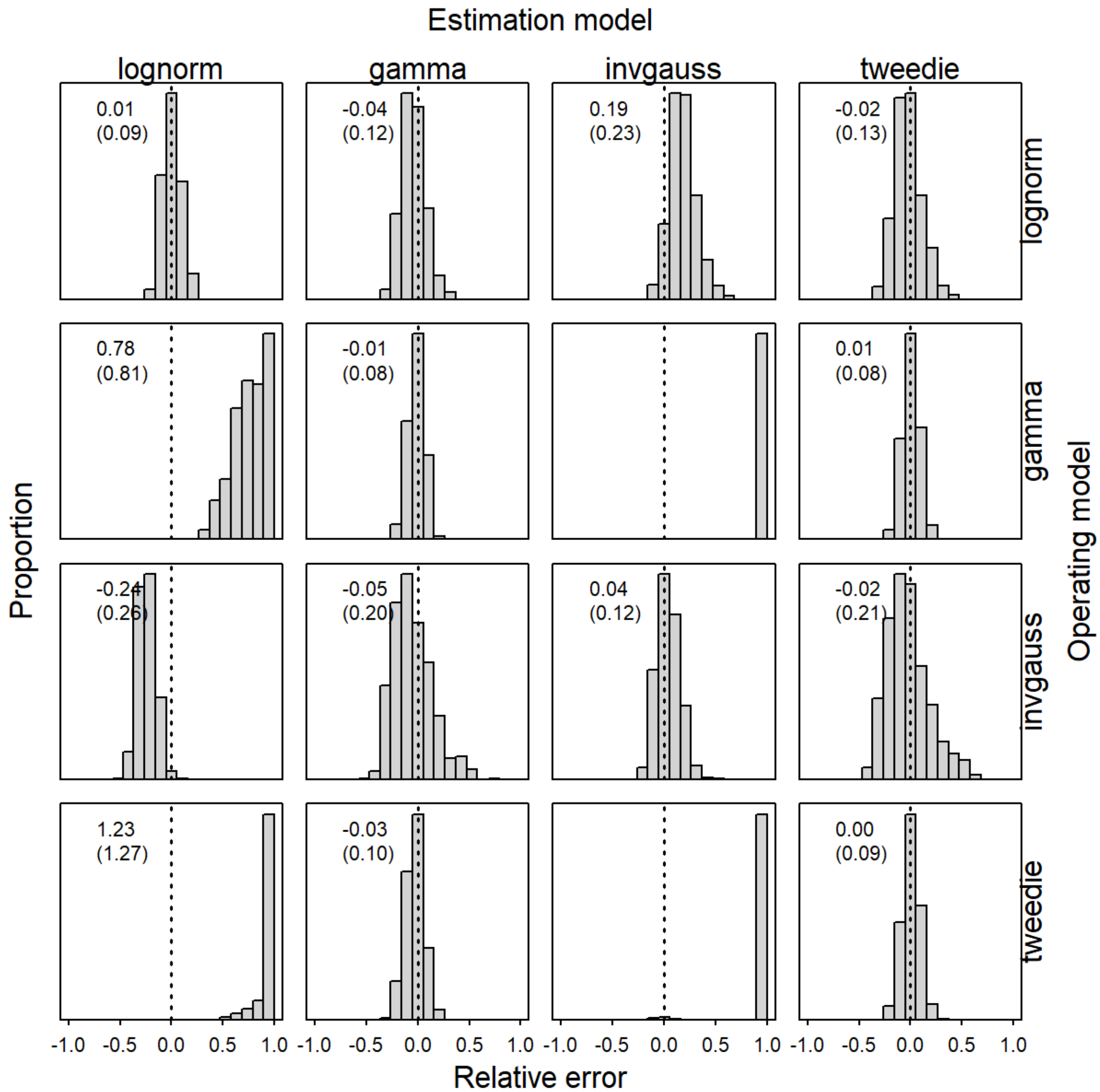
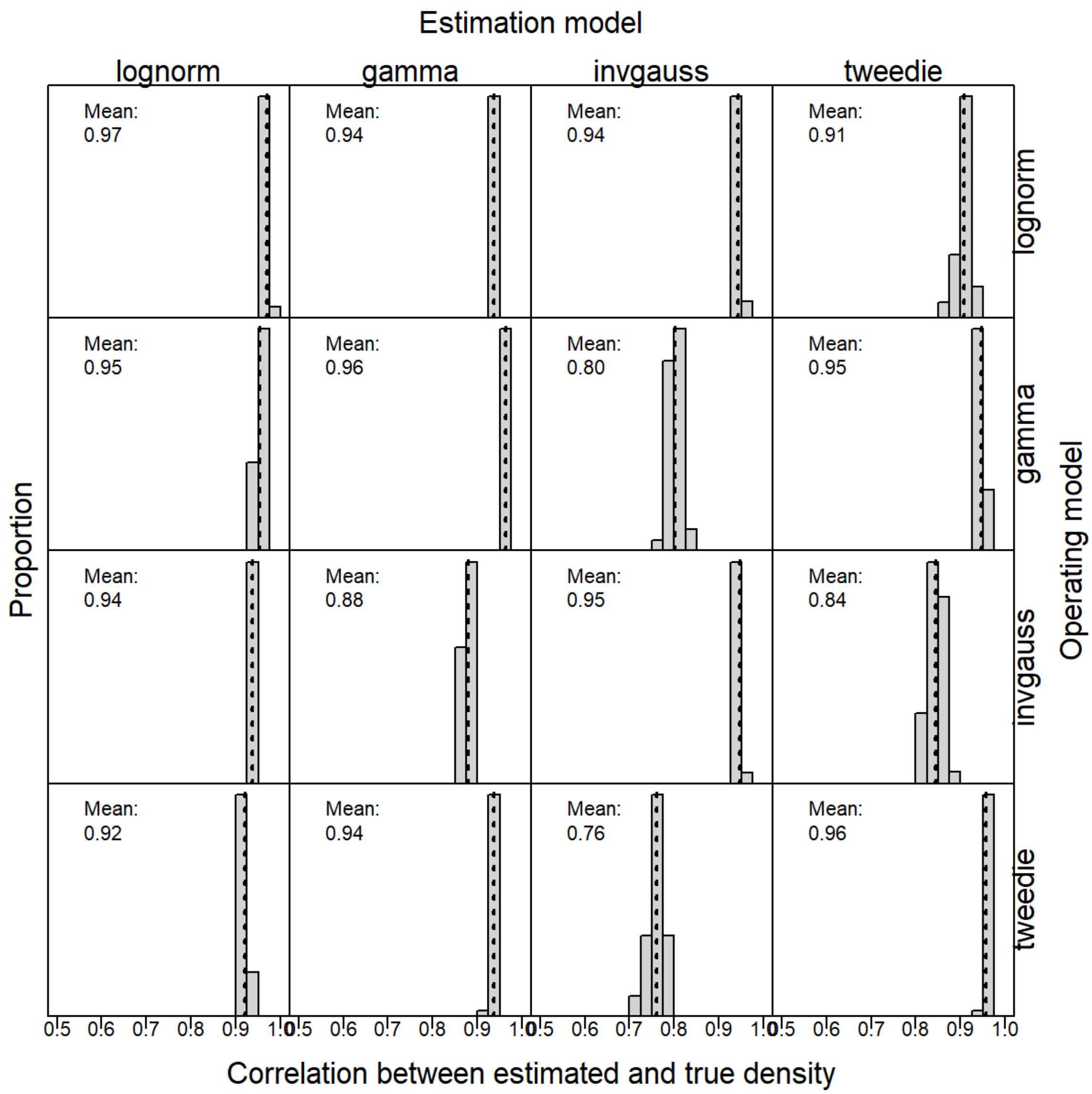

Estimation model
lognorm
gamma
invgauss
tweedie
Operating model
lognorm
gamma
invgauss
tweedie
Proportion
Relative error
0.01
(0.09)
-0.04
(0.12)
0.19
(0.23)
-0.02
(0.13)
0.78
(0.81)
-0.01
(0.08)
0.01
(0.08)
-0.24
(0.26)
-0.05
(0.20)
0.04
(0.12)
-0.02
(0.21)
1.23
(1.27)
-0.03
(0.10)
0.00
(0.09)
-1.0
-0.5
0.0
0.5
1.0

Fig. 6


Estimation model
lognorm
gamma
invgauss
tweedie
Mean:
0.97
Mean:
0.94
Mean:
0.94
Mean:
0.91
Mean:
0.95
Mean:
0.96
Mean:
0.80
Mean:
0.95
Mean:
0.94
Mean:
0.88
Mean:
0.95
Mean:
0.84
Mean:
0.92
Mean:
0.94
Mean:
0.76
Mean:
0.96
lognorm
gamma
invgauss
tweedie
Operating model
Proportion
0.5 0.6 0.7 0.8 0.9 1.0
Correlation between estimated and true density