Canadian Journal of
Fisheries and
Aquatic Sciences

Article

An empirical comparison of alternative likelihood formulations for composition data, with application to cobia and Pacific hake

Nicholas Fisch[a,c], **Robert Ahrens**[b], **Kyle Shertzer**[c], **and Ed Camp**[a]

[a]Fisheries and Aquatic Sciences, School of Forest Resources and Conservation, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL 32611, USA ; [b]National Marine Fisheries Service, Pacific Islands Fisheries Science Center, 1845 Wasp Blvd., Building 176, Honolulu, Hawaii 96818, USA; [c]National Marine Fisheries Service, Southeast Fisheries Science Center, 101 Pivers Island Road, Beaufort, NC 28516, USA

Corresponding author: **Nicholas Fisch** (email: nicholas.fisch@noaa.gov)

Abstract

Fitting composition data within stock assessment models has historically utilized the multinomial likelihood, often with iterative reweighting algorithms to account for overdispersion due to sampling and process error. Recently, the Dirichlet-multinomial has been increasingly incorporated into assessments as a composition likelihood that can be internally weighted using an estimated overdispersion parameter. There exist two popular formulations of the Dirichlet-multinomial. Recent research has also suggested improved performance in assessments using the logistic-normal for composition data, specifically when the composition sample size is large. We evaluated the performance of two Dirichlet-multinomial formulations and the logistic-normal by incorporating them into assessments that differed greatly in sample sizes for composition data: cobia (*Rachycentron canadum*) and Pacific hake (*Merluccius productus*). We compared the likelihoods against one another using various model diagnostic criteria common in stock assessments. Overall, the linear formulation of the Dirichlet-multinomial outperformed the saturating formulation. At small sample sizes of the cobia assessment, the logistic-normal performed poorly. The comparison was more robust at large sample sizes of the Pacific hake assessment; however on balance, it seems prudent to proceed with the Dirichlet-multinomial.

Key words: stock assessment, age-structured models, composition data, overdispersion, data-weighting, integrated models, model comparison, model diagnostics

Résumé

Le calage de données de composition dans des modèles d'évaluation de stocks a traditionnellement employé la probabilité multinomiale, souvent avec des algorithmes de repondération itérative pour tenir compte de la surdispersion due aux erreurs d'échantillonnage et de traitement. Ces dernières années, la loi multinomiale de Dirichlet est de plus en plus souvent intégrée aux évaluations en tant que probabilité de composition qui peut être pondérée au sein du modèle à l'aide d'un paramètre de surdispersion. Il existe deux formulations répandues de la loi multinomiale de Dirichlet. Des travaux récents indiqueraient aussi une performance améliorée dans les évaluations qui emploient la loi logistique-normale pour les données de composition, plus précisément quand la taille de l'échantillon de composition est grande. Nous évaluons la performance de deux formulations de la loi multinomiale de Dirichlet et de la loi logistique-normale en les incorporant à des évaluations qui diffèrent considérablement sur le plan de la taille des échantillons pour les données de composition, soit celles du cobia (*Rachycentron canadum*) et du merlu du Chili (*Merluccius productus*). Nous comparons les différentes probabilités en utilisant différents critères de diagnostic de modèles répandus dans les évaluations de stocks. Globalement, la formulation linéaire de la loi multinomiale de Dirichlet donne de meilleurs résultats que la formulation incrémentielle. Pour des échantillons de petite taille de l'évaluation du cobia, la loi logistique-normale ne donne pas de bons résultats. La comparaison est plus robuste pour des échantillons de grande taille de l'évaluation du merlu du Chili. Il semble toutefois prudent, en général, d'utiliser la loi multinomiale de Dirichlet. [Traduit par la Rédaction]

Mots-clés : évaluation de stocks, modèles structurés par âge, données de composition, surdispersion, pondération des données, modèles intégrés, comparaison de modèles, diagnostic de modèles

1. Introduction

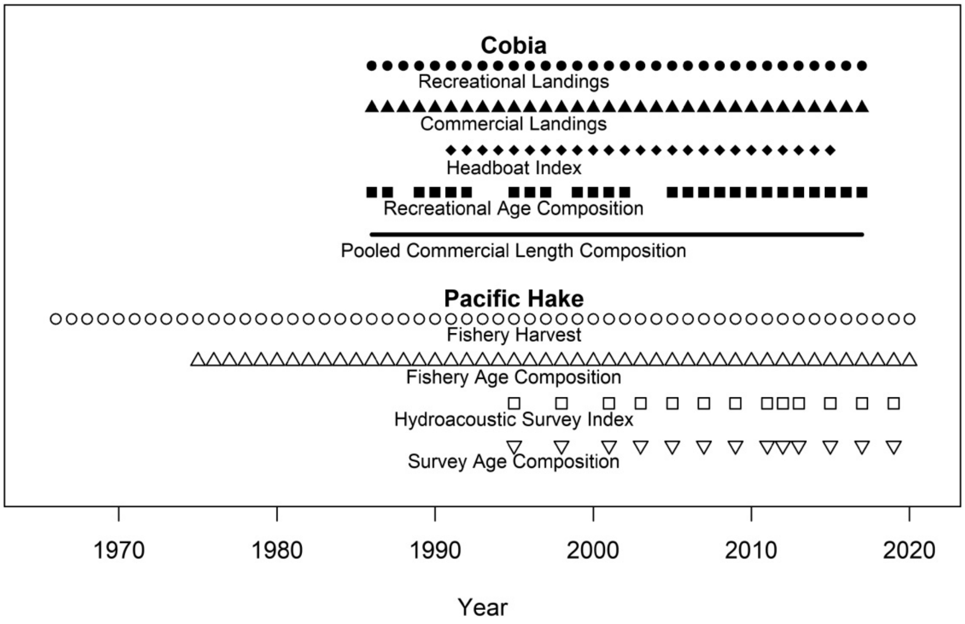
Modern fisheries management is largely facilitated by integrated stock assessments (Dichmont et al. 2016), which fit multiple data sources within a statistical model framework to estimate stock status and the influence of fishing on population dynamics. An invaluable data source often available in integrated assessments is composition data, or observations of the relative frequency of individuals in age or length groups, generally collected via sampling of the fishery harvest and by scientifically designed surveys. Composition data are critical to integrated assessments such as statistical catch-at-age (SCA) and catch-at-size (SCS) models as they make use of observations on relative differences in ages or size classes over time. Given that SCA and SCS models fundamentally track numbers of organisms through age and size classes, respectively, these relative differences are the key information source that allows for distinguishing recruitment strength, mortality-at-age, vulnerability to capture (Lee et al. 2011; Maunder and Piner 2015), and, for size-composition data, growth of organisms (Punt et al. 2016).

Composition data, as generally collected by fisheries sampling programs, suffer from two related issues: correlations and overdispersion (Pennington and Volstad 1994; Francis 2014; Thorson et al. 2017). Each stems from the general premise that there is rarely, if ever, a final pool of harvested fish at the end of the fishing season from which to take a random sample. Instead, samples must be taken throughout the season, from different ports and vessels, resulting in spatial and temporal autocorrelation between samples. Thus, composition likelihoods that assume independent and identically distributed (iid) draws from a population of interest, such as with the multinomial distribution, will be misspecified. The collection of multiple samples from a port or vessel results in clustered, and thus correlated, data. As an example, Pennington and Volstad (1994) put forth the premise of intrahaul correlation, suggesting that samples taken from the catch of one haul are more similar to each other in size and age than samples taken from different hauls. This relatedness can function at different levels such as the port of landing or the temporal grouping, and is likely present to some degree in every fishery. The relatedness between samples generally causes positive correlations between bins that are close together and negative correlations between bins that are far apart (Francis 2011, 2017). Not only are these correlations not accounted for in many composition likelihoods commonly utilized in stock assessment (Francis 2014; Fisch et al. 2021), there also exists a difference in the true amount of information conveyed in the sample compared with what is perceived. The variance of the data, because of clustered sampling, is greater than what would be expected had the same number of fish been sampled under iid assumptions. This is termed overdispersion and causes difficulty in appropriately weighting composition data within an integrated assessment, as the total number of fish measured or aged is not necessarily representative of the expected sampling error in the data. In addition, the process model of a stock assessment is theoretically guaranteed to be misspecified (as all models are approximations of reality). This not only causes additional variation in the model residuals above that expected from sampling error alone but depending on the nature of the misspecification, it can also induce correlation patterns in residuals similar to those that arise as a function of nonindependent sampling (Francis 2011). In concert, while composition data are invaluable for integrated SCA and SCS assessments, they also introduce some challenges.

Several studies have explored these general issues in stock assessments (Francis 2011, 2014; Maunder 2011; Albertsen et al. 2017; Thorson et al. 2017; Fisch et al. 2021). They generally focused on down-weighting the composition data in the total likelihood function using either iterative reweighting algorithms (Francis 2011; Truesdell et al. 2017) or alternative likelihoods that can be weighted within the assessment (Maunder 2011; Francis 2014; Albertsen et al. 2017; Thorson et al. 2017; Fisch et al. 2021). Fisch et al. (2021) utilized a spatially explicit operating model to generate overdispersed composition data, containing correlations similar to those described in real data and then fit those data using a variety of composition likelihoods, including those iteratively weighted, in spatially aggregated SCA assessment models. The examined likelihoods included the multinomial, the robust multinomial, the Dirichlet, the Dirichlet-multinomial, and the logistic-normal. They found that the likelihood which performed optimally depended on the sample size of the composition data and on the degree of misspecification between the operating model and the assessment model. Generally, the likelihoods with estimable weighting within the assessment outperformed the iteratively weighted likelihoods. More specifically, the Dirichlet-multinomial likelihoods performed optimally at small sample sizes or when there was little model misspecification, and the logistic-normal likelihoods performed optimally when there was significant model misspecification conditional on the composition sample size being at least moderate to large.

Herein, we aim to extend the analysis presented by Fisch et al. (2021) by empirically evaluating the performance of stock assessments under different likelihood formulations. We do this by incorporating alternative likelihoods for composition data within stock assessments of two managed species within the USA and Canada and compare their performance using various model evaluation criteria common to stock assessment (Carvalho at al. 2021; Kell et al. 2021). Specifically, we examine the best performing likelihoods from Fisch et al. (2021), which were the additive logistic-normal with a first-order autoregressive (AR(1)) parameterization of the variance–covariance matrix and two parameterizations of the Dirichlet-multinomial likelihood, one with an effective sample size (ESS) that linearly scales with the actual sample size and one with an ESS that saturates as the actual sample size increases. Given the findings of Fisch et al. (2021) with respect to the sample size of the composition data, we focused our efforts on two stock assessments at opposite ends of the spectrum with respect to the number of fish aged for composition data from the fishery: cobia (*Rachycentron canadum*; SEDAR 2020) whose age-composition sample size is comparatively small, and Pacific hake (*Merluccius productus*; Grandin et al. 2020), whose fishery age-composition surpasses

Fig. 1. Data availability in each year for each stock assessment. Each symbol denotes a year of data available for a different source, with the source identified below the symbols. The pooled length composition is depicted as an unbroken line to reflect the fact that it is a pooled data source across years.



thousands of fish aged over many years. This approach functions as an empirical evaluation of the Fisch et al. (2021) simulation study. Given the results of Fisch et al. (2021), we hypothesize that the logistic-normal likelihood will perform poorly in the cobia assessment and comparatively well in the Pacific hake assessment (assuming at least a moderate degree of process error).

2. Methods

Our general approach was to fit each stock assessment model using different composition likelihoods and to use various model comparison/diagnostic criteria to evaluate comparative performance. Given the logistic-normal likelihood is continuous and the Dirichlet-multinomial is discrete, the models cannot be compared using information criterion measures such as the widely applicable information criterion (WAIC; Watanabe 2010) or Pareto-smoothed importance sampling leave one out criterion (PSIS-LOO; Vehtari et al. 2017). For this reason, we chose to compare models via a suite of diagnostic/comparison criteria, by first evaluating point estimates and uncertainty, and subsequently fits to data, posterior profiles of data likelihood components, retrospective analyses, and hindcasting. Technical details of the cobia and the Pacific hake stock assessments can be found in SEDAR (2020) and Grandin et al. (2020), respectively. In the following sections, we briefly summarize these details. Our goal was to keep our assessment models as similar as possible to these respective, published benchmarks as these formulations have passed reviews by panels comprising independent experts, as well as by science and statistical committees that advise the management process.

2.1. Cobia

The cobia stock assessment (SEDAR 2020) was fit to commercial landings, general recreational landings, a recreational fishery headboat index, age-composition from the recreational fishery, and a pooled length composition from the commercial fishery (Fig. 1). The assessment was age structured, modeling ages 1–16+, from 1986 to 2017. The assessment model included two fleets (commercial and recreational) and estimated 120 parameters, in addition to composition weighting parameters (discussed below). Parameters estimated included initial fishing mortality (used to calculate an initial equilibrium age structure), initial abundance deviations from equilibrium for ages 2 through the plus group at 16 (15 parameters), initial fishing mortality, yearly recruitment deviations (30 parameters), mean fully selected fishing mortality for each fleet (two parameters), fully selected fishing mortality deviations in each year for each fleet (62 parameters), the coefficient of variation (CV) of length at age from the commercial landings, unfished recruitment, the standard deviation (SD) of log-recruitment, catchability for the headboat index, logistic selectivity parameters for the commercial fishery (two parameters), and two blocks of logistic selectivity parameters for the recreational fishery (four parameters). Natural mortality at age was assumed time-invariant and was fixed at estimates calculated from Charnov et al. (2013) using parameters from a von Bertalanffy growth function (von Bertalanffy 1957). The stock–recruitment relationship was modeled using the Beverton–Holt formulation (Beverton and Holt 1957) with steepness fixed at 0.99. Removals and index data were both fit with lognormal likelihoods. The removals from each fleet assumed CVs of 0.05 each and the relative variance between years for the index data was prespecified at estimates calculated during standardization. Informative

Table 1. Estimated parameters for each assessment model, including prior specification.

Description	Symbol	Prior
Cobia		
Unfished recruitment (log scale)	R_0	U[10, 16]
SD of recruitment	σ_R	N[0.6, 0.15^2]
Recruitment deviations (31, log scale)	$\delta_{y,\,R}$	N $\left[0, \sigma_R^2\right]$
Initial abundance deviations (15, log scale)	$\delta_{a,\,Ninit}$	U[−5, 5]
Initial fishing mortality	f_{init}	N[0.005, 0.00125^2]
Mean fully selected recreational fishing mortality (log scale)	$\bar{f}_{y,\text{rec}}$	U[−10, 0]
Recreational fishing mortality deviations (32, log scale)	$\delta_{y,\,\text{frec}}$	U[−10, 10]
Mean fully selected commercial fishing mortality (log scale)	$\bar{f}_{y,\text{comm}}$	U[−10, 0]
Commercial fishing mortality deviations (32, log scale)	$\delta_{y,\,\text{fcomm}}$	U[−12, 12]
Index catchability (log scale)	q	U[−16, −4]
Recreational selectivity age at 50% selected block 1	$m_{\text{rec},\,1}$	U[0.1, 15]
Recreational selectivity slope block 1	$k_{\text{rec},\,1}$	N[2, 1.1^2]
Recreational selectivity age at 50% selected block 2	$m_{\text{rec},\,2}$	U[0.1, 15]
Recreational selectivity slope block 2	$k_{\text{rec},\,2}$	N[2, 1.1^2]
Commercial selectivity age at 50% selected	m_{comm}	N[3.3, 1.155^2]
Commercial selectivity slope	k_{comm}	N[2, 1.1^2]
CV of length at age for commercial landings	CV_L	U[0.05, 0.5]
Pacific hake		
Unfished recruitment (log scale)	R_0	U[13, 17]
Steepness	h	Beta[9.76, 2.8]
Recruitment deviations (75, log scale)	$\delta_{y,\,R}$	N[0, 1.4^2]
Natural mortality (log scale)	M	N[−1.6, 0.1^2]
Fishing intensity (55, log scale)	f_y	U[−10, 0]
Fishery selectivity (5)	p_a	U[− 5, 9]
Selectivity deviations (150)	$\varepsilon_{a,\,y}$	N[0, 1.4^2]
Additive SD of index (log scale)	σ_{Index}	U[−3, 0.2]
Survey selectivity (4)	$p_{a,\,s}$	U[−5, 9]
Composition weighting parameters		
Dirichlet-multinomial overdispersion (2, log scale)	$\theta\|\beta$	Cobia—U[−10, 10] Hake—N[0, 1.813^2]
Logistic-normal SD (2, log scale)	σ_{LN}	U[−5, 5]
Logistic-normal Phi (2)	φ	U[−1, 1]

Note: Parentheses identify the number of parameters estimated for a group. Composition weighting parameters refer to both stock assessments.

priors were placed on recruitment variance, recruitment deviations, initial fishing intensity, and most selectivity parameters (Table 1).

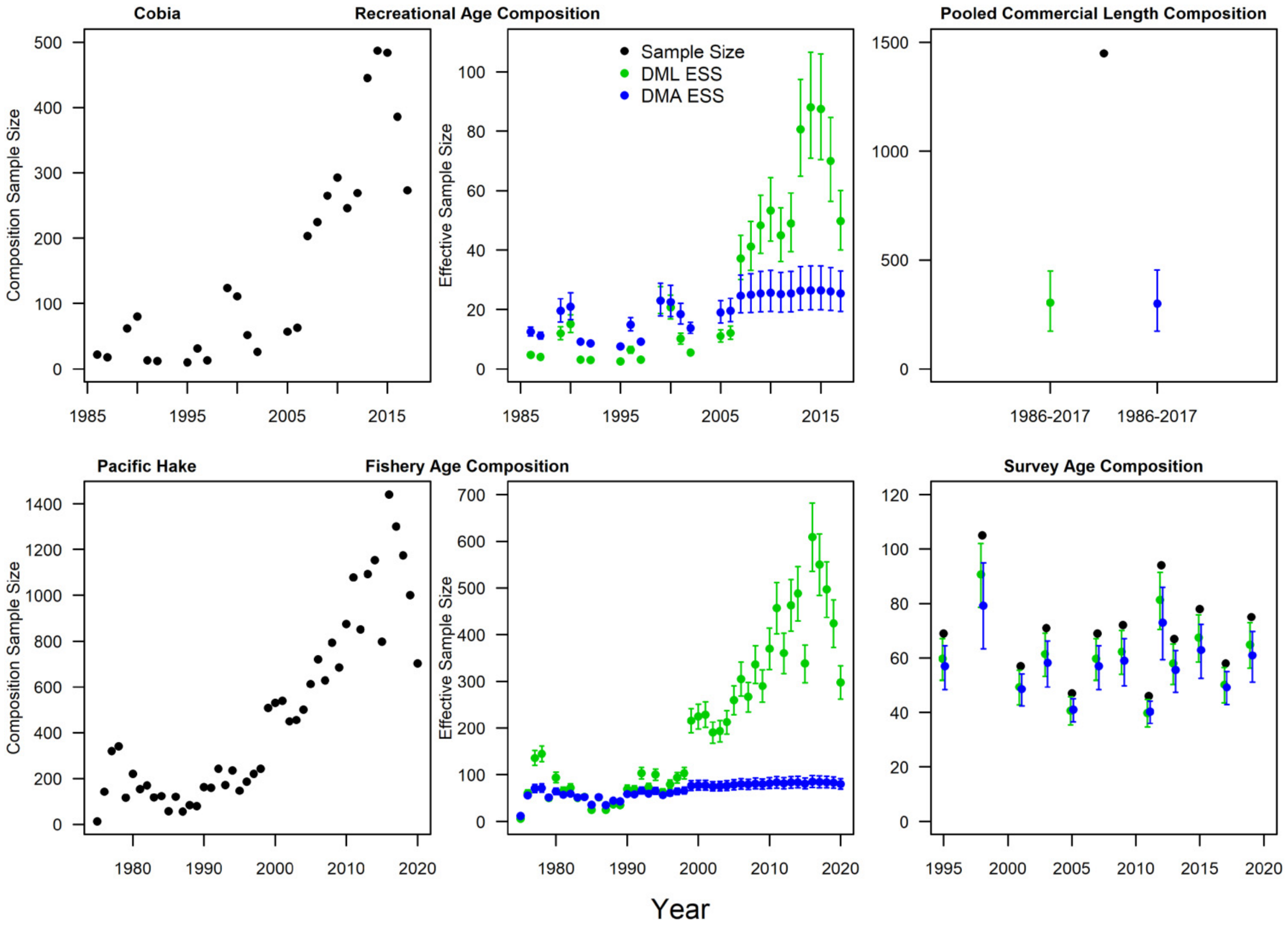
In the benchmark assessment (SEDAR 2020), the index data were weighted iteratively until the SDs of the normalized residuals were near 1. When incorporating alternative composition likelihoods into the assessment, we did not repeat this procedure as we wanted to observe how the fit to the index data might change when the composition likelihood changed. In addition, when incorporating the logistic-normal into the cobia assessment, the logistic selectivity for the commercial fleet was initially estimated at implausible selectivity values and resulted in nonconvergence of the model. To remedy this, we place a weakly informative prior on the age at 50% vulnerability according to a normal probability density with a mean of 3.3 and a CV of 0.35 (previously it was ~U[1, 10]). This is undesirable and a demerit against the logistic-normal likelihood model for cobia. For the purposes of consistency, we incorporated this prior into the Dirichlet-multinomial fits of the cobia assessment as well.

The sample size for composition data of the recreational fishery was fewer than 100 individuals for a large part of the first half of the time series. During latter years, it increased to a maximum of 484 although generally remained around 250 individuals aged (Fig. 2). The recreational age-composition data utilized in the assessment were expanded through weighting the age samples by state landings to provide an age-composition representative of the entire fleet across states. The pooled commercial length composition for the time series had a sample size of 1449 individuals.

2.2. Pacific hake

The Pacific hake stock assessment (Grandin et al. 2020) was fit to fishery harvest, the age-composition of the harvest, a hydroacoustic survey index, and age-composition sampled during the survey. It was an age structured assessment, model-

Fig. 2. Composition sample sizes and estimated ESSs for each assessment. The top row depicts the recreational age composition sample size and ESSs of the cobia assessment in the first two columns and the pooled length composition sample sizes in the third. The second row depicts the fishery age composition sample sizes and ESSs of the Pacific hake assessment in the first and second columns and the survey age composition sample sizes in the third. Shown are medians and 95% highest posterior density (HPD) intervals for the Dirichlet-multinomial linear formulation (DML; green) and the Dirichlet-multinomial saturating formulation (DMA; blue) models. [Colour online.]



ing ages 0–20+, from 1966 to 2020. The assessment modeled one fishery fleet and one survey fleet. In addition to composition weighting parameters, the model estimated 293 parameters including unfished recruitment, steepness, recruitment deviations (75 parameters), a time- and age-invariant natural mortality parameter, fishing intensity in each year (55 parameters), fishery selectivity parameters (five parameters), selectivity deviations (150 parameters), survey selectivity parameters (four parameters), and an additive SD term for the survey index. The model incorporated time-varying fishery selectivity by estimating year- and age-specific deviations in the years 1991–2020. The assessment also included year-specific ageing error matrices, calculated outside of the model. The only difference from the assessment of Grandin et al. (2020) and that done in this study is that a fully selected fishing mortality parameter (fishing intensity) was estimated for each year as opposed to employing the "hybrid" approach in Stock Synthesis (Methot and Wetzel 2013). The acoustic survey was fit using a lognormal likelihood with an estimated SD added to the observed sampling variability (obtained via kriging; Grandin et al. 2020). The harvest was also fit with a lognormal likelihood assuming a very small SD (0.01). Informative priors were placed on natural mortality, steepness, recruitment deviations, and selectivity deviations (and Dirichlet-multinomial composition parameters (defined below); Table 1).

The sample size for the fishery composition reported in the Pacific hake stock assessment (Grandin et al. 2020) was characterized by the number of trips sampled. This number averaged ∼200 trips sampled for the first half of the time series but surpassed 1000 trips sampled in many years toward the end of the time series (Fig. 2). An examination of tables 5–8 in Grandin et al. (2020) suggests that in the latter 10 years of the time series, the annual number of fish sampled fluctuated between 3000 and 6000 for the US fleets, which make up the largest proportion of the catch. This puts the number of fish sampled in each year at levels between the moderate and large treatments from Fisch et al. (2021). The sampled catch at age data for the fishery were preprocessed prior to incorporation of the assessment to consider the sampling protocols used to collect them. A full description of the analytical steps for expanding the age-compositions can be found in Taylor et al. (2014). For the acoustic survey, the composition sample size presented references the number of tows conducted by the integrated acoustic trawl survey and averages ∼70 tows in each year the survey was implemented.

2.3. Composition likelihoods

2.3.1. Dirichlet-multinomial

Each benchmark assessment utilized the Dirichlet-multinomial likelihood for composition data. Specifically, they each used the linear formulation. We abbreviate this formulation hereinafter in the text using “DML”. This formulation of the likelihood results in an ESS for composition data that scales linearly with the true sample size. The negative log-likelihood for this formulation is found using

$$\begin{aligned}(1)\quad &\text{NLL}\\ &= -\sum_{y}\Bigg[\log\left[\Gamma\left(N_y+1\right)\right]-\sum_{a}\left\{\log\left[\Gamma\left(N_yP_{a,y}+1\right)\right]\right\}\\ &+\log\left[\Gamma\left(\theta N_y\right)\right]-\log\left[\Gamma\left(N_y+\theta N_y\right)\right]\\ &+\sum_{a}\left\{\log\left[\Gamma\left(N_yP_{a,y}+\theta\,N_y\widehat{P}_{a,y}\right)\right]-\log\left[\Gamma\left(\theta\,N_y\widehat{P}_{a,y}\right)\right]\right\}\Bigg]\end{aligned}$$

where $\widehat{P}_{a,y}$ represents the predicted composition proportion for a given age (a) and year (y), $P_{a,y}$ the observed composition proportion, N_y the sample size in each year, and θ denotes a Dirichlet-multinomial overdispersion parameter estimated within the assessment. The Dirichlet-multinomial necessitates one overdispersion parameter per composition data source, and thus, each assessment estimated two Dirichlet-multinomial overdispersion parameters. The ESS for this formulation can be found using $\text{ESS}_y = (1 + \theta N_y)/(1 + \theta)$. Fisch et al. (2021) suggested that using the Dirichlet-multinomial saturating parameterization might improve performance; thus, we incorporated this likelihood into the assessments as our second evaluated likelihood. We abbreviate the saturating or asymptotic formulation hereinafter in the text using “DMA”. This formulation results in an ESS that asymptotes as true sample size increases.

$$\begin{aligned}(2)\quad &\text{NLL}\\ &= -\sum_{y}\Bigg[\log\left[\Gamma\left(N_y+1\right)\right]-\sum_{a}\left\{\log\left[\Gamma\left(N_yP_{a,y}+1\right)\right]\right\}\\ &+\log\left[\Gamma\left(\beta\right)\right]-\log\left[\Gamma\left(N_y+\beta\right)\right]\\ &+\sum_{a}\left\{\log\left[\Gamma\left(N_yP_{a,y}+\beta\widehat{P}_{a,y}\right)\right]-\log\left[\Gamma\left(\beta\widehat{P}_{a,y}\right)\right]\right\}\Bigg]\end{aligned}$$

Differences between the saturating and linear formulations simply include substituting the weighting parameter β for instances of θN_y and multiplying each term in the above ESS equation by N_y. This removes the linear scaling between the ESS and the true sample size and allows instead for a saturating function $\text{ESS}_y = (N_y + N_y\beta)/(N_y + \beta)$ (Thorson et al. 2017).

2.3.2. Logistic-normal

The third likelihood we evaluated was the logistic-normal (Schnute and Richards 1995; Aitchison 2003; Francis 2014). A composition conforms to a logistic-normal distribution with parameters [$\boldsymbol{P}$, $\mathbf{C}$] when $P_a = \left(\mathrm{e}^{X_a}/\sum_a \mathrm{e}^{X_a}\right)$. In this case, $\boldsymbol{X}$ conforms to a multivariate normal distribution with mean log($\boldsymbol{P}$) and covariance matrix $\mathbf{C}$. The logistic-normal is a continuous distribution (where the Dirichlet-multinomial is discrete) and is theoretically able to account for correlations between bins by specifically parameterizing the variance–covariance matrix to do so (although the correlations are on the original multivariate normal scale; Francis 2014). In this study, we explored the performance of a first-order autoregressive, AR(1), parametrization of the variance–covariance matrix. This parameterization necessitates two parameters per composition data source: σ_{LN} and φ. Incorporating this likelihood into each assessment required the estimation of two more parameters than each formulation of the Dirichlet-multinomial. Weighting between years based on composition sample size was achieved using $W_y = \sqrt{\bar{\text{N}}/\text{N}_y}$, where $\bar{\text{N}}$ denotes the mean sample size over the time series and $\sigma_{\text{LN},y} = \sigma_{\text{LN}}W_y$, as in Francis (2014). This allows the variance term, $\sigma_{\text{LN},y}$, to vary by year, which results in a unique variance–covariance matrix each year, while the correlations between bins, $\rho_{|a-a'|}$, are treated as constant over time. The variance–covariance matrix in each year, $\mathbf{C}_y$, is calculated using $C_{y,a,a'} = \sigma^2_{\text{LN},y}\rho_{|a-a'|}$, where $\rho_{|a-a'|} = \varphi^{\,|a-a'|}$ for an AR(1) process. The negative log-likelihood can then be found using eq. A9 in Francis (2014):

$$\begin{aligned}(3)\quad &\text{NLL}\\ &= \sum_{y}\Bigg[0.5\left(\text{Nb}-1\right)\log\left(2\pi\right)+\sum_{a}\left[\log\left(P_{a,y}\right)\right]\\ &+0.5\log\left(\left|\mathbf{V}_y\right|\right)+\left(\text{Nb}-1\right)\log\left(W_y\right)+\frac{\left(\mathbf{w}_y^T\mathbf{V}_y{}^{-1}\mathbf{w}_y\right)}{2W_y^2}\Bigg]\end{aligned}$$

where $\mathbf{V}_y = \mathbf{K}\mathbf{C}_y\mathbf{K}^T$, $\mathbf{K}$ is a matrix with dimensions [(Nb − 1), Nb] formed by adding a vector of −1s to the right side of an identity matrix with dimensions [Nb − 1, Nb − 1] , and $\mathbf{w}$ is a matrix where each row depicts a year and contains a vector of length (Nb − 1), filled using $w_{a,y} = \log\left(\frac{P_{a,y}}{P_{\text{Nb},y}}\right) - \log\left(\frac{\widehat{P}_{a,y}}{\widehat{P}_{\text{Nb},y}}\right)$ for a in 0, 1, 2,…, Nb – 1. The term Nb refers to the number of bins in a composition data set. Hereinafter, we abbreviate the logistic-normal models with an AR(1) variance–covariance matrix in the text using “LN”. Note that for the Pacific hake assessment, informative priors were placed on the Dirichlet-multinomial overdispersion parameters where diffuse uniform priors were placed on the parameters from the logistic-normal (Table 1). Diffuse uniform priors were placed on all composition weighting parameters for the cobia assessment.

2.4. Comparison criteria

We first examined sensitivity of estimated management quantities and their uncertainty to the specified likelihood formulation. We focused on estimates of depletion, spawning biomass, and exploitation rate for each assessment as these are generally of interest in making management decisions. Uncertainty was assessed via 95% HPD intervals in addition to using posterior CVs, defined as the SD of the posterior distribution divided by the median. These observations have little bearing on the diagnoses of which likelihood performed optimally, as a model with different point estimates or more/less uncertainty is not necessarily a better performing model than another. An exception to this may be if the point estimates or the bounds of uncertainty are implausible.

For model comparison or determination of an optimal composition likelihood for each assessment, we evaluated fits to the data sources, posterior profiles of data components, retrospective analyses, and hindcasting. Each of these diagnostics is detailed below.

2.4.1. Fits to data

We evaluated fits to the abundance indices for each assessment using the SD of the normalized residuals (SDNRs, Breen et al. 2003; Francis 2011; Carvalho et al. 2017) and a nonparametric runs test (Wald and Wolfowitz 1940). A relatively good model fit is characterized by a SDNR near 1 (Carvalho et al. 2017), although Francis (2011) notes that a value much less than 1 is not a cause for concern, but rather means that the data set is fitted better than was expected. A runs test is meant to assess whether the sign of the residuals is random with respect to time. A nonrandom pattern of residuals can potentially indicate model misspecification (Carvalho et al. 2017). For model comparison, we would interpret a model with a larger SDNR (specifically above 1) and (or) a model whose abundance index residuals indicate nonrandomness with respect to time as a worse-fitting model. Given that the Pacific hake assessment included an estimable parameter as an additive SD component meant to account for sources of process and sampling error in the acoustic index (Grandin et al. 2020), we also examined the magnitude of this parameter for each likelihood. We focused on abundance index data as it has been suggested that these data should have primacy when evaluating fits to multiple data sources given that they provide the most direct information about changes in abundance (Francis 2011, 2017).

We specifically examined fits to the composition data as a model diagnostic rather than a comparison tool for the different likelihoods, because the goal of incorporating different self-weighting composition likelihoods into the assessments was to allow each model to decipher how much combined sampling error and process error existed with respect to compositions. To achieve this, we calculated the root-mean-squared error (RMSE) of individual age bins for the fits to each of the composition data sets.

$$\text{RMSE}_a = \sqrt{\frac{\sum_y \left(P_{a,y} - \widehat{P}_{a,y}\right)^2}{n}} \tag{4}$$

where $P_{a,y}$ denotes an observed composition data point in year y at age a, $\widehat{P}_{a,y}$ a predicted composition, and n the number of years in the data set. We performed runs tests on the residuals of composition data using the mean observed and expected ages, $\bar{O}_y = \sum_a P_{a,y} \times a$ and $\bar{E}_y = \sum_a \widehat{P}_{a,y} \times a$, respectively. In addition, we visually examined the correlations in residuals between age bins for each age-composition data set and compared them with those that would be expected from each likelihood. Finally, we simulated prior and posterior predictive distributions for the index and composition data sources and examined the number and percentage of data points that were outside of 95% HPD intervals for each model.

We did not evaluate fits to removals because the assessment models were configured to fit them precisely, in effect treating removals as known.

2.4.2. Posterior profiles

Likelihood profiles are common model diagnostics employed in stock assessments fit in a maximum likelihood framework (Ichinokawa et al. 2014; Lee et al. 2014; Wang et al. 2014; Carvalho et al. 2021). They function by fixing a key parameter at various levels around its maximum likelihood estimate, re-estimating the remaining parameters, and examining the change in values of various likelihood components, most often those concerned with data. A large change in the likelihood value of a specific data component as the level of a key parameter varies is indicative of an informative data source for that specific parameter. In addition, the location of minima for the different individual likelihood values can suggest data conflicts if the minima are in very different locations for the parameter. Given our assessments were fit in a Bayesian framework using Markov chain Monte Carlo (MCMC), the model output already contained samples of key parameters across a range of values. For this reason, we examined posterior profiles by evaluating the posterior density of unfished recruitment with respect to the marginal posterior distributions of the negative log-likelihood values for each data component. To clarify, we did not fit any additional models as would be done in a maximum likelihood profile sense. We simply examined the posterior distributions of the negative log-likelihood values for each data component (and the prior for recruitment deviations) relative to MCMC values for unfished recruitment. We focused our posterior profiles on unfished recruitment, a key scaling parameter, which was estimated in each stock assessment. When comparing the performance of each likelihood, we anticipate large data conflicts as indicative of a worse performing model. In addition, we were interested in whether using different composition likelihoods changed the perception of information content in various data sources. For example, differences in the likelihood values of data components as unfished recruitment varies could suggest different levels of information content elucidated by the different likelihoods.

2.4.3. Retrospective analyses

We performed retrospective analyses by successively removing 1, 2, …, or 5 consecutive years of data from the end of the time series and refitting each assessment. This resulted in five "peels" fit to the reduced data sets for each stock assessment. Retrospective statistics were then calculated by comparing estimates from the terminal years of each peel to the respective year from the full assessment. We evaluated Mohn's rho (Mohn 1999) as the mean relative difference for the terminal year of each peel compared with the full assessment.

$$\rho = \frac{1}{p}\sum_{i=1}^{p}\frac{X_{T-i}^{(i)} - \widehat{X}_{T-i}}{\widehat{X}_{T-i}} \tag{5}$$

where p refers to the number of peels, T refers to the terminal year in the full model, $X_{T-i}^{(i)}$refers to a quantity in year $T - i$ from a peeled assessment fit with i number of years of data removed, and $\widehat{X}_{T-i}$ refers to a quantity from the full or reference assessment (in year $T - i$). It has been suggested that a large ρ can be indicative of model misspecification (Hurtado-Ferro et al. 2014). However, equal divergence from the full model over peeled years in positive and negative directions can result in the mean calculation producing a small ρ. Where this may allay concerns of severe model misspecification, one could argue a model whose peels diverge a greater amount from the full reference assessment is a worse performing model. For this reason, we also evaluated the mean absolute relative difference over peels (Fisch et al. 2019):

$$\lambda = \frac{1}{p}\sum_{i=1}^{p}\frac{|X_{T-i}^{(i)} - \widehat{X}_{T-i}|}{\widehat{X}_{T-i}} \tag{6}$$

This metric considers the difference in estimates in the final year of each peel compared with the reference assessment as opposed to whether or not there is a consistent pattern. Retrospective statistics were calculated for estimates of spawning biomass for each assessment. We interpret a model with smaller retrospective statistics as a preferential model over another.

2.4.4. Hindcasting

Similar to retrospective analysis, hindcasting refers to the process of leaving data points out at the end of a time series and refitting the assessment. The difference lies in that hindcasting focuses on predicting the data points left out to evaluate predictive skill (Kell et al. 2016; Kell et al. 2021). Hindcasting was implemented for index data, composition data from the fishery, and survey composition data (solely for hake), separately. We removed the final three data points of index data and the final 5 years of composition data from the fishery from each assessment. We also removed the final 3 years of survey composition data solely for Pacific hake. Thus, eight additional models were fit to the reduced data sets for cobia, and 11 for Pacific hake. The approach was to leave consecutive data points out and to assess prediction skill of the data, which were left out using the mean absolute scaled error (MASE; Hyndman and Koehler 2006). This has been called model-free hindcasting (Kell et al. 2016; Kell et al. 2021) and does not require a model forecast. Contrary to the retrospective models, the hindcasted assessments were still run for the full time series and estimated all parameters. The exception to this was for the Pacific hake model for the hindcasted fishery composition where, if a year of data was removed the fishery selectivity deviations for that year were not estimated. Commonly, hindcasting specifies a prediction horizon, h. In our hindcasts, the prediction horizon was simply equivalent to the number of data points left out. For example, in the model which was fit to a data set with the final 2 years of composition data removed, the prediction horizon was 2. The MASE score calculates an evaluation of prediction skill relative to a naïve baseline prediction, and for a given data source d and prediction horizon h, is calculated as

$$\mathrm{MASE}_h^{(d)} = \frac{\frac{1}{h}\sum_{t=T-h+1}^{T}|E_t^{(d)(h)} - O_t^{(d)}|}{\frac{1}{h}\sum_{t=T-h+1}^{T}|O_{t-1}^{(d)} - O_t^{(d)}|} \tag{7}$$

where T once again refers to the terminal year in the assessment (the same for hindcasts and the full assessment), $O_t^{(d)}$ denotes an observed data point of type d in time t and $E_t^{(d)(h)}$, calculated as a function of estimated parameters in the hindcasted assessment model, denotes a predicted data point of type d in time t from a hindcasted assessment fit with the terminal h data points of type d removed. The abundance index hindcast MASE calculations utilized the log of the predicted and observed data points where the age-composition hindcast MASE calculations were calculated using the mean observed and expected ages (defined in Section 2.4.1. Fits to data). Commonly, the naïve prediction used to calculate MASE is a random walk where $O_t = O_{t-1}$ (Hyndman and Koehler 2006; Carvalho et al. 2021; Kell et al. 2021), i.e., the next observed data point is the same as the previous. We employed this same naïve prediction in our analysis. A prediction is said to have skill if its MASE is less than 1, i.e., it improves the model prediction compared with the baseline. For example, a MASE score of 0.5 suggests that the model predicts twice as accurately as the naïve prediction. Used for model comparison, we interpret a model with lower MASE scores as a better performing model.

2.5. Fitting

Each assessment was developed and fit using Automatic Differentiation Model Builder (ADMB; Fournier et al. 2012). Assessments were run in a Bayesian framework with parameters and priors identified in Table 1. The cobia assessment models were run for 10 million iterations and the Pacific hake models were run for 50 million iterations. Each assessment consisted of a single chain and utilized the Metropolis–Hastings algorithm. Convergence was evaluated using Geweke's diagnostic (Geweke 1991) at an alpha level of 0.05. The first 15% of each chain was removed for burn in and every 1000th iteration was saved. Where calculations require point estimates from the models, we utilized the median of the MCMC chain. Finally, although we cannot compare the Dirichlet-multinomial to the logistic-normal models with information criteria, we can compare each parameterization of Dirichlet-multinomial to one another. Thus, we also compared the linear with the saturating parameterization using WAIC and PSIS-LOO.

2.6. Sensitivity runs

To explore whether particular specifications of each assessment affected our results, we explored a few different sensitivity analyses. For the cobia assessment, to test sensitivity to the iterative reweighting procedure of the index, we ran the assessment models employing iterative reweighting of the index data until its SDNR was approximately 1 (the procedure employed in the benchmark cobia assessment) for each likelihood. In addition, we ran the cobia assessment models with an additional estimated parameter meant to be an additive SD component for fitting to the index, similar to the parameterization of the index fit in the Pacific hake assessment.

To assess sensitivity to the prior specifications of the overdispersion parameters for the Dirichlet-multinomial models in the Pacific hake assessment, we ran the DML and DMA models for hake with uniform priors (U[−10, 10]) placed on the overdispersion parameters. We also ran the Pacific hake Dirichlet-multinomial assessments using a uniform prior solely for the fishery composition overdispersion parameter (where the informative prior remained for the survey parameter). To examine sensitivity to the input sample size for fishery composition in the Pacific hake assessment, we ran the DMA model with the input sample size for the fishery multiplied by 5 (in essence converting the number of trips sampled to approximate number of fish sampled). We also ran each assessment with full weight given to the input sample sizes, where the overdispersion parameters were not estimated and ESS equaled the input sample size. Finally, we examined the sensitivity to initial parameter values for each assessment by rerunning them starting at the posterior medians for alternative likelihoods. For example, we started the Pacific hake LN assessment at the median parameter estimates for the posterior distribution of the DML model and vice versa.

3. Results

3.1. Convergence, point estimates, and uncertainty

Most parameters (~95%) for each assessment model converged based on Gewekes' diagnostic at an alpha level of 0.05. Although estimates of depletion for the time series in the cobia assessment were consistent across likelihoods, estimates of spawning biomass and exploitation rate differed substantially as a function of composition likelihood specification (Fig. 3). The LN model led to consistently lower estimates of spawning biomass and larger estimates of exploitation rate throughout the time series compared with the other assessments, where the DMA model estimated a much larger spawning biomass and a much lower exploitation rate than both the DML and LN models. In addition, for spawning biomass and exploitation rate, the posterior CVs were similar for the LN and DML models where they were much larger for the DMA model (Fig. S1). Models had similar posterior CVs for depletion, although the LN model estimates were slightly larger than the others.

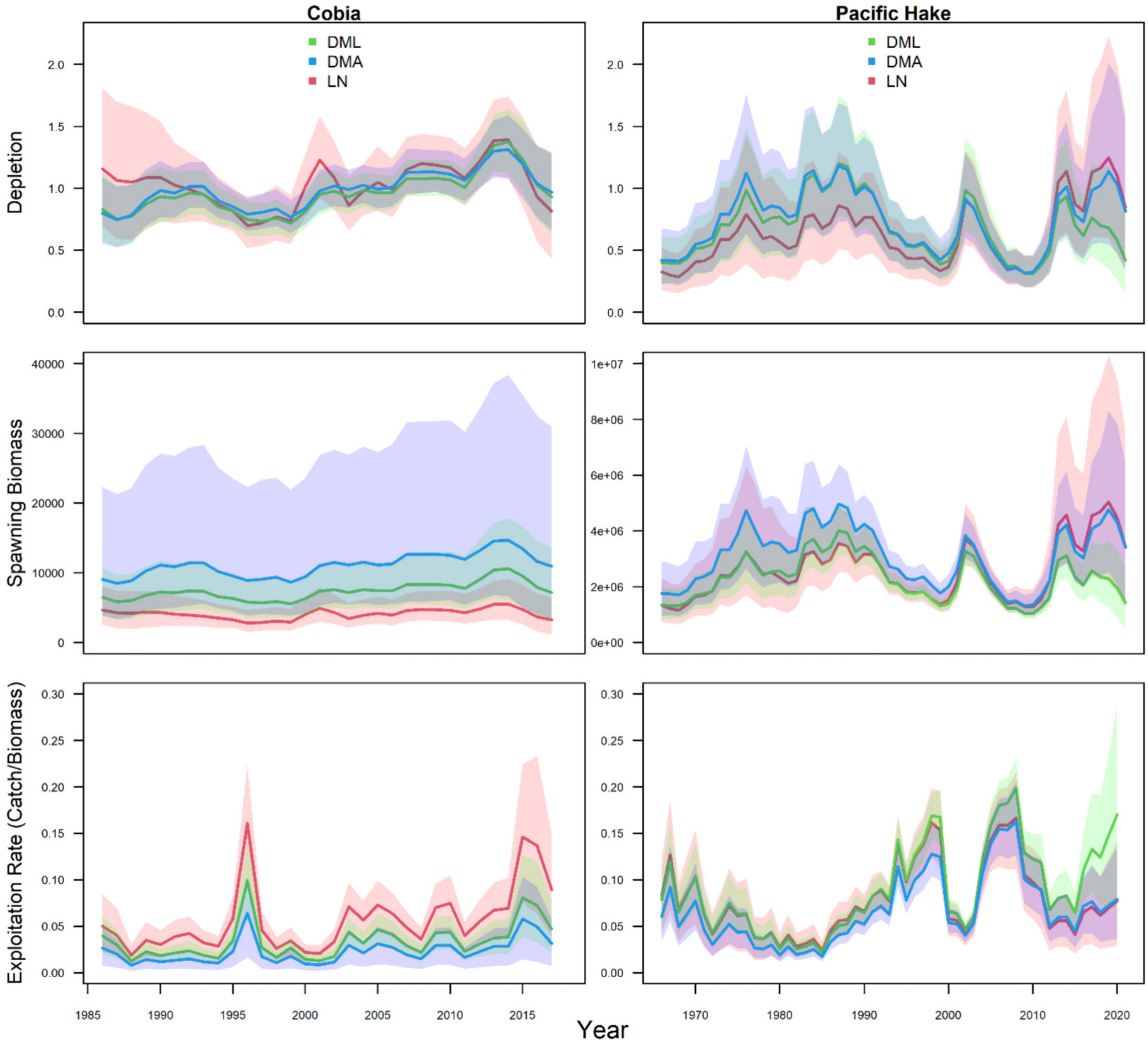
In the Pacific hake assessment, the LN and DML models produced very similar estimates throughout most of the time series but diverged considerably from one another in the final 10 years. In the terminal years, the LN model estimated a much larger degree of spawning biomass, a smaller exploitation rate, and a less depleted stock (Fig. 3). The DMA model estimates of depletion, spawning biomass, and exploitation rate differed from those of the other two models early in the time series, but converged with estimates of the LN model to produce very similar estimates at the end of the time series. This suggests two different solution spaces, with the LN flipping between them in 2010, as the LN matches DML estimates from 1966 to 2009 and DMA estimates from 2010 to 2020. Uncertainty in terms of the range of 95% HPD intervals was greater for the LN model at the end of the time series than for the DMA and the DML. Posterior CVs were similar, indicating that uncertainty was greatest for the LN model, followed by the DMA, and then the DML model (Fig. S1).

3.2. Fits to data

Fits to each data source were visually acceptable (reasonable fits with no clear residual patterns, Fig. 4, Figs. S2–S11), and all SDNR values for fits to abundance index data were approximately 1 (Table 2). In the cobia assessment, the lowest SDNR for index fits resulted from the DMA model, followed by the DML and then the LN. In the Pacific hake assessment, the lowest SDNR for index fits resulted from the LN model, followed by the DMA, and the DML. The estimated additive SD for the Pacific hake survey index was largest for the LN model, followed by the DMA, and then the DML model. The runs tests for each iteration of the MCMC chain indicated that only 1%–2% of iterations in the cobia assessment exhibited nonrandomness in residuals over the times series for fits to the abundance index. Conversely, in the Pacific hake assessment and specifically for the DML model, 19% of MCMC iterations exhibited nonrandomness in residuals for fits to the abundance index, where <4% of iterations exhibited nonrandomness for the DMA and LN models.

The RMSE for fits to the recreational age-composition data for the cobia assessment were generally lowest for the DML and DMA models, where the LN resulted in larger RMSEs (Fig. 4). The residuals for the fit to the pooled length composition were of similar magnitude across likelihoods. For the Pacific hake assessment, the RMSE for fits to the fishery age-composition was lowest for the DML, followed by the LN, and then the DMA. Conversely, for fits to the survey age-composition data, the RMSE was lowest for DMA, followed by DML and LN (Fig. 4). The runs tests for age-composition fit in the cobia assessment exhibited nonrandomness in residuals for 5%, 12%, and 5% of MCMC iterations for the DML, DMA, and LN models, respectively. In the Pacific hake assessment, runs tests for fits to the fishery age-composition identified nonrandomness in residuals for 22%, 5%, and 41% of MCMC iterations for the DML, DMA, and LN models, respectively. Similarly, for the survey age-composition, the same models exhib-

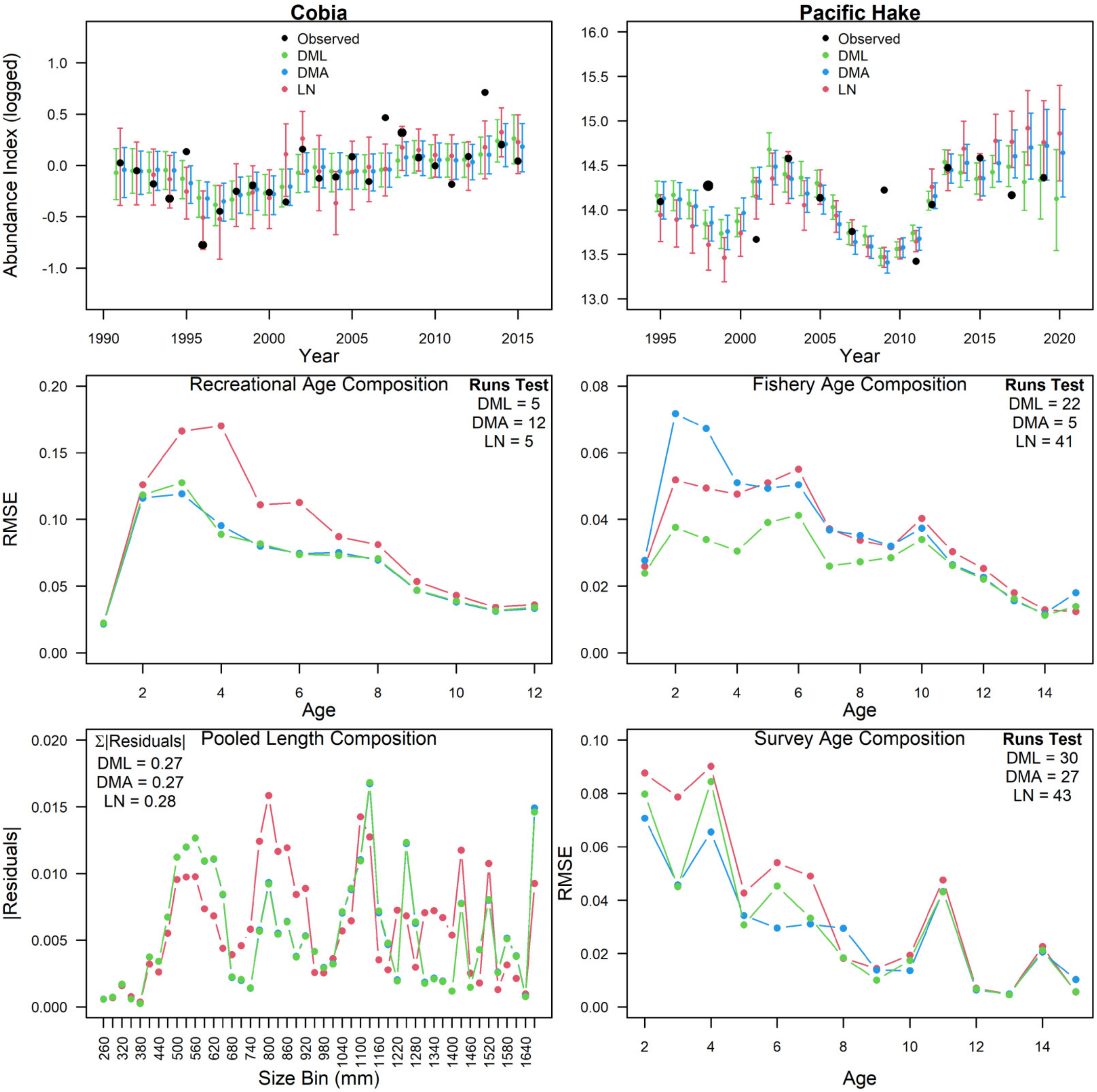
Fig. 3. Point estimates (medians) and 95% HPD intervals for depletion (spawning biomass/unfished), spawning biomass, and exploitation rate from each assessment fit using different likelihoods for composition data. Results for the logistic-normal are shown in red, the Dirichlet-multinomial linear formulation in green, and the Dirichlet-multinomial saturating formulation in blue. Spawning biomass is measured in units of metric tons of mature females for cobia and kilograms of mature biomass for Pacific hake. [Colour online.]



ited 30%, 27%, and 43% of MCMC iterations with nonrandomness in residuals per the runs tests. The observed residual correlation structure of each age composition data set showed no consistent pattern (Figs. S12–S14), and neither matched those expected from the Dirichlet-multinomial distributions or from the logistic-normal (Fig. S15).

Prior and posterior predictive distributions for the index and compositions data sources are presented in Figs. S21–S32. For cobia, no data points were outside of 95% HPD intervals of the prior predictive distributions for the headboat index. The number of data points that were outside of the 95% HPD intervals for the prior predictive distributions of the compositions was similar between the likelihoods, with 5, 2, and 7 for the DML, DMA, and LN for the pooled commercial length composition, and 7, 6, and 8 for the recreational age compositions, respectively. The number of data points that were outside of the 95% HPD posterior predictive distributions for the headboat index was 2, 2, and 0 for the DML, DMA, and LN, respectively. For the pooled commercial length composition, 7, 5, and 0 data points were outside of the 95% interval for the DML, DMA, and LN models, respectively (corresponding to 14.6%, 10.4%, and 0% of the data

Fig. 4. (Top row) Fits to abundance indices for each assessment. Black points represent observed data, where the size of the point is made relative to the inverse of the input standard error. Colored points denote medians and lines denotes 95% HPD intervals. (Bottom two rows) RMSE for fits to the composition data sets. The RMSE shown for each composition fit (outside of the pooled length composition) was calculated as the SD in residuals over years for each bin, and the median taken over MCMC samples. Given the length composition was pooled, we instead depict the absolute value of the median residual for each bin (and the sum of the absolute values of the residuals in legend text). The text in the upper right portion of each age composition plot presents the percentage of MCMC iterations whose *p*-value < 0.05 for the runs test (reject null hypothesis of randomness in residuals). A runs test was not performed for the pooled length composition. [Colour online.]



points for the length composition). For the recreational age composition, 17, 16, and 10 data points were outside of the 95% interval for the DML, DMA, and LN models, respectively (corresponding to 5.4%, 5.1%, and 3.2% of the data points for the fishery composition). Considering all data sources together (including landings), the DML and DMA models exhibited greater than 5% of data points outside of the 95% HPD intervals for posterior predictive distributions, with 5.8% of total data points outside of intervals for DML and 5.1% for DMA.

Table 2. SDNR values for fit to abundance indices for each assessment.

Likelihood	SDNR of index	Additive SD	Runs test
Cobia			
DML	1.001 (0.884, 1.141)	NA	1
DMA	0.981 (0.862, 1.125)	NA	1
LN	1.030 (0.876, 1.202)	NA	2
Pacific hake			
DML	1.044 (0.646, 1.446)	0.265 (0.141, 0.437)	19
DMA	1.037 (0.657, 1.453)	0.300 (0.154, 0.505)	3
LN	1.027 (0.646, 1.425)	0.338 (0.163, 0.579)	4

Note: Additive SD refers to the estimated additive SD term for the Pacific hake survey abundance index. Medians of the posterior distribution are reported with 95% HPD intervals in parentheses. The runs test column presents the percentage of MCMC iterations whose *p*-value < 0.05 (reject null hypothesis of randomness in residuals).

For Pacific hake, prior predictive distributions were visually similar between the different models (Figs. S21–S23), although the DML prior predictive distributions do appear more informative than the DMA and LN for the oldest ages (prior to the plus group) in each composition data source. The number of data points that were outside of the 95% HPD intervals for the prior predictive distributions was 0 for each model for the index, 17, 0, and 11 for the DML, DMA, and LN for the fishery composition, and 3, 0, and 1 for the survey composition, respectively. These results indicate that the DML priors were the most informative for the compositions, followed by the LN, and the DMA. The number of data points that were outside of the 95% HPD posterior predictive distributions was 0 for each model for the index data. For the fishery composition, 40, 33, and 18 data points were outside of the 95% interval for the DML, DMA, and LN models, respectively (corresponding to 5.7%, 4.8%, and 2.6% of the data points for the fishery composition). For the survey composition, 12, 8, and 6 data points were outside of the 95% interval for the DML, DMA, and LN models, respectively (corresponding to 6.5%, 4.4%, and 3.3% of the data points for the fishery composition). Considering all data sources together, only the DML model exhibited greater than 5% of data points outside of the 95% HPD intervals for posterior predictive distributions (with 5.5% of total data points outside of intervals).

3.3. Retrospective analysis

For cobia, retrospective patterns and statistics were closest to zero for the DML model (Fig. 5). This was followed by patterns and statistics for the DMA and subsequently the LN model. Each retrospective pattern indicated a positive bias in spawning biomass as successive years of data were omitted.

Retrospective patterns and statistics indicated mixed performance between the models for Pacific hake, where the LN model produced ρ nearest to zero; however, the DML produced the smallest estimate of λ (Fig. 5). Both the LN and the DML produced negative bias relative to the full assessment for spawning biomass as successive years of data were removed. The DMA model produced estimates farthest from zero for each retrospective statistic by a wide margin and instead estimated positive bias for spawning biomass peels compared with the full assessment.

3.4. Hindcasting

MASE values for the cobia index data were lowest for the DMA model across each prediction horizon (Table 3). This was followed by the DML model and then the LN model. Most MASE scores for the index data for cobia were below 1 (exception being $h \geq 2$ for LN), meaning those versions of the assessment are predicting the index more accurately than the naïve prediction. Conversely, for the fishery composition hindcasts, the MASE scores for each composition likelihood were all above 1. The lowest MASE score for the fishery composition hindcasts depended on the prediction horizon, with the DML having the lowest for $h \leq 2$, the DMA having the lowest for $2 < h \leq 4$, and the LN having the lowest for $h = 5$.

MASE scores for Pacific hake were the lowest with the DML model across each prediction horizon for the abundance index data. This was followed by the DMA model and then the LN model. The only model that predicted more accurately than the naïve index (MASE < 1) was the DML. For the fishery composition hindcast, the lowest MASE score depended on the prediction horizon, with the DML producing the lowest at $h = 1$, the LN at $h = 2$, and the DMA for $h \geq 3$. Only one MASE score was greater than 1 (DML when $h = 3$). Where the LN only had the lowest MASE for $h = 2$, it produced the second lowest for each of the other prediction horizons. For the survey composition hindcast, the DML produced the lowest MASE for $h = 1$, and the DMA produced the lowest for $h > 1$. The LN model had the largest MASE scores for the survey composition hindcast, predicting more accurately than the naïve prediction solely for $h = 3$.

3.5. Posterior profiles

Cobia profiles suggest that recreational fishery age-composition, the prior on recruitment deviations, and each of the landings data sources provided the most information on unfished recruitment for each of the models (Fig. 6). Differences were small between the different composition likelihoods in terms of changes in likelihood values as MCMC samples of unfished recruitment varied. However, the LN model experienced a greater change in likelihood value for the abundance index as unfished recruitment varied than did each of the Dirichlet-multinomial models. Similarly, for the recreational age-composition and recruitment deviations, the change in likelihood values as unfished recruitment varied was greatest for the DMA model, followed

Fig. 5. Retrospective figures and statistics for the cobia (top row) and Pacific hake (bottom row) assessments. The symbol rho (ρ) references Mohn's rho (eq. 5), and the symbol lambda (λ) references the mean absolute relative difference over peels (eq. 6). Spawning biomass is measured in units of metric tons of mature females for cobia and kilograms of mature biomass for Pacific hake. Shaded areas denote 95% HPD intervals for each full assessment. [Colour online.]

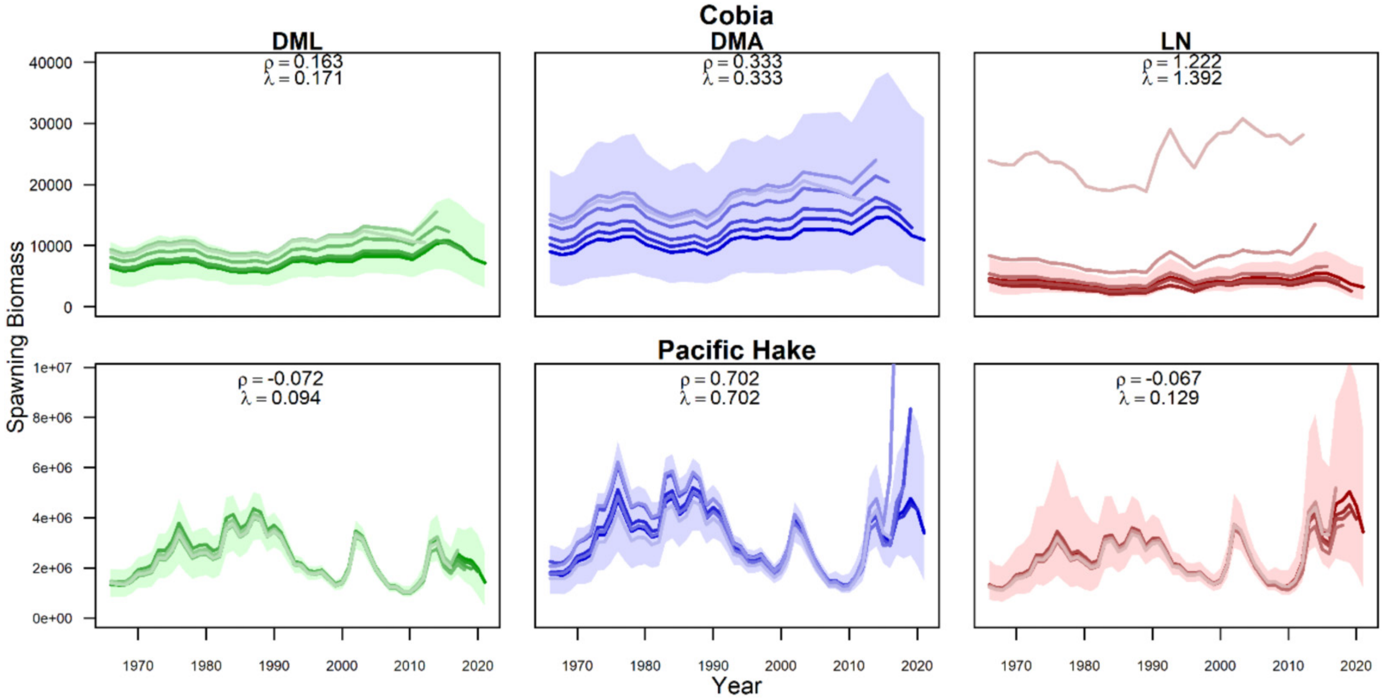
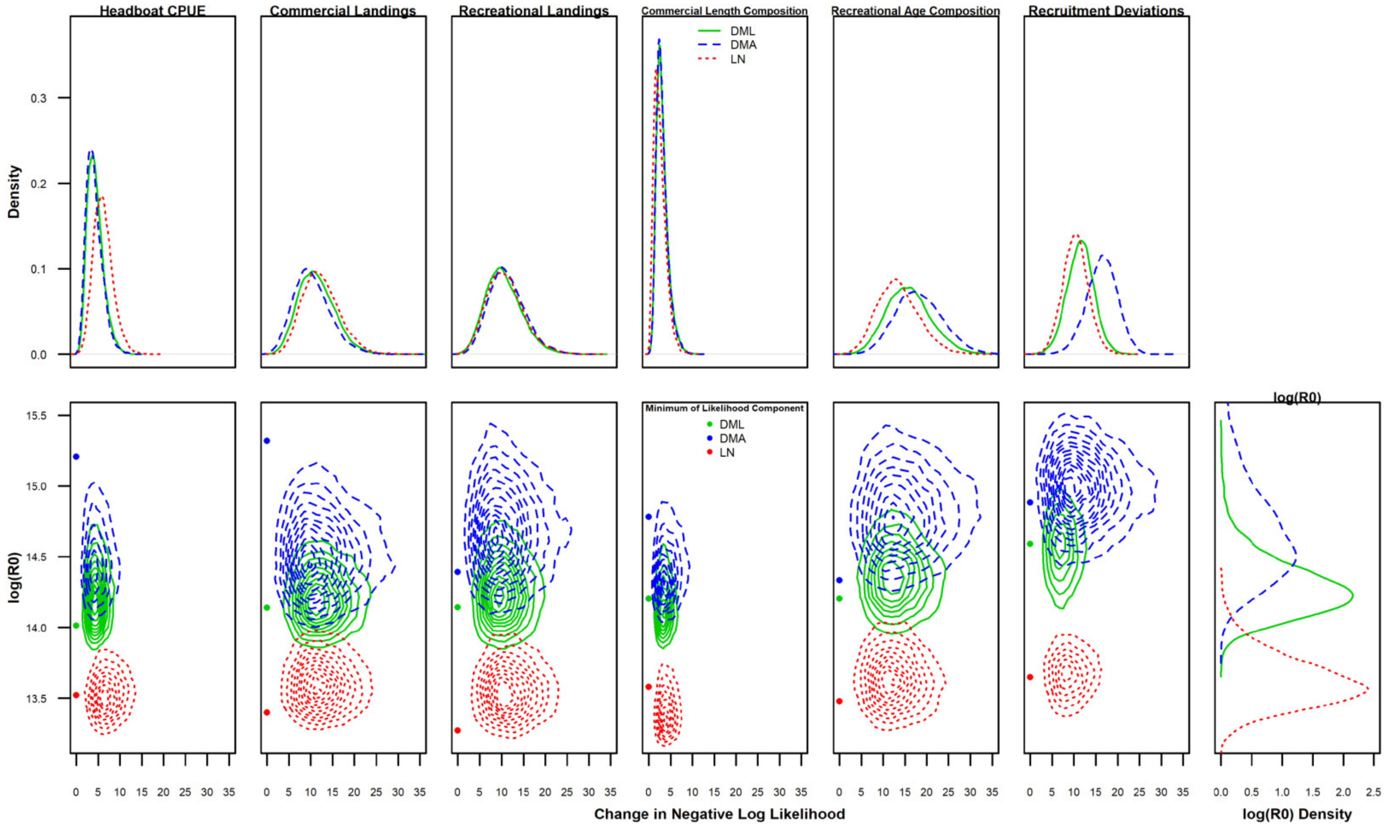


Table 3. Hindcasted MASE values for each assessment model, calculated using the median of the posterior distribution.

	h	DML	DMA	LN
Cobia				
Hindcast Index	1	0.701	0.600	0.798
	2	0.557	0.529	1.015
	3	0.863	0.857	1.093
Hindcast Fishery Composition	1	4.01	13.62	30.19
	2	1.38	1.53	2.65
	3	1.65	1.63	3.41
	4	1.57	1.41	2.71
	5	1.66	1.59	1.49
Pacific hake				
Hindcast Index	1	0.099	2.055	2.327
	2	0.737	2.010	2.235
	3	0.905	1.9	2.024
Hindcast Fishery Composition	1	0.41	0.68	0.6
	2	0.55	0.94	0.46
	3	1.08	0.4	0.86
	4	0.78	0.56	0.62
	5	0.88	0.67	0.88
Hindcast Survey Composition	1	0.08	1.67	1.14
	2	1.16	1.10	1.94
	3	0.44	0.27	0.76

Note: The symbol *h* denotes the number of data points, which were predicted for each hindcast.

Fig. 6. Cobia posterior profile. The top row depicts the marginal posterior distributions for each likelihood contribution, identified by the panel heading. The bottom row (with exception to the last column) depicts the joint posterior of each likelihood contribution and the log of unfished recruitment. Points denote locations of minima of each respective likelihood component. The last column of the final row depicts the marginal posterior distributions of the log of unfished recruitment for each likelihood. Results for DML are presented in green, for DMA in blue, and for LN in red. [Colour online.]



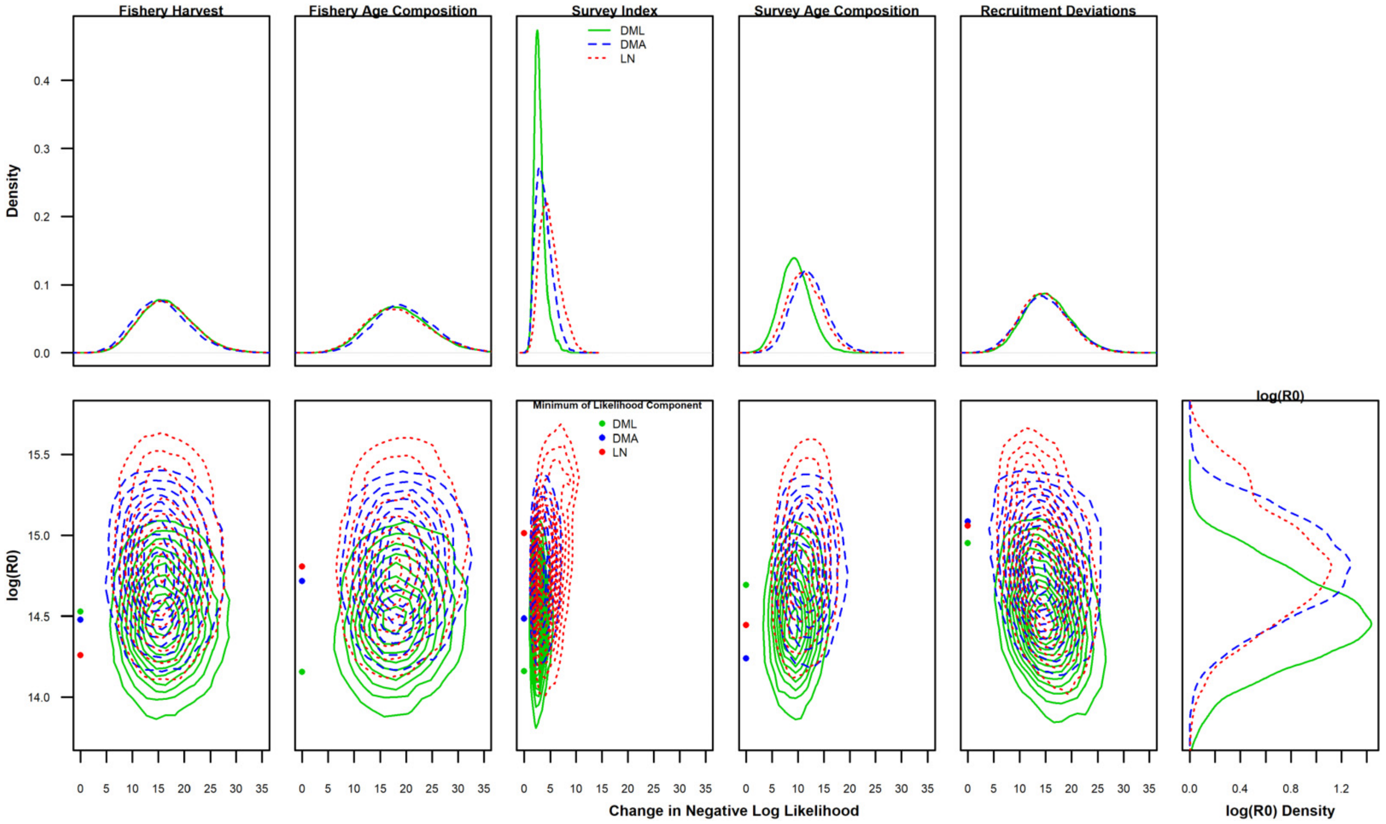
by the DML and the LN models. Variability regarding the locations of R_0 at each of the likelihood minima about the median estimate of R_0 was least for the LN model, followed by the DML and then the DMA model (Table S3).

Pacific hake profiles suggest the fishery composition data, recruitment deviations prior, harvest data, followed by the survey composition data provided the most information on unfished recruitment for each likelihood. Mostly, there was little difference in the change in individual likelihood scores across composition likelihoods, with exceptions for survey index and survey composition, which exhibited larger changes in likelihood scores for the LN and DMA compared with the DML as MCMC samples of unfished recruitment varied. Locations of R_0 at the individual likelihood component minima for each data source were quite variable across each of the composition likelihood models with the minima of fishery composition and abundance index data generally in similar locations and the minima of harvest and survey composition in similar locations (but different to index and fishery composition). This pattern seemed consistent across the DML and LN composition likelihoods (Fig. 7). Although the minima for harvest and survey composition data for DML suggested greater unfished recruitment (as compared with the minima for fishery composition and the index), it was the opposite for the LN model. For each of the models, the minima for the recruitment deviations prior suggested the largest estimates of unfished recruitment across all components examined. Overall, the CV of the likelihood minima calculated as a metric of variability of the locations of likelihood minima about the median estimate of R_0 was least for the LN model, followed by the DMA, and then the DML model (Table S3).

3.6. Information criteria and ESS

For the cobia assessment, the estimated ESS for the recreational fishery age-composition was much lower in the latter half of the time series for the DMA than for the DML (Fig. 2). A similar result occurred in the Pacific hake assessment, where for the fishery composition, the estimated ESS specifically for the latter half of the time series was much lower for the DMA than for the DML. Conversely, the ESSs of DML and DMA were similar to one another for the pooled length composition (in the cobia assessment) and for the survey age-composition (in the hake assessment). Information criteria measures WAIC and PSIS-LOO were each lower for the DML model in both the cobia and Pacific hake assessments. In the cobia assessment, the difference in criterion values between the likelihoods was 11 for WAIC and 10.4 for PSIS-LOO. In the Pacific hake assessment, these differences were 487 051 and

Fig. 7. Pacific hake posterior profile. The top row depicts the marginal posterior distributions for each likelihood contribution, identified by the panel heading. The bottom row (with exception to the last column) depicts the joint posterior of each likelihood contribution and the log of unfished recruitment. Points denote locations of minima of each respective likelihood component. The last column of the final row depicts the marginal posterior distributions of the log of unfished recruitment for each likelihood. Results for DML are presented in green, for DMA in blue, and for LN in red. [Colour online.]



5804 units, respectively. Although the diagnostic measures reported for each criterion suggested both WAIC and PSIS-LOO may be unreliable in this case (Vehtari et al. 2017).

3.7. Sensitivity

The cobia assessments fit under the different likelihoods were insensitive to both the iterative reweighting procedure for the index data and to estimating an additional parameter as an additive SD for the index, producing nearly equivalent results to those already presented in this study (Fig. S19). In addition, the DMA model for the Pacific hake assessment was largely insensitive to increases in the input sample sizes for composition data, producing results very similar to the model fit to the baseline input sample sizes (Fig. 2). The cobia assessment model fit with full weight given to the input sample sizes did result in spawning biomass point estimates that differed from both the DMA and the DML models, producing initial spawning biomass similar to the DML model; however, terminal estimates very similar to those from the DMA model (Fig. S20). Conversely, the Pacific hake model fit with full weight given to the input sample size produced nearly identical point estimates of spawning biomass as the DML model.

The Pacific hake assessment fit with uniform priors for the DML and DMA overdispersion parameters for the fishery and survey did not converge according to Geweke's diagnostic; however, when the uniform prior was solely placed on the overdispersion parameter for the fishery (and the informative prior remained for the survey), results were nearly identical to those presented in this study (and thus insensitive).

Each cobia assessment was insensitive to the alternative starting values, as was the Pacific hake DMA model. The alternative starting parameter values for the Pacific hake DML and LN assessments did result in large changes in model output compared with those presented in this study, with each model estimating larger spawning biomass levels at the end of the time series (Fig. S18). However, for each of the alternate solutions the negative log-likelihood values were larger than those for the respective baseline solution presented in this study (differences of 281 and 162 log-likelihood units at posterior medians, respectively), indicating that these alternate model solutions were suboptimal.

4. Discussion

This study compared likelihoods used for fitting composition data in stock assessment models through application to cobia and Pacific hake. A key finding was that, simply changing the likelihood, or even the formulation of the likelihood specified for composition data, led to considerable differences in model output in both the cobia and Pacific hake assessments.

Clearly, the DML was the best performing model for the cobia assessment. Compared with the LN model, it produced better retrospective and hindcasting metrics and did not necessitate the addition of a weakly informative prior to converge (on the age at 50% vulnerability for the commercial fleet selectivity). This can be interpreted as a failure to reject our hypothesis, that at small sample sizes, the Dirichlet-multinomial is a more suitable likelihood for composition data than the logistic-normal, consistent with Fisch et al. (2021). Conversely, the poor performance of the DMA compared with the DML parameterization in the cobia assessment contrasted with the results of Fisch et al. (2021). Compared with the DML, the DMA model produced larger information criterion values, a degree of uncertainty large enough (upper HPD interval ~3× the point estimate) to render management decisions difficult, a much greater degree of variability for likelihood minima with respect to unfished recruitment (more data conflict), and more unfavorable retrospective statistics. In fact, given that retrospective patterns for the LN and DMA models for cobia fall outside of the acceptable range proposed by Hurtado-Ferro et al. (2014) for longer-lived species ($-0.15 < \rho < 0.20$), it is likely these assessments would not pass a statistical review.

Independent of the composition likelihood chosen, the cobia assessments estimated a relatively unexploited population (depletion estimates were all ~1). However, these assessments differed greatly in scale, simply as a function of the likelihood chosen for fitting composition data. Although composition data are thought to mainly inform relative recruitment, mortality, and selectivity, it is acknowledged that they can indirectly inform estimation of the absolute scale of abundance (Maunder 2011; Maunder and Piner 2015), as fishing mortality in combination with known catch informs absolute abundance ($F \approx$ Catch/Biomass). The results of this study support that contention, as the choice of likelihood for fitting compositions led to different fishing mortality estimates in the cobia example, and thus different estimates of absolute abundance. This is further elucidated in the posterior profiles, where the age composition data were the most informative on the estimation of unfished recruitment.

For the Pacific hake assessment, the LN and the DML were much more similar in comparative performance (although not in output) than they were in the cobia assessment, where the DMA model again demonstrated concerning retrospective patterns and statistics. For the LN and DML, retrospective and hindcasting metrics were mixed, with the DML model outperforming the LN for the acoustic survey hindcasting (including the survey age-composition) and the opposite occurring for most fishery composition hindcasts. Both the LN model and the DMA model estimate an increasing abundance trend for the final 5 years of the time series, treating the terminal two data points for the survey index as underestimates (Fig. 4). Conversely, the data for the survey index suggest a decreasing trend in the index as does the fit from the DML model. It is likely for this reason the LN and DMA hindcasted models predicted the survey index poorly as indicated by the MASE scores where the terminal 3 data points were sequentially left out. Moreover, the LN and the DMA models allow for more variance in the fit to the survey index, indicated by their estimates of the additive survey SD (Table 2), although their SDNR values were still approximately 1 (and lower than DML). The LN and DMA models also allowed for more variance in fits to the composition data from the fishery, indicated by substantially lower estimates of ESS for the DMA (Fig. 2) and larger residual variance for the LN model (Fig. 4). The tradeoff lies in that they each estimate less variance in recruitment deviations than the DML model (Table S4) and suggest less data conflict in unfished recruitment (Table S3).

The maximum likelihood fitting process in integrated assessment models can be thought of as the partitioning of the total error among data sets (Francis 2017; and likelihood penalties if fit in a penalized likelihood context). Total error in this case includes both sampling error and process error or model misspecification (including unaccounted for random variation about biological processes, incorrect functional forms, or fixed parameters, etc.). In a Bayesian context, the error partitioning is expanded to also include deviations from prior distribution specifications, partitioning the total variance among all likelihood components specified. In the Pacific hake assessment, this included data sets and prior distributions, including those specified for time-varying processes and for estimated parameters. The differences in output between the Pacific hake model fit with the DML and that fit with the LN emerged in the last 10 years of the time series and can be explained with recourse to tradeoffs among likelihood components. It would appear that the LN and DMA models are foregoing closely fitting the last two data points in the abundance index (allowing for more variance in the index and composition data) in favor of less variability in recruitment deviations (more consistent with their prior specification) and less data conflict with respect to unfished recruitment. This same pattern was found in the suboptimal alternative DML solution, where the model was initialized at parameter estimates from the LN model posterior.

The choice between DML or LN for Pacific hake is consequential, as they output markedly different terminal estimates of depletion, spawning biomass, and exploitation rate, and given the stock is managed via a 40:10 harvest control rule (Grandin et al. 2020), the total allowable catch will substantially differ as a function of the composition likelihood chosen. If we were to follow the recommendations of Francis (2011, 2017), that abundance index data should be given primacy in integrated assessment model fitting, we might conclude that the DML performed better than the LN given that it fit the abundance index data more closely and provided more accurate predictions of the index in the hindcasted models. However, conversely, if the latter two data points of the survey are due to unaccounted for sampling error, or if the model is misspecified with respect to the survey index, an argument could be made that the LN model is accounting for that process or sampling error more appropriately (by somewhat ignoring the latter two points). A conservative interpretation would be that in the absence of significant evidence of additional sampling error or model misspecification with respect to the last two index points (e.g., the survey changed tactics or spatial coverage), it may be safer to move forward with the DML model. Although we note that the DML model exhibited the largest percentage of MCMC iterations indicating non-randomness in residuals for fits to the index, potentially indicating misspecification in the observation or system process (Carvalho et al. 2017). It may be that the sample size for the fishery composition data was simply not large enough for the LN to outperform the DML (as was found in Fisch et al. 2021), or through its parameterization of time-varying selectivity, the Pacific hake assessment is effectively minimizing process error. An alternative to choosing one optimal assessment would be an ensemble approach where the final output used to make management decisions is made up of estimates from each assessment likelihood configuration, potentially weighted by some of the model diagnostic/comparison criteria (Maunder et al. 2020).

An important condition to note is that, although most of the priors implemented for each composition likelihood were intended to be diffuse and uninformative, when converted to the positive scale (some were estimated on the log scale) and used in the different likelihoods, they actually provided different amounts of information. This is evidenced by our prior predictive distributions and the number of data points that were outside of the 95% HPD intervals for each composition likelihood. These suggested that for Pacific hake, the priors on the DML were the most informative, followed by the LN and the DMA, although the latter two were very similar. The justification for the priors placed on the Dirichlet-multinomial overdispersion parameters for the Pacific hake assessment was to avoid many MCMC samples for the survey overdispersion parameter occurring in the parameter space where $\theta/(1+\theta) \approx 1$, and the ESS converges on the true sample size (Grandin et al. 2020). The rationale for the prior on $\log(\theta)$, $N(0, 1.813^2)$, is that it provides an approximately uniform interval 0–1 for $\theta/(1+\theta\ )$ (the ESS scalar; Grandin et al. 2020). In reality, this prior is somewhat concave with a trough at $\theta/(1+\theta\ ) = 0.5$ and the greatest densities at 0.05 and 0.95 (Fig. S16). When this prior [$N(0, 1.813^2)$] was removed and uniform priors [$U(-10, 10)$] placed on the DML overdispersion parameters, the model suffered from convergence issues according to Geweke's diagnostic with ~20% of the parameters identified as not converged. Although the uniform prior is not necessarily wholly uninformative, given it is exponentiated and then utilized to determine ESS, it renders different amounts of prior information content for each formulation of the Dirichlet-multinomial. When the prior $N(0, 1.813^2)$ was removed solely for the fishery composition (replaced by the uniform prior, $U(-10, 10)$), the assessment model converged, and the results were nearly identical to those described in this study. This suggests that the ESS for the survey composition tends toward the input sample size in the absence of the informative prior and may simply indicate a well-designed survey (as the overdispersion parameter regarding the survey composition necessitates a prior to converge). An alternative in this case might be to use the actual number of fish aged as the survey input sample size. This presents a small computational concern with the Dirichlet-multinomial—if the composition data are effectively randomly sampled with replacement, the overdispersion parameter will tend to the upper bound (the Dirichlet-multinomial will effectively be collapsing to the multinomial), potentially causing convergence issues in the model. This was found to occur in Fisch et al. (2021) for the DML when there was little process error and at least close to random sampling, and in Cronin-Fine and Punt (2021) when composition data were generated using a multinomial distribution. Although this may simply represent a computational concern (and implies a good survey or close to random composition sampling), importantly the logistic-normal does not exhibit the same bound issues or require the specification of such an informative prior on its weighting parameters.

Comparisons between the DML and the DMA likelihoods in each assessment seem to suggest more optimal performance for the DML. This is in contrast to results from Fisch et al. (2021) who found marginally better performance for the DMA compared with the DML, in part, because of convergence issues similar to those discussed in the previous paragraph. In addition, it follows that there should be diminishing returns in decreased variance as the sample size approaches a census; thus, Fisch et al. (2021) recommended using the DMA. In this study, each assessment fit with the DMA seemed to estimate a steep saturating function with respect to ESS, causing the ESS to be relatively insensitive to large increases in the actual sample size. This affected both assessments as each time series of age-composition sample size had the characteristic of starting at small sample sizes for the first half of the time series and increasing linearly or even exponentially to conclude the time series (Fig. 2). Thus, the steep saturating functions estimated for each assessment under the DMA caused the ESS to increase little as the actual sample size doubled, tripled, even quintupled toward the end of each time series. This likely caused increased uncertainty in estimated quantities and larger degrees of variability and patterns in the retrospective analyses for the DMA assessments. Conversely, the DML parameterization, given its linear scaling of ESS with actual sample size, estimated an increase in the ESS at the end of the time series consistent with the actual sample size. This is likely an effect of the temporal pattern of sample size and not simply the scale of the overall sample size, as when we ran the Pacific hake assessment with the input fishery sample size multiplied by 5, the results were largely unchanged. We caution that the DML outperforming the DMA could be case specific, potentially a function of time series length, species life history, availability of auxiliary data, etc., although it may be prudent to proceed in operational assessments with the linear formulation of the Dirichlet-multinomial until these potential confounding factors are further elucidated. It seems the DMA parameterization under some circumstances (i.e., this study) causes ESS to saturate, or asymptote, at low values for the actual sample size early in the time series, resulting in a greater degree of estimated uncertainty and a poorer performing model.

There were some structural differences between the two stock assessment models that could have influenced our results, outside of simply the sample size for composition data. The Pacific hake assessment, in addition to having a larger sample size for fishery compositions, was privy to both more years of data and to fishery-independent data, given that it included a scientifically designed acoustic survey, where the cobia assessment only contained fishery-dependent information. The Pacific hake assessment also included both time-varying fishery selectivity and ageing error. Each of these factors likely decreased the probability of model misspecification and (or) the degree of process error in the Pacific hake assessment. Although neither likelihood replicated the correlations in composition residuals well (Figs. S12–S15), importantly, no structural pattern seemed evident in the residual correlations. In such instances, the Dirichlet-multinomial may be a more appropriate choice, as much of the rationale for using the logistic-normal arises from observing structural patterns in the composition residuals (Francis 2014, 2017), resulting from either process or sampling error and attempting to account for that error with a more flexible likelihood. Given what we might expect with composition likelihoods—that increased process error might favor the logistic-normal (Fisch et al. 2021)—the Pacific hake model, by minimizing process error, may have led to no obvious or characteristic structural patterns in residual correlations (often described in fisheries assessments as positive correlations between bins that are close together and negative between bins that are far apart; Francis 2011, 2014, 2017). The result was better performance of the less flexible model (being the DML in this case). In addition, the fact that the Dirichlet-multinomial greatly outperformed the logistic-normal in the cobia assessment (where the probability of misspecification or process error was increased) suggests that the sample size of the composition data may trump these other factors.

There also exists a major difference in the formulation of the two composition likelihoods, outside of their flexibility in residual correlation structure. The Dirichlet-multinomial is formulated to approximate the sampling error of discrete data, whereas the logistic-normal approximates that of continuous data. Each assessment did in fact utilize expanded compositions, meaning they were not measured in integers but some preprocessing was done prior to incorporation into each assessment, rendering them continuous data. In addition, each assessment model treated the composition data as though they were continuous, predicting proportions as a function of the total predicted catch, itself a continuous variable. The reality is that most fisheries assessments treat and model expected compositions as continuous, being a function of a variety of continuous variables/parameters (mortality, abundance, etc.). The process and observation models of fisheries assessments are simply approximations of reality, and the data-generating process is not captured fully by the Dirichlet-multinomial or the logistic-normal, and neither is it fully captured by the multinomial (or any other distribution). The DML outperforming the continuous LN despite the continuous composition data and model expectations suggests that the consideration of which likelihood to use based on whether the composition data are discrete or continuous is less important than other factors examined in this study, such as the sample size of the data or the amount of process error.

The sampling strategy used to collect composition data (Ono et al. 2015; Fisch and Bence 2020), sample size (He et al. 2016; Hulson et al. 2017), correlations and overdispersion (Pennington and Volstad 1994; McAlister and Ianelli 1997), and likelihood structure (Maunder 2011; Francis 2011, 2014; Thorson et al. 2017; Fisch et al. 2021) have long been relevant subjects in fisheries assessment literature. Recognizing that composition data do not comprise iid samples and that process error need be included in the residuals, stock assessment analysts have largely moved away from the multinomial likelihood for composition data weighted with the number of fish or trips sampled and are now starting to avoid iterative reweighting of composition data in favor of the self-weighting Dirichlet-multinomial likelihood, especially since its incorporation into software packages such as Stock Synthesis (Thorson et al. 2017) and the Beaufort Assessment model (BAM; Williams and Shertzer 2015). To our knowledge, few studies have compared the Dirichlet-multinomial to other likelihoods with estimable variances or compared different formulations of the Dirichlet-multinomial to one another (but see Fisch et al. 2021) as we have in this study. Although we note that in this study, we examined different assessment models fitted to real data and thus cannot claim to know that the estimated values for one model are more or less biased than another. Nonetheless, given what is known about model evaluation/diagnostic/comparison criteria (e.g., that retrospective patterns will tend to be worse for misspecified models, Hurtado-Ferro et al. 2014), we expect the more accurate model will tend to be one that performs best in the metrics evaluated in this study. Simulation studies within the field of stock assessment offer the distinct advantage of being able to compare true values of a system to those estimated; however, they are limited by the user-defined bounds of the derived system and could lead to overly optimistic results for simplistic simulators (Francis 2012). Frankly, there is no fully adequate substitute for real data, and fitting alternative models to empirical data can provide valuable information for comparison in search of optimal model formulations when information criterion measures are unavailable (Akselrud et al. 2017; Fisch et al. 2019; Fisch and Bence 2020). We find studies such as these offer an important analogue in fisheries assessment to explore replication of previous findings, particularly those from simulation studies.

In summary, results for each assessment indicate the saturating parameterization of the Dirichlet-multinomial is likely inferior to the linear formulation in at least these cases, although we encourage further simulation work on this topic with specific emphasis on composition sample size and on degree of model misspecification. Although all results were not as explicit as we would like, it does seem evident that the logistic-normal likelihood performs poorly compared with the linear formulation of the Dirichlet-multinomial when sample sizes are small for composition data from the fishery. The comparison at larger sample sizes is more robust, however, in the context of the Pacific hake assessment, without significant evidence of survey process misspecification or reason to disbelieve the last two data points of the acoustic index, on balance the DML approach seems preferable. It may be prudent to proceed in operational stock assessments with the linear formulation of the Dirichlet-multinomial; however, as our understanding of composition formulations continues to evolve, we encourage analysts to incorporate different composition likelihoods in the model fitting/development process in the interests of comparison.

Acknowledgements

We acknowledge the University of Florida Graduate School Preeminence Award and the NOAA National Marine Fisheries Service Sea Grant Population and Ecosystem Dynamics Fellowship for funding this research. We also acknowledge University of Florida Research Computing for providing computational resources that contributed to the research reported in this publication. We thank Erik Fitzpatrick for help with data regarding the cobia assessment, Nicholas Ducharme-Barth for a review on an earlier draft of this manuscript, and two anonymous reviewers for constructive comments.

Article information

History dates

Received: 14 February 2022
Accepted: 25 May 2022
Accepted manuscript online: 30 May 2022
Version of record online: 3 October 2022

Copyright

Author information

Author ORCIDs

Nicholas Fisch https://orcid.org/0000-0002-0505-4847

Author statements

The authors have no competing interests to declare. Data generated or analyzed during this study are available from the corresponding author upon reasonable request.

Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/cjfas-2022-0036.

References

Aitchison, J. 2003. The statistical analysis of compositional data. The Blackburn Press, Caldwell, New Jersey.

Akselrud, C.I.A., Punt, A.E., and Cronin-Fine, L. 2017. Exploring model structure uncertainty using a general stock assessment framework: the case of pacific cod in the Eastern Bering Sea. Fish. Res. **193**: 104–120. doi:10.1016/j.fishres.2017.03.016.

Albertsen, C.M., Nielsen, A., and Thygesen, U.H. 2017. Choosing the observational likelihood in state-space stock assessment models. Can. J. Fish. Aquat. Sci. **74**(5): 779–789. doi:10.1139/cjfas-2015-0532.

Beverton, R.J.H., and Holt, S.J. 1957. On the dynamics of exploited fish populations. Her Majesty's Stationery Office, UK. 533pp.

Breen, P.A., Kim, S.W., and Andrew, N.L. 2003. A length-based Bayesian stock assessment model for the New Zealand abalone *Haliotis iris*. Mar. Freshw. Res. **54**(5): 619–634. doi:10.1071/MF02174.

Carvalho, F., Punt, A.E., Chang, Y.J., Maunder, M.N., and Piner, K.R. 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. **192**: 28–40. doi:10.1016/j.fishres.2016.09.018.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L. Cardinale, M., et al. 2021. A cookbook for using model diagnostics in integrated stock assessments. Fish. Res. **240**: 105959. doi:10.1016/j.fishres.2021.105959.

Charnov, E.L., Gislason, H., and Pope, J.G. 2013. Evolutionary assembly rules for fish life histories. Fish Fish. **14**(2): 213–224. doi:10.1111/j.1467-2979.2012.00467.x.

Cronin-Fine, L., and Punt, A.E. 2021. Modeling time-varying selectivity in size-structured assessment models. Fish. Res. **239**: 105927. doi:10.1016/j.fishres.2021.105927.

Dichmont, C.M., Deng, R.A., Punt, A.E., Brodziak, J., Chang, Y.J. Cope, J.M., et al. 2016. A review of stock assessment packages in the United States. Fish. Res. **183**: 447–460. doi:10.1016/j.fishres.2016.07.001.

Fisch, N., Camp, E., Shertzer, K., and Ahrens, R. 2021. Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error. Fish. Res. **243**: 106069. doi:10.1016/j.fishres.2021.106069.

Fisch, N.C., and Bence, J.R. 2020. Data quality, data quantity, and its effect on an applied stock assessment of cisco in Thunder Bay, Ontario. N. Am. J. Fish. Manag. **40**(2): 368–382. doi:10.1002/nafm.10415.

Fisch, N.C., Bence, J.R., Myers, J.T., Berglund, E.K., and Yule, D.L. 2019. A comparison of age-and size-structured assessment models applied to a stock of Cisco in Thunder Bay, Ontario. Fish. Res. **209**: 86–100. doi:10.1016/j.fishres.2018.09.014.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A. Maunder, M.N., et al. 2012. AD model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim. Methods Softw. **27**(2): 233–249. doi:10.1080/10556788.2011.597854.

Francis, R.I.C.C 2012. The reliability of estimates of natural mortality from stock assessment models. Fish. Res. **119–120**: 133–134. doi:10.1016/j.fishres.2011.12.005

Francis, R.C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. **68**(6): 1124–1138. doi:10.1139/f2011-025.

Francis, R.C. 2014. Replacing the multinomial in stock assessment models: a first step. Fish. Res. **151**: 70–84. doi:10.1016/j.fishres.2013.12.015.

Francis, R.C. 2017. Revisiting data weighting in fisheries stock assessment models. Fish. Res. **192**: 5–15. doi:10.1016/j.fishres.2016.06.006.

Geweke, J. 1991. Evaluating the accuracy of sampling-based approaches to calculating posterior moments, *In*: Bayesian Statistics, vol. **4**. *Edited by* J.M. Bernado, J.O. Berger, A.P. Dawid and A.F.M. Smith. Clarendon Press, Oxford, UK.

Grandin, C.J., Johnson, K.F., Edwards, A.M., and Berger, A.M. 2020. Status of the Pacific Hake (whiting) Stock in U.S. and Canadian Waters in 2020. Prepared by the Joint Technical Committee of the U.S. And Canada Pacific Hake/Whiting Agreement. National Marine Fisheries Service and Fisheries and Oceans Canada. Available from https://media.fisheries.noaa.gov/dam-migration/hake-assessment-2020-final.pdf.

He, X., Field, J.C., Pearson, D.E., and Lefebvre, L.S. 2016. Age sample sizes and their effects on growth estimation and stock assessment outputs: three case studies from U.S. West Coast fisheries. Fish. Res. **180**: 92–102. doi:10.1016/j.fishres.2015.08.018.

Hulson, P.J.F., Hanselman, D.H., and Shotwell, S.K. 2017. Investigations into the distribution of sample sizes for determining age composition of multiple species. U.S. National Marine Fisheries Service. Fish. Bull. **115**: 326–342.

Hurtado-Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J. Johnson, K.F., et al. 2014. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES J. Mar. Sci. **72**(1): 99–110. doi:10.1093/icesjms/fsu198.

Hyndman, R.J., and Koehler, A.B. 2006. Another look at measures of forecast accuracy. Int. J. Forecast. **22**: 679–688. doi:10.1016/j.ijforecast.2006.03.001.

Ichinokawa, M., Okamura, H., and Takeuchi, Y. 2014. Data conflict caused by model mis-specification of selectivity in an integrated stock assessment model and its potential effects on stock status estimation. Fish. Res. **158**: 147–157. doi:10.1016/j.fishres.2014.02.003.

Kell, L. T., Kimoto, A., and Kitakado, T. 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fish. Res. **183**: 119–127. doi:10.1016/j.fishres.2016.05.017.

Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., and Cardinale, M., 2021. Validation of stock assessment methods: is it me or my model talking?. ICES J. Mar. Sci. **78**(6): 2244–2255. doi:10.1093/icesjms/fsab104. PMID: 33814897.

Lee, H.-H., Maunder, M.N., Piner, K.R., and Methot, R.D. 2011. Estimating natural mortality within a fisheries stock assessment model: an evaluation using simulation analysis based on twelve stock assessments. Fish. Res. **109**: 89–94. doi:10.1016/j.fishres.2011.01.021.

Lee, H.-H., Piner, K.R., Methot, R.D., and Maunder, M.N. 2014. Use of likelihood profiling over a global scaling parameter to structure the population dynamics model: an example using blue marlin in the Pacific Ocean. Fish. Res. **158**: 138–146. doi:10.1016/j.fishres.2013.12.017.

Maunder, M.N. 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: estimating the effective sample size. Fish. Res. **109**(2-3): 311–319. doi:10.1016/j.fishres.2011.02.018.

Maunder, M.N., and Piner, K.R. 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. **72**: 7–18. doi:10.1093/icesjms/fsu015.

Maunder, M.N., Xu, H., Lennert-Cody, C.E., Valero, J.L., Aires-da-Silva, A., and Minte-Vera, C. 2020. Implementing reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses (No. SAC-11- INF-F). Scientific Advisory Commitee, Inter-American Tropical Tuna Commission, San Diego.

McAllister, M.K., and Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling - importance resampling algorithm. Can. J. Fish. Aquat. Sci. **54**: 284–300. doi:10.1139/f96-285.

Methot, R.D., and Wetzel, C.R. 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. **142**: 86–99. doi:10.1016/j.fishres.2012.10.012.

Mohn, R. 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES J. Mar. Sci. **56**(4): 473–488. doi:10.1006/jmsc.1999.0481.

Ono, K., Licandeo, R., Muradian, M.L., Cunningham, C.J., Anderson, S.C. Hurtado-Ferro, F., et al. 2015. The importance of length and age composition data in statistical age-structured models for marine species. ICES J Mar Sci. **72**: 31–43. doi:10.1093/icesjms/fsu007.

Pennington, M., and Volstad, J.H. 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics, **50**: 725. doi:10.2307/2532786.

Punt, A.E., Haddon, M., and McGarvey, R. 2016. Estimating growth within size-structured fishery stock assessments: what is the state of the art and what does the future look like? Fish. Res. **180**: 147–160. doi:10.1016/j. fishres.2014.11.007.

Schnute, J.T., and Richards, L.J. 1995. The influence of error on population estimates from catch-age models. Can. J. Fish. Aquat. Sci. **52**: 2063–2077. doi:10.1139/f95-800.

SEDAR. 2020. SEDAR 58 – Atlantic Cobia Stock Assessment Report. SEDAR, North Charleston, SC.Available from http://sedarweb.org/sedar-58.

Taylor, N., Hicks, A.C., Taylor, I.G., Grandin, C., and Cox, S. 2014. Status of the Pacific Hake (whiting) stock in U.S. and Canadian waters in 2014 with a management strategy evaluation. International Joint Technical Committee for Pacific Hake. Available from http://www.pcouncil.org/groundfish/stock-assessments/by-species/pacific-whiting-hake.

Thorson, J.T., Johnson, K.F., Methot, R.D., and Taylor, I.G. 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fish. Res. **192**: 84–93. doi:10.1016/j.fishres.2016.06.005.

Truesdell, S.B., Bence, J.R., Syslo, J.M., and Ebener, M.P. 2017. Estimating multinomial effective sample size in catch-at-age and catch-at-size models. Fish. Res. **192**: 66–83. doi:10.1016/j.fishres.2016.11.003.

Vehtari, A., Gelman, A., and Gabry, J. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. **27**(5): 1413–1432. doi:10.1007/s11222-016-9696-4.

von Bertalanffy, L. 1957. Quantitative laws in metabolism and growth. Q. Rev. Biol. **32**(3): 217–231. doi:10.1086/401873. PMID: 13485376.

Wald, A., and Wolfowitz, J. 1940. On a test whether two samples are from the same population. Ann. Math. Stat. **11**: 147–162. http://www.jstor.org/stable/2235872. doi:10.1214/aoms/1177731909.

Wang, S.-P., Maunder, M.N., Piner, K.R., Aires-da-Silva, A., and Lee, H.-H. 2014. Evaluation of virgin recruitment profiling as a diagnostic for selectivity curve structure in integrated stock assessment models. Fish. Res. **158**: 158–164. doi:10.1016/j.fishres.2013.12.009.

Watanabe, S. 2010. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. **11**(116):3571–3594.

Williams, E.H., and Shertzer, K.W. 2015. Technical documentation of the Beaufort Assessment Model (BAM). NOAA Technical Memorandum NMFS-SEFSC-671.