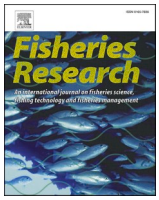
Fisheries Research 243 (2021) 106069

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error

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

[a] Fisheries and Aquatic Sciences, School of Forests, Fisheries, and Geomatics Sciences, Institute of Food and Agricultural Sciences, University of Florida, USA
[b] National Marine Fisheries Service, Pacific Islands Fisheries Science Center, 1845 Wasp Blvd., Building 176, Honolulu, HI 96818, USA
[c] National Marine Fisheries Service, Southeast Fisheries Science Center, 101 Pivers Island Road, Beaufort, NC 28516, USA

ARTICLE INFO

Handled by A.E. Punt

Keywords:
Stock assessment
Age-structured models
Composition data
Overdispersion
Data-weighting
Integrated models

ABSTRACT

Fisheries stock assessments have traditionally modeled age and size composition data using the multinomial likelihood, however the multinomial cannot appropriately account for the correlations and overdispersion that exist in the observed data or in the model residuals. Not accounting for these phenomena can affect assessment performance. Methods to remedy this have included down-weighting composition data within assessments either arbitrarily or by using iterative re-weighting algorithms, and by using alternative likelihoods to the multinomial that can be weighted within the assessment. Iteratively re-weighting composition data in stock assessments is inefficient and does not ultimately account for correlations in the residuals, and alternative likelihoods for composition data have not all been evaluated using stock assessment simulations. To evaluate the performance of alternative likelihoods in fitting composition data, we first developed a spatially explicit age-structured operating model to simulate correlation structure observed in real composition data. We then fit spatially aggregated assessment models to the simulated data and assessed the performance of various formulations of composition likelihoods (Multinomial, Robust Multinomial, Dirichlet, Dirichlet-multinomial, and Logistic-normal) in estimating stock dynamics and quantities of management interest. Results suggest that the degree of process error (combining both process variation and model misspecification) and the sample size of the composition data have a larger effect on the relative performance of different likelihoods than the degree of overdispersion and correlations in composition data. When the composition sample size was moderate to large and there existed at least a moderate amount of process error, the Logistic-normal likelihood performed best. When the sample size was small, or when process error was non-existent or negligible, the Dirichlet-multinomial likelihood performed best.

1. Introduction

Fisheries management is largely facilitated by stock assessments (Dichmont et al., 2016), providing managers with estimates of, and uncertainty regarding stock abundance and dynamics on which to base management decisions. The term "stock assessment" within fisheries generally refers to a suite of formal population models that have been developed for the purposes of estimating past and current stock dynamics, such as abundance and exploitation rates, using fishery and/or research survey data within a statistical modeling framework. A "statistical modeling framework" refers to the principle that these models compare expected values (predicted by the model) to observations (data) using statistical distributions. These distributions are referred to as likelihoods when used to estimate parameters conditioned on data, as they output a measure of how likely a specific set of model parameters are, given observations and assumed model structure.

Stock assessments vary in complexity, however contemporarily almost all involve at least one likelihood function (Maunder and Punt, 2013). When multiple data sources are used within a stock assessment (and thus multiple likelihoods), the model is considered an "integrated" assessment (Francis, 2017; Maunder and Punt, 2013). The main types of data integrated within assessments include annual removals (total catch from a fishery), abundance indices, and distributions of lengths and ages in the catches, referred to as composition data. Abundance indices and composition data often derive from both the sampling of fisheries and research surveys. Contemporarily, most stock assessments are integrated, and most "data-rich" stock assessments are based on age- or size-structure. In age- or size-structured assessments, the model tracks

* Corresponding author.
E-mail address: nfisch@ufl.edu (N. Fisch).

https://doi.org/10.1016/j.fishres.2021.106069
Received 15 March 2021; Received in revised form 5 July 2021; Accepted 6 July 2021
Available online 4 August 2021

the number of fish in age or size classes over time, thus greatly benefiting from the incorporation of age- or size-composition data from the harvest and/or survey within its total likelihood function. These data provide critical information to the assessment on relative year-class strength, mortality, selectivity, and in the case of size-composition, growth of organisms (Lee et al., 2011; Maunder and Piner, 2015; Punt et al., 2016). For integrated assessments, a biologist must make a decision, either explicit or implicit, regarding how much weight to assign to each data set in the total likelihood function (Francis, 2011). The focus of this study is on the treatment and weighting of composition data within stock assessment.

Composition data sources have historically been modeled within age- or size-structured assessments using the multinomial likelihood (Francis, 2014), however this approach poses some notable issues (Francis, 2014; Thorson et al., 2017). These issues largely stem from the premise that the multinomial distribution expects independent and identically distributed (iid) samples from a population. This assumption is almost always violated in fisheries, resulting in large correlations between age and length bins, as well as overdispersed composition data. Not accounting for correlative error structure and overdispersion in composition data can affect assessment point estimates (Francis, 2014) and precision (Maunder, 2011) of parameter values important to ecology and management. This occurs because, if a statistical distribution describing the error structure about an expected value does not conform to the distribution of the observed residuals, the maximum likelihood estimate is not guaranteed to converge on the true parameter values as the sample size approaches infinity (i.e., the consistency property of maximum likelihood estimation). Similarly, the estimates of the standard errors of parameters are not guaranteed to satisfy the asymptotic efficiency property of maximum likelihood estimation, which assures that the variance of unbiased estimators is at least as small as the inverse of the Fisher Information, as the second order derivatives of the Fisher Information Matrix (the Hessian) will be improper.

The correlations and overdispersion in composition data can result from the behaviors of fishers and target species, in addition to the way in which the data are sampled. Aquatic organisms often aggregate based on age or size, or associate with different habitats at different ages and sizes, and fishers tend not to distribute and fish randomly. A simple example is that of intrahaul correlation (Pennington and Volstad, 1994), which is based on the premise that fish in the same catch (e.g., from the same gear haul) are generally more similar to each other in length and age than fish from different catches. Thus, multiple samples taken from the same catch will not be independent, instead amounting to a cluster or batch sample, correlated spatially and temporally. The correlations observed in the data tend to be positive between bins that are close together, and negative between bins that are far apart (Francis, 2011; see Fig. 1. in Francis (2017) for an example). Having substantial correlations in composition data can cause overdispersion, or a larger variance observed in the data than would be expected under the multinomial. Conceptually, this results in the actual sample size not being indicative of the true information content contained in the data, and thus multinomial models will assume a smaller variance than is actually present in the data. Given that the multinomial distribution expects negative correlations between bins (the correlations are usually small, Francis, 2011), it cannot mimic the positive correlations present in fisheries composition data, nor can it account for overdispersion given the true sample size of the data.

Process error, which we broadly define as the mismatch between the estimation model and the true dynamics of the system (or operating model), can also create and may exacerbate this issue (Francis, 2011, 2017), by causing similar correlations in the residuals once the model is fit. Consider an example from Francis (2011): take an assessment model that specifies a constant fishery selectivity, however the true selectivity actually varies about the model predicted selectivity from year to year. In a year in which the fishery selects more larger fish and fewer smaller fish, the residuals will be structured such that the model is underestimating the catch composition of larger fish and overestimating the catch composition of smaller fish, or positive residuals for large fish and negative for small (see Fig. 3. in Francis, 2011). The pattern would be reversed in a year where the fishery selects fewer larger fish and more smaller fish. This can create similar correlations to those observed as a function of cluster or batch sampling (i.e., positive correlations between bins that are close together, and negative for bins that are far apart). Such process error can also lead to overdispersion, where model residuals have a greater variance than would be expected under multinomial (iid) sampling given the sample size.

Considering the issues presented above, a few notable studies have proposed some solutions (Francis, 2011, 2014, 2017; Maunder, 2011; Thorson et al., 2017). Collectively these studies focused on two potential solutions to the overdispersion-correlation issue in composition data. The first involves down-weighting composition data within an assessment using an effective sample size (ESS) that is less than the true sample size to reflect both the decreased information content in the data and error due to any model misspecification. This is often done using iterative methods, which involve fitting an assessment, adjusting ESS based on some formula (such as those in Table A1 of Francis (2011)), then re-fitting the assessment with the new ESS. This process is repeated until the expected variance of the residuals given an ESS matches (or is near) that of the observed residual variance (although the formulas differ in how they calculate the variances). This iterative approach suffers from several drawbacks. First, down-weighting does not account for correlations in the error structure; it solely accounts for the fact that the data contain less information than is expected with iid sampling and attempts to down weight the data to account for model misspecification. This addresses the overdispersion issue, however not the correlative issue. The second, and potentially more important drawback with this approach is that iterative re-weighting is impractical and inefficient. Stock assessments are rarely done without sensitivity analyses or retrospective analyses, each of which involve re-running assessments multiple times under different parameters or data. This problem is further compounded when an assessment is run in a computationally intensive Bayesian framework where a single run could take multiple days, or in an ensemble framework (Stewart and Martell, 2015; Jardim et al., 2020) that combines multiple (perhaps thousands of) model fits.

The second solution to the overdispersion-correlation problem that the aforementioned publications explore is utilizing a likelihood that is able to be weighted within a stock assessment. This approach also has the advantage of directly incorporating uncertainty in data weighting within the assessment (through uncertainty in the self-weighting parameter). Maunder (2011) tested several likelihood formulations and found the Dirichlet produced the most unbiased estimates of effective sample size. Francis (2014) described both the Logistic-normal and the Dirichlet as promising candidates for modeling composition data, with specific emphasis given to the Logistic-normal as it is theoretically able to incorporate residual correlation structure other than that of the multinomial within its likelihood formulation. Thorson et al. (2017) proposed utilizing a compound Dirichlet-multinomial likelihood as a self-weighting alternative to the multinomial in stock assessment. The Dirichlet-multinomial has since been increasingly incorporated into more assessments. Unfortunately, to our knowledge the Logistic-normal has received little simulation testing within stock assessment (as Francis (2014) did not fit the likelihoods within stock assessments), and the Dirichlet as well as the Dirichlet-multinomial are unable to model positive correlations in composition data given their correlation structure is the same as that of the multinomial.

One issue that few studies have addressed is the challenge of simulating realistic composition data that includes the correlations theorized to be a result of schooling behavior, fisher distribution, or other processes. Most simulation studies in the field of stock assessment simulate a generic population and generate age- or size-composition data using the multinomial distribution, thus assuming iid sampling and generating data lacking the overdispersion and correlations that are so often present

in real data (notable exceptions being Maunder (2011) and Hulson et al. (2012), who simulated schooling). These studies then commonly fit the data generated with the multinomial distribution using the multinomial or other likelihood (in the case of Thorson et al. (2017), the Dirichlet-multinomial). This unrealistic simulation of the data-generating process could be expected to produce overly optimistic results (Francis, 2012). In this interest, herein we explore how different composition likelihoods perform when fit to simulated data containing correlations and overdispersion. We achieve this by first building a spatially explicit operating model that is able to simulate the correlational structure often seen in composition data. We then fit spatially aggregated assessment models to data generated from the operating model to assess the performance of various likelihoods used for fitting compositional data. We specifically explore how the degree of observation error, process error, and the sample size of the composition data influence the relative performance of composition likelihoods. In what follows, when referencing process error we mean to denote the total mismatch between an estimation model and the true dynamics of a system, consisting of both white noise deviations about quantities (termed **process variation**), and systematic bias (termed **model misspecification**) which can come in the form of incorrect sub model structure (e.g., functional form of selectivity or stock-recruit relationship), fixing parameters at incorrect values, or not accounting for time variation in processes (e.g., directional variation in selectivity). In contrast when referencing observation error, we mean to denote the error in data if the sampling process were repeated (i.e., sampling error).

2. Methods

2.1. Overview

The spatially explicit operating model is based on the life history of red snapper (*Lutjanus campechanu*) in the US Gulf of Mexico (GOM), and many life history parameters were taken from the most recent stock assessment (SEDAR, 2018; parameter values provided in Table 1). Red snapper exhibit ontogenetic movement offshore (Grüss et al., 2017; Karnauskas et al., 2017), which makes them an ideal candidate species to develop a spatially explicit simulation model where fish segregate to different habitats by age. The red snapper stock assessment (SEDAR, 2018) spans the entire Gulf of Mexico contained within the US Exclusive Economic Zone. To make computation feasible, the spatial extent of the spatially explicit operating model was decreased to span the Florida Gulf of Mexico coastline from 10 to 500 meters in depth (Fig. 1). The model is divided into 0.1 decimal degree areas, resulting in 1559 individual spatial cells. The western spatial extent of the model was cut off at -87.5° longitude (roughly the border of Florida), while the southern extent was cut off at 24.5° latitude.

2.2. Operating model (OM)

The operating model is structured by age and space. The model runs for 150 years, which includes 50 years of initialization and 100 years of fishing. The ages modeled start at age 0 and include a plus group at age 20. The abundance at age within a spatial cell is calculated using

$$N_{y,a,c} = \begin{cases} R_y X_c^R & if \;\; a = 0 \\ \sum_{c'} X_{y,a,c',c}\left(N_{y-1,a-1,c'} e^{-\left(F_{y-1,a-1,c'}+M_{a-1}\right)}\right) & if \;\; 0 < a < 20+ \\ \sum_{c'} X_{y,a,c',c}\left(N_{y-1,a-1,c'} e^{-\left(F_{y-1,a-1,c'}+M_{a-1}\right)} + N_{y-1,a,c'} e^{-\left(F_{y-1,a,c'}+M_{a}\right)}\right) & if \;\; a = 20+ \end{cases}$$

Where $N_{y,a,c}$ is the abundance at age a in year y that is in spatial cell c (a cell in the grid), R_y is the global recruitment of age-0 fish in a given year, X_c^R is the proportion of recruits allocated to cell c, $F_{y-1,a-1,c'}$ is the instantaneous fishing mortality in a given cell for an age and year, M_a is the natural mortality for an age, and $X_{y,a,c',c}$ denotes the proportion of individuals of a given age that move from cell c' to c (age 0 s do not move). Movement is assumed to occur instantaneously at the start of the year. Global recruitment is calculated using the steepness parameterization of the Beverton-Holt stock recruitment function (Mace and Doonan, 1988)

$$R_y = \frac{4hR_0 SB_y}{SB_0(1-h) + SB_y(5h-1)}$$

Where h denotes steepness, SB_y denotes spawning biomass ($SB_y = \sum_a N_{y,a} * Fec_a$, where Fec_a represents a combined fecundity/maturity ogive), R_0 denotes unfished recruitment, and SB_0 unfished spawning stock biomass. Recruits are allocated to spatial cells based on their depth and substrate preference (see movement section) using X_c^R . Note that the recruitment spatial distribution is independent of year (and thus density).

2.2.1. Parameterizing movement

The movement matrix was calculated based on a probability function of cell attributes, including depth, substrate type, distance to a cell, and density of fish in a spatial cell. We based our movement modeling on preference-type movement from the spatially explicit stock assessment platform/program Spatial Population Model (Dunn et al., 2012). Movement of this type has been conducted for Ross Sea Antarctic Toothfish (Mormede et al., 2017), which exhibit a similar ontogenetic movement offshore to red snapper. To formulate movement, the preference for each spatial attribute type (i) is defined based on some function $f_i(\theta_i, A_{i,c})$, where θ_i are the parameters of a function for a given attribute type and $A_{i,c}$ is the value of the specific attribute for that type and spatial cell. Given four attribute types chosen in our model (depth, distance, density, substrate), the total preference of each cell is then the product of the individual preference functions

$$P_c = \prod_i f_i(\theta_i, A_{i,c})$$

The probability of moving from cell c' to any other cell c is then defined as the preference of moving to cell c divided by the sum preference of all the cells.

$$X_{y,a,c',c} = \frac{P_{y,a,c}}{\sum_c P_{y,a,c}}$$

Note the preferences in the above formula are year- and age-specific. The spatial distribution of recruits was calculated solely using the depth and substrate preference functions

$$X_c^R = \frac{P_c^R}{\sum_c P_c^R}$$

Variation was added in movement and the spatial distribution of recruits using the multinomial distribution. Further details on the

Table 1
Parameter table for the spatially explicit operating model (OM) and the sampling model (SM). OMP references an operating model parameter and SMP a sampling model parameter.

Parameter	Value	Source
Operating Model		
Ages (bins = 21)	0−20+	
Natural mortality – M_a	See Supplemental Table 2	SEDAR (2018)
Fecundity -Fec_a	See Supplemental Table 2	SEDAR (2018)
Growth		
OMP.1: Asymptotic length – L_∞	85.64cm	SEDAR (2018)
OMP.2: Brody growth coefficient – K	0.19	SEDAR (2018)
OMP.3: Age at size 0 – t_0	−0.39	SEDAR (2018)
Weight-Length		
OMP.4: a	1.7E-5	SEDAR (2018)
OMP.5: b	3	SEDAR (2018)
Recruitment		
OMP.6: Steepness – h	0.99	SEDAR (2018)
OMP.7: Proportion of recruits allocated east of Mississippi River	0.23	SEDAR (2018)
OMP.8: Proportion of cells in Florida (out of eastern GOM)	~0.90	Preliminary Calcs
OMP.9: Unfished recruitment (GOM wide)	1.63E8	SEDAR (2018)
OMP.10: Unfished recruitment (Florida) – R_0	33,277,765	Preliminary Calcs
OMP.11: Florida unfished spawning stock biomass (Eggs) – SB_0	9.25E14	Preliminary Calcs
OMP.12: SD recruitment (ln scale)	0.3	SEDAR (2018)
Depth Preference		
OMP.13: Asymptote of mean depth	94.1	Preliminary Calcs
OMP.14: Growth coefficient of mean depth	0.199	Preliminary Calcs
OMP.15: Age at mean depth 0	−2.59	Preliminary Calcs
OMP.16: Asymptote of variance in depth	1164.5	Preliminary Calcs
OMP.17: Growth coefficient of variance in depth	0.259	Preliminary Calcs
OMP.18: Age at variance in depth 0	−0.85	Preliminary Calcs
Distance Preference		
OMP.19: Negative exponential rate - λ_D	0.026	Preliminary Calcs
Density Dependent Preference		
OMP.20: Threshold density - D^*	75 % quantile of unfished density	Arbitrary
OMP.21: Decay rate - λ_{DD}	0.5	Arbitrary
Fishing		
OMP.22: Catchability (within cell) - cq	0.005	Arbitrary
OMP.23: Contact selectivity midpoint (logistic)	2	Arbitrary
OMP.24: Contact selectivity growth rate (logistic)	2	Arbitrary
Fishing Distance Preference		
OMP.25: Negative exponential rate - λ_{FD}	0.03	Arbitrary
Fishing Exploitable Biomass Preference		
OMP.26: Logistic growth rate - γ	0.00025	Arbitrary
OMP.27: Logistic midpoint	Median Exploitable Biomass in Current Year	Arbitrary
Total Effort		
OMP.28: Mean effort logistic growth rate	0.15	Arbitrary
OMP.29: Mean effort logistic midpoint	25	Arbitrary
OMP.30: Effort scalar	100,000	Arbitrary
OMP.31: Percentage of maximum effort	0.75	Arbitrary
OMP.32: CV of effort	0.25	Arbitrary
Sampling Model		
Fishery Data		
SMP.1: Percentage of total effort sampled	Treatment	Arbitrary
SMP.2: Percentage of catch sampled per sampled unit of effort	Treatment	Arbitrary
SMP.3: CV of fishery harvest	0.05	Arbitrary
SMP.4: CV of fishery CPUE	0.25	Arbitrary
Survey Data		
SMP.5: Number of cells sampled per year	50	Arbitrary
SMP.6: Survey catchability (within cell)	0.001	Arbitrary
SMP.7: Survey contact selectivity midpoint (logistic)	2	Arbitrary
SMP.8: Survey contact selectivity growth rate (logistic)	2	Arbitrary

formulation of each preference function can be found in Appendix B.

2.2.2. Fishing

Total effort in each year was simulated by first defining the average effort as a logistic increase from year 51 to the midpoint of the fishing time series (year 100), followed by constant effort fixed at a value that equaled 75 % of the asymptote of the logistic (year 100 effort level) for the final 50 years of the time series. This was done to simulate the early development of a fishery followed by a management regime starting in the second half of the fishing time series. Variability was then added to the average effort timeseries by drawing from a normal distribution with a specified CV (operating model parameter (OMP). 32, set at 0.25), to obtain the total effort in each year (Supplemental Fig. 1).

Effort was assumed to originate from a port at the coastline center point for the 23 coastal counties in Florida that boarder the Gulf of Mexico (Supplemental Table 1, Supplemental Fig. 2). Units of effort (from the total effort expended in the fishery that year) were allocated to each individual port based on relative population size within each coastal county (Supplemental Table 1) with variation added by drawing from a multinomial distribution.

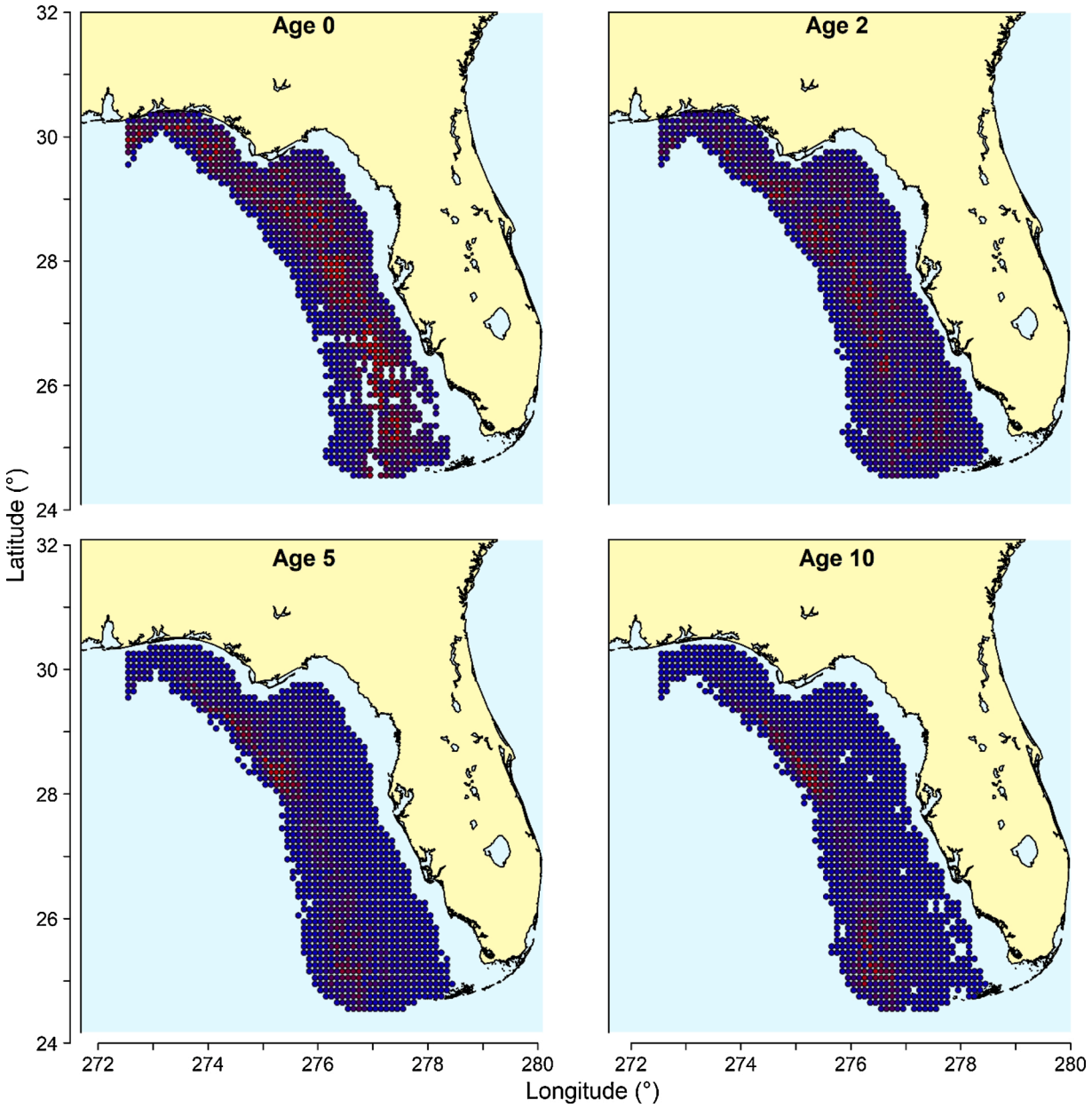


Fig. 1. Abundance at select ages (0, 2, 5, and 10) across spatial cells for 1 year (year 100) in the spatially explicit operating model and for one individual simulation iteration. Red indicates cells of higher abundance and blue lower.

The amount of fishing effort each spatial cell received in each year from a given port was modeled using a gravity model, which assumes the share of the total effort allocated to each spatial cell is proportional to the relative economic "attractiveness" of that cell, where attractiveness is proportional to the expected profitability of a cell based on resource availability and cost (Caddy, 1975; Walters and Bonfil, 1999). We assumed resource availability or profit as a function of exploitable biomass of a cell and that cost was a function of the distance to a cell. This allowed us to model the effort allocated to spatial cells from a port as positively associated with the exploitable biomass of cells and negatively associated with the distance of cells from the port. This is conceptually identical to preference movement in that fishers have preference probabilities for fishing in each cell depending on the exploitable biomass in that cell at the start of the year, and the distance of that cell from their port. The distance function was modeled as a negative exponential decay (with one parameter λ_{FD}, OMP.25). The exploitable biomass profit function was modeled as a logistic function where the midpoint was adjusted each year to be the median exploitable biomass across all cells, so as to account for shifting baselines. The probability a unit of effort was allocated to cell c from port p was then calculated as

$$P\left(E_{p,c}\right)_y = \frac{e^{-\lambda_{FD}*km_{p,c}}*1\Big/\left(1+e^{\left(-\gamma*\left(EB_{y,c}-median_c\left(\mathbf{EB}_y\right)\right)\right)}\right)}{\sum_c\left[e^{-\lambda_{FD}*km_{p,c}}*1\Big/\left(1+e^{\left(-\gamma*\left(EB_{y,c}-median_c\left(\mathbf{EB}_y\right)\right)\right)}\right)\right]}$$

Where $P\left(E_{p,c}\right)$ denotes the probability a unit of effort will go from port p to cell c, $km_{p,c}$ denotes the kilometers from a port to a cell, and $EB_{y,c}$ denotes the exploitable biomass in a cell ($\sum_a N_{y,a,c}*cs_a*w_a$, where cs_a and w_a refer to contact selectivity and the weight at age, respectfully). Variation in effort allocation is added using the multinomial distribution with probabilities calculated from the gravity model and sample size as the total amount of effort originating from a port in a given year. This allocation process resulted in spatial cells not receiving effort and would

complicate the calculation of true fishery catch-per-unit-effort (CPUE) in each year due to a non-random effort distribution and the need to impute CPUE for areas not fished (Walters, 2003). To avoid having to impute CPUE for spatial cells that experienced no effort, a baseline of 1 unit of effort was assumed for all cells during active fishing years (years 51–150). Given the focus of this analysis, this option was preferred to assuming known relative abundances in unfished cells in some other manner.

Fishing mortality for each cell was simulated using a fishery catchability parameter which defined the proportion of abundance in a cell caught per unit effort (thus it is the catchability within that cell, *cq*, OMP.22) and a logistic fishery selectivity which modeled the selectivity of fish at age within a cell (cs_a, OMPs 23–24). The logistic fishery selectivity simulates a contact selectivity, given within a spatially explicit model the need for a spatial availability component of selectivity is removed (as this is represented in effort and movement dynamics).

$$F_{y,a,c} = cq^{*}cs_a^{*}E_{y,c}$$

2.2.3. Process variation

Process variation, or random variation about biological and fishery processes, is included in recruitment using a lognormal distribution (log scale standard deviation set at 0.3; OMP. 12), total effort time series using a normal distribution (with CV; OMP. 32), and fish movement, the spatial distribution of recruits, the proportion of total effort allocated to each port in each year, and the probability of fishing spatial cells from a port using draws from multinomial distributions. Process variation, as noted earlier, is distinct from what we refer to as process error, defined as the combination of both process variation and model misspecification (similar to Francis, 2014), the latter meant to encompass aspects such as incorrect sub-model structure (such as incorrect functional forms of selectivity or stock-recruit), fixing parameters at incorrect values, or not accounting for time variation in processes.

2.3. Sampling model (SM)

The sampling model simulated collecting the following data from the true fishery and population quantities of the operating model.

2.3.1. Fishery age composition

The observed fishery catch age composition was simulated by sampling the catch at age from a subset of units of effort. This is analogous to sampling the catch of a subset of trips or from a subset of fishery operators.

The SM sampled a pre-specified percentage of the total effort in the fishery each year (sampling model parameter (SMP). 1). Sampling effort at each port was proportional to its effort allocation with variation added by drawing from a multinomial distribution. The spatial cells sampled from a given port were drawn with replacement with probabilities equal to the probability of fishing spatial cells from that port (in a given year). This is equivalent to sampling units of effort from each port. The catch at age in a *sampled* spatial cell was sampled using the multinomial distribution with sample size equal to a pre-specified proportion of the total number of fish caught for one unit of effort in the cell (SMP. 2). This was designed to allow for proportional sampling of the catch at age in sampled cell (to account for different abundances in cells). To obtain the total age composition for the year, the catch at age samples were aggregated across spatial cells and ports for a year. A small constant (1E-5) was added to suppress zeroes in the aggregated age composition data (and renormalized).

Overdispersion in the pooled age composition sample was assessed by repeating the sampling process 100 times (and only for 100 simulations of OM-SM combinations). This generated 100 age composition data sets for each year, to facilitate the calculation of variance within age bins across the replicates. This variation was compared to the sampling error that would be expected had the samples come from a multinomial distribution with the same sample size (and expected proportions from the true catch at age).

2.3.2. Fishery harvest

Observed harvest in each year of the time series was simulated by drawing from a normal distribution with mean as the true harvest (aggregated across space) and a CV of 5%.

2.3.3. Fishery catch-per-unit-effort (CPUE)

Fishery CPUE was simulated in each year by calculating the CPUE in each spatial cell (catch in that cell/effort expended in that cell) and summing these values across space. This was done as opposed to taking the total catch divided by the total effort in each year because of known biases in CPUE when effort is not equally distributed across space (Walters, 2003). To obtain observed fishery CPUE, observation error was added to the true CPUE by drawing from a normal distribution with a CV of 25 %.

2.3.4. Fishery independent survey age composition

Fishery independent surveys were simulated by randomly sampling spatial cells in the matrix at the start of the year. The final 60 years of the time series were subject to fishery independent surveys. Within a year, a pre-specified number of spatial cells were to be sampled (50, SMP.5). The cells chosen to be sampled were randomly drawn with replacement from all spatial cells. Once a cell was chosen, the vulnerable numbers at age in a cell $\left(cs_a^{*}N_{y,a,c}\right)$ were sampled using the multinomial with sample size equal to the total vulnerable numbers at age in a cell multiplied by a fishery independent catchability parameter defining the proportion of abundance in a cell caught per unit survey effort (SMP.6). The same contact selectivity was used for the survey as for the fishery (simple logistic, SMP. 7–8). The observed survey age composition for each year was then calculated by aggregating samples across cells within a year. Zeroes were suppressed using the same procedure described for fishery dependent compositions.

2.3.5. Fishery independent CPUE

Fishery independent survey CPUE was simulated by summing survey catches for each year and dividing by the total number of cells sampled in that year (or the total survey effort, i.e., 50).

2.4. Estimation model (EM)

The estimation models are age-structured assessment models that run for 100 years (fishing time series from OM) using ages 0-20+. The models are fit to 5 sources of data; (1) the fishery harvest, (2) the fishery CPUE, (3) the fishery age composition, (4) the fishery independent survey age composition, and (5) the fishery independent survey CPUE.

The models are initialized by estimating unfished recruitment (as a parameter) and projecting it forward using a known natural mortality ogive to calculate unfished abundance at age (Table 2). Recruitment is estimated each year using the Beverton-Holt stock recruitment function (Table 3; Eq. 1.2), with annual lognormal recruitment deviations about the median value penalized in the likelihood (using a prespecified recruitment standard deviation of 0.3 from SEDAR, 2018). Recruitment deviations are also estimated for cohorts that make up the initial abundance at age. Recruitment deviations were estimated on the log scale and summed to zero.

Fishery selectivity is estimated as either a two-parameter logistic function (Eq. 1.4b) or as a five-parameter double-logistic function (Eq. 1.4a). This fishery selectivity option was used as a treatment in the simulation design (see treatment section). The five-parameter double-logistic functional form was chosen as it closely resembles the true selectivity pattern in the OM (Fig. 2), although it is still misspecified due to the time varying nature of true selectivity in the OM (discussed below). Fishing mortality for each age in each year is calculated as the

Table 2
Descriptions of parameters/symbols for the estimation models, and whether they were estimated, fixed, etc. If they were estimated the bounds of the estimation are identified in parentheses.

Symbol	Description	Estimated or Fixed (bounds)
h	Steepness	Fixed – 0.99
R_0	Unfished recruitment	Estimated ln scale (10,25)
ε_y	Recruitment deviations (120 parameters)	Estimated ln scale (-10,10)
ϖ	Descending limb asymptote of double-logistic	Estimated (0,0.999)
κ	Growth rate of ascending limb of double-logistic	Estimated (-2,5)
τ	Growth rate of descending limb of double-logistic	Estimated (-2,5)
θ_1	Double-logistic ascending limb midpoint	Estimated (0,20)
θ_2	Double-logistic descending limb midpoint	Estimated (0,20)
k	Fishery logistic selectivity growth rate	Estimated (-2,5)
x_0	Fishery logistic selectivity midpoint	Estimated (0,20)
v	Survey logistic selectivity growth rate	Estimated (-2,5)
y_0	Survey logistic selectivity midpoint	Estimated (0,20)
q	Fishery catchability	Estimated ln scale (-20, 1)
z	Survey catchability	Estimated ln scale (-20, 1)
CV_H	Fishery harvest CV	Fixed (0.05)
σ_R	Recruitment deviations SD	Fixed (0.3)
CV_I	Fishery CPUE CV	Fixed (0.25)
CV_Q	Survey index CV	Estimated ln scale (-5,2)
f_y	Fishing intensity (100 parameters)	Estimated ln scale (-20,0)
α	Weighting parameter for the Dirichlet (2 parameters; 1 fishery, 1 survey)	Estimated ln scale (-10,20)
θ	Weighting parameter for the DML (2 parameters; 1 fishery, 1 survey)	Estimated ln scale (-10,20)
β	Weighting parameter for the DMA (2 parameters; 1 fishery, 1 survey)	Estimated ln scale (-10,20)
σ_{AR1}	Logistic-normal AR1 SD (2 parameters; 1 fishery, 1 survey)	Estimated ln scale (-5,5)
φ	Logistic-normal AR1 Phi (2 parameters; 1 fishery, 1 survey)	Estimated (-1,1)
σ_{AR2}	Logistic-normal AR2 SD (2 parameters; 1 fishery, 1 survey)	Estimated ln scale (-5,5)
φ_1	Logistic-normal AR2 Phi1 (2 parameters; 1 fishery, 1 survey)	Estimated (-2,2)
ω	Logistic-normal AR2 Omega (2 parameters; 1 fishery, 1 survey)	Estimated logit scale (-10,10)
φ_2	Logistic-normal AR2 Phi2 (2 parameters; 1 fishery, 1 survey)	$\varphi_2 = -1 + (2 - \lvert\varphi_1\rvert)\omega$
σ_{ARMA}	Logistic-normal ARMA SD (2 parameters; 1 fishery, 1 survey)	Estimated ln scale (-5,5)
φ_{ARMA}	Logistic-normal ARMA Phi (2 parameters; 1 fishery, 1 survey)	Estimated (-1,1)
ψ	Logistic-normal ARMA Psi (2 parameters; 1 fishery, 1 survey)	Estimated (-100,100)
M_a	Natural mortality at age	Fixed – OM values (Supp. Table 2)
SB_0	Unfished spawning biomass	Function of R_0, M_a, Fec_a and ε_y
Fec_a	Fecundity at age	Fixed – OM values (Supp. Table 2)
w_a	Weight at age	Fixed – OM values (Supp. Table 2)
$\overline{N}_{a,y}$	Mean numbers at age over a given year	Function of $N_{a,y}$ and $Z_{a,y}$ (Eq. 2.4)
H_y	Observed harvest in a given year	Data
I_y	Observed fishery CPUE in a given year	Data
Q_y	Observed survey CPUE in a given year	Data
N_y	Sample size, or the number of fish aged in a year. Can also reference the effective sample size.	Data
$P_{a,y}$	Observed proportion in a fishery composition data set for a given age and year	Data
$G_{a,y}$	Observed proportion in a survey composition data set for a given age and year	Data
Nb	Number of bins in a composition data set	Fixed - 21

*Further symbols used in the LN likelihoods are given in Appendix A.

product of fishery selectivity and fishing intensity (fully selected fishing mortality, Eq. 1.5). Fishing intensity in each year is estimated as a log scale vector.

Much of the observation model is quite standard in stock assessment and its equations are presented in Table 3. Of note is that predicted fishery CPUE is calculated by multiplying a fishery catchability parameter, which is estimated, by the mean exploitable biomass over the year (sum of the mean numbers at age over the year, $\overline{N}_{a,y}$ multiplied by a weight at age ogive and fishery selectivity, Eq. 2.4). Fishery catchability is solely used in the observation model and not for the calculation of fishing mortality.

2.4.1. Likelihoods

The fishery harvest, fishery CPUE, and survey CPUE were fit with normal likelihoods using CVs (Table 3, Eqs. 3.1–3.3). The CV for survey CPUE was estimated while the CVs for harvest and fishery CPUE were pre specified (fixed at their OM values). Recruitment deviations were penalized using a lognormal likelihood (Eq. 3.4) with a pre-specified standard deviation on the log scale (fixed at OM value of 0.3).

2.4.2. Composition likelihoods

A total of ten likelihoods were tested as treatments for fitting composition data within this study. These included; (1) the multinomial weighted with the true sample size in each year (MN, Eq. 4.1), (2) the multinomial with effective sample sizes calculated using Francis (2011) weighting method TA1.8 (MNFr), (3) the robust multinomial weighted with true sample sizes (MNR, Eq. 4.2), (4) the robust multinomial with ESSs calculated using Francis (2011) weighting method TA1.8 (MNRFr), (5) the Dirichlet (D, Eq. 4.3), (6) the Dirichlet-multinomial with linearly parameterized ESS (DML, Eq. 4.4), (7) the Dirichlet-multinomial with saturating ESS (DMA, Eq. 4.5), and (8–10) three parameterizations of the Logistic-normal likelihood (Eq. 4.6) including an AR(1), an AR(2), and an autoregressive moving average (ARMA) parameterization of the variance-covariance matrix (LNAR1, LNAR2, LNARMA). Each of these likelihoods was included because they showed both theoretical and applied support from the literature (Francis, 2011, 2014, 2017; Hulson et al., 2011; Maunder, 2011; Thorson et al., 2017). We placed a specific emphasis on 3 desired qualities; whether their weighting is estimable within an assessment (i.e., whether they are able to account for over-dispersion within the assessment), whether they allow for positive correlations in composition data, and the number of extra parameters to estimate (Table 4). The DMA, DML, D, and LN likelihoods all estimate weighting within the assessment, the MNFr and the MNRFr estimate it iteratively, and the MN and MNR fix the weighting at the true sample size. The MN, MNFr, DMA, DML, and D all have similar variance-covariance structure only allowing for negative correlations between bins which are usually small. The LN allows for different correlation structure according to the parameterization of the covariance (AR(1), AR(2), or ARMA), the degree of which is estimated. The MNR model uses the normal distribution with a multinomial variance but does not explicitly model correlations between bins. However, some correlation may arise by nature of compositions given the proportion for one age will influence the proportion for another age. The likelihoods are described in further detail within Appendix A. Each likelihood was used for both fishery and survey composition data and was not crossed

Table 3
Estimation model, or spatially aggregated age-structured assessment model equations. Further description on the formulation of the Logistic-normal can be found in Appendix A.

Quantity		Equation
Process Model		
1.1	Abundance at age	$N_{a,y} = \begin{cases} R_y & \text{if } a = 0 \\ N_{a-1,y-1}e^{-(F_{a-1,y-1}+M_{a-1})} & \text{if } 1 \le a < 20+ \\ N_{a-1,y-1}e^{-(F_{a-1,y-1}+M_{a-1})} + N_{a,y-1}e^{-(F_{a,y-1}+M_a)} & \text{if } a = 20+ \end{cases}$
1.2	Recruitment	$R_y = \frac{4hR_0SB_y}{SB_0(1-h)+SB_y(5h-1)}e^{\varepsilon_y} \quad \varepsilon_y \sim N(0,\sigma_R^2)$
1.3	Spawning Biomass	$SB_y = \sum_a N_{a,y}Fec_a$
1.4a	Fishery Selectivity (Double Logistic)	$s_a = \left[\left(1 - \frac{\varpi}{(1+e^{(-\tau(a-\theta_2))})}\right)\Big/\left(1+e^{(-\kappa(a-\theta_1))}\right)\right]\Big/\max(s_a)$
1.4b	Fishery Selectivity (Logistic)	$s_a = \frac{1}{(1+e^{(-k(a-x_0))})}\Big/\max(s_a)$
1.5	Fishing Mortality	$F_{a,y} = s_a f_y$
Observation Model		
2.1	Predicted Catch-at-age	$\widehat{C}_{a,y} = \frac{F_{a,y}}{F_{a,y}+M_a}N_{a,y}\left(1-e^{-(F_{a,y}+M_a)}\right)$
2.2	Predicted Fishery Harvest	$\widehat{H}_y = \sum_a \widehat{C}_{a,y}w_a$
2.3	Predicted Composition	$\widehat{P}_{a,y} = \frac{\widehat{C}_{a,y}}{\sum_a \widehat{C}_{a,y}}$
2.4	Predicted Fishery CPUE	$\widehat{I}_y = q^*\sum_a \overline{N}_{a,y}{}^*w_a{}^*s_a$ where $\overline{N}_{a,y} = \frac{N_{a,y}(1-e^{-(F_{a,y}+M_a)})}{F_{a,y}+M_a}$
2.5	Survey Selectivity	$g_a = \frac{1}{\left(1+e^{(-\nu(a-y_0))}\right)}$
2.6	Predicted Survey CPUE	$\widehat{Q}_y = z^*\sum_a g_a{}^*N_{a,y}$
2.7	Predicted Survey Composition	$\widehat{G}_{a,y} = \frac{g_a{}^*N_{a,y}}{\sum_a g_a{}^*N_{a,y}}$
Negative Log Likelihoods (excluding compositions)		
3.1	Fishery Harvest	$\sum_y \ln\left(CV_H{}^*\widehat{H}_y\right) + 0.5\left(\frac{H_y-\widehat{H}_y}{CV_H{}^*\widehat{H}_y}\right)^2$
3.2	Fishery CPUE	$\sum_y \ln\left(CV_I{}^*\widehat{I}_y\right) + 0.5\left(\frac{I_y-\widehat{I}_y}{CV_I{}^*\widehat{I}_y}\right)^2$
3.3	Fishery-Independent Survey CPUE	$\sum_y \ln\left(CV_Q{}^*\widehat{Q}_y\right) + 0.5\left(\frac{Q_y-\widehat{Q}_y}{CV_Q{}^*\widehat{Q}_y}\right)^2$
3.4	Recruitment Deviations	$\sum_y \ln(\sigma_R) + 0.5\left(\frac{\varepsilon_y}{\sigma_R}\right)^2$
Composition Negative Log Likelihoods		
4.1	Multinomial (MN)	$-\sum_y N_y \sum_a P_{a,y}\ln\left(\widehat{P}_{a,y}\right)$
4.2	Robust Multinomial (MNR)	$\sum_y\sum_a 0.5\log\left((1-P_{a,y})P_{a,y}+\frac{0.1}{Nb}\right) - \log\left(\exp\left(\frac{-\left(P_{a,y}-\widehat{P}_{a,y}\right)^2}{2\left((1-P_{a,y})P_{a,y}+\frac{0.1}{Nb}\right)/N_y}\right)+0.01\right)$
4.3	Dirichlet (D)	$\sum_y\left[-\log(\Gamma(\alpha_y)) + \sum_a \log\left(\Gamma\left(\alpha_y\widehat{P}_{a,y}\right)\right) - \left(\alpha_y\widehat{P}_{a,y}-1\right)\log\left(P_{a,y}\right)\right]$
4.4	Dirichlet-multinomial Linear (DML)	$-\sum_y\left[\begin{array}{c}\log(\Gamma(N_y+1)) - \sum_a\left(\log\left(\Gamma(N_y{}^*P_{a,y}+1)\right)\right) + \log(\Gamma(\theta N_y)) \\ -\log(\Gamma(N_y+\theta N_y)) + \sum_a\left(\log\left(\Gamma\left(N_y{}^*P_{a,y}+\theta N_y{}^*\widehat{P}_{a,y}\right)\right) - \log\left(\Gamma\left(\theta N_y{}^*\widehat{P}_{a,y}\right)\right)\right)\end{array}\right]$
4.5	Dirichlet-multinomial Saturating (DMA)	$-\sum_y\left[\begin{array}{c}\log(\Gamma(N_y+1)) - \sum_a\left(\log\left(\Gamma(N_y{}^*P_{a,y}+1)\right)\right) + \log(\Gamma(\beta)) \\ -\log(\Gamma(N_y+\beta)) + \sum_a\left(\log\left(\Gamma\left(N_y{}^*P_{a,y}+\beta{}^*\widehat{P}_{a,y}\right)\right) - \log\left(\Gamma\left(\beta{}^*\widehat{P}_{a,y}\right)\right)\right)\end{array}\right]$
4.6	Logistic-Normal (LN)	$\sum_y\left[0.5(Nb-1)^*\log(2\pi) + \sum_a\left[\log\left(P_{a,y}\right)\right] + 0.5^*\log\left(\lvert\mathbf{V}_y\rvert\right) + (Nb-1)^*\log(W_y) + \frac{\left(\mathbf{w}_y^T\mathbf{V}_y^{-1}\mathbf{w}_y\right)}{2W_y^2}\right]$

*Note that composition likelihoods are also used for the survey compositions, thus $\widehat{G}_{a,y}$ and $G_{a,y}$ can be substituted for $\widehat{P}_{a,y}$ and $P_{a,y}$.

with any other likelihood (i.e., if the fishery composition was fit with the Dirichlet, the survey composition was also fit with the Dirichlet).

2.5. Additional treatments

In addition to the choice of likelihood for composition data within the EM, three further treatments were included in the simulation design; the sample size of the fishery composition data, the degree of overdispersion in the fishery composition data, and the degree of process error between the estimation model and the operating model. These treatments were created by changing the OM fishing scenario, the EM form of fishery selectivity, and the SM fishery composition sampling formulation (Table 5). Each treatment is described in detail below.

2.5.1. OM fishing specification

We explored performance of estimation models related to two specific formulations of the spatially explicit OM with respect to fishing; a gravity model fishing formulation, as described in section 2.2.2, and an OM where fishing occurs randomly over space; however, the sampling still occurs as previously described. When fishing is random with respect to space, the fishery selectivity of a spatially explicit OM collapses to what is specified as the contact selectivity. Thus, an estimation model parameterized with a logistic fishery selectivity function fit to the OM that employs random fishing will be nearly correctly specified in process; however, it should be misspecified in the observation error of the composition data sets (as the composition is still made up of cluster samples). We use the term "nearly correctly specified" because the spatially aggregated EM will be unable to account for process variation in both fish movement and effort distribution from the OM, which will cause some variation in fishery selectivity (not accounted for in the functional form, Fig. 2). Conversely, estimation models fit to the gravity model OM will be misspecified in process in terms of fishery selectivity and in observation error. We structured the simulation design in this fashion such that we could observe how the performance of the different composition likelihoods compared when there was almost exclusively observation error compared to when there was both process and observation error (and different degrees of process error). An additional version of the random fishing OM was run where composition data was sampled from the true catch at age and the survey catch at age using the multinomial (thus resulting in iid samples), and process variation in fish movement and effort distribution were turned off. An EM with logistic fishery selectivity which uses the multinomial likelihood for composition data should be correctly specified when fit to this random fishing operating model with iid catch and survey composition sampling (i.e., a data-generating model).

2.5.2. EM fishery selectivity

We varied the functional form of fishery selectivity for estimation models fit to the gravity model OM, from double-logistic to logistic. This was done to examine the effect of different degrees of process error on the performance of the likelihoods. The true selectivity pattern which emerged from the OM (gravity model OM) was time-varying, due to the nature of a spatially explicit model where the spatial availability component of selectivity becomes an emergent property of the model caused by fish movement and the dynamic fisher effort distribution. We chose a 5-parameter double-logistic functional form as this seemed a reasonable approximation of the true selectivity pattern which emerged from the OM (Fig. 2). A model with a double logistic fishery selectivity fit to the gravity model OM will be less misspecified than an estimation model using a simple logistic selectivity. Additionally, as noted previously, an estimation model with logistic fishery selectivity will be nearly correctly specified in its process when fit to the random fishing OM (and correctly specified when fit to the random fishing OM with iid sampling). Hereafter we refer to EMs with double logistic fishery selectivity fit to the gravity model OM as the **baseline** scenario. We call this scenario the baseline as we think it most closely approximates reality, that in truth selectivity will vary year to year due to unmodeled processes (in this case fish movement and effort distribution), and these processes can be time-varying (Sampson and Scott, 2011, 2012). EMs with logistic fishery selectivity fit to the gravity model OM are referred to as the **Max PE** (maximum process error) scenario. Finally, we refer to EMs with logistic fishery selectivity fit to the random fishing batch OM as the **Min PE** (minimum process error) scenario.

2.5.3. SM fishery composition sampling

Within the OM scenarios, using the sampling model we explored how varying the fishery composition sample size and the degree of overdispersion affected the performance of the estimation models. To do this we varied the percentage of total units of effort sampled each year (SMP. 1) and the percentage of the catch sampled per unit of effort sampled (SMP. 2). We were specifically interested in how, independent of sample size, the performance of the composition likelihoods changed with different degrees of overdispersion in the data, and how performance changed as the sample size of the composition increased or decreased. We explored three levels of composition sample size (hereafter called small, medium, and large composition sample sizes), and four levels of overdispersion within each (Supplemental Fig. 5). These treatments are identified by the percentage of total effort sampled in the fishery, followed by the percentage of the catch sampled per unit of effort sampled. For example, sampling 2% of the total units of effort each year and 10% of the catch of sampled units had a very similar fishery composition sample size to 1% of the total units of effort sampled and 20% of the catch of sampled units sampled. We refer to these treatments as "2% 10 %" and "1% 20 %", respectively. Although they had similar sample sizes, the overdispersion was more severe for the "1% 20 %" sampling scenario. Overdispersion was measured by examining the variance in proportions of age bins when the sampling process was repeated compared to the expected sampling variance from the multinomial with a similar sample size (Fig. 3). The small sample sizes included SM treatments of "0.25 % 10 %", "0.125 % 20 %", "0.0625 % 40 %", and "0.03125 % 80 %". The medium sample sizes included SM treatments of "2% 10%", "1% 20 %", "0.5 % 40 %", and "0.25 % 80 %". The large sample sizes included SM treatments of "20% 10 %", "10 % 20 %", "5% 40 %", and "2.5 % 80 %". The sample size treatments aged approximately 200 fish/year, 2000 fish/year, and 20,000 fish/year for the small, medium, and large sample size levels respectively (for latter 75 % of the time series, Supplemental Fig. 5). These sample size levels were on average ~0.02 %, ~0.2 %, and ~2% of the total numbers of fish caught in each year (Supplemental Fig. 6). Holding sample size constant, treatments with fewer total units of effort sampled and larger batch sizes (in terms of % of catch sampled per unit of sampled effort) had more severe overdispersion (Fig. 3). We compared within each sample size level as overdispersion increased, and across these three levels as the total age composition sample size increased. The full treatment design can be found in Table 5.

2.5.4. Correlations and overdispersion observed from sampling models

As noted above, overdispersion across age bins increased as batch sizes increased within each sampling model. The degrees of overdispersion remained similar as sample size of the composition increased (across batch sizes; Fig. 3). However, degrees of increase in overdispersion with batch sampling were not uniform across age bins or OM scenarios. For the gravity model OM, younger ages experienced greater degrees of overdispersion, which decreased to approximately age 4 for each SM, then increased once again to approximately age 8, and decreased until the plus group where it once again increased. For the random fishing OM, a similar pattern emerged where the young age bins experienced greater overdispersion (however the degree of overdispersion was less than that of the gravity model OM). Although in the random fishing OM, the plus group did not see an increase in overdispersion as batch size increased (as it did in the gravity model OM).

With respect to correlations, the observation error residuals

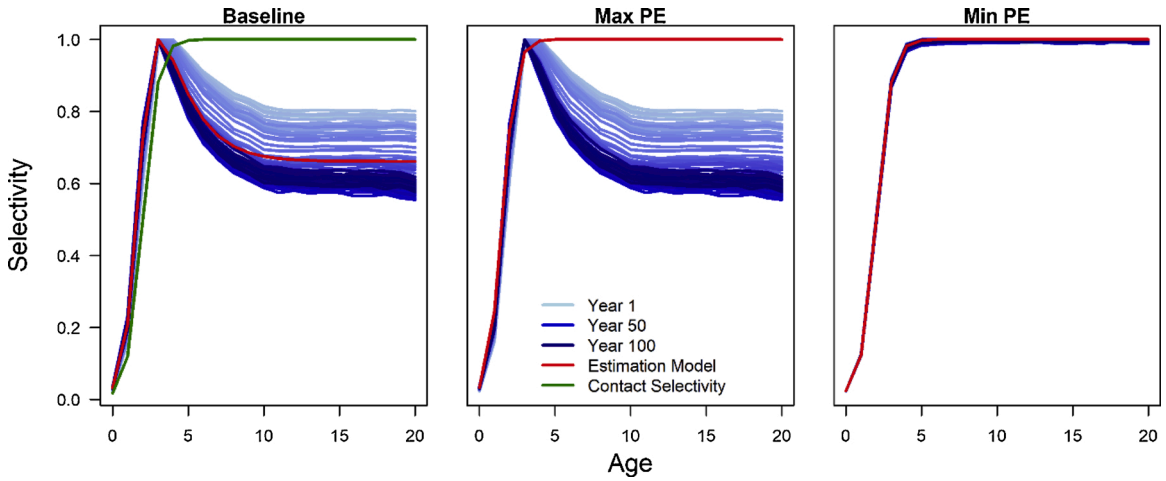


Fig. 2. Depiction of process error regarding fishery selectivity between the operating model and the estimation models. The true fishery selectivity in each year for the OMs is depicted in shades of blue (gravity model OM for the first two panels, and the random fishing OM for the third), which progressively get darker with the timeseries. The time-invariant fishery selectivity for the estimation models for each scenario is depicted in red. The contact selectivity for the spatially explicit operating model is specified on the first panel in green. This figure depicts the true fishery selectivity in each year for one individual OM simulation, and the estimated fishery selectivity using the MNFr likelihood and the 2% 10 % sampling model.

Table 4
Likelihoods used as treatments for fitting composition data within this study in addition to select qualities of each. The ability to account for correlations refers to correlation structure other than that of the multinomial (negative correlation which are generally small).

Likelihood	Acronym	Estimable Weighting Parameter?	Able to Account for Correlations?	Extra Parameters
Multinomial	MN	No	No	0
Multinomial (Iteratively weighted)	MNFr	Iterative	No	0
Robust multinomial	MNR	No	No	0
Robust multinomial (Iteratively weighted)	MNRFr	Iterative	No	0
Dirichlet	D	Yes	No	1
Dirichlet-multinomial Linear	DML	Yes	No	1
Dirichlet-multinomial Saturating	DMA	Yes	No	1
Logistic-normal AR(1)	LNAR1	Yes	Yes	2
Logistic-normal AR(2)	LNAR2	Yes	Yes	3
Logistic-normal ARMA	LNARMA	Yes	Yes	3

(observed age composition – true age composition) correlated across years show a similar pattern to that described in the literature, with positive correlations between bins that are close together and negative correlations between bins that are far apart. The strength of these correlations increased both as overdispersion increased (holding sample size constant), and as the sample size of the composition increased (for both OMs, Supplemental Figs. 7–8). Most of the correlations were stronger for the gravity model OM than for the random fishing OM (Supplemental Fig. 9).

Table 5
Simulation experiment design. DGM = data generating estimation model, correctly specified in both process and observation (when the multinomial is used). Min PE scenario is nearly correctly specified in process, misspecified in observation. The Max PE scenario is misspecified in both process and observation. The Baseline scenario is misspecified in both process and observation however less so than the Max PE scenario. For the random fishing model with iid composition sampling, sample size was equivalent to 0.25 % 10 %, 2% 10 %, and 10 % 20 % random fishing OM-SM scenarios. The term "X10" refers to the ten likelihoods tested.

Scenario	Operating Model	Sampling Model	Estimation Model	
	OM Fishing Type	SM Fishery Composition Sampling	EM Fishery Selectivity	EM Composition Likelihood
DGM	Random Fishing[a]	iid sampling	Logistic	X10
Min PE	Random Fishing	0.25 % 10 % | 0.125 % 20 % | 0.0625 % 40 % | 0.03125 % 80 %	Logistic	X10
	Random Fishing	2% 10 % | 1% 20 % | 0.5 % 40 % | 0.25 % 80 %	Logistic	X10
	Random Fishing	20 % 10 % | 10 % 20 % | 5% 40 % | 2.5 % 80 %	Logistic	X10
Baseline	Gravity Model	0.25 % 10 % | 0.125 % 20 % | 0.0625 % 40 % | 0.03125 % 80 %	Double-Logistic	X10
	Gravity Model	2% 10 % | 1% 20 % | 0.5 % 40 % | 0.25 % 80 %	Double-Logistic	X10
	Gravity Model	20 % 10 % | 10 % 20 % | 5% 40 % | 2.5 % 80 %	Double-Logistic	X10
Max PE	Gravity Model	0.25 % 10 % | 0.125 % 20 % | 0.0625 % 40 % | 0.03125 % 80 %	Logistic	X10
	Gravity Model	2% 10 % | 1% 20 % | 0.5 % 40 % | 0.25 % 80 %	Logistic	X10
	Gravity Model	20 % 10 % | 10 % 20 % | 5% 40 % | 2.5 % 80 %	Logistic	X10

[a] Process variation in fish movement and fisher effort distribution were turned off for this version of the random fishing OM.

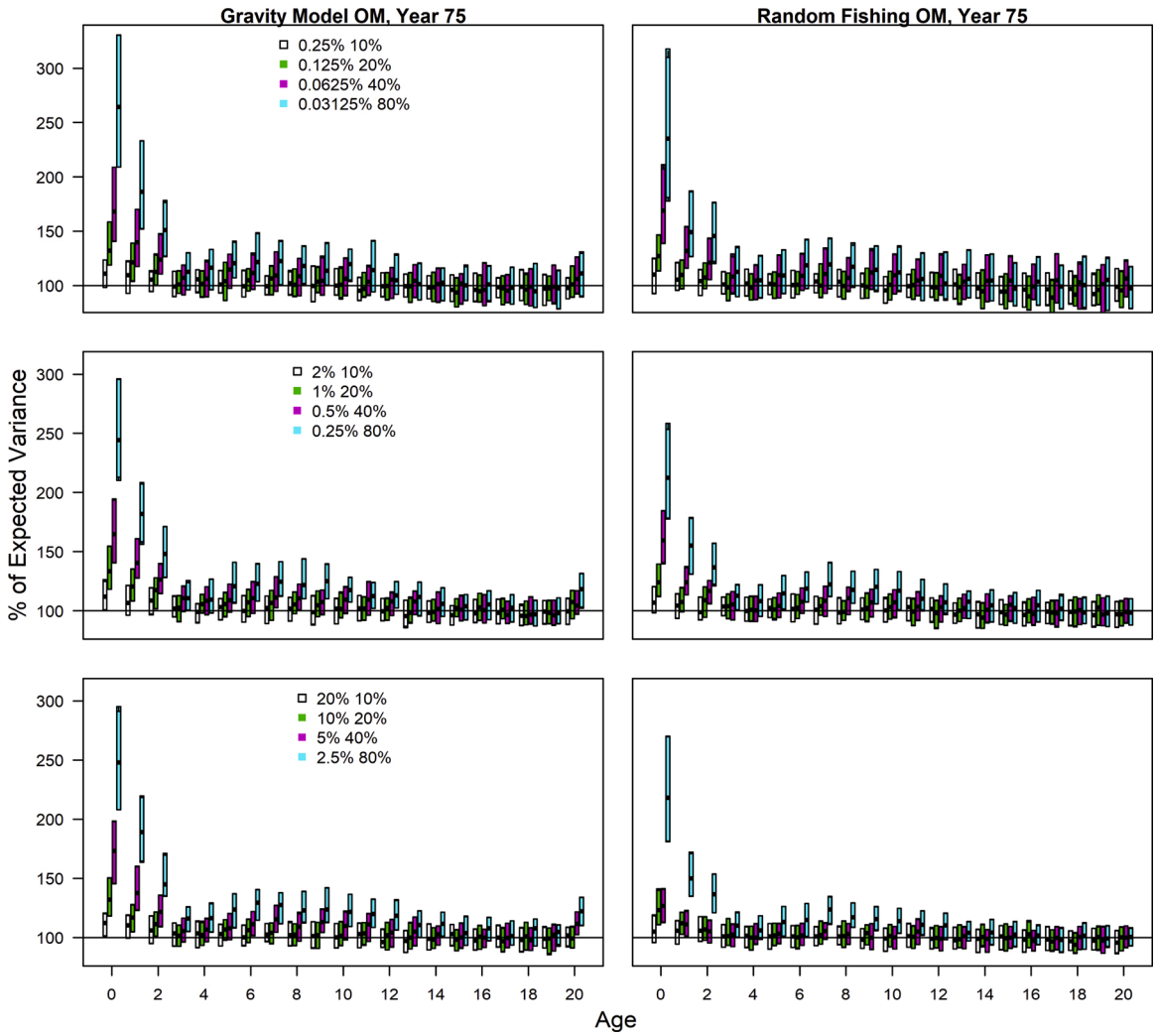


Fig. 3. Overdispersion in age composition for each OM scenario (columns) and sampling models (levels of sample size as rows and different levels of overdispersion as colored boxplots within each panel). These metrics describe the variance, across re-sampling, in the observed age composition as a percentage of the variance that would be expected from multinomial sampling given the number of samples taken, for a reference year in the fishing time series (Year 75). The resampling was done for 100 of the 1000 OM simulations, and 100 resampling replicates were run. Resampling refers to repeating the sampling process for fishery composition which includes the number of units of effort sampled from each port, the specific spatial cells that will be sampled (the cell that the unit of effort fished), and the sampling of the catch-at-age from those cells. Shown are the interquartile ranges across simulations.

2.6. Experimental design and performance metrics

Estimation model performance was evaluated focusing on quantities of management interest such as the ratio of the terminal year spawning stock biomass to the first year (unfished level including recruitment deviations, hereafter termed depletion) and the exploitation rate in the final year of the assessment (Harvest / Biomass). Results of the performance metrics across simulations were summarized by relative error (RE). We refer to these metrics as management metrics. We also evaluated the absolute relative error (ARE) for the full time series of stock abundance. This was done in an effort to examine how well the estimation models recreated operating model quantities as opposed to simply looking at performance metrics of management interest.

$$RE = (E - T)/T$$

$$ARE = |E_y - T_y|/T_y$$

Where E denotes the point estimate from the EM and T the true value from the OM.

2.6.1. Model running and convergence

One thousand replicates of each operating and sampling model were run, fit by each estimation model identified in Table 5. The operating models, across different sampling models, shared the same random number seed for population dynamics (e.g., the 2% 10 % and 1% 20 % model have the same population and fishery dynamics for an individual simulation but different fishery composition data). Estimation models were fit in Automatic Differentiation Model Builder (ADMB; Fournier et al., 2012) using penalized maximum likelihood. A model was considered converged if it achieved a gradient of 1E-4 (default setting in ADMB) and a positive definite hessian matrix. If an estimation model did not converge, initial parameter values in the. pin file were jittered and the estimation model was fit again. This process was repeated a maximum of 10 times, after which a model was considered not converged. For likelihood treatments which included iterative reweighting, effective sample sizes were considered converged once the mean deviation between runs was less than 5 (i.e., mean(ESS(i) - ESS (i-1)) < 5). This convergence rule for ESS was arbitrarily chosen as it has been noted that there is technically no correct way to do this and that it usually takes large changes in ESS to result in appreciable changes in model outputs (Maunder, 2011; Francis, 2017). Maximum likelihood

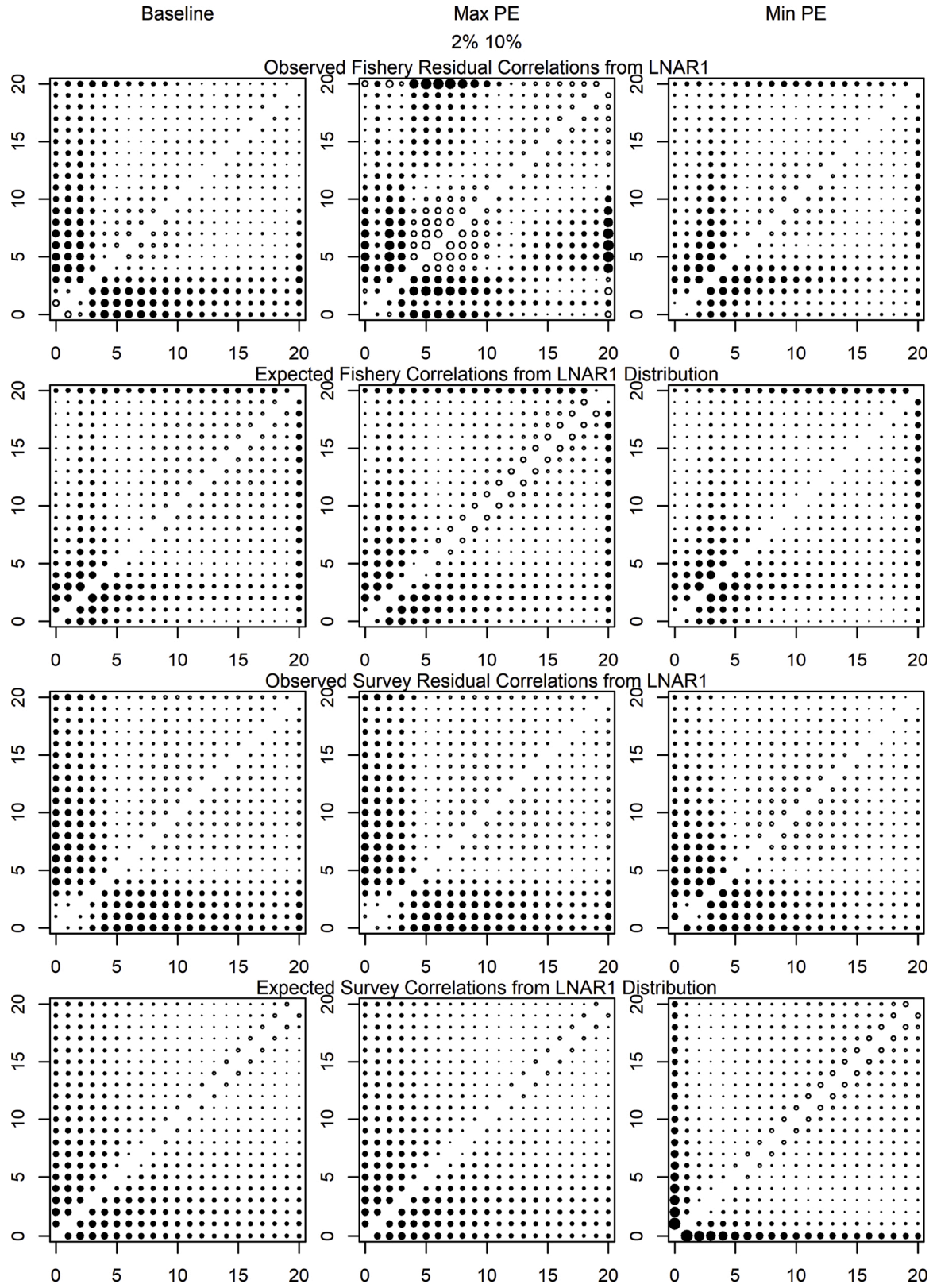


Fig. 4. Residual correlations and expected correlations for an estimation model fit using the LNAR1 likelihood. Residual correlations describe correlations in residuals across years for a given simulation, and the median taken across simulations. The expected correlations are simulated correlations that would be expected from a Logistic-normal AR(1) distribution if the sampling process were repeated. They were generated by drawing, for each year within a simulation, a composition from the LNAR1 distribution parameterized using parameter estimates (σ, φ, and expected composition) from that year and specific simulation, and subtracting from the resulting composition draw the expected proportions for that year. These draws were then correlated across years and the median taken across simulations. This figure represents results from the 2% 10 % EMs fit using the LNAR1 likelihood.

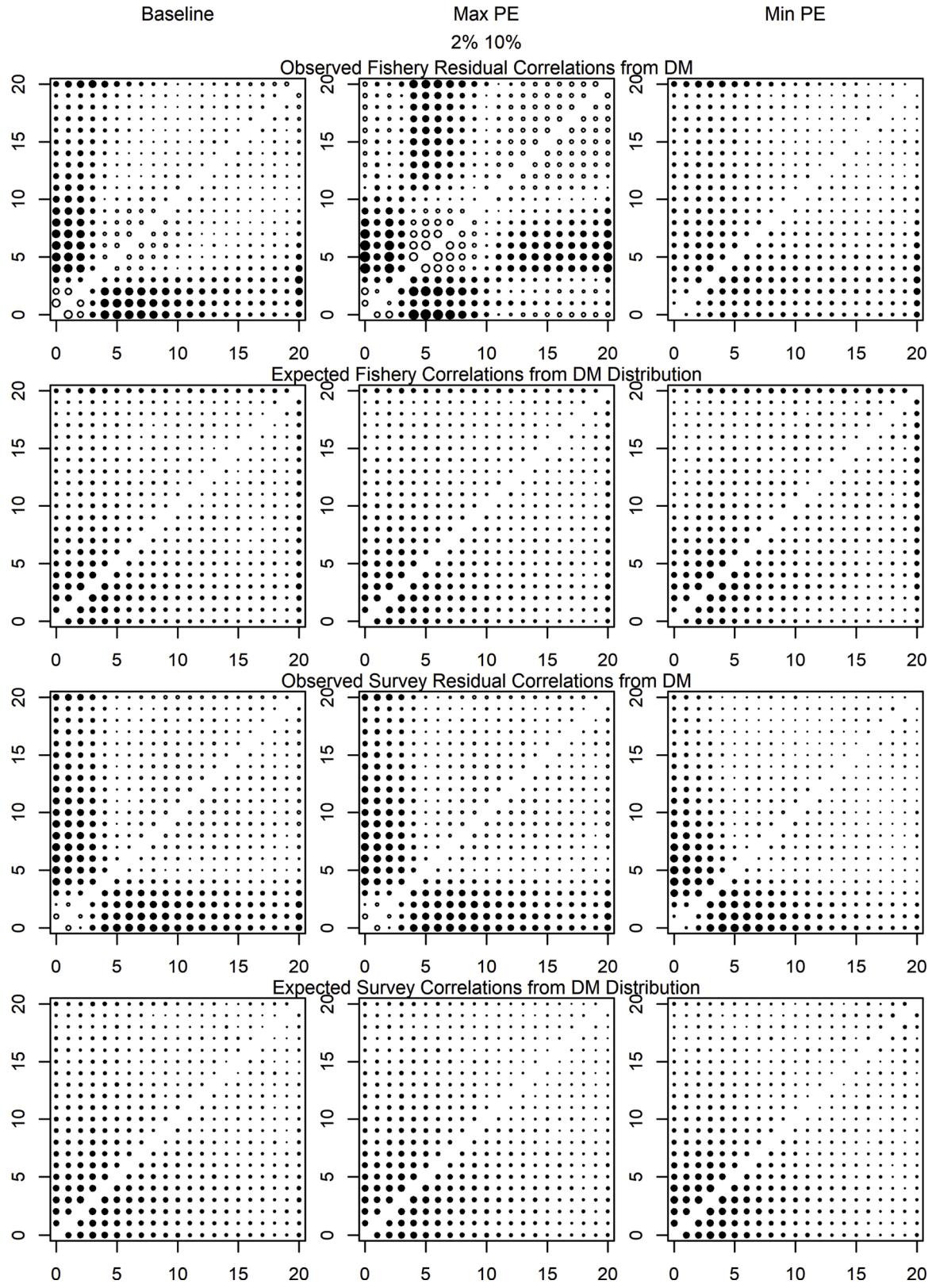


Fig. 5. Residual correlations and expected correlations for an estimation model fit using the DMA likelihood. Residual correlations describe correlations in residuals across years for a given simulation, and the median taken across simulations. The expected correlations are simulated correlations that would be expected from a Dirichlet-multinomial distribution (saturating parameterization) if the sampling process were repeated. They were generated by drawing, for each year within a simulation, a composition from the DMA distribution parameterized using parameter estimates from that year and specific simulation, and subtracting from the resulting composition draw the expected proportions for that year. These draws were then correlated across years and the median taken across simulations. This figure represents results from the 2% 10 % EMs fit using the DMA likelihood.

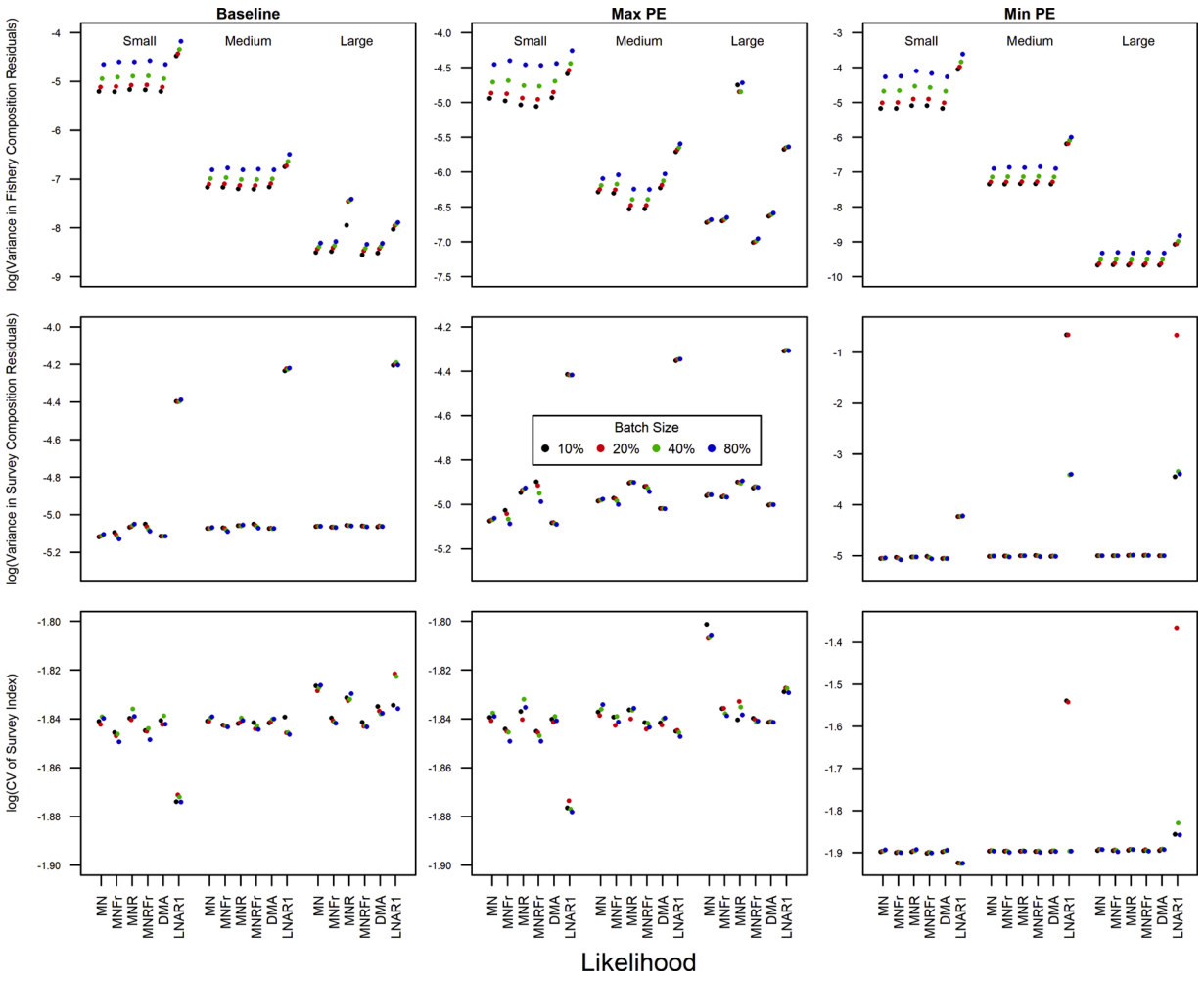


Fig. 6. Variance in the residuals for fishery composition (row 1), survey composition (row 2), and the estimated (as a parameter) CV of the survey index (row 3). Shown are the medians across simulations. The variances of compositions were summed across bins and logged to aid in visualization. Colors denote different batch sizes. Columns of the figure denote different scenarios. Within each individual panel, the 3 distinct groups of points represent different levels of sample size (small, medium, large).

estimates for each estimation model were saved and compared to operating model values. In addition to performance metrics/statistics, we also considered the computational intensity, practicality (in terms of iterative re-weighting), and percentage convergence in assessing model performance.

3. Results

The different parameterizations of the Logistic-normal models produced very similar results to one another as did the Dirichlet and the parameterizations of the Dirichlet-multinomial models. To be concise, we only present results regarding management metrics and ARE of abundance for the DMA and the LNAR1 parameterizations. Results for the Dirichlet, DML, LNAR2, and LNARMA can be found in Supplemental files (Supplemental Figs. 10–14).

3.1. Computational intensity

The estimation models which included the Logistic-normal likelihoods (LNs) were the most computationally intense. Of the LNs, the ARMA parameterization was the most computationally intense, which frequently took ~80 h for 1000 EMs. The AR(1) and AR(2) parameterizations were also very computationally intense, taking approximately 50 h for 1000 EMs (with the AR(2) model taking slightly longer). The computational intensity of the estimation models that utilized the Logistic-normal likelihoods was likely due to the need to automatically differentiate the variance covariance matrix which has dimensions year by age by age (100 × 21 × 21). The iteratively weighted multinomial and robust multinomial each took averages of ~7 h to complete 1000 EMs. The DMA models took ~6 h, the DML ~4.5 h, while the Dirichlet models took averages of ~3 h. The multinomial and robust multinomial models that were weighted using the sample size of the composition data were the least computationally intense with average run times of ~2.5 h for 1000 EMs, respectively.

3.2. Convergence

Almost all EMs converged within the Baseline and Max PE scenarios (Supplemental Fig. 13), with the most nonconvergence attributed to the DML models (40/1000 for 0.25 % 10 % for the Baseline, 30/1000 for 0.5 % 40 % for Max PE). For the Min PE and data generating model scenarios, many EMs which used the DML did not converge. Convergence for the DML increased as sample sizes decreased. Further, within each level of sample size for the Min PE, convergence increased as the batch size increased (as overdispersion increased). As an example, for large sample sizes, convergence increased from ~65 % of simulations for batch sizes of 10 %, ~75 % batch sizes of 20 %, >90 % for batch sizes of 40 %, to >99 % for batch sizes of 80 % (Supplemental Fig. 13).

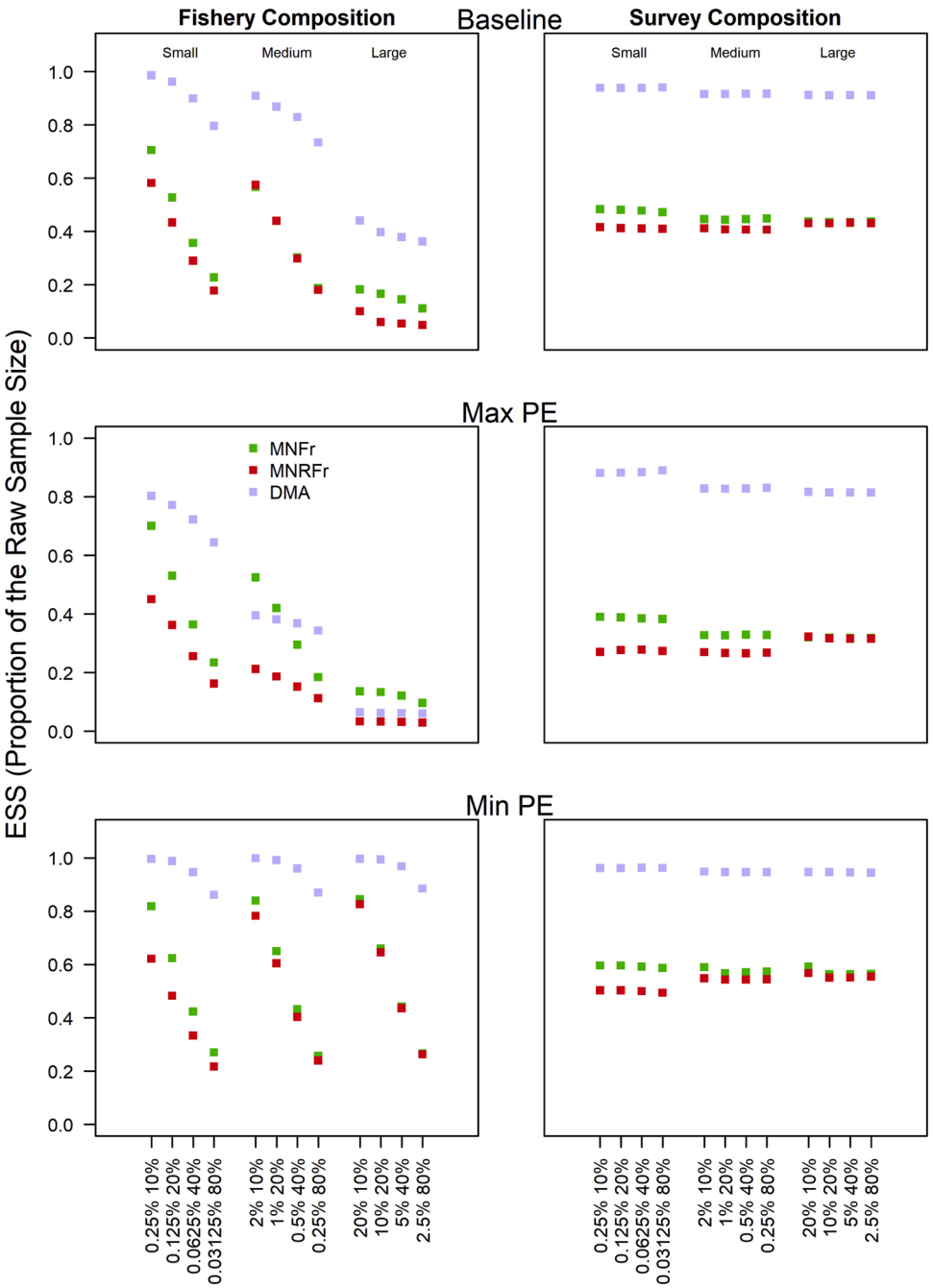


Fig. 7. Effective sample sizes, as a proportion of the raw sample size, for the Dirichlet-Multinomial and iteratively-weighted multinomial models. Depicted are the medians across simulations and the medians then taken across years in the time series. The first column presents effective sample sizes for fishery composition data and the second for survey composition, where the rows depict different scenarios. The three groups of points within each panel (groups of 4) depict the three levels of sample size, which increases from left to right. Within each group the points on the left contain smaller batch sizes and to the right larger batch sizes (more overdispersed).

3.3. Residual correlations and effective sample size

The only estimation models that allowed for correlation structure of the residuals different from that of the multinomial (negative correlations) and the MNR (no apparent correlations) were the estimation models using the Logistic-normal likelihoods. They did not recreate the correlations in fishery composition residuals exactly, however, they did show a similar pattern, much more so than the other likelihoods (Figs. 4, 5).

For each scenario, the variance of the fit to the fishery composition data was much greater for the LN models than for the others, as was the variance of the fit to the survey composition data (Fig. 6). With respect to ESS, for Min PE the DMA models only down-weighted the fishery composition a small amount at the most extreme batch sizes. Conversely, the MNFr and MNRFr models did down-weight the fishery composition data in the Min PE scenario, and ESS as a proportion of the raw sample size decreased as batch size increased (Fig. 7). The ESS as a proportion of the raw sample size did not seem to decrease as the sample sizes increased in the Min PE scenario. For the Baseline, fishery ESS as a proportion of the true sample size for each model (DMA, MNFr, and MNRFr) decreased as sample size increased and also decreased as batch sizes increased. The DMA models estimated larger fishery ESS than did the MNFr and MNRFr models. For Max PE, the same occurred where fishery ESS as a proportion of the true sample size decreased as sample size and batch sizes increased. However, where the DMA models estimated larger fishery ESSs than did MNFr at small sample sizes they estimated smaller ESSs than the MNFr at large sample sizes. The MNRFr model estimated the smallest ESS for fishery composition for each sample size level in the Max PE scenario.

With respect to the ESSs estimated for the survey compositions, for

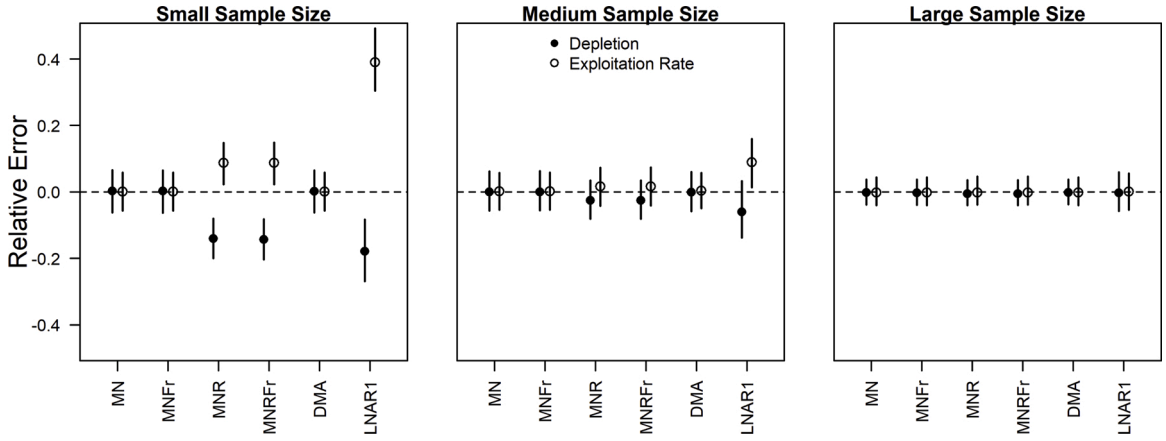


Fig. 8. Results for fit to Data-Generating model. This describes a random fishing OM with multinomial sampling of fishery catch-at-age and survey catch-at-age (and no process variation in fish movement or fisher effort distribution) fit by EMs with logistic fishery selectivity. The composition sample sizes are equivalent to the random fishing sampling model scenario 0.25 % 10 % for the small level, 2% 10 % for the medium level, and 10 % 20 % for the large level. Points denote medians where lines denote interquartile ranges across simulations.

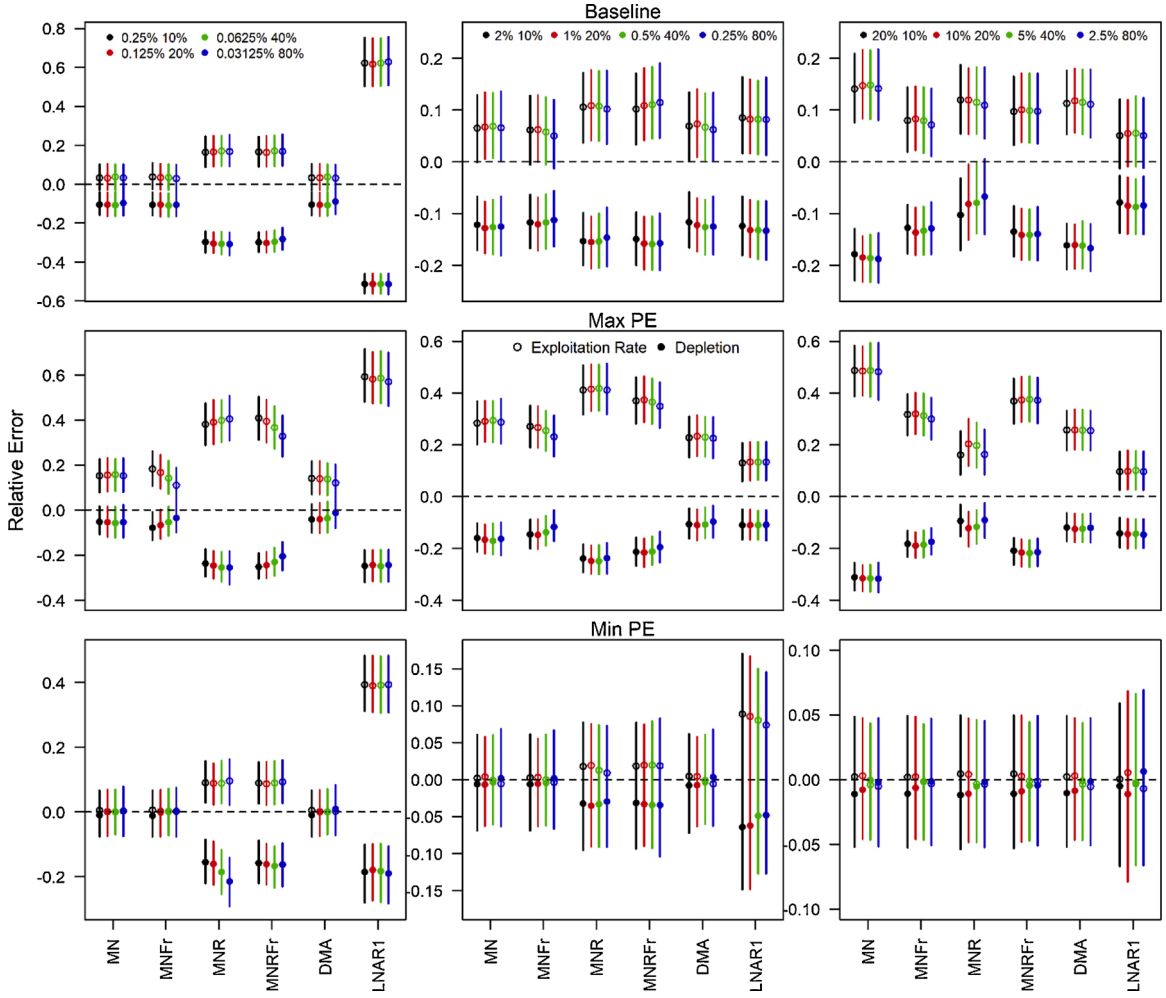


Fig. 9. Management metric results for each scenario (rows) and each sampling model, where the composition sample size increases by column from left to right. Points denote medians where lines denote interquartile ranges. Filled circles refer to depletion and open circles to exploitation rate. Within each panel, groups of 4 denote the levels of overdispersion, which increases from left to right (as batch size increases). Note that the y-axes can differ between the panels.

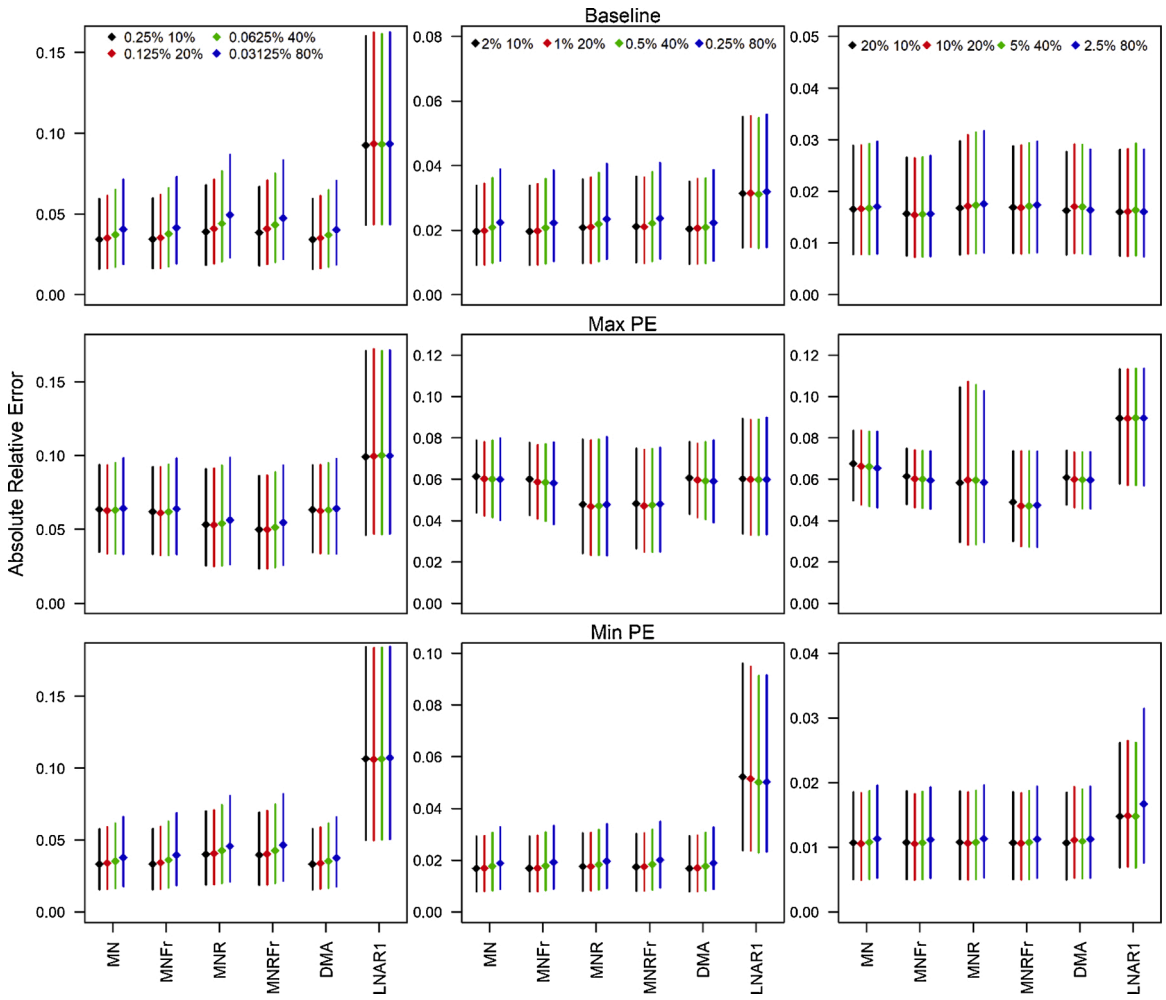


Fig. 10. Absolute relative error for abundance in each year of the time series. Rows depict scenarios whilst columns depict different fishery composition sample sizes, where the sample size increases from left to right with each column of plots. Points denote medians where lines denote the interquartile range of simulations. Within each panel, groups of 4 denote the levels of overdispersion, which increases from left to right (as batch size increases). Note that the y-axes can differ between the panels.

each scenario the DMA ESS was near ~90 % of the true sample size, where it was between 30–50 % of the true sample size for the MNFr and MNRFr models (and decreased from Min PE to Baseline to Max PE).

3.4. Data generating model

The MN, MNFr, and DMA models were each effectively unbiased when fit to data simulated from the random fishing operating model with iid catch composition sampling (and no process variation in fish movement and effort distribution) at each level of sample size (Fig. 8). The LN model was the most biased at small and medium sample sizes followed by the MNR and MNFr models. As sample size increased each of the LN, MNR, and MNRFr models improved in performance such that they were effectively unbiased at large sample sizes.

3.5. Minimum PE

For the scenarios where the EMs specified a simple logistic fishery selectivity and were fit to data simulated from an OM with random fishing (where the process model is nearly correctly specified), the MN, MNFr, and DMA models were least biased in management metrics and most accurate for abundance across most sampling models, followed by the MNR and MNRFr, and finally the LN models (Figs. 9, 10). As the sample size of the composition data increased, the performance of all models improved, such that the LN models were nearly unbiased at large sample sizes, and almost as accurate for the abundance metric. There did not seem to be a strong effect of increasing the degree of overdispersion while controlling for sample size on the rankings of the likelihoods, however across many models, especially at small sample sizes, accuracy decreased as batch sizes increased.

3.6. Baseline scenario

For the scenario where the EMs that specified a 5-parameter double-logistic fishery selectivity were fit to data simulated from the gravity model OM, the best performing likelihoods (in terms of relative error) depended on the composition sample size. When the sample size was small, the MN, MNFr, and DMA were least biased in estimating management metrics. As sample size increased, the LN improved in performance relative to the other likelihoods such that it was the least biased at large sample sizes (for most sampling models). As the sample size of the composition data increased, the bias in management metrics of the MN model increased, such that it was the most biased at large sample sizes. There did not seem to be any large change in the relative ranking of management metrics for the different EMs when controlling for sample size and increasing overdispersion.

All EMs had similar ARE of abundance in each year of the time series, with exception to the LN model at small and medium sample sizes. The

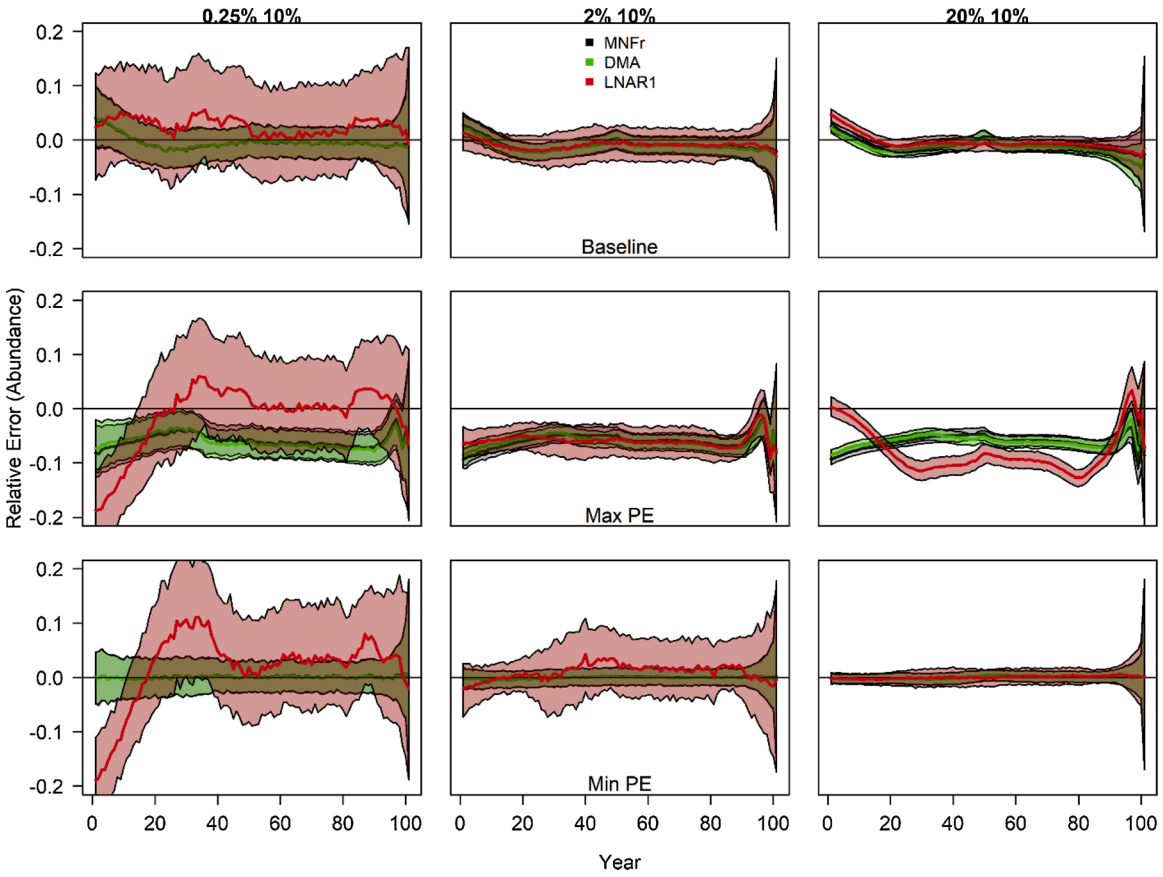


Fig. 11. Relative error results for abundance in each year of the time series. Each row depicts a different scenario and each column a different composition sample size. The lines depict medians whereas the shaded region denotes the interquartile range across simulations. The only EMs shown in this graphic are the MNFr, DMA, and LNAR1 models. Solely the 10 % batch sizes are shown in this figure. Note that in each of the plots the DMA and MNFr models mostly overlap and thus the MNFr is largely invisible behind the DMA.

MN, MNFr, and DMA were as accurate as one another at small and medium samples sizes, however at large sample sizes the MNFr was slightly more accurate. The MNR and MNRFr models were slightly less accurate than the MN, MNFr, and DMA for each level of sample size. The accuracy of the LN model improved as sample size increased, such that it was ranked second only to MNFr at large sample sizes. There did not seem to be a strong effect in the ranking of the different EMs when controlling for sample size and increasing overdispersion, however across most models, accuracy decreased as batch sizes increased.

3.7. Maximum PE

For the scenarios where the EMs that specified a simple logistic fishery selectivity were fit to simulated data from the gravity model OM, the DMA was least biased in estimating management metrics at small sample sizes, and the LN was least biased at medium and large sample sizes. An exception to this was for depletion at medium and large sample sizes, where the LN was slightly outperformed by the DMA (and MNR for large sample sizes). The DMA model was the next least-biased EM for both management metrics at medium and large sample sizes. As sample size of the fishery composition increased, the performance of the MN and MNFr models decreased, however the decrease in performance was much more pronounced for the MN. There did not seem to be a large change in the ranking of the different EMs when controlling for sample size and increasing overdispersion. However, the MNFr model did experience less bias as overdispersion increased for each level of sample size.

When examining the results with respect to the ARE of abundance in each year of the time series, the MNR and MNRFr models outperformed all other models for each sample size formulation. This was followed by the MNFr and DMA models, which had similar accuracy, and then the MN and LN models. The accuracy of the LN models improved from small to medium sample sizes, however worsened from medium to large sample sizes (Figs. 10, 11). There did not seem to be an observable effect on ARE for each EM when controlling for sample size and increasing overdispersion.

4. Discussion

4.1. Overview

Within our study, the degree of process error and the sample size of the composition data had a much greater effect on the relative ranking of the likelihoods than did the degree of overdispersion in the data. Specifically, with substantial process error present and a medium to large sample size, the LN models performed relatively well with regards to management metrics. Conversely, when the process was correctly specified (or nearly so), or the sample size was small, the DMA models performed comparatively well. However, we did find some discrepancies between results of performance criteria in the final year of the model (management metrics) and results of stock abundance in each year encompassing the whole time series.

4.2. Overdispersion effect

We did not see a substantial effect of increasing the degree of overdispersion in the data while controlling for sample size in the rankings of the different likelihoods. Although one can observe consistent minute

differences in the medians of performance metrics when overdispersion increased, these results did not appreciably change the relative rankings of the different likelihoods. When the EMs were fit without survey composition (but the survey index remained), the results remained largely the same (Supplemental Figs. 20–21). Thus, it is unlikely that this effect was due to informative survey composition data overcoming the overdispersion in the fishery composition data. Rather, it may be that even the largest batch sizes are sufficiently informative on the age composition of the catch, and/or coupled with informative indices and catch data, do not negatively affect assessment point estimates. Maunder (2011) notes that estimators of model parameters can often remain unbiased in the presence of overdispersion. However, increased overdispersion as a result of observation error is likely to affect uncertainty estimates from the stock assessment, and thus confidence intervals, hypothesis tests, and management strategy evaluations could be impacted.

4.3. Why are Likelihoods performing the way they are?

One of the main effects of changing likelihoods within the estimation models is to change how the model weights composition data coming from the fishery and the survey relative to other components. The effect of a smaller/larger weight given to a composition data set can be thought of as an increase/decrease in the composition data's share of the total error (Francis, 2017). When a composition data set is increased in relative weight, the EM will emphasize more closely fitting those data (i.e., less variation in residuals), at the expense of fitting other data. This manifests into estimation model performance by affecting the ability to estimate parameters accurately. For all models with the exception of LN, the weighting of composition data sources can be examined with recourse to ESS. In an effort to compare the weighting of the LN likelihood with the others, we examined the variance of the residuals in addition to the correlation structure of the residuals with what would be expected given the likelihood.

What seems to occur is that the LN model down-weights both the fishery composition and the survey composition more so than the other likelihoods (per the variance in the residuals, Fig. 6). One pattern that seems evident is that the model with the least amount of bias in fishing mortality generally achieves the best ranking among management metrics (Supplemental Figs. 18–19). We suspect that, when there is substantial process error, as in the Baseline and Max PE scenarios, the LN model is able to target the influence of individual composition data points more effectively through the parameterization of its variance-covariance matrix. It may then be able to better account for influential data points through recruitment deviations and fishing mortality. In contrast, the DMA and MNFr models must target or weight the entire composition at once. The exception to this is with the Min PE scenario, where process error was minimized and thus the pattern of the residuals changed (more similar to those expected from the DMA likelihood) and degree of correlations in residuals decreased (Figs. 4, 5). In this case, the DMA and MNFr models likely weighted fishery composition data more appropriately. It seems the LN also requires a large enough sample size for sufficient characterization of the variance-covariance matrix, as its performance relative to the other likelihoods improved as sample size increased for each scenario. This was especially evident in the Min PE and data-generating model scenarios, where the LN model was effectively unbiased given a sufficiently large sample size.

Although the LN models performed best with moderate to large degrees of process error and sufficient sample size with regards to management performance metrics, when examining results for abundance of the full time series of the assessment they performed more poorly relative to EMs which employed some of the other likelihoods. Their performance improved with sample size for the Baseline and Min PE scenarios to a point where they were as accurate as the MNFr and DMA models in the Baseline and nearly as accurate for Min PE, however decreased in performance from moderate to large sample sizes for Max PE. The decrease in performance for the Max PE scenario may be occurring because the estimate of unfished recruitment for LN becomes more biased relative to other models as the sample size of the composition increases (Supplemental Fig. 15). This is also the reason why the LN models did not improve in their estimation of depletion relative to the other EMs as the sample size of the composition increased (where they did for exploitation rate).

4.4. Process error

Given that fishery selectivity in the estimation models is misspecified, many of the effects on model performance manifested themselves through bias in fishing mortality and fishery catchability, and to a lesser extent, other parameters of the EMs (unfished recruitment, survey catchability, etc.). Misspecification of fishery selectivity has long been known to be consequential in stock assessment (Martell and Stewart, 2014; Thorson and Taylor, 2014; Punt et al., 2014), and some recent studies have specifically examined the role that weighting composition data plays with regards to misspecification of fishery selectivity (Stewart and Monnahan, 2017; Xu et al., 2020). Stewart and Monnahan (2017) examined how under-, over-, or right-weighting composition data affected stock assessment performance in the face of different degrees of process error and whether steepness and natural mortality were estimated. Stewart and Monnahan (2017) did not attempt to estimate composition weights within assessments but rather fixed them a priori. They found that when variation in selectivity was not accounted for, underweighting the composition led to less bias in spawning biomass than right- or overweighting the data. Our study suggests a similar finding, where decreased effective sample size led to less bias in management metrics and abundance over the time series for scenarios with significant process error in selectivity. Another study, Xu et al. (2020), compared the performance of iteratively weighting the multinomial using algorithm TA1.8 from Francis (2011) and the McAllister and Ianelli (1997) method to using the Dirichlet-multinomial likelihood for composition data within stock assessment. They examined performance across degrees of overdispersion and model misspecification with regard to time-varying selectivity. They found that, for a case similar to ours where the estimation model assumed constant selectivity however the operating model specified time-varying selectivity, each method performed similarly in mean ARE for final-year spawning biomass. However, assessments that attempted to estimate time-varying selectivity demonstrated that the Dirichlet-multinomial outperformed both iterative methods when time varying selectivity or overdispersion was present. In addition, Xu et al. (2020) found that data weighting in many cases had a large impact on the estimation performance of assessment models with correctly specified selectivity. Our findings with regard to the Min PE scenario and the data generating scenario support this finding, that even with a correctly or nearly correctly specified model, likelihood choice is consequential. In our study this was likely due to the observation error structure in the data being most similar to the multinomial (iid), given each model which did not include this sampling structure (outside of MN, MNFr, and DM) performed worse in these scenarios.

Another notable study with respect to how model misspecification affects composition likelihood choice, Maunder (2011), simulated a variety of aspects of misspecification such as ageing error, variation in natural mortality, variation in selectivity, misspecified age-specific natural mortality, and fish schooling. Although Maunder (2011) did not examine the Dirichlet-multinomial or the Logistic-normal likelihoods, he did find that methods which estimated ESS, either iteratively or within the model, improved results only when there was annual variability in selectivity (and notes that the error was still large). Maunder (2011) further found that it takes large changes in ESS to materially affect assessment model fits, noting that only when ESSs were 1/5 of the true sample size did models that estimated the ESS improve results. Direct comparisons between our study and Maunder (2011) are

difficult to make, since our likelihoods were misspecified in all but the data generating model scenario, given the nature of how we sampled from the spatially explicit OM (observation error will not match exactly with what any of the likelihoods expect). However, we do note that for the methods that estimated ESS, when there was substantial process error, performance of management metrics improved relative to models with fixed sample size as the sample size of the fishery composition increased.

4.5. Implications

Data weighting and composition likelihood choice cannot fix model misspecification. Nonetheless, our study suggests that some composition likelihoods perform better than others depending on the details of the scenario and the metric used for evaluation.

What does this imply? The answer depends on which scenario mimics a real-world situation best. We would always like to think that our models are correctly or nearly correctly specified, and in that case, it seems the Dirichlet-multinomial likelihood or the multinomial weighted with TA1.8 is the most appropriate. However, in reality, there will always be unmodeled processes not accounted for, and these processes are unlikely to be symmetric white noise deviations (as they were in Min PE case). It may be most practical to choose a likelihood that is robust across a suite of likely scenarios. In that sense, we find our study pares the choice down to the DM or the LN. The LNs were robust to scenarios with process error for management metrics and performed reasonably well with minimum process error and large sample sizes. However, with small and moderate sample sizes for composition data and a correctly specified process, the performance of the LN was poor for both management criteria and abundance over the time series. This is worrisome, as the goal is to specify the process model correctly, and we may not always have large sample sizes to overcome this hurdle. A close contender is the Dirichlet-multinomial likelihood, which was somewhat robust to process error and nearly unbiased when the process was correctly specified.

Previous research (Thorson et al., 2017) had found that Dirichlet-multinomial likelihoods performed similarly to an iterative reweighting procedure initially developed by McAllister and Ianelli (1997), although it has been suggested that the McAllister and Ianelli (1997) method overweights the composition data, and that method TA1.8 usually produces a lower effective sample size (Francis, 2017), as it was formulated to make the weights consistent with the size of the errors in mean age as opposed to individual composition proportions (thus implicitly accounting for correlations). In our study, it seems that the Dirichlet-multinomial was down-weighting mostly for process error and not as much for observation error (overdispersion and correlations in data) compared to TA1.8, evidenced both by fishery ESS and survey ESS, where there was little error specified in the survey process. In contrast TA1.8 was accounting for both, however not down-weighting as much as the DM when there was severe process error. We believe this may be occurring as TA1.8 could be losing some information on covariation between bins by calculating the error in mean age, and thus residuals in mean age may be consistent with ESS where residuals in individual bins imply more overdispersion for the DM. The chances of this occurring likely increase as process error increases, given expected values may be less likely to match with observed, and depending on the residual structure the errors in mean age may remain consistent with ESS where the residuals in individual bins are overdispersed, leading TA1.8 to estimate a larger ESS than the DM.

Although the MNFr did outperform the DM with moderate process error (Baseline scenario) and seems to be better able to account for overdispersion and correlations in the data (as evidenced by its ESS across scenarios), the DM likelihood offers several advantages over any iteratively weighted likelihood (Thorson et al., 2017), the most important of which, in our view, is eliminating the need for the iterative procedure. This should allow for more efficient exploration of alternative models, retrospective analyses, sensitivity analyses, and a more efficient run of the model itself (e.g., Bayesian estimation). We find these advantages outweigh benefits of iterative reweighting using algorithm TA1.8. Overall our study suggests that with small to moderate sample sizes, the Dirichlet-multinomial is the best choice. With larger composition sample sizes, it may be prudent to consider the Logistic-normal. The Logistic-normal likelihood may also be valuable to consider when modeling length composition data, given they often exhibit very strong correlation structure (Hrafnkelsson and Stefansson, 2004; Miller and Skalski, 2006), and are generally much more numerous than age data. In other systems/stocks, an assessment scientist may consider sample sizes as they relate to this study using the raw sample size (Supplemental Fig. 5) and the percentage of the total catch sampled in each year (Supplemental Fig. 6). Another potential use of the Logistic-normal may be as a diagnostic tool to identify significant process error. Given that the LN and the DM performed similarly at large sample sizes when there was little process error, and the LN outperformed the DM at large sample sizes when there was substantial process error, it stands to reason that if one has a large enough sample size for composition data, differences in model fit between the LN and the DM may suggest a large degree of process error.

4.6. Comparison among the parameterizations

Within our study, the different parameterizations of both the Logistic-normal and the Dirichlet-multinomial likelihoods (including the Dirichlet) performed similarly to one another. Given the difference in performance between LN parameterizations was effectively negligible (Supplemental Figs. 10–12), we recommend the LNAR1 parameterization as it is more parsimonious than the others and was the least computationally intense. The LN formulations we tested solely differed in how they modeled covariance between bins, and it should be noted that the LN can be parameterized to allow σ to vary with age or length bins (e.g., making σ a linear function of bin, Francis (2014)). This may improve performance given variance of compositions has been shown to be bin dependent (Fig. 3; Crone and Sampson, 1997; Hrafnkelsson and Stefansson, 2004; Miller and Skalski, 2006). In addition, we explored but three parameterizations of the variance-covariance matrix for the LN, and these are by no means exhaustive. We encourage the exploration of other parameterizations with respect to correlations between ages/lengths, sex, and/or years.

As for the Dirichlet and Dirichlet-multinomial parameterizations, the DMA slightly outperformed the DML when there was a large amount of process error and did not run into the convergence issues that plagued the DML. Thorson et al. (2017) notes that the Dirichlet-multinomial likelihood converges to the value of the multinomial likelihood as β or θN (weighting parameters) approach infinity. This parameter was not identified well when the DML estimation models were applied to the Min PE scenario (in which the process was nearly correctly specified), which likely led to the convergence issues noted in the results. The same convergence issues occurred in the data-generating model scenario. This suggests that the DML may have been collapsing to the multinomial distribution. The Dirichlet likelihood also produced very similar results to those of the Dirichlet-multinomial. Between these two likelihoods, we recommend use of the Dirichlet-multinomial likelihood, as within our study when the process was nearly (Min PE scenario) or completely correctly specified (Data generating model scenario), the Dirichlet-multinomial outperformed the Dirichlet likelihood. In addition, the Dirichlet performed worse in all scenarios at small sample sizes. The Dirichlet-multinomial likelihood also experiences a theoretical advantage over the Dirichlet in that the weighting parameter of the Dirichlet (α) is unbounded (Thorson et al., 2017), which could lead to overweighting (meaning the weight could theoretically be larger than the sample size), where the DM models are bounded by the sample size of the composition data. However, this property of the Dirichlet could be an advantage or used as a diagnostic tool when methods to set input

sample size prior to the assessment are used (discussed below). If the ESS for the Dirichlet were estimated to be greater than the input sample size, this may suggest error in the method used for setting input sample size. In these cases, the property of the DM models where the ESS cannot exceed the input sample size may be a disadvantage.

As for the robust multinomial, its performance varied widely across treatment levels, and thus patterns in performance are difficult to glean. One pattern that seems to have emerged is that, where the MNFr was almost always an improvement on the MN weighted with the raw sample size, the same was not true for the MNR and MNRFr. Iteratively weighting the robust multinomial may not have the same desired effect as doing so on the multinomial. This may be because the MNR is already formulated to be robust to certain deviations in composition data, as it was originally formulated to aid in keeping outlier composition data points from unduly influencing model fit (Fournier et al., 1990), and thus is already down-weighting the composition to a certain degree. The performance of the MNR and MNRFr models were, in all but a few cases, worse than either the MNFr or the DM; for this reason we suggest the latter likelihoods. Robust multinomials may perform better with length data as opposed to age data, or simply in case studies with more outlier composition data points. We did not explicitly model outliers in our simulations, and performance of the MNR models may have improved relative to the others had we done so, as robust likelihoods have been shown to improve estimates in their presence (Chen and Paloheimo, 1995; Chen and Fournier, 1999; Chen et al., 2000).

4.7. Caveats and alternative modeling

We tested 10 likelihoods and associated parameterizations for composition data in this study, 6 versions of which estimated weighting within the stock assessment. This list is by no means exhaustive, and there are additional candidate likelihoods with estimable weighting that can be found in Maunder (2011), and likely among other publications. The likelihoods and associated parameterizations tested in this study reflect the authors' conception of the most promising candidates for modeling composition data with emphasis on correlations and over-dispersion in observed data and model residuals. It is unlikely that the likelihoods tested by Maunder (2011), but not tested in this study, would provide substantial improvements over the Dirichlet or Dirichlet-multinomial used in this study, as Maunder (2011) found that the Dirichlet was least biased in estimating the true effective sample size.

It is important to note that within our study, we chose to suppress zeros in the composition data. Of the likelihoods tested in this study, the Logistic-normal and the Dirichlet are not able to incorporate zero observations. We take the position that most zero observations arise by chance (happenstance zeros; Francis, 2014), and that these observations could have been non-zero (for example taking another sample, i.e., arising via observation error). We follow the recommendations of Francis (2014) that happenstance zeros (those arising via observation error) be replaced by a small constant, and that zeros on the ends of the composition (zeros that could conceivably be true zeros) be dealt with by tail compression. We encourage further research on the effect of suppressing or allowing zeros within composition data.

Another issue brought forth in the literature with regard to likelihood choice for composition data is that of sexed compositions (Francis, 2014). This mainly presents an issue for the Logistic-normal likelihood as the correlation structure between bins becomes more complicated with the age and sex dynamic. Francis (2014) suggests two potential approaches to deal with sexed compositions. The first is to find a two-dimensional approach to including correlations within the variance-covariance matrix of the Logistic-normal, a potential candidate being the multivariate normal (Maunder, pers comm). The second is to separate the data sets such that there is an age composition and also a sex composition data set, and to choose a likelihood with appropriate structure for each (although consideration should be given to avoid double counting of samples). Further research may again be warranted on each of these approaches.

In our study, we chose to focus on effects of likelihood using a fixed effect approach. An emerging topic in the fisheries modeling literature has been the introduction of state-space modeling to separate process variation and observation error (Aeberhard et al., 2018). Accounting for process variation would seem to tilt the estimation model closer to correct model specification, which from the results of our study would imply the Dirichlet-multinomial may be a prudent likelihood choice. However, as Francis (2014) notes, we can hope to account for some process variations such as year to year variation in recruitment and selectivity, but we may not be accounting for model misspecification such as incorrect functional forms, fixed parameters, etc. In addition, in this study we chose not to model time varying selectivity. This is another recent topic in stock assessment (Martell and Stewart, 2014; Punt et al., 2014; Xu et al., 2019, 2020; with potential for applying state-space methods, see Nielsen and Berg, 2014). The resulting selectivity emerging from the spatially explicit operating model confirms that spatial structure in the population and regarding how the fishery operates can create the realized selectivity to be dome shaped, even if the contact selectivity is asymptotic, and that time-invariant selectivity is likely an exception rather than the norm. These results suggest, as many other studies previously have (Sampson, 2014; Sampson and Scott, 2011, 2012; Waterhouse et al., 2014), that time invariant and asymptotic selectivity should be modeled with caution in fisheries stock assessments. Incorporating a suitable formulation of time-varying selectivity may have pushed the estimation models closer to correct specification, which would again imply the Dirichlet-multinomial may be a prudent likelihood choice. However, the incorrect specification of time varying selectivity may also lead to increased model misspecification and process error and thus could change the choice of which composition likelihood to employ.

Composition data within stock assessment are known to influence the estimation of natural mortality (Lee et al., 2011), and estimating natural mortality within an assessment model may have cascading effects on the estimation of other parameters such as unfished recruitment, fishing mortality, and selectivity. For simplicity, we chose to fix natural mortality in this study. It is conceivable that results of our study may have differed had we attempted to estimate natural mortality. More research is needed to determine whether estimating natural mortality within stock assessment influences which composition likelihood would perform best with respect to process error, composition sample size, and overdispersion in composition data.

One might be hard-pressed to find a contemporary assessment model that weighted the multinomial likelihood using the raw sample size of fish aged. We used this approach for one of our treatments as we found this provided a good theoretical baseline for comparing likelihoods. In addition, if an assessment biologist specifies the multinomial as the likelihood for composition data, that inherently assumes the samples were independent and identically distributed and thus should be weighted using the raw sample size. Any other method to weight the multinomial is ad-hoc and admits to violating the iid sampling assumption. Some common methods of setting input sample size or ESS a priori include bootstrapping (Stewart and Hamel, 2014), using the number of sets or hauls sampled (Pennington and Volstad, 1994; Helle and Pennington, 2004), setting fixed values (Methot, 2000; Fournier and Archibald, 1982), or simply using the number of fish sampled, but not exceeding a cap (Methot, 1989). Although these methods would have likely outperformed the multinomial weighted with the true sample size in our study, none of these methods has the ability to account for process error, hence the need for iterative re-weighting or a likelihood that is able to be weighted within the assessment (and thus use of these methods is unlikely to have altered our conclusions). There has also been some recent research on spatiotemporal standardization of composition data (Thorson, 2014; Thorson and Haltuch, 2019; Thorson et al., 2020; Maunder et al., 2020). This would likely have a similar effect of providing an input sample size lower than the true sample size, in

addition to more appropriately characterizing the annual aggregated composition of the catch. However, these methods are still unable to account for process error. We encourage further research on how spatiotemporal standardization of composition data may influence the relative performance of different likelihoods. It may be that appropriately standardized composition data results in the multinomial performing better than, or as well as the other likelihoods examined in this study.

In addition, we note that we were fitting two different composition data sets (fishery and survey) that arose from two different sampling processes within each EM using the same likelihood. We did not implement a full factorial design where all likelihoods would be crossed for fishery and survey data due to time and computational constraints. Our results with respect to composition sample size and process error suggest that EM performance may have improved had we chose to model each using different likelihoods. A mixed approach could improve assessment performance, such as utilizing the Logistic-normal for fishery composition data with sufficient sample size that exhibits strong correlation structure and less confidence in correct process specification, and the Dirichlet-multinomial for survey composition data where the correct process may be specified more confidently and the sample size smaller.

Finally, we solely examined one species life history and movement pattern (and simulated exploitation pattern), in a relatively data rich scenario. It is possible that the results of our study could have differed with more sparse auxiliary data or with different life histories/movement patterns, and we encourage future research to explore these phenomena. We find that the spatially explicit operating model structure, to be fit by spatially aggregated estimation models, offers an ideal framework to test each of the issues noted above.

4.8. Conclusion

Overall, our study suggests that the Logistic-normal and Dirichlet-multinomial likelihoods are both prudent choices with respect to modeling composition data in stock assessment. Based on our simulations, the choice of which to employ should depend on the sample size of the data and the biologists' conception of the potential degree of process error. When the composition sample size is moderate to large and there exists at least a moderate amount of process error, the Logistic-normal likelihood may be the best estimator. When the sample size is small, when process error is negligible, or when observations of zeros are prevalent, the Dirichlet-multinomial is a reasonable choice.

CRediT authorship contribution statement

Nicholas Fisch: Conceptualization, Methodology, Formal analysis, Writing - original draft, Investigation, Visualization. **Ed Camp:** Supervision, Writing - review & editing. **Kyle Shertzer:** Writing - review & editing. **Robert Ahrens:** Supervision, Conceptualization, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

We would like to 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. The authors acknowledge University of Florida Research Computing for providing computational resources that contributed to the research reported in this publication. For constructive reviews, we thank Lee Cronin-Fine and Mark Maunder, who suggested the potential for using the Logistic-normal and the Dirichlet as diagnostic tests, among many other productive comments.

Appendix A

Likelihood Formulations

Multinomial

The multinomial likelihood is given in Eq. 4.1. Parameter N_y in this formula denotes the sample size collected (number of fish aged) in that year. This is the weighting parameter for the multinomial and cannot be estimated within the assessment model as its integral decreases as sample size increases (and thus the negative log likelihood decreases as sample size decreases), resulting in the effective sample size tending to zero if estimated (Francis, 2017). This likelihood was used in two different treatments; weighted using the true sample size and iteratively weighted from ESSs calculated using equation TA1.8 in Francis (2011). Effective sample sizes were considered converged once the mean deviation between runs was less than 5 (i.e., mean(ESS(i) - ESS(i-1)) < 5). This convergence rule for ESS was arbitrarily chosen as it has been noted that there is technically no correct way to do this and that it usually takes large changes in ESS to result in appreciable changes in model outputs (Francis, 2017). The multinomial cannot account for positive correlations in composition data, as the correlation between expected proportions in bins k and j, $\widehat{P}_k$ and $\widehat{P}_j$, is found using $-\left(\widehat{P}_k\widehat{P}_j \Big/ (1-\widehat{P}_k)(1-\widehat{P}_j)\right)^{0.5}$, which results in negative correlations between bins, which are usually small. Another important aspect of the multinomial distribution is that it is a discrete distribution.

Robust Multinomial

The robust multinomial (Eq. 4.2), which has also been called the multivariate normal (Francis, 2014) and the normal approximation (Maunder, 2011), was originally formulated in Fournier et al. (1990) as a normal approximation with a multinomial variance that includes two robustifying constants (0.1 and 0.01, see Eq. 4.2) meant to aid in keeping a small number of outlier composition data points from unduly influencing model fit. We chose to include this likelihood within our study as it is commonly used within age- and size-structured assessments (Francis, 2011; Maunder, 2011). The weighting parameter is the same as in the standard multinomial, N_y, and cannot be estimated within an assessment as its integral depends on N_y (Francis, 2014). The robust multinomial does not allow for correlation structure, although some correlation may be inferred given the nature of composition (if you are in one bin, you cannot be in another). We chose the Starr et al. (1999) parameterization of the robust multinomial likelihood, which uses observed proportions instead of expected proportions in some components of the formula, as we found it performed better than the original parameterization described in Fournier et al. (1990) and Francis (2011). This parameterization was also recommended in Maunder (2011). The robust multinomial was also used in two treatments; weighted using the true sample size and iteratively weighted from ESSs calculated using equation TA1.8 in Francis (2011). Effective sample sizes were considered converged once the mean deviation between runs was less than 5 (i.e., mean(ESS(i) - ESS (i-1)) < 5).

Algorithm TA1.8 from Francis (2011):

$N_y = \widetilde{N}_y w_a$

Where N_y references the effective sample size used in the likelihood formula and the tilde ($\sim$) references a value from the previous iteration (first iteration uses the raw sample size). w_a is found using $w_a = \frac{1}{Var\left(\left[\overline{O}_y - \overline{E}_y\right] \Big/ \sqrt{\left(v_y / \widetilde{N}_y\right)}\right)}$

Where $\overline{O}_y$ is the observed mean age from the composition data set for a given year, found using $\overline{O}_y = \sum_a a^* P_{a,y}$, and $\overline{E}_y$ the expected mean age, $\overline{E}_y = \sum_a a^* \widehat{P}_{a,y}$. The symbols $P_{a,y}$ and $\widehat{P}_{a,y}$ in these formulas denote an observed and expected composition data point, respectively. The variance of the expected age distribution, v_y, is found using $v_y = \sum_a (a^2 \widehat{P}_{a,y}) - \overline{E}_y^2$.

Dirichlet

The Dirichlet likelihood (Eq. 4.3), as described in Francis (2014), contains weighting parameter α, which is estimable within an assessment model. A composition, **P**, conforms to a Dirichlet distribution with parameters $[\widehat{\mathbf{P}}, \alpha]$ where $\widehat{\mathbf{P}}$ denotes a vector of expected composition. Each individual proportion is found by $P_a = \frac{X_a}{\sum_a X_a}$, where X_a are independent gamma variates with shape parameters $\alpha \widehat{P}_a$ and common scale parameter α. As in Francis (2014), we chose to allow the weighting of the Dirichlet likelihood to vary as a function of sample size each year using

$$\alpha_y = \alpha \left[\frac{N_y}{\sum_y N_y \Big/ Nyrs} \right]$$

Where *Nyrs* denotes the number of years and thus $\sum_y N_y / Nyrs$ denotes the mean sample size over the time series. In this formulation α_y is analogous to ESS, and α is estimable within the stock assessment. The correlation between proportions is the same as the multinomial (Francis, 2014), however the Dirichlet is a continuous distribution.

Dirichlet-multinomial

The Dirichlet-multinomial likelihoods used herein were parameterized as described in Thorson et al. (2017); one formulated with a weighting parameter proportional to the sample size (linear parameterization), and one formulated with a weighting parameter that saturates (reaches an asymptote) at large sample sizes. A composition is said to be Dirichlet-multinomially distributed with parameters $[\widehat{\mathbf{P}}, \alpha, N]$ if it has a multinomial distribution with parameters [**P'**, N], where **P'** is Dirichlet distributed with parameters $[\widehat{\mathbf{P}}, \alpha]$. The weighting parameters θ and β (see Eqs. 4.4 & 4.5) can be estimated within the assessment. It is a discrete distribution and has the same correlational structure as the multinomial. Effective sample sizes for each parameterization of the DM can be calculated using $N_y^{ESS} = \frac{N_y + N_y \beta}{N_y + \beta}$ for the saturating parameterization and $N_y^{ESS} = \frac{1}{1+\theta} + N_y \frac{\theta}{1+\theta}$ for the linear parameterization.

Logistic-normal

The Logistic-normal likelihood (Eq. 4.6) was parameterized as in Francis (2014). The first use of the logistic normal within stock assessment is attributed to Schnute and Richards (1995). A composition is said to conform to a Logistic-normal distribution with parameters $\left[\widehat{\mathbf{P}}, \mathbf{C}\right]$ when $P_a = \frac{e^{X_a}}{\sum_a e^{X_a}}$.

In this case, **X** conforms to a multivariate normal distribution with mean $\log(\widehat{\mathbf{P}})$ and covariance matrix **C**. The Logistic-normal is a continuous distribution 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 and second order autoregressive (AR(1), AR(2)) parametrization of the variance-covariance matrix in addition to an autoregressive moving average (ARMA) parameterization. The weighting parameters for these parameterizations are σ_{AR1} and φ for AR(1); σ_{AR2}, φ_1, and φ_2 for AR(2); and σ_{ARMA}, φ_{ARMA}, and ψ for the ARMA. Different weighting between years based on composition sample size was achieved using $W_y = \sqrt{\frac{\sum_y (N_y)/Nyrs}{N_y}}$ where $\sigma_y = \sigma W_y$ (with σ as a stand in for either σ_{AR1}, σ_{AR2}, or σ_{ARMA}), as in Francis (2014). This allows σ_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 for each formulation of the Logistic-normal using $C_{y,a,a'} = \sigma_y^2 \rho_{|a-a'|}$, where

$\rho_{|a-a'|} = \varphi^{|a-a'|}$ for an AR(1) process.

$\rho_0 = 1 \quad \rho_1 = \frac{\varphi_1}{(1-\varphi_2)} \quad \rho_k = \varphi_1 \rho_{k-1} + \varphi_2 \rho_{k-2}$ for an AR(2) process

$\rho_0 = 1 \quad \rho_1 = \varphi + \frac{\psi}{\left(1 + (\varphi_{ARMA} + \psi)^2 \Big/ (1 - \varphi_{ARMA}^2)\right)} \quad \rho_k = \varphi_{ARMA}^{k-1} \rho_1$ for an ARMA process.

The negative log likelihood can then be found using equation A9 in Francis (2014)

$$NLL = \sum_y \left[0.5(Nb-1)^* \log(2\pi) + \sum_a \left[\log\left(P_{a,y}\right)\right] + 0.5^* \log\left(\left|\mathbf{V}_y\right|\right) + (Nb-1)^* \log\left(W_y\right) + \frac{\left(\mathbf{w}_y^T \mathbf{V}_y^{-1} \mathbf{w}_y\right)}{2W_y^2} \right]$$

Where $\mathbf{V}_y = \mathbf{K}\mathbf{C}_y\mathbf{K}^T$, **K** is an $[(Nb-1), Nb]$ matrix formed by adding a vector filled with -1 to the right side of an identity matrix with dimensions $[Nb-1, Nb-1]$, and **w** is a matrix where each row depicts a year and contains a vector of length $(Nb-1)$, filled using $w_{a,y} = \log\left(\frac{P_{a,y}}{P_{Nb,y}}\right) -$

$\log\left(\frac{\widehat{P}_{a,y}}{\widehat{P}_{Nb,y}}\right)$ *for a in* $0, 1, 2, \ldots, \ldots, Nb-1$. The term *Nb* refers to the number of bins in a composition dataset. The σ parameter for each likelihood treatment was estimated on the log scale, φ was bound between (-1,1) for the AR(1) and ARMA, and ψ was effectively unbounded. For the AR(2) process estimation, bounds of (-2,2) were placed on φ_1 and a new parameter, ω, was estimated to keep φ_2 within proper bounds. This parameter was estimated on the logit scale, keeping it between (0,1) and used to calculate φ_2 using $\varphi_2 = -1 + (2 - |\varphi_1|)\omega$. This was done to ensure that the parameters of the AR(2) process lie within the triangle defined by $-1 \leq \varphi_2 < 1 - |\varphi_1|$ (for stability of AR(2) process; Francis, 2014).

Appendix B

Operating Model Parameterization and Preference Functions

Due to the differences in the spatial extent between the spatially explicit operating model in this study and the Gulf of Mexico red snapper assessment, some population parameters from the assessment had to be adjusted to account for the smaller geographic area. The GOM red snapper assessment allocates total recruits each year to the western GOM (mean ~ 64 %) and eastern GOM (mean ~ 36 %), split by the Mississippi River. In 2016 and for the assessment projection to 2076, this apportionment was 23 % to the eastern GOM and 77 % to the Western GOM. To obtain the unfished recruitment for Florida waters, we calculated the proportion of recruits in the eastern GOM (using the 2016 estimate) that would occur in Florida waters based on availability of habitat for recruitment. We did this by dividing the spatial cells with depths from 10−70 m in Florida waters by the total spatial cells with depths from 10−70 m in the eastern GOM (eastern GOM longitudinal cutoff -89°). The depth cutoff of 70 m was chosen as this is roughly equivalent to the mean + 2SD of depth preference of age 0 red snapper (see movement section). Unfished recruitment of red snapper for Florida (and thus the spatial model) was then calculated as the product of unfished recruitment for the entire Gulf of Mexico (1.63E8), the proportion of recruits allocated east of the Mississippi (23 %), and the proportion of recruits in the east zone that are allocated to Florida (~90 %). Equilibrium spawning biomass for Florida was then calculated by projecting this new unfished recruitment to a plus group at age 20 using a natural mortality ogive, multiplying each value by its age specific fecundity, and summing across the values.

Preference Functions

Data

Spatially referenced red snapper catch at age data was compiled from the US Gulf of Mexico reef fish bottom longline and vertical line observer database. This database contained captures-at-age for red snapper age 0−10. We compiled catches at age across gears and classified them into the 0.1 decimal degree grids. These data were not standardized for effort. We cross referenced capture locations with depth and substrate shapefiles to create movement preference functions.

Depth

The preference function for depth was age specific. Depth information for the GOM was collected from Becker et al. (2009; https://topex.ucsd.edu/cgi-bin/get_srtm30.cgi). This data set was at a more fine resolution than the spatially explicit grid, in 30-arc seconds (30 arc seconds = 0.0083 decimal degrees). For this reason, depth values for each spatial cell within the model were calculated as the mean depth (of 30 arc second data) within the cell. The mean and variance in depth of capture for each age was calculated and a Von-Bertalanffy function was fit through these values (with age as the explanatory variable) so as to capture the asymptotic nature of these two relationships as fish aged (Supplemental Fig. 22). The depth preference function for each age was then characterized as a normal distribution using the mean and variance of capture depth (from the Von-Bertalanffy functions).

Substrate Type

The preference function for substrate type was calculated as the percentage of red snapper at age that were captured on a specific substrate type (Supplemental Fig. 23). Bottom substrate data were collected from the NOAA Gulf of Mexico Data Atlas (https://www.ncddc.noaa.gov/website/DataAtlas/atlas.htm). Substrate classes included rock, gravel, sand, and mud, and are divided into dominant classifications if the most abundant fractions of the substrate classes are greater than 66 %, and subdominant classifications if the most abundant fraction is greater than 33 %, resulting in 8 substrate classes (Supplemental Fig. 24).

Distance

The distance preference function, referencing the distance from one cell to another cell, was modeled as an exponential decay by Euclidean distance in km from the midpoint of cell (Supplemental Fig. 25).

$$e^{-\lambda_D * km}$$

The decay rate (λ_D) was parameterized using tag-recapture data on red snapper. We fit an exponential decay model using maximum likelihood to the distance red snapper traveled in a year, corrected for time at liberty. We omitted all recaptures where fish spent less than 200 days at liberty, as the daily movement rates were much higher for fish that spent less than 200 days at liberty. Tag-recapture data on red snapper was obtained from Addis et al. (2013).

Density

The density preference function (Supplemental Fig. 26) was modeled as an exponential decay below a density threshold (Bentley et al., 2004).

$$\begin{cases} 1 & \text{if} \quad D \leq D^* \\ 1 \Big/ \left(\frac{D}{D^*}\right)^{\lambda_{DD}} & \text{if} \quad D > D^* \end{cases}$$

The quantity D describes density of fish in a cell, and was characterized as the sum of the squared lengths of fish within a cell ($\sum_a N_{y,a,c} L_a^2$, where L_a refers to the length of a fish age a). The density threshold D^* (Table 1; OMP.20) was arbitrarily set at the 75 % quantile of the unfished densities (year 1 of the model) in each cell. The decay rate, λ_{DD} (OMP. 21), was arbitrarily set at 0.5.

Appendix C. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fishres.2021.106069.

References

Addis, D.T., Patterson III, W.F., Dance, M.A., Ingram Jr., G.W., 2013. Implications of reef fish movement from unreported artificial reef sites in the northern Gulf of Mexico. Fish. Res. 147, 349–358.

Aeberhard, W.H., Mills Flemming, J., Nielsen, A., 2018. Review of state-space models for fisheries science. Annu. Rev. Stat. Appl. 5, 215–235.

Becker, J.J., Sandwell, D.T., Smith, W.H.F., Braud, J., Binder, B., Depner, J., Fabre, D., Factor, J., Ingalls, S., Kim, S.-H., Ladner, R., Marks, K., Nelson, S., Pharaoh, A., Trimmer, R., Von Rosenberg, J., Wallace, G., Weatherall, P., 2009. Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. Mar. Geod. 32, 355–371.

Bentley, N., Davies, C.R., McNeill, S.E., Davies, N.M., 2004. A framework for evaluating spatial closures as a fisheries management tool. New Zealand Fisheries Assessment Report 2004/25, 25 p.

Caddy, J.F., 1975. Spatial model for an exploited shellfish population, and its application to the georges bank scallop fishery. J. Fish. Res. Bd. Can. 32, 1305–1328. https://doi.org/10.1139/f75-152.

Chen, Y., Fournier, D., 1999. Impacts of atypical data on Bayesian inference and robust Bayesian approach in fisheries. Can. J. Fish. Aquat. Sci. 56 (9), 1525–1533.

Chen, Y., Paloheimo, J.E., 1995. A robust regression analysis of recruitment in fisheries. Can. J. Fish. Aquat. Sci. 52 (5), 993–1006.

Chen, Y., Breen, P.A., Andrew, N.L., 2000. Impacts of outliers and mis-specification of priors on Bayesian fisheries-stock assessment. Can. J. Fish. Aquat. Sci. 57 (11), 2293–2305.

Crone, Sampson, 1997. Evaluation of assumed error structure in stock assessment models that use sample estimates of age composition. International Symposium on Fishery Stock Assessment Models for the 21st Century.

Dichmont, C.M., Deng, R.A., Punt, A.E., Brodziak, J., Chang, Y.-J., Cope, J.M., Ianelli, J. N., Legault, C.M., Methot, R.D., Porch, C.E., Prager, M.H., Shertzer, K.W., 2016. A review of stock assessment packages in the United States. Fish. Res. 183, 447–460. https://doi.org/10.1016/j.fishres.2016.07.001.

Dunn, A., Rasmussen, S., Mormede, S., 2012. Spatial Population Model User Manual, SPM v1.1-2012-09-06 (rev 4806). CCAMLR; WG-FSA-12/46, Hobart, 164 pp.

Fournier, D., Archibald, C.P., 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39 (8), 1195–1207.

Fournier, D.A., Sibert, J.R., Majkowski, J., Hampton, J., 1990. MULTIFAN a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (Thunnus maccoyii). Can. J. Fish. Aquat. Sci. 47 (2), 301–317. https://doi.org/10.1139/f90-032.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A., Sibert, J., 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim. Methods Softw. 27, 233–249.

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

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

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

Francis, R.C., 2017. Revisiting data weighting in fisheries stock assessment models. Fish. Res. 192, 5–15.

Grüss, A., Thorson, J.T., Sagarese, S.R., Babcock, E.A., Karnauskas, M., Walter, J.F., Drexler, M., 2017. Ontogenetic spatial distributions of red grouper (Epinephelus morio) and gag grouper (Mycteroperca microlepis) in the U.S. Gulf of Mexico. Fish. Res. 193, 129–142. https://doi.org/10.1016/j.fishres.2017.04.006.

Helle, K., Pennington, M., 2004. Survey design considerations for estimating the length composition of the commercial catch of some deep-water species in the northeast Atlantic. Fish. Res. 70 (1), 55–60.

Hrafnkelsson, B., Stefansson, G., 2004. A model for categorical length data from groundfish surveys. Canadian J. Fish. Aquat. Sci. 61 (7), 1135–1142.

Hulson, P.J.F., Hanselman, D.H., Quinn, T.J., 2011. Effects of process and observation errors on effective sample size of fishery and survey age and length composition using variance ratio and likelihood methods. ICES J. Mar. Sci. 68 (7), 1548–1557.

Hulson, P.J.F., Hanselman, D.H., Quinn, T.J., 2012. Determining effective sample size in integrated age-structured assessment models. ICES J. Mar. Sci. 69 (2), 281–292.

Jardim, E., Azevedo, M., Brodziak, J., Brooks, E.N., Johnson, K.F., Klibansky, N., Millar, C.P., Minto, C., Mosqueira, I., Nash, R.D., Vasilakopoulos, P., 2020. Operationalizing Ensemble Models for Scientific Advice to Fisheries Management.

Karnauskas, M., Walter, J.F., Campbell, M.D., Pollack, A.G., Drymon, J.M., Powers, S., 2017. Red snapper distribution on natural habitats and artificial structures in the Northern Gulf of Mexico. Mar. Coast. Fish. Dyn. Manag. Ecosyst. Sci. 9, 50–67. https://doi.org/10.1080/19425120.2016.1255684.

Lee, H.-H., Maunder, M.N., Piner, K.R., 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. https://doi.org/10.1016/j.fishres.2011.01.021.

Mace, P.M., Doonan, I.J., 1988. A Generalised Bioeconomic Simulation Model for Fish Population Dynamics. MAFFish, NZ Ministry of Agriculture and Fisheries.

Martell, S., Stewart, I., 2014. Towards defining good practices for modeling time-varying selectivity. Fish. Res. 158, 84–95.

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, 311–319. https://doi.org/10.1016/j.fishres.2011.02.018.

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

Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74. https://doi.org/10.1016/j.fishres.2012.07.025.

Maunder, M.N., Thorson, J.T., Xu, H., Oliveros-Ramos, R., Hoyle, S.D., Tremblay-Boyer, L., Lee, H.H., Kai, M., Chang, S.K., Kitakado, T., Albertsen, C.M., 2020. The need for spatio-temporal modeling to determine catch-per-unit effort based indices of abundance and associated composition data for inclusion in stock assessment models. Fish. Res. 229, 105594.

McAllister, M.K., 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. https://doi.org/10.1139/f96-285.

Methot, R.D., 1989. Synthetic estimates of historical abundance and mortality for northern anchovy. Am. Fish. Soc. Symp. 6, 66–82.

Methot, R.D., 2000. Technical description of the Stock Synthesis assessment program. U. S. Dept. Commer., NOAA Tech. Memo. NWFD-NWFSC-43, 46 p.

Miller, T.J., Skalski, J.R., 2006. Integrating design-and model-based inference to estimate length and age composition in North Pacific longline catches. Canadian J. Fish. Aquat. Sci. 63 (5), 1092–1114.

Mormede, S., Dunn, A., Parker, S., Hanchet, S., 2017. Using spatial population models to investigate the potential effects of the Ross Sea region Marine Protected Area on the Antarctic toothfish population. Fish. Res. 190, 164–174. https://doi.org/10.1016/j.fishres.2017.02.015.

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. 158, 96–101.

Pennington, M., 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. https://doi.org/10.2307/2532786.

Punt, A.E., Hurtado-Ferro, F., Whitten, A.R., 2014. Model selection for selectivity in fisheries stock assessments. Fish. Res. 158, 124–134. https://doi.org/10.1016/j.fishres.2013.06.003.

Punt, A.E., Haddon, M., 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? Fisheries Research, Growth: Theory, Estimation, and Application in Fishery Stock Assessment Models, 180, pp. 147–160. https://doi.org/10.1016/j.fishres.2014.11.007.

Sampson, D.B., 2014. Fishery selection and its relevance to stock assessment and fishery management. Fish. Res. 158, 5–14.

Sampson, D.B., Scott, R.D., 2011. A spatial model for fishery age-selection at the population level. Can. J. Fish. Aquat. Sci. 68 (6), 1077–1086. https://doi.org/10.1139/f2011-044.

Sampson, D.B., Scott, R.D., 2012. An exploration of the shapes and stability of population–selection curves. Fish Fish. 13 (1), 89–104.

Schnute, J.T., Richards, L.J., 1995. The influence of error on population estimates from catch-age models. Can. J. Fish. Aquat. Sci. 52, 2063–2077.

SEDAR, 2018. SEDAR 52 - Gulf of Mexico Red Snapper Stock Assessment Report, 434 pp. available online at:. SEDAR, North Charleston, SC http://sedarweb.org/sedar-52.

Starr, P.J., Bentley, N., Maunder, M.N., 1999. Assessment of the NSN and NSS stocks of red rock lobster (Jasus edwardsii) for 1998. New Zealand Fisheries Assessment Research Document 99/34. Ministry of Fisheries, Wellington, New Zealand.

Stewart, I.J., Hamel, O.S., 2014. Bootstrapping of sample sizes for length- or age-composition data used in stock assessments. Can. J. Fish. Aquat. Sci. 71 (4), 581–588. https://doi.org/10.1139/cjfas-2013-0289.

Stewart, I.J., Martell, S.J., 2015. Reconciling stock assessment paradigms to better inform fisheries management. ICES J. Mar. Sci. 72 (8), 2187–2196.

Stewart, I.J., Monnahan, C.C., 2017. Implications of Process Error in Selectivity for Approaches to Weighting Compositional Data in Fisheries Stock Assessments. Fisheries Research, Data Conflict and Weighting, Likelihood Functions, and Process Error, 192, pp. 126–134. https://doi.org/10.1016/j.fishres.2016.06.018.

Thorson, J.T., 2014. Standardizing compositional data for stock assessment. Ices J. Mar. Sci. 71 (5), 1117–1128.

Thorson, J.T., Haltuch, M.A., 2019. Spatiotemporal analysis of compositional data: increased precision and improved workflow using model-based inputs to stock assessment. Can. J. Fish. Aquat. Sci. 76 (3), 401–414.

Thorson, J.T., Taylor, I.G., 2014. A comparison of parametric, semiparametric, and non-parametric approaches to selectivity in age-structured assessment models. Fish. Res. 158, 74–83. https://doi.org/10.1016/j.fishres.2013.10.002.

Thorson, J.T., Johnson, K.F., Methot, R.D., 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. https://doi.org/10.1016/j.fishres.2016.06.005.

Thorson, J.T., Bryan, M.D., Hulson, P.J.F., Xu, H., Punt, A.E., 2020. Simulation testing a new multi-stage process to measure the effect of increased sampling effort on effective sample size for age and length data. ICES J. Mar. Sci. 77 (5), 1728–1737.

Walters, C., 2003. Folly and fantasy in the analysis of spatial catch rate data. Can. J. Fish. Aquat. Sci. 60, 1433–1436. https://doi.org/10.1139/f03-152.

Walters, C.J., Bonfil, R., 1999. Multispecies spatial assessment models for the British Columbia groundfish trawl fishery. Can. J. Fish. Aquat. Sci. 56, 601–628. https://doi.org/10.1139/f98-205.

Xu, H., Thorson, J.T., Methot, R.D., Taylor, I.G., 2019. A new semi-parametric method for autocorrelated age-and time-varying selectivity in age-structured assessment models. Can. J. Fish. Aquat. Sci. 76 (2), 268–285.

Waterhouse, L., Sampson, D.B., Maunder, M., Semmens, B.X., 2014. Using areas-as-fleets selectivity to model spatial fishing: asymptotic curves are unlikely under equilibrium conditions. Fish. Res. 158, 15–25.

Xu, H., Thorson, J.T., Methot, R.D., 2020. Comparing the performance of three data-weighting methods when allowing for time-varying selectivity. Can. J. Fish. Aquat. Sci. 77, 247–263. https://doi.org/10.1139/cjfas-2019-0107.