

Process and sampling variance within fisheries stock assessment models: estimability, likelihood choice, and the consequences of incorrect specification

N. Fisch^{1,*}, K. Shertzer², E. Camp¹, M. Maunder³, and R. Ahrens⁴

¹Fisheries and Aquatic Sciences, School of Forest, Fisheries, and Geomatics Sciences, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL 32611, USA.

²National Marine Fisheries Service, Southeast Fisheries Science Center, 101 Pivers Island Road, Beaufort, NC 28516, USA

³Inter-American Tropical Tuna Commission, 8901 La Jolla Shores Dr, La Jolla, CA 92037, USA

⁴National Marine Fisheries Service, Pacific Islands Fisheries Science Center, 1845 Wasp Blvd., Building 176, Honolulu, HI 96818, USA

*Corresponding author: tel: +001 727-798-8424; e-mail: nickcfisch@gmail.com.

Increasingly, mixed-effect fishery stock assessment models are being developed where deviations about functional forms of different processes are modelled as random effects and the extent of variance is estimated internal to the model. Concurrently, sampling variance parameters associated with likelihoods for fitting composition data within fisheries assessments are more often being estimated internal to the model as well. We examine the performance of stock assessment models when multiple process variance and sampling variance terms are simultaneously estimated within assessment models. We specifically examine how assessment performance is affected by the choice of composition likelihood, the degree of overdispersion in composition data, overparameterization, and modelling variation on the wrong process. In doing so, we build a simulation containing overdispersion and correlations in composition data, directional variation in catchability and/or selectivity, and estimation models which include random effects and composition likelihoods with theoretically estimable variances. Results suggest that with standard data available in fisheries assessments, process variance parameters associated with some commonly employed methods and sampling variance parameters can be simultaneously estimated internal to an assessment, and performance greatly improves with increased composition data. Our results also suggest little downside to overparameterization of selectivity and catchability when the true process is not time-varying, which largely agrees with previous research. However, when a process is truly time-varying and the assessment models time-variation on a different process, namely when selectivity is time-varying and instead natural mortality is modelled as potentially time-varying, we find a risk of severe increases in bias and decreases in confidence interval coverage for assessed quantities. This bias and decrease in coverage could, however, be partially mitigated by also modelling time-variation on the correct process.

Keywords: Stock Assessment, Composition Data, Age-structured models, Overdispersion, Mixed-effects, Process Error, Sampling Error, Process Variation.

Introduction

The field of fisheries assessment develops population models of fish and invertebrate stocks informed by various data sources in an effort to help inform management decisions (Dichmont *et al.*, 2016). These data sources commonly include the total catch, the age or length composition of the catch, the catch rate of fishers, and similar data collected from research surveys. The population models are developed by statistically estimating key parameters to come up with the most plausible population and exploitation model of the specific fishery. The statistical estimation process compares predictions from the population model to observed data (under assumed observation models), searching for the set of model parameters that maximize the likelihood of observing the data. The predictions will never exactly match the data due to several different sources of variability, the distinction of which is important.

With a perfectly specified population and observation model, provided it is not feasible to census each data source, the sole source of variability between model predictions and the observed data would be attributed to sampling variability, which describes the expected variability from a statistical distribution if the sampling process were repeated. This variability arises due to the act of taking a sample and not a census.

Considering random sampling variability as constituting the sole source of variability between model predictions and data is idealized and unrealistic. In reality, models are imperfect representations and discrepancies between models and observations will vary beyond those attributed to sampling variability alone. One reason for this is because population models are simply, and necessarily, approximations of reality. They often characterize quantities or rates using mean processes and, in the limit, are at least approximating continuous values. For example, fisheries assessments often characterize the natural mortality rate of fish as constant with respect to time, which does not account for a myriad of unmodelled ecosystem and environmental processes that can cause this rate to vary from year to year. Although it may be a reasonable and necessary approximation if the variation is small (and stationary in time), it will result in additional variation between model predictions and observations.

A related but somewhat different source of variability between model predictions and data relates to the fact that models can simply be incorrect representations of a system. This can occur in fisheries assessments as mischaracterizations of the functional form of a relationship between two variables (e.g. stock and recruitment), inaccurate assumed values for pa-

Received: 21 February 2023; Revised: 13 July 2023; Accepted: 12 August 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of International Council for the Exploration of the Sea. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

rameters (e.g. natural mortality and growth), incorrect probability models for process or sampling variances, or misrepresenting the process of observing the data (e.g. unmodelled density dependence in the relationship between an index in abundance and population size, or measurement error such as ageing and sizing error).

In this investigation, we characterize the variability between predictions and data associated with models being incorrect representations of a system as model misspecification and that associated with models approximating mean processes as process variability. There is necessarily some overlap between these two terms as one could argue that not accounting for process variation makes a model an incorrect description of the system. Hence together, model misspecification and process variability collectively make up what we refer to as model error, or all of the variation between model predictions and data that is not due to random sampling variability. Model error, in addition to sampling variability (sampling error) make up the total error, or the total amount of variability expected between the model predictions and the observations. Importantly, there have been many different characterizations and categorizations of the components making up total error (Francis and Shotton, 1997; Hulson *et al.*, 2012; Maunder and Piner, 2015), and each have their merit. The general structure is similar, often differing in the categorization of components that make up model error to suit the specific study in question (e.g. categorizing into process model components and observation model components, and potential error sources of each).

The distinction of error is important because when fitting a statistical model, an analyst must decide how much of the total variance between model predictions and data to allocate to sampling error and to process variation, while acknowledging one or both variance components may be affected by some level of model misspecification. Further an analyst must decide how to model process variation. The paradigm of mixed-effects and state-space modelling has proved increasingly useful in fisheries assessment as a means to model process variation using random effects and potentially even estimate the extent of the variation internal to the model (Thorson, 2019). Although random effects are not novel to the field of statistics and fisheries assessment, recent advances in computation and approximation have allowed much more efficient development and implementation of such models within the Template Model Builder (TMB) programme (Kristensen *et al.*, 2016). In addition, research has suggested that for some datasets, namely those pertaining to the proportions caught at age in fishery and survey samples (compositions), the sampling error variance can also be estimated internal to an assessment model using specific likelihoods (Maunder, 2011; Albertsen *et al.*, 2016; Thorson *et al.*, 2017; Fisch *et al.*, 2021).

The implications of getting either the total variance or its constituent components wrong are non-trivial. The relative contribution of the different components of total error, even if the amount of total variance is approximately correct, can have important implications with respect to predicted population responses to management. Consider a case where process variation in the model (e.g. recruitment variation) is incorrectly attributed to increased sampling variation, or vice versa. In such a case, not only would standard assessment output used to make catch advice be affected, but also analyses of risk, including population viability (Maunder, 2004) and management strategy evaluations could be impacted.

Being able to estimate sampling error variance for composition datasets and variances associated with process variation using mixed-effect models presents an opportunity for stock assessment modelling to discontinue iterative reweighting procedures (Thorson, 2019). However, a number of factors concerning this opportunity have not been fully evaluated, such as the estimability of multiple variance terms with standard data available in fisheries assessments, the performance of general methods to account for process variation, the consequences of misspecification, and the effect of the choice of sampling error likelihood for composition data.

The choice of the sampling error likelihood relates to the ability of different likelihoods used for fitting composition data to account for some degree of model error in addition to sampling error. It has been shown that estimating the variance of sampling error for composition data internal to an assessment using the Dirichlet-multinomial or logistic-normal likelihoods can account for some degree of model error when the model contains unmodelled process variation or model misspecification (Fisch *et al.*, 2021). That is, additional misfit between the predictions and data, which is actually due to unmodelled processes, results in an increase in the estimated sampling variability for composition data. The degree of inflated estimated sampling variance can differ based on the observational likelihood chosen and the nature of the model error. For example, the logistic-normal was found to inflate sampling variance more when a large degree of model error existed compared to the Dirichlet-multinomial (conditional on adequate sample sizes), and this led to less biased estimates of quantities important to management (Fisch *et al.*, 2021). For this reason, it has been argued that the estimation of sampling variance for composition data internal to an assessment can make a model more robust to model error (Francis, 2014, 2017; Fisch *et al.*, 2021). However, an alternative, although not a mutually exclusive approach, is to explicitly model additional variation in specific influential processes, which may be giving rise to some of the model error, using random effects. Explicitly modelling process variation allows the effect of the variation to propagate through the population dynamics model and influence the fit to other data sources. In contrast, attributing unaccounted model error as sampling error in the likelihood function only impacts predictions of the data related to the likelihood function (except, of course, through the influence on the parameter estimates). Some studies suggest little downside to estimating time-varying processes when they are not present, and large upside when they are present (Stewart and Monnahan, 2017; Cronin-Fine and Punt, 2021). However, others have suggested more caution, as Privitera-Johnson *et al.* (2022) found that modelling time variation in a process (selectivity in this case) led to worse performance when there was no time variation in that process within the operating model, and Szuwalski *et al.* (2018) showed that modelling variation on the wrong process can lead to poor management advice. Few studies have also considered the variance of the estimated quantity (confidence interval coverage) or overdispersion and correlations in sampling error for composition data, each of which may affect one another (not accounting for overdispersion will lead to overconfident intervals) and could lead to identifiability issues (will one variance account for another).

We pay particular interest to the modelling of composition data since in addition to having a theoretically estimable sampling variance (when fit with certain likelihoods), it provides information on the length/age of fish caught and information

on the abundance of individual cohorts. Therefore, composition data provides information on temporal variation in processes such as selectivity, recruitment, and mortality. Length composition data also provides information on temporal variation in growth. Consequently, accounting for the misfit due to not modelling process variation in the likelihood function may be most relevant for composition data. In addition, the sampling process for composition data is often plagued by pseudo replication due to fish aggregation (e.g. schooling), leading to effective sample sizes that can be much less than the actual sample size even in the absence of model misspecification and unmodelled process variation, requiring the sampling variance to be estimated. For this reason, we were particularly interested in the dynamic between the simultaneous estimation of process variation and sampling error for composition data internal to an assessment model when there exists overdispersion and correlations in the data on top of variation in system's processes and model misspecification.

We structure a general simulation experiment in an attempt to provide insight on the topics presented by assessing the performance of stock assessment models when multiple process variance and sampling variance terms are simultaneously estimated within the model. We specifically examine how assessment performance is affected by the choice of composition likelihood, overdispersion and correlations in composition data, overparameterization, and modelling variation on the wrong process. In doing so, we build a simulation operating and sampling model containing overdispersion and correlations in composition data, directional variation in fishery catchability and/or fishery selectivity, and estimation models, which include random effects (some of which are misspecified) and composition likelihoods with theoretically estimable variances.

Methods

In this study, data are simulated from fine-scale spatially explicit operating models that can produce composition data that are overdispersed and correlated, and fit using spatially aggregated assessment/estimation models. The spatially explicit operating models create variable (potentially time-variable) processes in selectivity and catchability using mechanistic means, and the estimation models attempt to account for the variation in processes using general methods while being blind to the mechanism, in addition to accounting for overdispersion and correlations in sampling error using either the Dirichlet-multinomial or the logistic-normal likelihood for composition data. The goal was to simulate model error based on a mechanism in a way that is not captured in an assessment, to avoid simulating data using some process and fitting them using the same process (Francis, 2012).

We utilize the Red Snapper spatially explicit simulation operating model developed in Fisch *et al.* (2021) with minor adjustments, and develop estimation models using TMB. A full description of the operating model can be found in Appendix A. We briefly summarize the treatments related to the operating model (OM), the sampling model (SM), and those related to the estimation model (EM). We use the word “realized” to describe spatially aggregated qualities of the spatially explicit system. For example, realized catchability refers to the catchability of the whole stock, or the parameter that is commonly used in non-spatial assessments to describe the proportion of the stock caught per unit effort.

Operating model

The spatially explicit age-structured operating model with a spatially dynamic fishery is similar to that presented in Fisch *et al.* (2021) and used many life history characteristics presented in the 2018 Gulf of Mexico red snapper assessment (SEDAR, 2018). The exception to this was for natural mortality (M) of fish from ages 2–20, which was a function of length-at-age (L_a) and set equal to $M_a = 0.099 \frac{L_a}{L_\infty}$ (ages 0 and 1 used M values of 2 and 1.2). This exception was made to allow for its potential estimation within the estimation models. Time series were generated from 40 years of fishing following a 50-year unfished period. The effort time series, which roughly correlates with fishing mortality, increased logistically for the first 75% of the fishing time series length, followed by a linear decrease in effort for the last 25% of the fishing time series.

Operating model treatments

Operating model treatments included whether each of selectivity and catchability was directionally time-varying or not (Table 1). Directionally time-varying selectivity arises from the OM as a function of the gravity model, which determines the spatial allocation of effort. To make selectivity time-invariant, effort distribution in the OM is random with respect to space. Note that the selectivity from year to year will still vary as a function of stochastic fisher effort distribution over space; however, this variability will be random with respect to time (hence, we often make use of the term “directional” variation to distinguish this). The apparent functional form of realized fishery selectivity, which emerged from the OM with a gravity model effort distribution is dome-shaped (Figure 1), due to the dynamic combination of contact selectivity and spatial availability. When a random effort distribution is implemented, this functional form collapses to the contact selectivity, which is logistic, as all fish are made spatially available. When necessary, the dome-shaped functional form is preserved under a random effort distribution scenario in the OM by changing contact selectivity to be dome-shaped according to a six-parameter double-normal selectivity function (Figure 1, Appendix B).

Realized catchability (q) per unit effort, which emerges from the OM where the effort distribution is a function of a gravity model is directionally time-varying if nominal catch-per-unit-effort (CPUE) is calculated. When the effort distribution is made random with respect to space, the realized catchability is not directionally time-varying (Figure 1). To develop a full factorial design involving selectivity and catchability in the OM, we needed to create time-invariant catchability when the effort distribution was a function of a gravity model, and directionally time-varying catchability when the effort distribution was random. To achieve this, we utilized the realized q values from the gravity model OM to simulate a CPUE time series for the random effort OM, and realized q values from the random effort OM to simulate a CPUE time series for the gravity model OM. In total, this resulted in four different OM treatments (Figure 1, Table 1).

Sampling models

Data that were generated from the OM included the annual catch, a fishery-dependent index of abundance, an age composition of the catch, and an index and age composition from a survey available for the second half of the fishing time series. Observed catch (removals) in each year of the time series was simulated by drawing from a normal distribution with mean

Table 1. Simulation experimental design key.

Notation	Description
Operating models	
OM1	Time-invariant selectivity and time-invariant catchability
OM2	Time-invariant selectivity and time-varying catchability
OM3	Time-varying selectivity and time-invariant catchability
OM4	Time-varying selectivity and time-varying catchability
Sampling models	
SM11	Annual sample size ~100, randomly sampled from catch at age
SM12	Annual sample size ~100, low level of cluster sampling (0.0625% of annual effort sampled and 20% of the catch at age sampled from each sampled unit of effort)
SM13	Annual sample size ~100, high level of cluster sampling (0.015625% of annual effort sampled and 80% of the catch at age sampled from each sampled unit of effort)
SM21	Annual sample size ~1000, randomly sampled from catch at age
SM22	Annual sample size ~1000, low level of cluster sampling (0.5% of annual effort sampled and 20% of the catch at age sampled from each sampled unit of effort)
SM23	Annual sample size ~1000, high level of cluster sampling (0.125% of annual effort sampled and 80% of the catch at age sampled from each sampled unit of effort)
SM31	Annual sample size ~10000, randomly sampled from catch at age
SM32	Annual sample size ~10000, low level of cluster sampling (5% of annual effort sampled and 20% of the catch at age sampled from each sampled unit of effort)
SM33	Annual sample size ~10000, high level of cluster sampling (1.25% of annual effort sampled and 80% of the catch at age sampled from each sampled unit of effort)
Estimation models	
EM1111	Constant selectivity, constant q , constant M_a , Dirichlet-multinomial
EM1112	Constant selectivity, constant q , constant M_a , logistic-normal
EM2111	White noise selectivity deviations, constant q , constant M_a , Dirichlet-multinomial
EM2112	White noise selectivity deviations, constant q , constant M_a , logistic-normal
EM4111	2D AR(1) selectivity deviations, constant q , constant M_a , Dirichlet-multinomial
EM4112	2D AR(1) selectivity deviations, constant q , constant M_a , logistic-normal
EM1211	Constant selectivity, white noise q deviations, constant M_a , Dirichlet-multinomial
EM1212	Constant selectivity, white noise q deviations, constant M_a , logistic-normal
EM1311	Constant selectivity, random walk q deviations, constant M_a , Dirichlet-multinomial
EM1312	Constant selectivity, random walk q deviations, constant M_a , logistic-normal
EM1131	Constant selectivity, constant q , random walk \hat{M} deviations, Dirichlet-multinomial
EM1132	Constant selectivity, constant q , random walk \hat{M} deviations, logistic-normal
EM1141	Constant selectivity, constant q , 2D AR(1) M_a deviations, Dirichlet-multinomial
EM1142	Constant selectivity, constant q , 2D AR(1) M_a deviations, logistic-normal

The first digit of the operating model notation describes the combination of whether selectivity and/or catchability were time-varying or not. The first digit of the sampling model notation describes the sample size of composition data and the second digit describes the degree of clustering. For the cluster treatments, the percentage of the total effort sampled and the percentage of the catch-at-age from each individual unit of effort sampled is presented in parentheses. The first digit of the estimation model notation describes the treatment for selectivity in the EM; the second digit describes the treatment for catchability; the third describes the natural mortality treatment; and the fourth describes the composition likelihood.

as the true catch in weight (aggregated across space) and a CV of 5%. Variation was included in the observed fishery index time series by drawing from a normal distribution using the index time series described above as the mean with a CV of 25%. The process that generated the composition and index of abundance from the survey was largely independent of spatial structure, included a constant asymptotic selectivity, and is described in further detail within [Appendix A](#).

Sampling model treatments

The SM included two treatments related to sampling the age composition of the catch: the annual sample size of the composition data and the degree of clustering in the composition data ([Table 1](#)). The two SM treatments (sample size and clustering) were created by varying the percentage of the annual total effort sampled and the percentage of the catch-at-age sampled per unit of effort sampled. More detail on this process is given in [Appendix A](#) and the specific percentages for each sampling model treatment are presented in [Table 1](#). Each treatment included three levels. For the degree of clustering in composition sampling, levels varied from no clustering to low and high degrees of clustering ([Figure 2](#)). The “no cluster-

ing” level for each sample size treatment level was simulated by randomly sampling from the annual catch-at-age using the multinomial distribution with the annual sample size equal to that from the low clustering level. For the sample size treatment, levels varied from ~100 to 1000 to 10000 annual fish aged for most of the time series to make up fishery compositions ([Figure 2](#)). These levels were chosen to cover a range of orders of magnitude. At the largest level of clustering, the effective sample size equaled ~20% of the true sample size using the Francis (2011)-TA1.8 algorithm (Fisch *et al.*, 2021). Thus, the high clustering treatment still contained a greater information content than the previous sample size level with “no clustering”. In total the sampling model yielded nine unique treatments and levels ([Table 1](#)). One hundred simulation replicates of each OM–SM combination were run and fit by each estimation model described in the following section. Zeroes were suppressed in the age composition data by adding a small constant (1E-5) and renormalizing the vector to sum to one.

Estimation models

We begin by describing the baseline estimation model with process variation only included in recruitment, followed by a

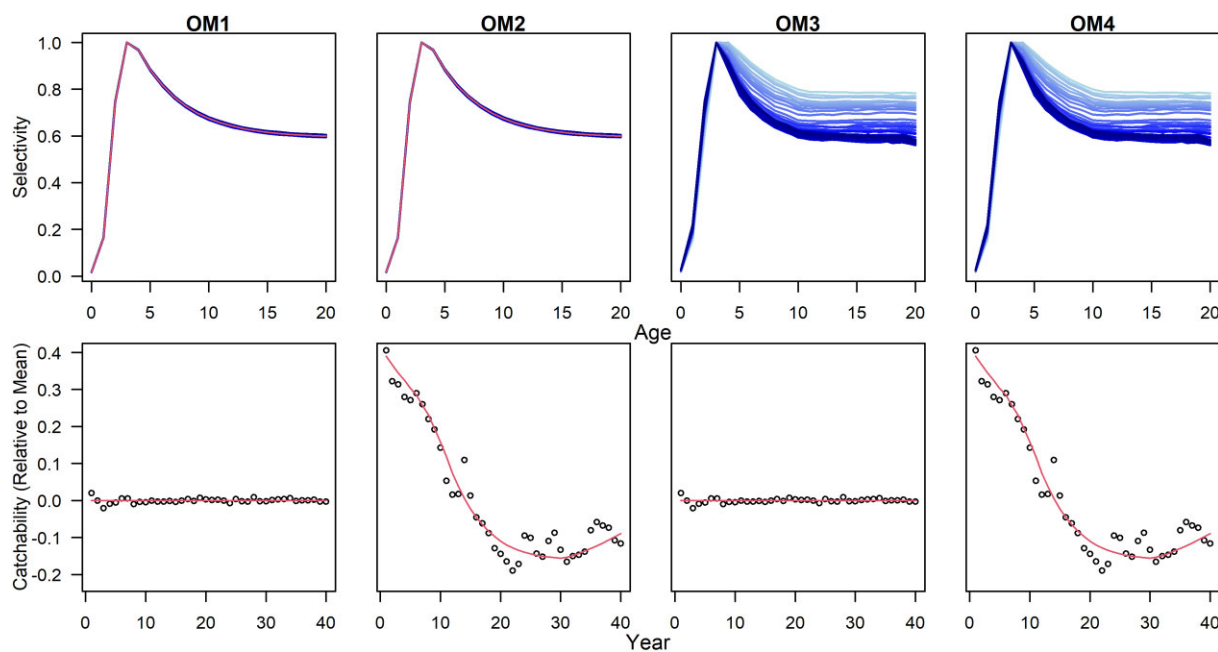


Figure 1. Depiction of OM treatments. The top row depicts the realized selectivity, columns one and two showing results from an OM with a random effort distribution where the contact selectivity is changed to a dome shape. The top row of columns three and four depicts the realized selectivity from an OM with a gravity model effort distribution. These plots represent a single simulation iteration. Lighter blue signifies years at the beginning of the time series, which become darker in later years. The bottom row shows the catchability for each OM treatment [normalized to the mean of each time series $(q_y - \text{mean}(q_y)) / \text{mean}(q_y)$]. Depicted in red are results from OMs with no stochasticity in fish movement and fisher effort distribution in each year, to illustrate the distinction between process variation and directional time variation. For the purposes of visualization, the selectivity results from an OM with no stochasticity in fish movement and fisher effort distribution are not depicted for OMs 3–4 (top right two panels).

full description of the implementation of the process variation treatments.

The estimation models are age-structured assessments that run for 40 years (fishing time series from OM) modelling ages 0–20+. The models are fit to five sources of data: (1) the fishery catch, (2) a fishery index of abundance, (3) a fishery catch age composition, (4) a survey index of abundance, and (5) a survey age composition. The models begin the time series at an unfished state (through an estimated unfished recruitment parameter and known natural mortality at age). Recruitment is estimated each year using the Beverton–Holt stock recruitment function, with steepness fixed at 0.99 (consistent with OM) and annual lognormal recruitment deviations about the median value estimated as random effects on the log scale with the variance in recruitment estimated. Recruitment deviations are also estimated for cohorts that make up the initial abundance at age.

The functional form of fishery selectivity is estimated using an age-based six-parameter double-normal function [Methot and Wetzel, 2013—Supplementary Material Equations (A1.30–1.34)]. Fishing mortality for each age in each year is calculated as the product of fishery selectivity and a fully selected fishing mortality parameter estimated for each year.

Symbology is described in Table 2 and the observation model is presented in Table 3. The predicted fishery index is calculated by multiplying a fishery catchability parameter, which is estimated, by the exploitable biomass over the year [Equation (2.4)]. Fishery catchability is solely used in the observation model and not for the calculation of fishing mortality (the estimation models are not given an effort time series).

Likelihoods

The fishery catch, index, and survey index were fit with normal likelihoods using CVs on their original scale (i.e. not log-transformed; Table 2). The CVs for the fishery catch and index were fixed at their correct values from the OM, while the survey index CV was estimated. The composition data for both the fishery and the survey were either modelled with a Dirichlet-multinomial likelihood or a logistic-normal likelihood, presented in detail in the EM treatments section.

Estimation model treatments

Selectivity

Two additional treatments beyond the time-invariant functional form of selectivity were considered. First, we implemented white-noise iid lognormal deviations about the functional form of selectivity for each age and year.

$$s_{a,y} = s_a \times e^{(\varepsilon_{s,a,y})}, \quad \varepsilon_{s,a,y} \sim N(0, \sigma_s^2),$$

where s_a denotes the estimated values from the double-normal selectivity function, $\varepsilon_{s,a,y}$ denotes the estimated age and year-specific fishery selectivity deviation, and σ_s^2 denotes the estimated process variance for the fishery selectivity deviations. As a simplifying assumption to reduce computation time, we only estimate selectivity deviations for ages 2–10 and assume that $\varepsilon_{s,a,y} = \varepsilon_{s,2,y}$ for $a < 2$ and $\varepsilon_{s,a,y} = \varepsilon_{s,10,y}$ for $a > 10$, similar to Xu *et al.* (2019).

For the second treatment, we implemented a 2-dimensional (2D) AR(1) selectivity formulation developed in Xu *et al.* (2019) where $\varepsilon_{s,a,y}$ are correlated by age (a) and year (y) ac-

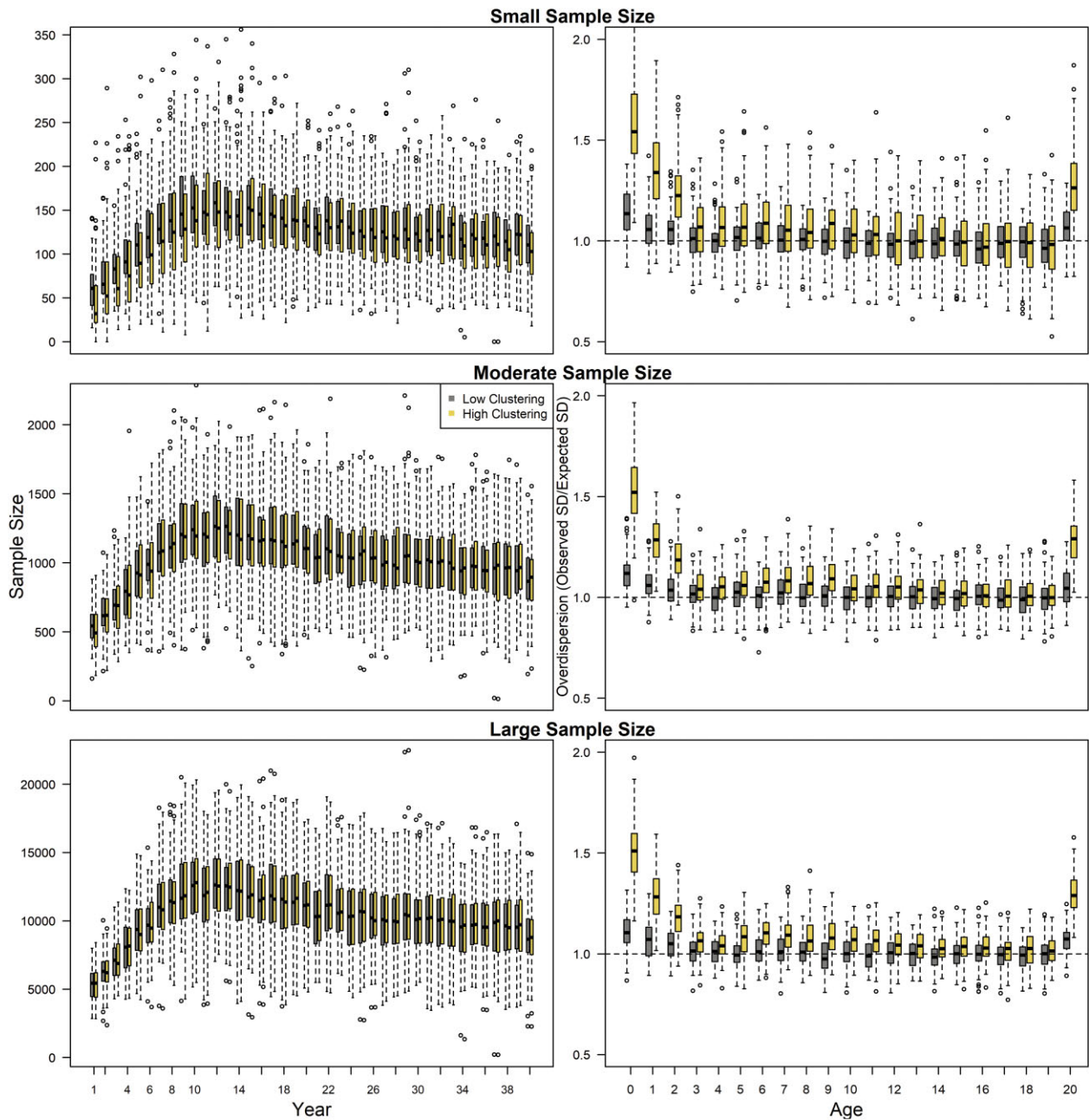


Figure 2. Sampling model treatments, including the sample size and the degrees of overdispersion. The left column depicts annual sample sizes for the three levels of sample size and two levels of clustering (low and high). The right column depicts the degrees of overdispersion as the observed standard deviation for each age bin across bootstrapped sampling replicates, divided by the expected standard deviation, both regarding year 20 in the time series. The boxplots depict results over simulation iterations for the gravity model fishing scenario. The “no clustering” treatment had the same sample size as the “low clustering” treatment.

cording to

$$\varepsilon_s \sim \text{MVN}(0, \sigma_s^2 \mathbf{R}_{s\text{Total}}),$$

where $\mathbf{R}_{s\text{Total}}$ represents a correlation matrix and is equal to the Kronecker product of two separate correlation matrices, one for among age correlations (\mathbf{R}) and another for among year correlations ($\hat{\mathbf{R}}$), each calculated using AR(1) processes.

$$\mathbf{R}_{s\text{Total}} = \mathbf{R} \otimes \hat{\mathbf{R}} \text{ where } R_{a,a'} = \rho^{|a-a'|} \text{ and } \hat{R}_{y,y'} = \hat{\rho}^{|y-y'|},$$

where ρ and $\hat{\rho}$ represent estimated correlation parameters for an AR(1) process for difference in ages ($|a - a'|$) and

years ($|y - y'|$), respectively. Deviations were also only estimated for ages 2–10 ($\varepsilon_{s,a,y} = \varepsilon_{s2,y}$ for $a < 2$ and $\varepsilon_{s,a,y} = \varepsilon_{s10,y}$ for $a > 10$) as in the white noise fishery selectivity treatment.

The first form of accounting for time-varying selectivity was chosen based on its generality and lack of directional variation. The second was chosen as this formulation has performed well in simulation analyses (Xu *et al.*, 2019, 2020) and has been recommended as a best practice for incorporating time-varying selectivity in assessment models (Privitera-Johnson *et al.*, 2022).

Table 2. Descriptions of parameters/symbols regarding the estimation models, and whether they were estimated or fixed.

Symbol	Description	Estimated or Fixed (bounds)
a	Age subscript	NA
y	Year subscript	NA
b	Steepness	Fixed – 0.99
\hat{M}	Natural mortality scalar	Fixed – 0.099
R_0	Unfished recruitment	Estimated ln scale (10, 25)
εR_y	Recruitment deviations	Estimated ln scale (–10, 10)
B_1	Double-normal selectivity parameter 1	Estimated (–10, 20)
B_2	Double-normal selectivity parameter 2	Estimated (–10, 20)
B_3	Double-normal selectivity parameter 3	Estimated (–10, 20)
B_4	Double-normal selectivity parameter 4	Estimated (–10, 20)
B_5	Double-normal selectivity parameter 5	Estimated (–10, 20)
B_6	Double-normal selectivity parameter 6	Estimated (–10, 20)
v	Survey logistic selectivity slope	Estimated (–2, 5)
u	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	Estimated ln scale (–10, 1)
CV_I	Fishery index CV	Fixed – 0.25
CV_Q	Survey index CV	Estimated ln scale (–5, 2)
f_y	Fully selected fishing mortality	Estimated ln scale (–20, 0)
θ	Weighting parameter for the DM (two parameters; one fishery, one survey)	Estimated ln scale (–10, 20)
σ_{AR1}	Logistic-normal AR1 SD (two parameters; one fishery, one survey)	Estimated ln scale (–5, 5)
φ	Logistic-normal AR1 Phi (two parameters; one fishery, one survey)	Estimated logit (–1, 1) scale (–1e3, 1e3)
σ_s	SD of white noise or 2D AR(1) age- and year-correlated fishery selectivity deviations	Estimated ln scale (–10, 1)
ρ	2D AR(1) selectivity correlation parameter for age	Estimated logit (0,1) scale (–5, 5)
$\hat{\rho}$	2D AR(1) selectivity correlation parameter for year	Estimated logit (0,1) scale (–5, 5)
$\varepsilon s_{a,y}$	Fishery selectivity deviations	Estimated ln scale (–10, 10)
σ_q	SD of white noise or random walk catchability	Estimated ln scale (–10, 2)
$\varepsilon q_{a,y}$	Fishery catchability deviations	Estimated ln scale (–10, 10)
σ_M	SD of random walk or 2D AR(1) natural mortality deviations	Estimated ln scale (–10, 1)
εM_y	Natural mortality scalar deviations	Estimated ln scale (–10, 10)
ρ_M	2D AR(1) natural mortality correlation parameter for age	Estimated logit (0,1) scale (–5, 5)
$\hat{\rho}_M$	2D AR(1) natural mortality correlation parameter for year	Estimated logit (0,1) scale (–5, 5)
$\varepsilon M_{a,y}$	Age- and year specific natural mortality deviations regarding the 2D AR(1) formulation	Estimated ln scale (–10, 10)
Fec_a	Fecundity at age	Fixed—OM values (Supplementary Table 2)
SB_0	Unfished spawning biomass	Function of R_0 , $M_{a,y}$ and Fec_a
L_a	Length at age	Fixed—OM values (Supplementary Table 3)
Wt_a	Weight at age	Fixed—OM values (Supplementary Table 3)
$\tilde{N}_{a,y}$	Mean numbers at age	Function of $N_{a,y}$ and $Z_{a,y}$ [Equation (2.4)]
H_y	Observed harvest	Data
I_y	Observed fishery index	Data
Q_y	Observed survey index	Data
SS_y	The number of fish aged in a given year	Data
$P_{a,y}$	Observed fishery composition	Data
$G_{a,y}$	Observed survey composition	Data
b	Number of bins in a composition data set	21

If estimated the bounds of estimation are identified in parentheses. Fixed parameters were assumed known at their specified values in the OM.

Fishery catchability

We considered two additional fishery catchability treatments in addition to a constant fishery catchability. The first estimated annual lognormal white noise iid deviations about the mean fishery catchability (an estimated parameter).

$$q_y = q \times e^{(\varepsilon q_y)}, \quad \varepsilon q_y \sim N(0, \sigma_q^2),$$

where εq_y denotes the fishery catchability deviation for a given year and σ_q^2 denotes the process variance of the fishery catchability deviations. The second treatment modelled q as a random walk.

$$q_{y+1} = q_y \times e^{(\varepsilon q_y)}, \quad \varepsilon q_y \sim N(0, \sigma_q^2),$$

where the initial year fishery catchability value is estimated.

The two forms of accounting for time-varying fishery catchability were chosen again because of their generality (white noise deviations), and the random-walk formulation has performed well in simulation studies compared to a suite of other methods (Labelle, 2005; Wilberg and Bence, 2006), leading to a recommendation as a default approach in the absence of prior information (Wilberg *et al.*, 2009). Note that time-varying fishery catchability is applied only to the index of abundance from the fishery. The survey catchability was assumed constant in all models (z in [Tables 2](#) and [3](#)).

Table 3. Estimation model equations.

Quantity		Equation
Process model		
1.1	Abundance at age	$N_{a,y} = \begin{cases} R_y & \text{if } a = 0 \\ N_{a-1,y-1} e^{-(F_{a-1,y-1} + M_{a-1})} & \text{if } 1 \leq a < 20 \\ N_{a-1,y-1} e^{-(F_{a-1,y-1} + M_{a-1})} + N_{a,y-1} e^{-(F_{a,y-1} + M_a)} & \text{if } a = 20+ \end{cases}$
1.2	Recruitment	$R_y = \frac{4bR_0SB_y}{SB_0(1-b) + SB_y(Sb-1)} e^{\varepsilon_y} \quad \varepsilon_y \sim N(0, \sigma_R^2)$
1.3	Spawning biomass	$SB_y = \sum_a N_{a,y} Fec_a$
1.4	Natural mortality	$M_a = \hat{M} \frac{L_z}{L_\infty}$
1.5	Fishery selectivity (double-normal)	$s_{a,y} \text{—see Appendix B}$
1.6	Fishing mortality	$F_{a,y} = s_{a,y} f_y$
Observation model		
2.1	Predicted catch-at-age	$\hat{C}_{a,y} = \frac{F_{a,y}}{F_{a,y} + M_a} N_{a,y} [1 - e^{-(F_{a,y} + M_a)}]$
2.2	Predicted fishery harvest	$\hat{H}_y = \sum_a \hat{C}_{a,y} W t_a$
2.3	Predicted fishery composition	$\hat{P}_{a,y} = \frac{\hat{C}_{a,y}}{\sum_a \hat{C}_{a,y}}$
2.4	Predicted fishery index	$\hat{I}_y = q \sum_a \tilde{N}_{a,y} W t_a s_a \text{ where } \tilde{N}_{a,y} = \frac{N_{a,y} [1 - e^{(-Z_{a,y})}]}{Z_{a,y}}$
2.5	Survey selectivity	$g_a = \frac{1}{(1 + e^{(-v(a-u))})}$
2.6	Predicted survey index	$\hat{Q}_y = z \sum_a g_a N_{a,y}$
2.7	Predicted survey composition	$\hat{G}_{a,y} = \frac{g_a N_{a,y}}{\sum_a g_a N_{a,y}}$
Negative log likelihoods		
3.1	Fishery harvest	$\sum_y \ln(CV_H \times \hat{H}_y) + 0.5 \left(\frac{H_y - \hat{H}_y}{CV_H \times \hat{H}_y} \right)^2$
3.2	Fishery CPUE	$\sum_y \ln(CV_I \times \hat{I}_y) + 0.5 \left(\frac{I_y - \hat{I}_y}{CV_I \times \hat{I}_y} \right)^2$
3.3	Fishery-independent survey CPUE	$\sum_y \ln(CV_Q \times \hat{Q}_y) + 0.5 \left(\frac{Q_y - \hat{Q}_y}{CV_Q \times \hat{Q}_y} \right)^2$
3.4	Recruitment deviations	$\sum_y \ln(\sigma_R) + 0.5 \left(\frac{\varepsilon_R}{\sigma_R} \right)^2$
3.5	Process variations—normal distribution	$\sum_y \ln(\sigma_j) + 0.5 \left(\frac{\varepsilon_j}{\sigma_j} \right)^2,$
3.6	Process variations—multivariate normal [2D AR(1) selectivity or natural mortality deviations]	where j denotes a process variation treatment $0.5 [\ln(\Sigma_s) + \varepsilon_s' \Sigma_s^{-1} \varepsilon_s] \text{ where } \Sigma_s = \sigma_s^2 \mathbf{R} s_{Total}$ $0.5 [\ln(\Sigma_M) + \varepsilon_M' \Sigma_M^{-1} \varepsilon_M] \text{ where } \Sigma_M = \sigma_M^2 \mathbf{R} M_{Total}$
3.7	Dirichlet-multinomial (DML) ¹	$- \sum_y \left[\log(\Gamma(SS_y + 1)) - \sum_a (\log(\Gamma(SS_y P_{a,y} + 1))) + \log(\Gamma(SS_y \theta)) \right. \\ \left. - \log(\Gamma(SS_y + SS_y \theta)) + \sum_a (\log(\Gamma(SS_y P_{a,y} + SS_y \theta \hat{P}_{a,y})) - \log(\Gamma(SS_y \theta \hat{P}_{a,y}))) \right]$
3.8	Logistic-normal (LN) ^{1,2}	$\sum_y \left[0.5(b-1) \log(2\pi) + \sum_a [\log(P_{a,y})] + 0.5 \log(V_y) \right. \\ \left. + (b-1) \log(W_y) + \frac{(x_y^T V_y^{-1} x_y)}{2W_y^2} \right]$

¹Note that composition likelihoods are also used for the survey compositions, thus $\hat{G}_{a,y}$ and $G_{a,y}$ can be substituted for $\hat{P}_{a,y}$ and $P_{a,y}$.

²Additional details regarding the formula for the negative log-likelihood of the logistic-normal can be found in [Appendix C](#).

Natural mortality

We chose to include process variation in natural mortality within the estimation models to provide additional insight on the consequences of specifying variation on the incorrect process. We incorporated two formulations of process variance on natural mortality. The first was a random walk parameterization of annual deviations on the scalar parameter of natural mortality (\hat{M}), similar to that used for catchability (where the initial year $\hat{M}_1 = \hat{M}$).

$$\hat{M}_{y+1} = \hat{M}_y \times e^{(\varepsilon_{M_y})}, \quad \varepsilon_{M_y} \sim N(0, \sigma_M^2).$$

The second was a 2D AR(1) formulation of age- and year-specific deviations on natural mortality at age similar to that

modelled on fishery selectivity.

$$M_{a,y} = M_a \times e^{(\varepsilon_{M_{a,y}})}, \quad \varepsilon_M \sim MVN(0, \sigma_M^2 \mathbf{R} M_{Total}),$$

where $\mathbf{R} M_{Total}$ is again the Kronecker product of individual correlation matrices for age and year. We again chose to only model deviations for ages 2–10 as in the fishery selectivity formulation and thus $\varepsilon_{M_{a,y}} = \varepsilon_{M_{2,y}}$ for $a < 2$ and $\varepsilon_{M_{a,y}} = \varepsilon_{M_{10,y}}$ for $a > 10$.

Note that the scalar parameter (\hat{M}) in each natural mortality formulation was not estimated and was fixed at its OM value. Thus, for unfished calculations within the estimation model, the correct natural mortality ogive was used.

Composition likelihoods

Two levels of composition likelihood were tested: the Dirichlet-multinomial and the logistic-normal. We utilized the linear parameterization of the Dirichlet-multinomial (Thorson *et al.*, 2017), and the first-order autoregressive parameterization of the additive logistic-normal (Aitchison, 2003; Francis, 2014). We abbreviate the linear formulation of the Dirichlet-multinomial hereafter in the text using “DM”. This formulation of the likelihood results in an effective sample size for composition data that scales linearly with the true sample size and estimates a single additional parameter per composition dataset meant to increase the expected variation about residuals (above that implied by the sample size alone) to account for overdispersion. The true number of age samples was used for SS_y in the Dirichlet-multinomial likelihood [Table 3, Equation (3.7)].

We abbreviate the first-order autoregressive parameterization of the logistic-normal hereafter in the text using “LN”. A composition conforms to a logistic-normal distribution with parameters $[P, C]$ when $P_a = \frac{e^{T_a}}{\sum_a e^{T_a}}$. In this case, T conforms

to a multivariate normal distribution with mean $\log(P)$ and covariance matrix C . For this reason, the logistic-normal cannot accommodate zeroes in composition data. However, the logistic-normal is theoretically able to account for correlations in residuals between bins by virtue of the variance-covariance matrix regarding the multivariate normal distribution. The formulation tested in this study implements a first-order autoregressive [AR(1)] parametrization of the variance-covariance matrix. This parameterization necessitates two parameters per composition data source: σ_{LN} and φ . Differential weighting between years based on composition sample size was achieved $W_y = \sqrt{\bar{SS}_{SS_y}}$, where \bar{SS} denotes the mean sample size over the time series and $\sigma_{LN,y} = \sigma_{LN} \times W_y$, as in Francis (2014). The variance-covariance matrix in each year, C_y , was calculated according to $C_{y,a,a'} = \sigma_{LN,y}^2 \rho_{LN|a-a'|}$, where $\rho_{LN|a-a'|} = \varphi^{|a-a'|}$ for an AR(1) process.

Note that each likelihood assumes overdispersion is constant across ages, whereas the sampling model clustering in composition data simulates overdispersion that does vary between ages (Figure 2).

Model running

A full factorial of OM-SM treatments resulted in 36 combinations (data files) with 100 simulation replicates of each, to be fit by 14 unique EMs (Table 1). We did not implement a full factorial cross of EM process variation treatments due to time and computational constraints. Each model was fit in TMB with estimated parameters and bounds identified in Table 2. TMB first calculates the marginal negative log-likelihood given the fixed effect parameters using the Laplace approximation to integrate over random effects. Fixed effect parameters are then estimated via minimizing the marginal negative log-likelihood within the programme R using the nlminb function. Random effects and derived quantities are then predicted using empirical Bayes (Kristensen *et al.*, 2016). All deviations from process variation treatments, in addition to those from recruitment variation were treated as random effects. Variance parameters for process variations and composition likelihoods were estimated for all treatments. The two correlation parameters associated with each 2D AR(1) param-

eterization were constrained between 0 and 1 using a logit transformation. This constraint was implemented similar to Xu *et al.* (2019) largely based in biological plausibility, that we would expect selectivity and natural mortality deviations to be positively correlated with age and year (H. Xu, pers. comm), as fish of similar ages tend to behave similarly, fishing practices are likely to gradually evolve over time, and environmental or ecosystem conditions affecting natural mortality are highly likely to be autocorrelated. We decided not to implement restricted maximum likelihood estimation (REML) to correct for negative biases when estimating variances using maximum likelihood (Cheang and Reinsel, 2000; Miller *et al.*, 2018; Thorson, 2019) as an iterative approach accompanies this method (Xu *et al.*, 2019) and our goal was not to fix variances (and ideally to move away from iterative approaches). Thus, there is likely to be a small negative bias in our variance estimates. We do implement the epsilon bias correction method of Thorson and Kristensen (2016) for point estimates of derived quantities resulting from each model. We assessed convergence based on whether the hessian matrix for a given model run was positive-definite. If a model did not converge, initial parameter values were jittered, and the estimation model was fit again. This process was repeated a maximum of five times, after which a model was considered not converged. For models that converged for <25% of the simulation iterations, we do not present results other than noting convergence.

Performance metrics

We focused on the ability of the estimation models to estimate depletion (SB_{41}/SB_0) and spawning stock biomass in the terminal year of each EM by assessing the relative error over simulations [Relative error = (Estimated-True)/True]. We also examined the confidence interval coverage of point estimates for each using 95% Wald confidence intervals calculated with standard errors obtained from the generalized delta method in TMB. Coverage was assessed as the percentage of iterations from converged models where the true value was within the estimated interval.

Questions addressed by case comparisons

In the interest of brevity, we focus our synthesis of the results on a few key comparisons across treatments.

What is the impact of the composition likelihood chosen?

Here we examine all cases of OM-SM-EM combinations, comparing the performance of the DM to the LN. This included situations where the estimation models only considered variation in recruitment (the baseline EM). We pay particular attention to the impact of fishery composition sample sizes, degrees of clustering in fishery composition data, and potential differences in parameter confounding between the likelihoods.

How do the general methods for process variation perform?

Here we compare the methods implemented that attempt to account for process variation and pay particular attention to cases where processes from the OM were directionally time-

varying and the EMs attempted to account for it by specifying variation on the correct process.

Are process and sampling error variances identifiable and estimable?

Here we examine the estimation of the process variance terms for all model formulations but pay particular attention to cases where the true process was variable, and the correct estimation process was specified as variable. We examine whether the sampling variance and the process variation were confounded. We also note the number of converged models.

What if we model variation on the wrong process?

Here we examine situations where there was time-variation in a process within the OM; however, a different process included the variation within the EM, including the EMs with variation specified on natural mortality.

What if we model variation and there is none?

Here we pay attention to cases where variation in the OM was minimal in selectivity and catchability (OM1); however, the estimation models attempted to estimate variation in each. Were they successful in minimal estimates of the process variances? We also pay particular attention to the estimation models which considered variation in natural mortality as there was no variation specified for this in the OM.

Sensitivity analyses

We evaluated some additional models as sensitivities. These included baseline estimation models utilizing the multivariate Tweedie likelihood (Thorson *et al.*, 2022) for fitting the composition data and baseline estimation models fit to OM–SM combinations where the time series of data was extended from 40 to 80 years. The multivariate Tweedie likelihood used was as formulated in Thorson *et al.* (2022). We also ran sensitivity EMs that estimated both 2D AR(1) selectivity deviations and process variation in natural mortality [using either the random walk or 2D AR(1) formulation]. These models were only run for OMs 3–4 at the largest sample sizes to evaluate whether time-varying selectivity could be teased apart from time-varying natural mortality. We also ran EMs using the logistic-normal for fitting composition data however without relative weighting between years based on the composition sample size ($W_y = 1$). Finally, we ran EMs with 2D AR(1) process variance on either selectivity or natural mortality where we allowed the correlation parameters of the 2D AR(1) to be bounded between -1 and 1 (as opposed to $0-1$). Results for the sensitivity models can be found in the [supplemental material \(Supplementary Figures 3–33\)](#) and are briefly presented and discussed in the results and discussion.

Results

In what follows, we describe general results and patterns from each estimation model treatment regarding the prompts listed above. In general, EMs most often exhibited a negative bias in depletion and spawning stock biomass in the terminal year, and very few treatments exhibited a positive bias in confidence interval coverage.

What was the impact of the composition likelihood chosen?

The likelihood chosen for composition data was consequential, and the only situations in which the LN outperformed the DM were at the largest sample sizes, and for only a subset of metrics and OM treatments (Figures 3–6). Consequently, we focus most of our synthesis of results on the EMs that fit composition data with the DM. When we do reference results of EMs fit with the LN, we mostly focus on the largest sample size level.

There were, however, a few notable differences between the two likelihoods for different OM–SM–EM treatments. The LN tended to improve in performance as the sample size increased across all EM–OM combinations, where the DM performance tended to worsen specifically for OMs 3–4 [a notable exception being the 2D AR(1) selectivity EM, which improved as sample size increased]. When fishery selectivity directionally varied in the OM (OMs 3–4) and was unaccounted for in the EM, the LN outperformed the DM in most metrics at the largest sample size level. Conversely, when the process variation was modelled on selectivity in the EM and it directionally varied in the OM, the DM outperformed the LN. Lastly, when process variation was placed on natural mortality in the EM and the OM exhibited directional variation in selectivity (OMs 3–4), the LN was more robust to this misspecification than the DM at the largest sample size level.

In addition, overall, the degree of clustering in composition data did not seem to greatly affect results regarding bias, although generally, more clustering led to increased negative bias in confidence interval coverage for the DM models.

How do the methods for modelling process variation perform?

For the OM treatment that included directionally time-varying catchability and constant selectivity (OM2), the random walk parameterization of catchability variation had the best performance, estimating the least amount of bias for terminal year estimates and some of the least negatively biased confidence interval coverage (Figures 3–6). The white noise parameterization of catchability process variance improved performance compared to the baseline; however, it was more biased than that which modelled catchability with a random walk.

For the OM treatment that included directionally time-varying selectivity and a constant catchability (OM3), the EM with 2D AR(1) selectivity deviations was least biased in terminal year estimates and confidence interval coverage when the sample size was greater than the lowest level. At the lowest sample size level, all EMs had similar levels of bias.

When the OM included both directionally time-varying selectivity and catchability (OM4), the EM with random walk catchability was the least biased in terminal year estimates and confidence interval coverage for all but the largest sample sizes. At the largest sample size level, the EM with 2D AR(1) selectivity deviations was least biased in terminal year estimates and confidence interval coverage.

Each of the above points is true for the DM models. For LN models, the points made at the largest sample size level remain true with the exception that the random walk catchability outperformed the 2D AR(1) selectivity for OM4.

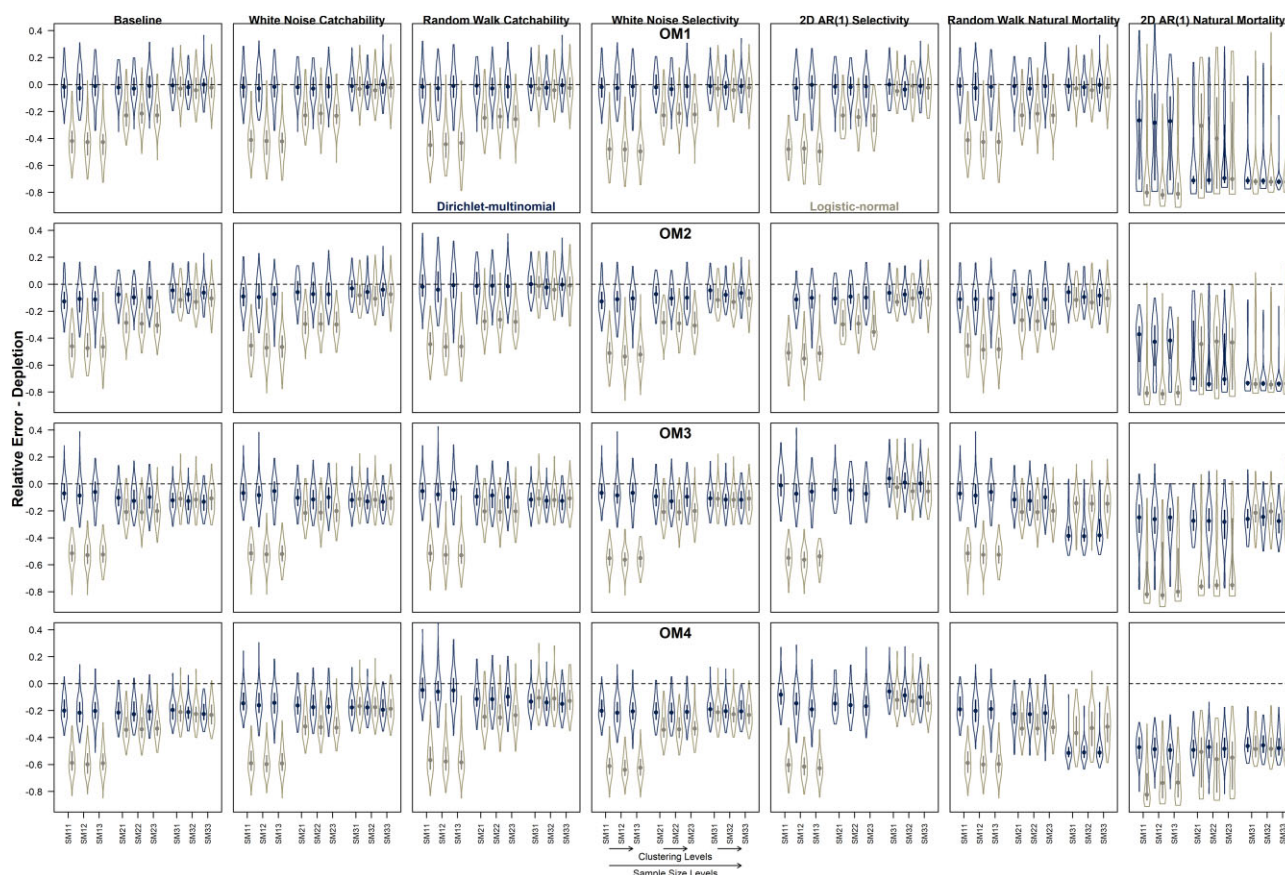


Figure 3. Relative error of terminal-year estimates of depletion from the estimation models. The columns depict different EM process variance treatments where the rows depict different OM treatments. Shown in blue are the results from the DM and in tan those from the LN. The x-axis is presented as three groups of increasing sample size, within which are the three levels of increasing clustering. The violin plots depict an approximation of the density of the distributions across simulations, and the point denotes the median.

Were multiple process and sampling error variances identifiable and estimable?

For the most part, the process variance terms were estimable, although it was dependent on the process variation modelled, the sample size of the composition data, and the choice of likelihood for fishery composition data. Recruitment variation was estimated well in nearly all cases for the DM models. The exceptions to this were when process variation was also modelled on M . EMs with random walk M at the largest sample sizes overestimated recruitment variation for OM3–4, and EMs with a 2D AR(1) parameterization of M variation had highly variable and poor estimates of recruitment variation for fits to OM1–2 (Figure 7). For models fit with the LN, at small sample sizes the estimates of the standard deviation of recruitment were positively biased across all OM-SM-EM treatments. As expected, when the standard deviation of recruitment was estimated well there was a minute negative bias.

For OM treatments with directionally time-varying catchability (OM2 and OM4), estimates of the standard deviation of catchability for the EMs with either white-noise catchability or random walk catchability were consistent across sample sizes and greater than zero for models fit with the DM (Figure 8). For EMs fit with the LN, the estimates were similar to the DM; however, only at sample sizes above the lowest level.

For OM treatments with directionally time-varying selectivity (OMs 3–4) fit with the DM, estimates of the standard deviation parameter for the white noise selectivity EM treat-

ment were not identified well at small sample sizes; however, precision across simulations improved as the sample size of the composition data increased. The estimates also increased as sample clustering in composition data increased (and more pronounced for small and moderate sample sizes than at large). At the small sample size level, the estimates tended to zero (exception being the high clustering level), where at moderate and large sample sizes the estimates were consistently greater than zero (Figure 8). The similarly parameterized LN models only estimated the standard deviation parameter greater than zero at the largest sample sizes, and the estimates were similar across levels of clustering. For the EMs parameterized with 2D AR(1) selectivity deviations fit with the DM, the SD estimates for the fishery selectivity deviations were greater than zero and generally larger for greater levels of clustering when fit to OM3–4. The precision of the estimates across simulations again improved with sample size. For EMs fit with the LN, this parameter was greater than zero again only at the largest sample size levels. Estimates of the correlation parameters in the 2D AR(1) selectivity were more variable, and for the DM EMs, the age correlation parameter was not identified well at smaller sample sizes and approached one as the sample size increased to the highest level. The year correlation parameter estimates tended to decrease as the level of clustering increased and for small and moderate sample sizes were not identified very well (variable across simulations). At the largest sample sizes, this parameter was estimated near one

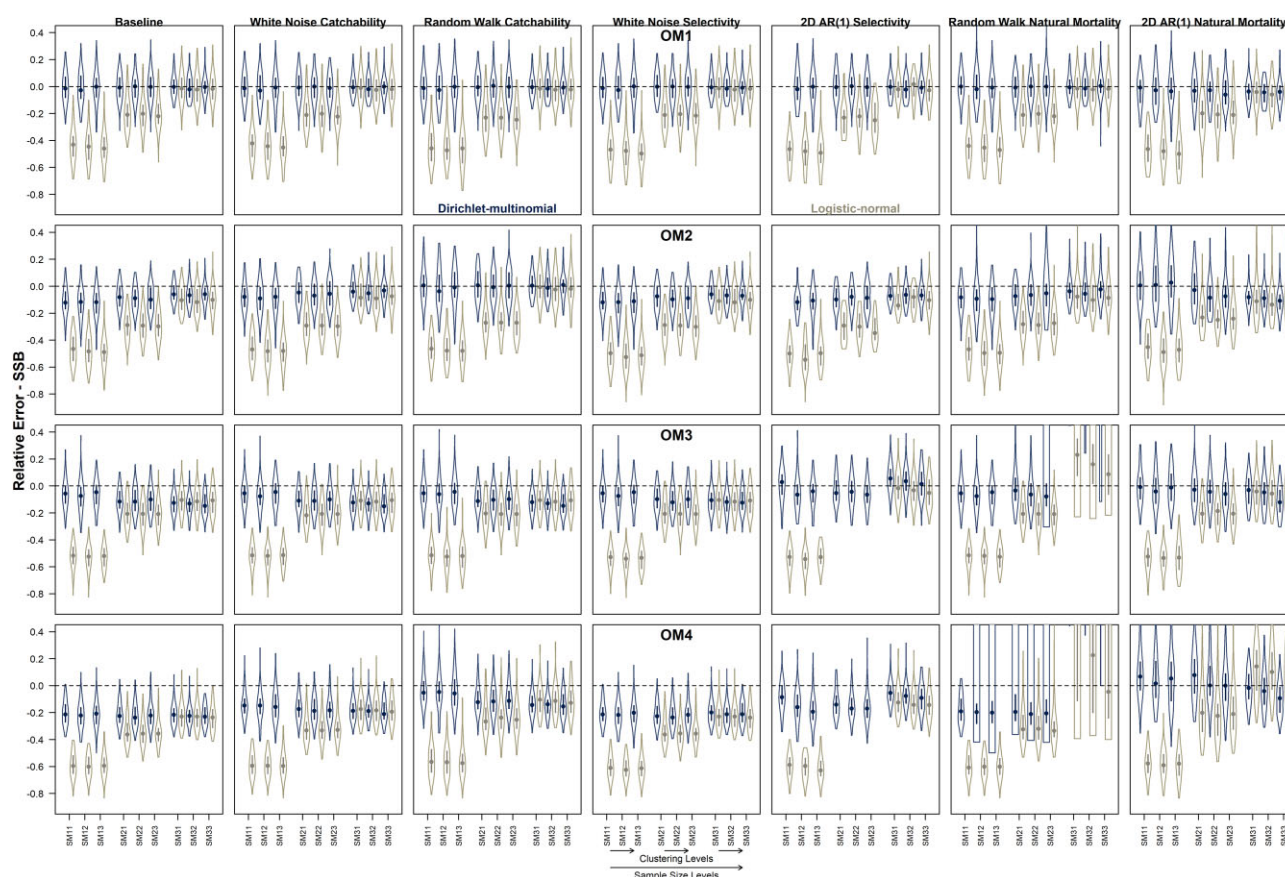


Figure 4. Relative error of terminal-year estimates of spawning stock biomass from the estimation models. The columns depict different EM process variance treatments where the rows depict different OM treatments. Shown in blue are the results from the DM and in tan those from the LN. The x-axis is presented as three groups of increasing sample size, within which are the three levels of increasing clustering. The violin plots depict an approximation of the density of the distributions across simulations, and the point denotes the median.

indicating the model is estimating a large degree of correlation in selectivity over years (coupled with *SD* estimates greater than zero suggests the model is accounting for time-variation in selectivity). For the LN, the year correlation parameter approached one at all sample sizes; however, the age correlation parameter approached one only for the largest sample sizes. The number of converged models for the 2D AR(1) selectivity parameterization was generally the lowest across all treatments and was only at levels comparable to the other EMs at the largest sample sizes, for both EMs fit with the DM and LN (Figure 9). Notably for models fit with the DM, the number of converged models for the 2D AR(1) selectivity parameterization increased as clustering in composition data increased for all OM and increased as the sample size increased specific to OM with directionally time-varying selectivity (OMs 3–4).

The sampling variance parameters for each likelihood were estimable (for converged models), although estimates differed depending on the EM process variation treatment, the sample size of the composition data, and the degree of clustering in composition data. For example, in the baseline EM, estimates of the fishery composition overdispersion parameter for the DM likelihood decreased (i.e. estimating more overdispersion) as the degree of clustering increased and also as the sample size of compositions increased, particularly for fits to OM with directional variation in selectivity (Supplementary Figure 1). Similarly, the estimated correlation parameter for the AR(1) parameterization of the LN

likelihood increased as the degree of clustering in composition data increased and as sample size increased to its highest level when there was directionally time-varying selectivity in the OM (OMs 3–4). Conversely, for EMs with 2D AR(1) selectivity deviations, estimates of the overdispersion parameter for fishery composition data were now much more consistent across sample sizes for OMs 3–4 and the correlation parameter for LN AR(1) did not increase as the sample size increased for OMs 3–4 (although both still included clustering effects, Supplementary Figure 2). This suggests in the former case that the sampling variance parameters are accounting for process variation in selectivity.

What if we model variation on the wrong process?

When there was minimal variation in catchability and directional time-variation in selectivity in the OM (OM3), there was little downside to specifying catchability as a random walk process in the EM (Figures 3–4). In some cases, negative bias in confidence interval coverage decreased compared to the baseline EM, which did not include process variation outside of recruitment variation. The same can be said for the white noise catchability deviations.

When there was minimal variation in fishery selectivity and directional time-variation in fishery catchability in the OM (OM2), modelling variation in fishery selectivity using either the white noise deviations or the 2D AR(1) parameterization led to roughly equivalent bias compared to modelling no vari-

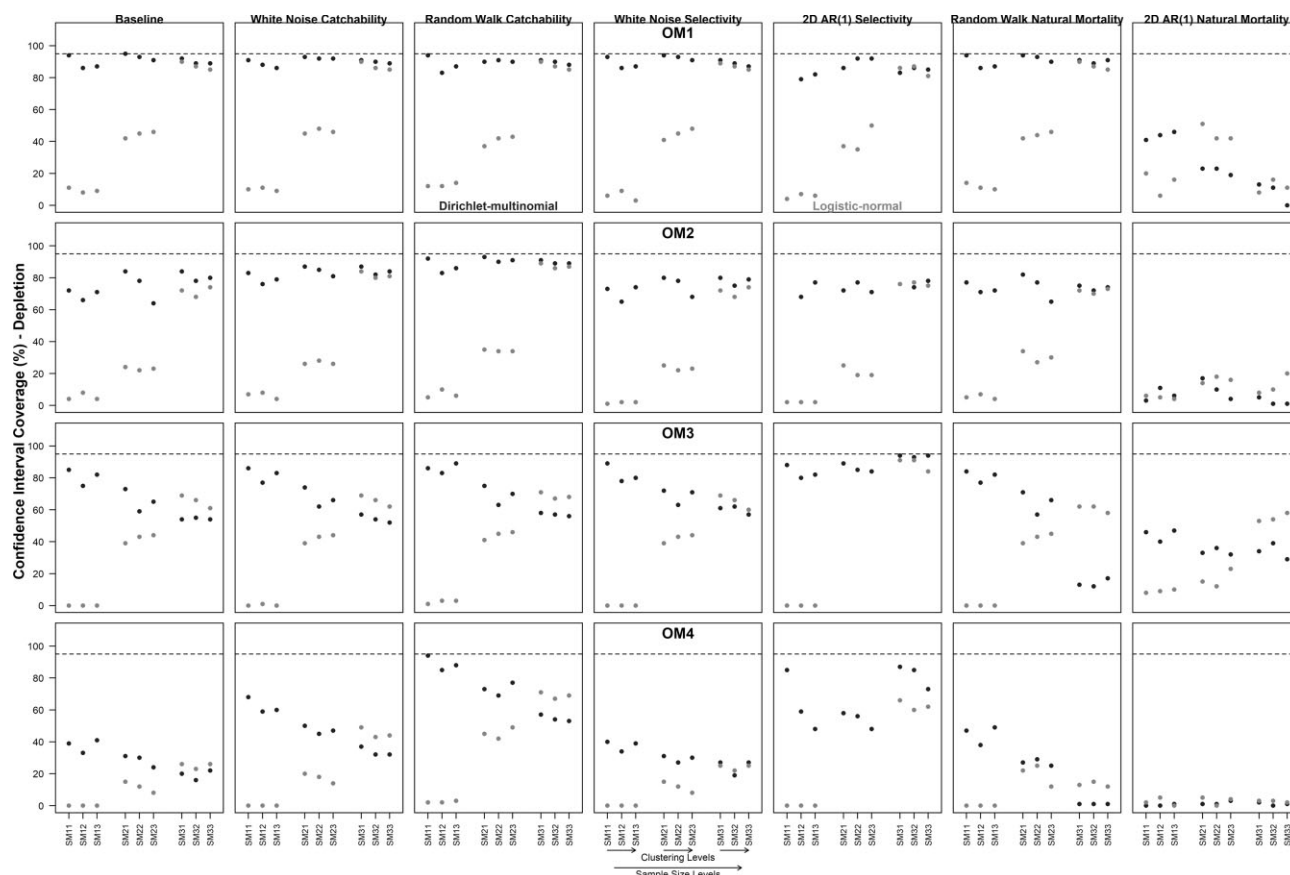


Figure 5. Confidence interval coverage for terminal-year estimates of depletion. The dotted line depicts the 95% line. Perfect coverage would equal 95%. The columns depict different EM process variance treatments where the rows depict different OM treatments. Shown in blue are the results from the DM and in tan those from the LN. The x-axis is presented as three groups of increasing sample size, within which are the three levels of increasing clustering. The violin plots depict an approximation of the density of the distributions across simulations, and the point denotes the median.

ation in fishery catchability and selectivity (although exhibited more bias than either EM which allowed fishery catchability to vary), conditional on using the DM. For EMs fit with the LN at large sample sizes, specifying variation on fishery selectivity led to no more bias than the baseline models and a minor decrease in the negative bias exhibited by confidence interval coverage.

When process variation was modelled on natural mortality in estimation models fit to OMs 2–4, depletion estimates were generally more biased than specifying variability on the correct process, and typically more so for larger sample sizes. For EMs with a random walk M fit to OMs 3–4, the bias at the largest sample sizes was substantially greater than that of all other EMs, and much greater for the DM models compared to the LN in these cases. The confidence interval coverage for depletion in these cases was also more severely negatively biased for the DM (for SSB, the DM overestimates coverage). The standard deviation for the random walk parameterization of natural mortality was effectively zero for all but the largest sample sizes for EMs fit to OMs 3–4 (although many estimates were greater than zero at moderate sample sizes for the DM). At the large sample size level for EMs fit to OMs 3–4, the median (across simulations) estimate of SD for the random walk parameterization of natural mortality was ~ 0.12 for the DM and ~ 0.05 for the LN (Figure 8). For the 2D AR(1) parameterization of deviations on natural mortality, the bias for depletion was also greater than other EMs and notably increased with sample size for fits to OMs 1–2; however, re-

mained consistent across sample sizes for OMs 3–4. The increase in bias with sample size was not as severe for estimates of terminal spawning stock biomass and at small sample sizes this metric was unbiased for the DM (although the precision of estimates was much greater than the baseline EMs). The bias for models fit with the LN at the largest sample sizes was roughly equivalent to that from the DM; however, confidence interval coverage was generally less negatively biased for the LN. The SD parameter for the 2D AR(1) formulation on natural mortality was consistently estimated greater than zero, the exception being at the largest sample sizes of OMs 1–2 (notably where selectivity did not vary in the OM). The correlation parameters for the 2D formulation on natural mortality were highly variable and generally the age correlation term increased with increased clustering and sample size, where the year correlation term approached one across sample sizes and clustering for OMs 3–4 (indicating natural mortality is estimated as highly correlated by year) and was not identified well but approached zero for OMs 1–2.

What if we model variation and there is none?

When there was minimal variation in catchability and selectivity in the OM (OM1), with exception to the 2D AR(1) natural mortality EMs, specifying any of the other EM variation treatments led to similar bias levels compared the baseline EM with no process variation modelled (Figure 3). This was true for all DM models and only LN models at the largest sam-

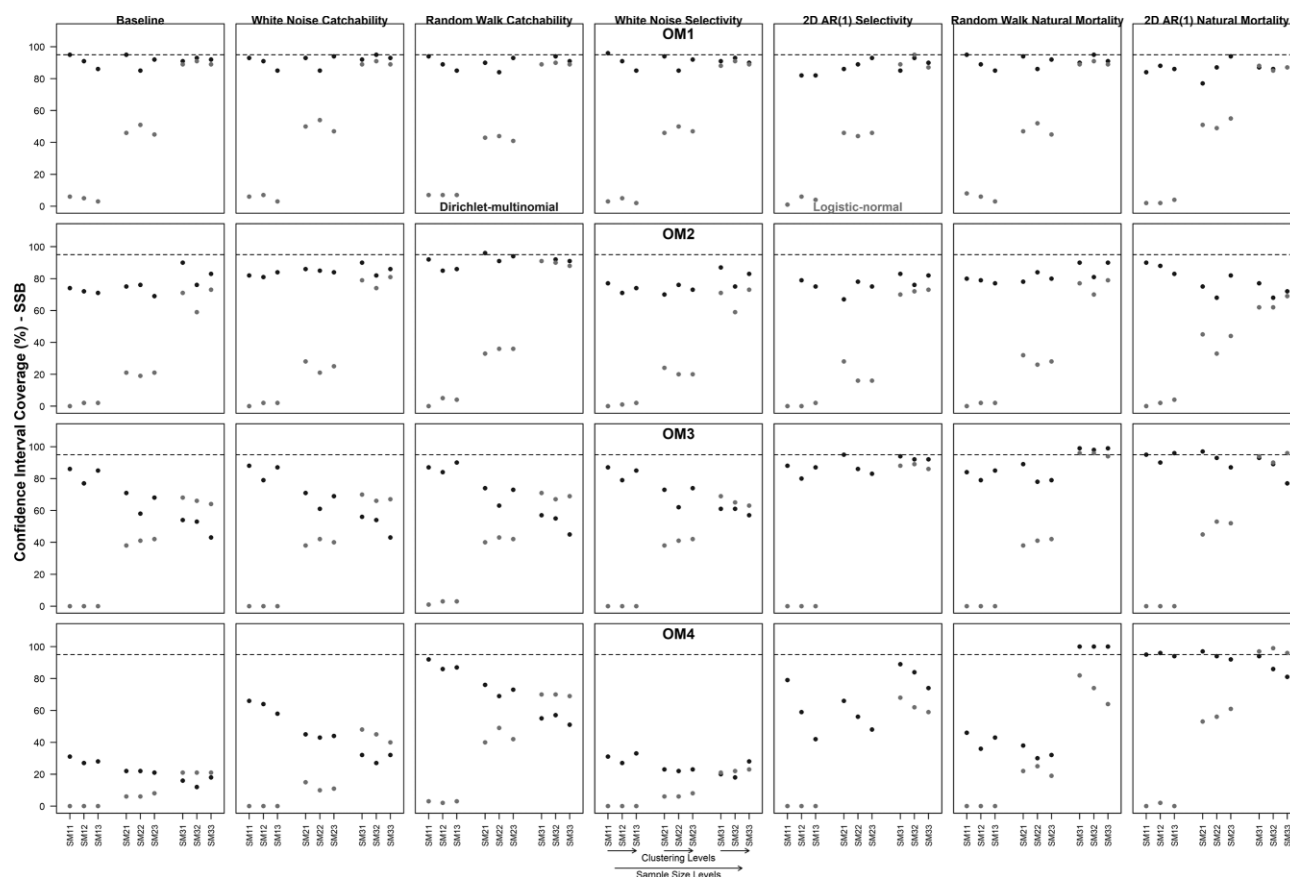


Figure 6. Confidence interval coverage for terminal-year estimates of spawning stock biomass. The dotted line depicts the 95% line. Perfect coverage would equal 95%. The columns depict different EM process variance treatments where the rows depict different OM treatments. Shown in blue are the results from the DM and in tan those from the LN. The x-axis is presented as three groups of increasing sample size, within which are the three levels of increasing clustering. The violin plots depict an approximation of the density of the distributions across simulations, and the point denotes the median.

ple size level. In these cases, negative biases in confidence interval coverage did increase for many EMs compared to the baseline, although this effect was small (1–5%, [Figures 5–6](#)). EMs with 2D AR(1) deviations on natural mortality performed considerably worse than the baseline (and all other EMs) in terms of bias in estimates and confidence interval coverage of depletion, and this bias increased as sample size increased. The levels of bias for spawning stock biomass estimates and confidence interval coverage were not as severe.

The median estimated variance parameters for the process variation treatments on catchability and random walk natural mortality were approximately zero when using the DM for fits to OM1 ([Figure 8](#)). For catchability, this was only true for the LN models at the largest sample size level. The process variation parameters for selectivity variation treatments were approximately zero across sample sizes when there was no clustering in composition data for the DM. As clustering increased the standard deviation estimates for either selectivity parameterization did as well. At the largest sample size level, the LN estimated these terms at zero across clustering levels. For the 2D AR(1) parameterization of variation on natural mortality, estimates of the standard deviation parameter were consistently greater than zero across clustering treatments for small sample sizes and approached zero as the sample size increased. Estimates of the age correlation parameter for the 2D AR(1) parameterization of variation on natu-

ral mortality increased both as clustering and as sample size increased, where estimates of the year correlation parameter were highly imprecise across simulations, although the mode of the distribution decreased as sample size increased. At the largest sample sizes for the LN, the standard deviation parameter for the 2D AR(1) parameterization of variation on natural mortality approached zero.

Sensitivity models

Results of the EMs fit with the multivariate Tweedie likelihood were very similar to those of the DM EMs ([Supplementary Figures 3–5](#)). No notable improvements or deteriorations in model performance appear evident across OM–SM treatments. For EMs fit to an 80-year time series of data, patterns in results of baseline EMs were largely similar to the 40-year time series, although at the largest sample size levels the LN now consistently exhibited less bias in depletion and spawning stock biomass and was less negatively biased confidence interval coverages for fits to OMs 3–4 (a more pronounced improvement than the baseline exhibited; [Supplementary Figures 6–8](#)).

For estimation models that specified both 2D AR(1) variation in selectivity and random walk variation in the natural mortality scalar parameter (\hat{M}) fit to the largest sample size SMs from OM3 to OM4, little to no variation in \hat{M} and some variation in selectivity was estimated

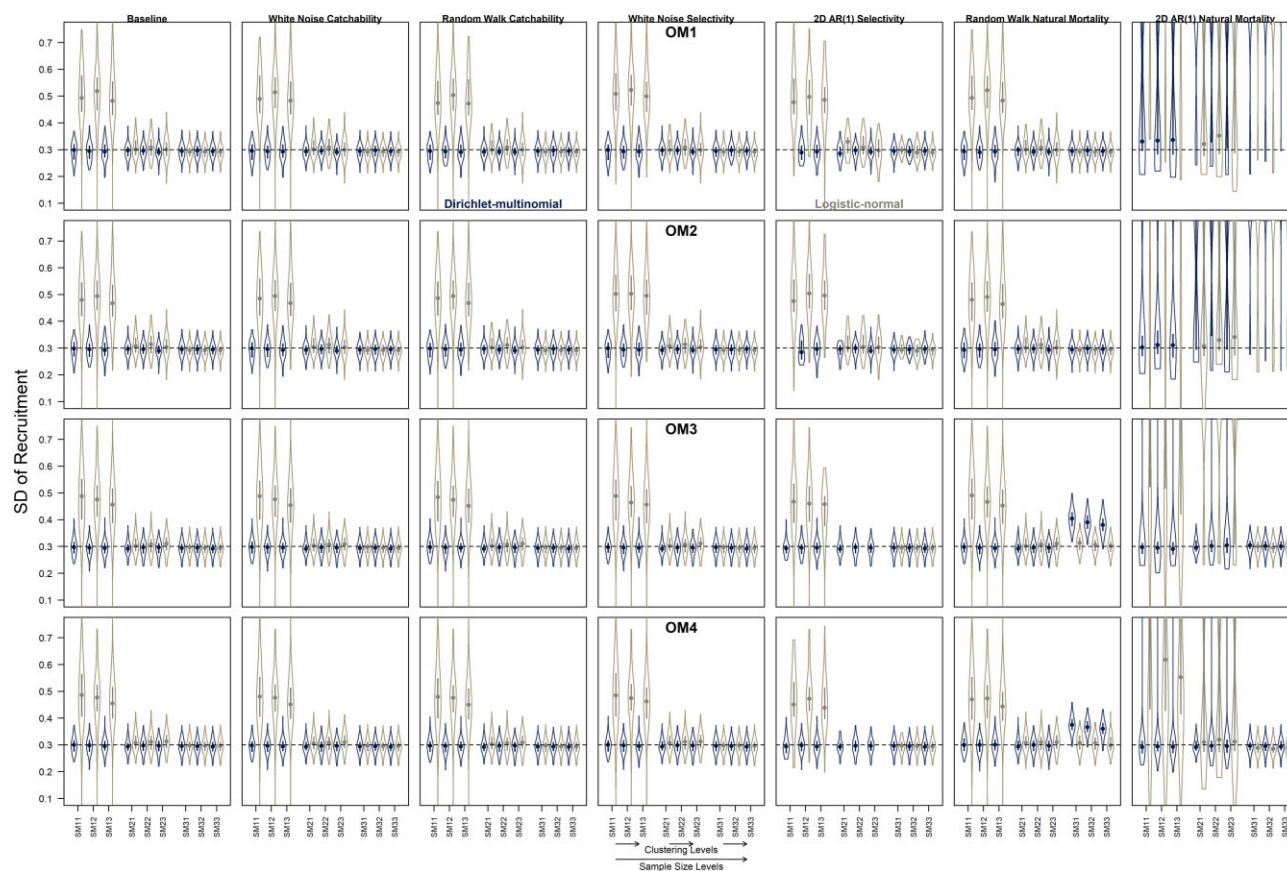


Figure 7. Estimates of recruitment variation for each OM-SM-EM treatment. The columns depict different EM process variance treatments where the rows depict different OM treatments. Shown in blue are the results from the DM and in tan those from the LN. The x-axis is presented as three groups of increasing sample size, within which are the three levels of increasing clustering. The violin plots depict an approximation of the density of the distributions across simulations, and the point denotes the median.

(Supplementary Figure 9), leading to less bias in spawning biomass and depletion and less negatively biased confidence interval coverage compared to the baseline estimation models or when simply \hat{M} was modelled as a random walk (Supplementary Figures 11–12). These bias and coverage levels were approximately equivalent to the estimation models that only estimated 2D AR(1) deviations in selectivity. For these sensitivity models, those fit with the DM exhibited less bias and underestimation of coverage than those fit with the LN.

For estimation models that specified both 2D AR(1) variation in selectivity and natural mortality fit to the largest sample size SMs from OM3 to OM4, some variation in selectivity was estimated [however, less than that from the preceding sensitivity model and the model that only included 2D AR(1) deviations] in addition to some variation in natural mortality (Supplementary Figure 14). This led to increased bias in terminal year estimates and confidence interval coverage for depletion compared to the baseline EM; however, less biased estimates and coverage for depletion compared to the EM that only estimated 2D AR(1) deviations on natural mortality (Supplementary Figures 16–17). Estimates and coverage levels for spawning stock biomass were less biased than both the baseline EM and the EM that only estimated 2D AR(1) deviations on natural mortality. For each of the sensitivity models that attempted to estimate 2D AR(1) deviations in both fishery selectivity and natural mortality, those

fit with the LN generally exhibited less bias in terminal year estimates and confidence interval coverage than those fit with the DM.

The sensitivity EMs fit with the LN likelihood formulated without relative weighting between years based on annual composition sample size exhibited very similar output to the baseline EMs using the LN as described in the methods, with similar bias levels in depletion and spawning stock biomass (Supplementary Figures 19–20) and marginally more biased confidence interval coverage (Supplementary Figure 21). The sensitivity EMs with 2D AR(1) process variance on either selectivity or natural mortality where the correlation parameters of the 2D AR(1) were bounded between -1 and 1 (as opposed to 0 – 1) rarely improved upon and were often more biased with worse coverage than the original estimation models that constrained the correlation terms of the 2D AR(1) from 0 to 1 (Supplementary Figures 22–33).

Discussion

Getting the process variability correct

Getting the process variability “correct”, in the sense that variance is modelled on a process that is directionally varying in the operating model (noting we are not modelling the actual mechanism), proved in large part possible and led to improvements in model performance compared to not doing so.

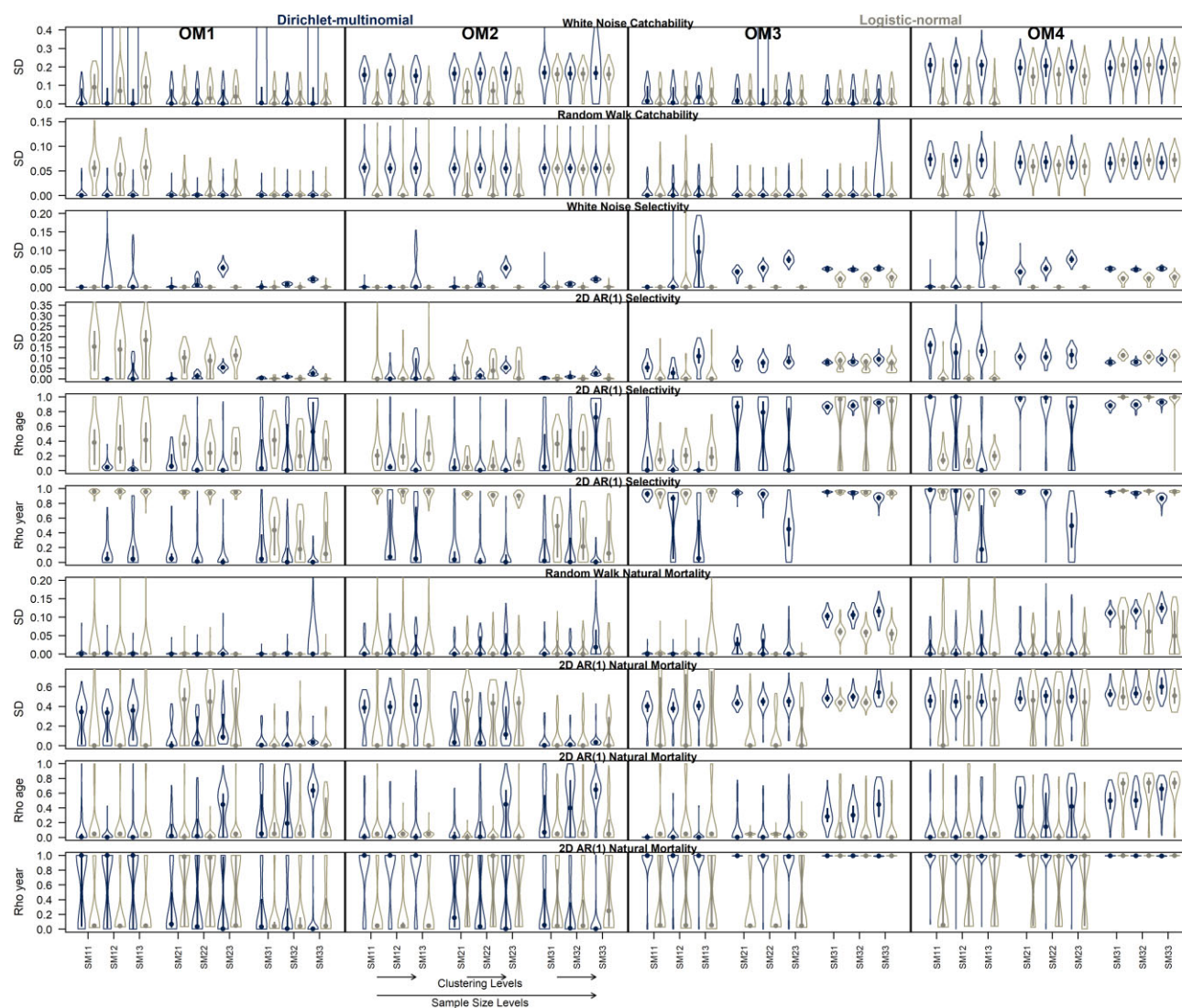


Figure 8. Estimates of hyperparameters for each process variation treatment. The columns depict the different OM treatments where the rows depict different process variance hyperparameters for each EM treatment. The EM treatment is identified in each row title and the y-axis for each row denotes the specific hyperparameter. Shown in blue are the results from the DM and in tan those from the LN. The x-axis is presented as three groups of increasing sample size, within which are the three levels of increasing clustering. The violin plots depict an approximation of the density of the distributions across simulations, and the point denotes the median.

Accounting for white-noise variability did improve performance compared to not accounting for variation in the process at all; however, the formulations that allowed for directional variation [random walk, 2D AR(1)] performed best given the true variation was directional. The variance parameters in these processes were largely estimable along with those relating to the sampling variance. The performance of the 2D AR(1) selectivity formulation greatly improved as the sample size increased for fishery composition data, as it led to roughly equivalent bias levels compared to not modelling variation in fishery selectivity at the smallest sample sizes, and accounting for variation in catchability proved more important for all but the largest sample sizes in OM4 (when there was variation in both catchability and selectivity in the operating model). The hyperparameters of the 2D AR(1) selectivity formulation do appear to require a large sample size of composition data for reasonably precise estimation. The study that first presented the method, Xu *et al.* (2019), do note that the 2D AR(1) method for selectivity variation is

more important for assessments with high-quality composi-

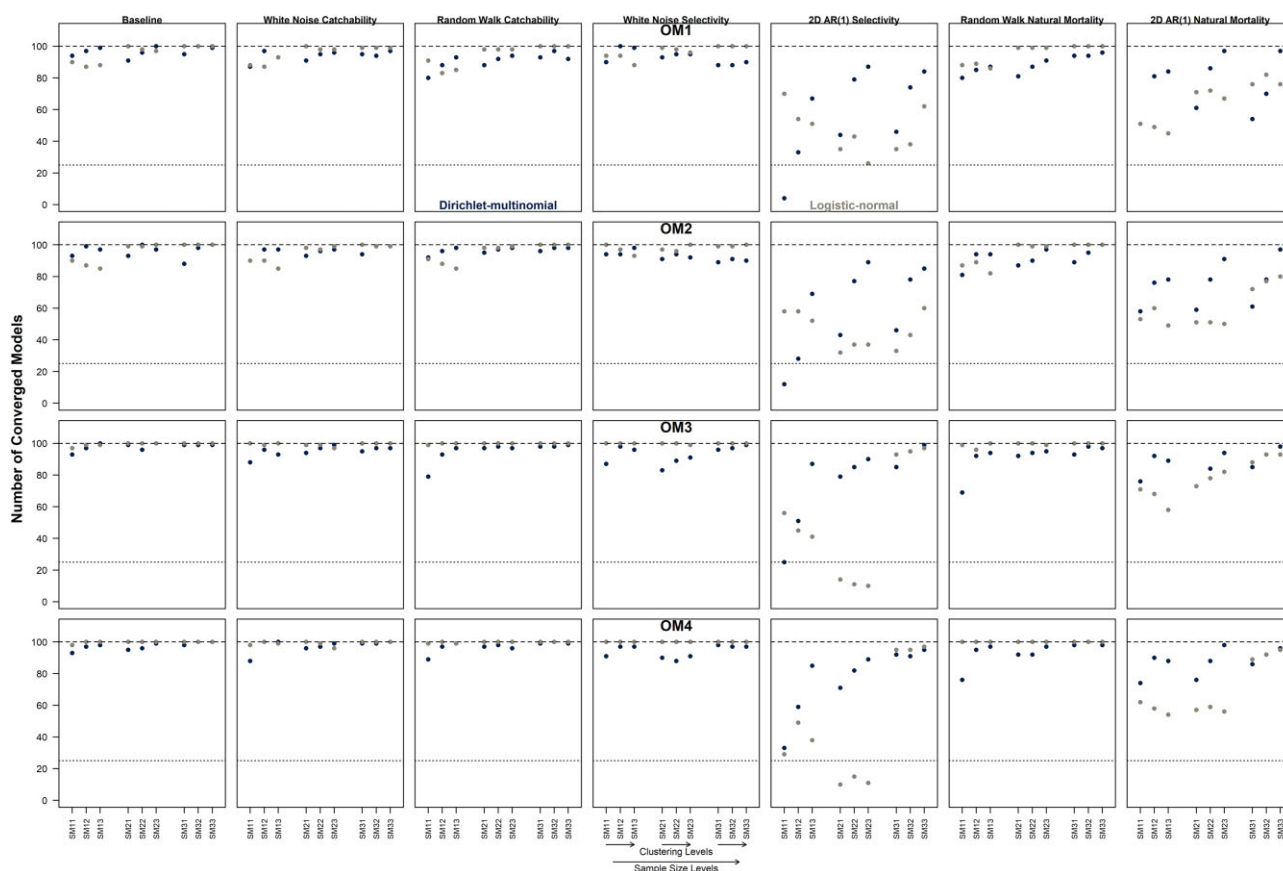


Figure 9. Convergence, depicted as the number of iterations where an EM was considered converged. The columns depict different EM process variance treatments where the rows depict different OM treatments. Shown in blue are the results from the DM and in tan those from the LN. The x-axis is presented as three groups of increasing sample size, within which are the three levels of increasing clustering. The violin plots depict an approximation of the density of the distributions across simulations, and the point denotes the median. The dotted line at 25 depicts a level below which we do not present results of an EM as too few simulation iterations converged.

tion data and that the precision of parameter estimates improved when the sample size increased, similar to our results. In fact, within Xu *et al.* (2020), where both the 2D AR(1) selectivity parameterization and fitting composition data with the DM were implemented, although they concluded that the median estimates of the three hyperparameters were encouragingly >50% of their true level, the distributions for the parameters (particularly the correlation terms) were highly variable across simulations similar to the results from our analysis.

Notably, the estimation models were availed of a fishery-independent survey for the second half of the time series that included constant catchability and constant, asymptotic selectivity, which may have aided in both the estimation of random walk fishery catchability and 2D AR(1) fishery selectivity. It is likely these treatments would have performed worse without a fishery-independent survey, as has been found in other studies on time-varying catchability (Wilberg and Bence, 2006).

Overparameterization

With exception to the 2D AR(1) natural mortality formulation, including process variation in the estimation model (other than recruitment variation) when there was no directional time-variation in any process within the operating model, or overparameterizing the EM, did not result in appreciable decreases in model performance. There did ap-

pear to be an effect of clustering in composition data particular to modelling process variability on selectivity and natural mortality, where estimates of the process variance terms increased with increased clustering in composition data (Figure 8), suggesting the process variation was beginning to fit to noise in the composition data. For fishery selectivity, however, this effect did not seem to greatly diminish our performance metrics. In these instances for the 2D AR(1) parameterization, the yearly correlation term tended to zero, resulting in strictly age-correlated variation in selectivity accounting for clustering in composition data. Concurrently, as the variation estimated in selectivity was accounting for increased clustering in sampling, the DM estimated less overdispersion in fits to the fishery compositions (Supplementary Figures 1–2). The estimation of effectively zero annual correlation in selectivity variation is likely what led to similar bias levels, while the estimation of age-correlated variation in fishery selectivity likely led to similar estimates of the total variance compared to the baseline treatments, thus not having a great effect on confidence intervals relative to the baseline models [evidenced by coverage increasing with clustering for 2D AR(1) fishery selectivity fit to OMs 1–2]. The lack of consequence for overparameterized estimation models could also be partly due to the EMs not being wholly correctly specified in the sense that the simulation operating model was a spatially explicit

population model and contained stochastic fish movement and fisher effort distribution. Consequently, there was already some background level of variation unaccounted for in the spatially aggregated estimation models. The lack of consequence to specifying variability on selectivity when it is not time-varying in the operating model has been noted in other studies (Stewart and Monnahan, 2017; Cronin-Fine and Punt, 2021). In our case, the variance terms approached zero in many cases if that specific process was not directionally time-varying within the operating model, which in a practical application would likely lead to the analyst removing the variation. In addition, we bounded the variance terms at levels close to zero (e.g. -10 on \ln scale). Had the variance terms been unbounded, this may have led to convergence issues as the log SD approached negative infinity, and thus would hopefully also lead to the analyst removing the variation.

Specific to catchability, the lack of consequence to specifying variation in the estimation model when the operating model contained little or minimal variation in this parameter may be due to the fact that catchability was solely used within the observation model and had no direct effect on the process model (outside of its effect on parameter estimates). Had a time series of effort been used in combination with catchability to estimate fishing mortality within the estimation models (i.e. conditioning on effort), it is possible that the specification of variation on this parameter would have been of more consequence when there was no variation present in the operating model.

When variation was modelled on the wrong process

When variation was modelled on the wrong process, i.e. catchability or selectivity was directionally varying in the operating model; however, variation was modelled on a different process in the estimation model, the consequences were large but were specific to modelling variation in natural mortality. Notably, at the largest sample sizes in OMs 3–4, coincidentally the same cases where including process variation in selectivity led to large improvements, there was great consequence to incorrectly specifying variability on natural mortality using a random walk (and to a smaller degree at moderate sample sizes), as terminal year estimates were substantially more biased than all other EMs that did not include process variation on natural mortality. In these cases, variation in natural mortality accounts for what is actually variation in fishery selectivity and is confounded with parameters defining recruitment (recruitment variation and unfished recruitment), in so doing resulting in greater bias. The same confounding was present in all the 2D AR(1) natural mortality models, even in cases where the SD term for the 2D AR(1) formulation was close to zero (largest sample sizes for OMs 1–2), highlighting very small changes in $M_{a,y}$ have a large effect on the model. However, the bias could be reduced and coverage improved in the random walk case and to some degree in the 2D AR(1) case when variation was also modelled on the correct process (selectivity in this case), evidenced by the sensitivity model runs. The LN outperforming the DM at large sample sizes in cases where variation in natural mortality was estimated, by not leading to either as severe bias in relative error or coverage for depletion, is likely due to the modelling of correlations in composition residuals [with the LN AR(1)] and allowing for

more residual variation in fits to the composition data (Fisch *et al.*, 2021), leading to less variability being modelled with deviations in natural mortality. This is evidenced by distributions of the estimated standard deviation parameter for the random walk and 2D AR(1) formulations of natural mortality, which were greater for DM than LN (Figure 6).

Notably, we did not examine operating models where natural mortality was time-varying. Had we modelled time-variation in natural mortality within the OM, it is conceivable that an EM with process variation modelled on fishery selectivity would have performed poorly by attempting to account for what was actually variation in natural mortality through selectivity deviations, similar to what was observed in the opposite case.

These results suggest that specifying variation on processes other than recruitment should be modelled with caution within fisheries assessments, especially for variables likely to have a large impact on population processes and projections (e.g. natural mortality compared to an observation model catchability parameter). Szuwalski *et al.* (2018) outline a few ways that plausible time-varying processes may be narrowed down using model diagnostics, such as applying assessment methods that allow each process to vary in turn and removing from consideration models that do not eliminate retrospective patterns or whose composition residuals indicate model misspecification. Importantly, other than observing whether the hessian matrix was positive-definite, given the scale of our simulations, we were not able to examine additional model diagnostics for each individual simulation iteration, such as those suggested by Szuwalski *et al.* (2018). Doing so could have led to removing a process variation treatment that performed poorly in an additional convergence metric [e.g. 2D AR(1) natural mortality].

Likelihoods

The performance of EMs fit with the logistic normal was only comparable to those fit with the Dirichlet-multinomial at the largest sample sizes, similar to Fisch *et al.* (2021). At these levels, however, it did outperform in the DM in many instances and notably was more robust to incorrectly modelling variation in natural mortality within the EM. In short, conditional on a large sample size, getting the process model wrong favoured the LN, particularly for processes closely linked to composition data, whereas a more correct process model favoured the DM.

Although the patterns were similar to Fisch *et al.* (2021), in comparing likelihoods we did not see the logistic-normal show less bias than the Dirichlet-multinomial in the OM treatments when there was significant model error unaccounted for in the estimation model, as was observed in Fisch *et al.* (2021) when the composition sample size was large. The logistic-normal improved in performance to a point where it was as biased as the Dirichlet-multinomial; however, not clearly less-biased, although it did have better confidence interval coverage. This could be a function of differences in the time series length (40 years compared to 100 in Fisch *et al.*, 2021). In fact, the baseline EMs fit to OMs and SMs with 80-year time series of data using the same composition sample size levels resulted in the LN exhibiting less bias than the DM for models with directional variation in selectivity unmodelled in the assessment (OMs 3–4), and as much bias when selectivity was not directionally varying in the operating model (OMs 1–2). This

suggests that there may be a multiplicative effect of sample size coupled with time series length where more data leads to better characterization of the residual variance–covariance matrix and consequently more optimal performance of the LN likelihood compared to the DM when used for fishery composition data. It also suggests there is likely an effect regarding which process is misspecified, where the LN did not outperform the DM when the misspecification was on catchability. This makes intuitive sense as composition data have a great influence on selectivity and a more indirect impact on the fishery catchability. There is also likely an effect regarding the degree of model misspecification specific to a process. The EMs developed in this study were most similar to the “Baseline” scenario from Fisch *et al.* (2021), where the shape of selectivity was approximately correct; however, the time variation was not accounted for. In the maximum process error (i.e. “Max PE”) scenario within Fisch *et al.* (2021), the shape of the selectivity was also incorrect (logistic when the true selectivity was dome-shaped), and this led to further improved performance of the LN compared to the DM even at moderate sample sizes (~2000 annually). We did not include such misspecified estimation models within this study, and it remains to be evaluated whether this type of error is best allocated to gaussian-based random effects or generalized sampling distributions such as the logistic-normal.

Although the variance parameters for each composition likelihood estimated more overdispersion as clustering in composition data increased, controlling for sample size the degree of clustering had little impact on performance metrics. In Fisch *et al.* (2021), it was suggested that overdispersion may not affect point estimates; however, confidence intervals would be affected if the composition likelihood did not adequately account for the increased sampling variance regarding the clustering treatments. There was indeed an effect on confidence interval coverage, as was suggested, and this effect was particular to the estimation models fit with the Dirichlet-multinomial, suggesting that the DM variance is not fully accounting for the increased sampling variance and/or structure of the variance regarding the clustering treatments. Although we do emphasize that this effect was minimal. It could be that levels of correlation and overdispersion in the data we simulated are less than those in some real systems, in which case there may be both a greater effect on confidence interval coverage when using the Dirichlet-multinomial and, as alluded in the overparameterization section, a greater effect of process variance formulations fitting to noise in composition data (rather than actual variation in processes).

Estimability

Mixed effects and state-space population and assessment models where both process and sampling error variances are estimated are a growing body of literature (Nielsen and Berg, 2014; Cadigan, 2015; Albertsen *et al.*, 2016; Miller *et al.*, 2016; Xu *et al.*, 2020; Stock *et al.*, 2021), and this study is certainly not the first to estimate sampling variance of composition data and process variance simultaneously within an assessment. However, few to our knowledge have examined the effect that multiple degrees of correlation and overdispersion in composition data and the choice of observational likelihood used might have on estimation and identifiability within a simulation framework. Nonetheless, their findings offer important insight to these topics.

Albertsen *et al.* (2016) examined the performance of various observation likelihoods for fitting either composition data (as proportions) or catch-at-age data (as real numbers) within state-space assessment models of four different European stocks. They utilized AIC to compare the performance of the likelihoods and found that for assessments fit to composition data, the logistic-normal (both additive and multiplicative formulations) had better model fits than the Dirichlet distribution. Their parameterization of the state-space assessments fit to each stock estimated process variation parameters associated with additive deviations on log fishing mortality at age and log numbers at age (Nielsen and Berg, 2014), and sampling variation parameters associated with the Dirichlet and logistic-normal. Two of the four assessments for which we were able to find sample sizes (ICES, 2020; De Oliveira and Walker, 2021) suggest that sample sizes were consistently >10000 annually for each, which likely aided in parameter estimation and identification and could have resulted in the logistic-normal outperforming the Dirichlet.

Xu *et al.* (2020) evaluated the ability of an estimation model to estimate sampling variation of composition data using the Dirichlet-multinomial while accounting for time-varying fishery selectivity. They simulated composition data under different sample sizes using the multinomial and found that the estimation model was able to estimate the level of effective sample size and account for time variation in fishery selectivity. However, in an effort to replicate an assessment model process in Stock Synthesis, they did not simultaneously estimate the variance parameters internal to the assessment. They instead utilized an iterative approach (EM3 of Xu *et al.*, 2019), where first the variance term of the age and year autocorrelated selectivity was tuned using the method proposed by Methot and Taylor (2011), and correlation parameters were then estimated by fitting an external model to the selectivity deviations from the model fit of the tuning approach using the multivariate normal distribution. All three parameters were then fixed in the actual assessment model run. They concluded that although the three hyperparameters were estimated with a negative bias when correlation existed within the operating model, it was encouraging that the median values were >50% of the true correlation level. Also, when there was no correlation within the operation model, they were able to produce minimal estimates of the correlation parameters. Using an iterative approach such as this may have helped us alleviate some of the imprecision we experienced at smaller sample sizes; however, our goal was to estimate variances within assessments and ideally improve upon iterative estimation methods.

Stock *et al.* (2021) implemented 2D AR(1) random effect deviations either on natural mortality or survival with an assessment of southern New England–mid-Atlantic yellowtail flounder and concluded that including the deviations on either process considerably improved model fit and reduced retrospective patterns. They estimated a sampling variance parameter for the logistic-normal likelihood for fitting composition data (however, they did not include correlation in the logistic-normal parameterization), the variance in recruitment, and the three hyperparameters of the 2D AR(1) parameterization simultaneously within a mixed-effects framework using TMB. In addition to the commercial catch and age composition, the assessment was fit to indices of abundance and age compositions from three bottom trawl surveys, suggesting a non-trivial amount of informative data to supplement what was collected

from the fishery. The amount of different composition data sources likely aided in the estimation of the process variation hyperparameters.

Cadigan (2015) developed a state-space, age-structured assessment model for Northern Cod and estimated a suite of variance terms related to both process and sampling variation. He specified 2D AR(1) deviations on natural mortality and parameterized fishing mortality as in Nielsen and Berg (2014). He also estimated a variance term for fits to the composition data, using what he called a continuation-ratio logit likelihood. He ultimately found that without the inclusion of tagging data, the process variance parameter was highly confounded with sampling error variance.

Collectively, these studies highlight that incorporating both process variation as random effects and estimating the extent of the variation internal to the model, in addition to sampling variation, is possible provided high-quality data from compositions and other sources (e.g. tagging) is available. Our study supports this, but also adds that the implications of placing variation on the incorrect process are serious for management advice.

We have presented a data-rich and somewhat ideal scenario. Consequently, there may have been some specifications in our simulations that could have influenced results. We only simulated a two-way trip effort time series and a single fish species life history with a single degree of recruitment variation, which was a reasonably low degree compared to other species (Thorson *et al.*, 2014). We also fixed the natural mortality scalar parameter, steepness, the variance of annual catch, and the variance of the fishery index at correct values within our estimation models, and the estimation or misspecification of these parameters could have influenced results. The estimation models were also privy to an informative survey with constant catchability and constant, asymptotic selectivity for the second half of the time series, which may have aided in both the estimation of time-varying fishery catchability and time-varying fishery selectivity. The estimation models were initialized from an unfished state (a correct assumption given operating models), with fishery catch, composition, and index data available in the first year. This likely aided in parameter estimation. We did suppress zeroes in composition data, which the logistic-normal requires; however, the Dirichlet-multinomial does not, although this is unlikely to have influenced results as zeroes did not occur commonly in the data prior to suppression. We encourage further research exploring the impact of the mentioned variables with respect to the performance assessments when both process variances and sampling variances of composition likelihoods are estimated. We also encourage additional evaluation of the multivariate Tweedie (Thorson *et al.*, 2022), which in our simulations performed no worse than the Dirichlet-multinomial, although we did not include it within estimation models that incorporated process variation in quantities other than recruitment.

Summary

We grant that modelling random effects to account for process variation is suboptimal compared to a thorough understanding of the fishery coupled with the availability of data regarding an environmental or mechanistic link to model the actual mechanism giving rise to the variation (Cao and Chen, 2022). However, in the absence of such data and understanding, we have shown that under the circumstances evaluated in

this study, parameters defining the extent of process variation in selectivity or catchability can be estimated concurrent with recruitment variation and multiple sampling variation parameters. Notably, the autoregressive process variation methods performed very well when fit to what was actually variation emergent from a spatially explicit operating model as a function of stochastic fish movement and fisher effort distribution, rather than simulated from the same autoregressive process in a non-spatial context. Process variation should, however, be modelled with caution in these circumstances, as our study as well as others (Szuwalski *et al.*, 2018) suggest estimates and management advice could be heavily biased if placed on the wrong process.

In sum, it does seem risky to utilize the LN for fitting composition data within stock assessments, and the Dirichlet-multinomial appears a prudent choice in most circumstances. However, if a stock assessment includes a lengthy time series of large sample sizes for composition data (albeit a high bar), an analyst may be better off fitting composition data with the LN as opposed to the DM given it performs equivalently when the model specification is largely correct and often better when there is unmodelled variation in fishery selectivity (and this is likely to be extended to other processes that are closely linked to composition data). Moreover, the LN was more robust to incorrectly placing process variance on natural mortality instead of fishery selectivity. In addition, in real-world situations, the degree of model misspecification is likely to be greater than that within simulation experiments, which we have shown further favours the logistic-normal compared to the Dirichlet-multinomial. Herein, we focused on modelled, unmodelled, or incorrectly modelled process variation and we encourage additional research on this topic to include systematic model misspecification and its effect on the estimation of both sampling and process variance within fisheries assessments.

Funding

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.

Acknowledgements

We would also like to thank Tim Miller, Amy Schueller, and one anonymous reviewer for providing revisions of this manuscript. The authors acknowledge the University of Florida Research Computing for providing computational resources that contributed to the research reported in this publication.

Supplementary data

[Supplementary material](#) is available at the *ICESJMS* online version of the manuscript.

Author contributions

NF: Conceptualization, Methodology, Formal analysis, Writing—original draft, Investigation, Visualization. EC: Supervision, Writing—review and editing. KS: Methodology, Writing—review and editing. MM: Methodology, Writing—

review and editing. RA: Supervision, Conceptualization, Methodology, Writing—review and editing.

Conflict of interest

The views and opinions expressed or implied in this article are those of the authors and do not necessarily reflect the position of the National Oceanic and Atmospheric Administration. The authors declare no conflicts of interest.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

- Addis, D.T., Patterson, W.F., Dance, M.A., and Ingram, G.W. 2013. Implications of reef fish movement from unreported artificial reef sites in the northern Gulf of Mexico. *Fisheries Research*, 147: 349–358.
- Aitchison, J. 2003. *The Statistical Analysis of Compositional Data*. The Blackburn Press, Caldwell, NJ.
- Albertsen, C. M., Nielsen, A., and Thygesen, U. H. 2016. Choosing the observational likelihood in state-space stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 74: 779–789.
- Becker, J. J., Sandwell, D. T., Smith, W. H. F., Braud, J., Binder, B., Depner, J., Fabre, D. *et al.* 2009. Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. *Marine Geodesy*, 32: 355–371.
- Bentley, N., Davies, C. R., McNeill, S. E., and Davies, N. M. 2004. A framework for evaluating spatial closures as a fisheries management tool. *New Zealand Fisheries Assessment Report*, 25: 25.
- Caddy, J. F. 1975. Spatial model for an exploited shellfish population, and its application to the Georges Bank scallop fishery. *Journal of the Fisheries Research Board of Canada*, 32: 1305–1328.
- Cadigan, N. G. 2015. A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. *Canadian Journal of Fisheries and Aquatic Sciences*, 73: 296–308.
- Cadrin, S. X. 2003. Stock Assessment of Yellowtail Flounder in the Southern New England–Mid-Atlantic Area NEFSC Reference Document 03-02. Northeast Fisheries Science Center. <https://repository.library.noaa.gov/view/noaa/5363>, Last accessed: 08/25/23.
- Cao, J., and Chen, Y. 2022. Modeling time-varying natural mortality in size-structured assessment models. *Fisheries Research*, 250: 106290.
- Cheang, W. K., and Reinsel, G. C. 2000. Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood. *Journal of the American Statistical Association*, 95: 1173–1184.
- Cronin-Fine, L., and Punt, A. E. 2021. Modeling time-varying selectivity in size-structured assessment models. *Fisheries Research*, 239: 105927.
- De Oliveira, J., and Walker, N. 2021. Stock Annex: Cod (*Gadus morhua*) in Subarea 4 and divisions 7.d and 20 (North Sea, eastern English Channel, Skagerrak). ICES Stock Annexes. https://ices-library.figshare.com/articles/report/Stock_Annex_Cod_Gadus_morhua_in_Subarea_4_and_divisions_7_d_and_20_North_Sea_eastern_English_Channel_Skagerrak/18622184/1, Last accessed: 08/25/2023.
- Deriso, R. B., Maunder, M. N., and Skalski, J. R. 2007. Variance estimation in integrated assessment models and its importance for hypothesis testing. *Canadian Journal of Fisheries and Aquatic Sciences*, 64: 187–197.
- Dichmont, C. M., Deng, R. A., Punt, A. E., Brodziak, J., Chang, Y.-J., Cope, J. M., Ianelli, J. N. *et al.* 2016. A review of stock assessment packages in the United States. *Fisheries Research*, 183: 447–460.
- Dunn, A., Rasmussen, S., and Mormede, S. 2012. *Spatial Population Model User Manual, SPM v1.1-2012-09-06 (rev 4806)*. CCAMLR, Hobart. 164pp.
- Fisch, N., Camp, E., Shertzer, K., and Ahrens, R. 2021. Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error. *Fisheries Research*, 243: 106069.
- Francis, R. C. 2017. Revisiting data weighting in fisheries stock assessment models. *Fisheries Research*, 192: 5–15.
- Francis, R. I. C. C. 2012. The reliability of estimates of natural mortality from stock assessment models. *Fisheries Research*, 119–120: 133–134.
- Francis, R. I. C. C. 2014. Replacing the multinomial in stock assessment models: a first step. *Fisheries Research*, 151: 70–84.
- Francis, R. I. C. C., and Shotton, R. 1997. Risk in fisheries management: a review. *Canadian Journal of Fisheries and Aquatic Sciences*, 54: 1699–1715.
- Hulson, P. J. F., Hanselman, D. H., and Quinn, T. J. 2012. Determining effective sample size in integrated age-structured assessment models. *ICES Journal of Marine Science*, 69: 281–292.
- ICES. 2020. Blue whiting (*Micromesistius poutassou*) in subareas 27.1–9, 12, and 14 (northeast Atlantic). *ICES Scientific Reports*, 2: 82.
- Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. 2016. TMB: automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70: 1–21.
- Labelle, M. 2005. Testing the MULTIFAN-CL assessment model using simulated tuna fisheries data. *Fisheries Research*, 71: 311–334.
- Mace, P. M., and Doonan, I. J. 1988. A Generalised Bioeconomic Simulation Model for Fish Population Dynamics. MAFFish, NZ Ministry of Agriculture and Fisheries, Wellington.
- Maunder, M. N. 2004. Population viability analysis, based on combining integrated, Bayesian, and hierarchical analyses. *Acta Oecologica*, 26: 85–94.
- Maunder, M. N. 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: estimating the effective sample size. *Fisheries Research*, 109: 311–319.
- Maunder, M. N., and Piner, K. R. 2015. Contemporary fisheries stock assessment: many issues still remain. *ICES Journal of Marine Science*, 72: 7–18.
- Maunder, M. N., and Watters, G. M. 2003. A general framework for integrating environmental time series into stock assessment models: model description, simulation testing, and examples. *Fishery Bulletin*, 101: 89–99.
- Methot, R. D., and Taylor, I. G. 2011. Adjusting for bias due to variability of estimated recruitments in fishery assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 68: 1744–1760.
- Methot, R. D., and Wetzel, C. R. 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research*, 142: 86–99.
- Miller, T. J., Hare, J. A., and Alade, L. A. 2016. A state-space approach to incorporating environmental effects on recruitment in an age-structured assessment model with an application to southern New England yellowtail flounder. *Canadian Journal of Fisheries and Aquatic Sciences*, 73: 1261–1270.
- Miller, T. J., O'Brien, L., and Fratanoni, P. S. 2018. Temporal and environmental variation in growth and maturity and effects on management reference points of Georges Bank Atlantic cod. *Canadian Journal of Fisheries and Aquatic Sciences*, 75: 2159–2171.
- Mormede, S., Dunn, A., Parker, S., and 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. *Fisheries Research*, 190: 164–174.
- Nielsen, A., and Berg, C. W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research*, 158: 96–101.

- Privitera-Johnson, K. M., Methot, R. D., and Punt, A. E. 2022. Towards best practice for specifying selectivity in age-structured integrated stock assessments. *Fisheries Research*, 249: 106247.
- SEDAR. 2018. SEDAR 52 —Gulf of Mexico Red Snapper Stock Assessment Report. SEDAR, North Charleston, SC. 434pp.
- Stewart, I. J., and Monnahan, C. C. 2017. Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. *Fisheries Research*, 192: 126–134.
- Stock, B. C., Xu, H., Miller, T. J., Thorson, J. T., and Nye, J. A. 2021. Implementing two-dimensional autocorrelation in either survival or natural mortality improves a state-space assessment model for southern New England–mid Atlantic yellowtail flounder. *Fisheries Research*, 237: 105873.
- Szuwalski, C. S., Ianelli, J. N., and Punt, A. E. 2018. Reducing retrospective patterns in stock assessment and impacts on management performance. *ICES Journal of Marine Science*, 75: 596–609.
- Thorson, J. T. 2019. Perspective: let's simplify stock assessment by replacing tuning algorithms with statistics. *Fisheries Research*, 217: 133–139.
- Thorson, J. T., Jensen, O. P., and Zipkin, E. F. 2014. How variable is recruitment for exploited marine fishes? A hierarchical model for testing life history theory. *Canadian Journal of Fisheries and Aquatic Sciences*, 71: 973–983.
- Thorson, J. T., Johnson, K. F., Methot, R. D., and Taylor, I. G. 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. *Fisheries Research*, 192: 84–93.
- Thorson, J. T., and Kristensen, K. 2016. Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. *Fisheries Research*, 175: 66–74.
- Thorson, J. T., Miller, T. J., and Stock, B. C. 2022. The multivariate-tweedie: a self-weighting likelihood for age and length composition data arising from hierarchical sampling designs. *ICES Journal of Marine Science*, fsac159, 0: 1–12. [CrossRef]
- Walters, C. J., and Bonfil, R. 1999. Multispecies spatial assessment models for the British Columbia groundfish trawl fishery. *Canadian Journal of Fisheries and Aquatic Sciences*, 56: 601–628.
- Wilberg, M. J., and Bence, J. R. 2006. Performance of time-varying catchability estimators in statistical catch-at-age analysis. *Canadian Journal of Fisheries and Aquatic Sciences*, 63: 2275–2285.
- Wilberg, M. J., Thorson, J. T., Linton, B. C., and Berkson, J. 2009. Incorporating time-varying catchability into population dynamic stock assessment models. *Reviews in Fisheries Science*, 18: 7–24.
- Xu, H., Thorson, J. T., and Methot, R. D. 2020. Comparing the performance of three data-weighting methods when allowing for time-varying selectivity. *Canadian Journal of Fisheries and Aquatic Sciences*, 77: 247–263.
- Xu, H., Thorson, J. T., Methot, R. D., and Taylor, I. G. 2019. A new semi-parametric method for autocorrelated age-and time-varying selectivity in age-structured assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 76: 268–285.

Appendix A

The spatially explicit operating model is based on Gulf of Mexico red snapper, as such most life history parameters were taken from the most recent stock assessment (SEDAR, 2018). The operating model is structured by year, age, and space. The model runs for a desired number of years which experience fishing after an unfished initialization period of 50 years. As noted in the methods, 40 years of fishing were simulated. The ages modelled start at age 0 and include a plus group at age 20. The abundance at age within a spatial cell is calculated

using

$$N_{y,a,c} = \begin{cases} R_y \hat{X}_c & \text{if } a = 0 \\ \sum_c X_{y,a,c'} \left[N_{y-1,a-1,c'} e^{-(F_{y-1,a-1,c'} + M_a)} \right] & \text{if } 0 < a < 20+ \\ \sum_c X_{y,a,c'} \left[N_{y-1,a-1,c'} e^{-(F_{y-1,a-1,c'} + M_a)} + N_{y-1,a,c'} e^{-(F_{y-1,a,c'} + M_a)} \right] & \text{if } a = 20+ \end{cases}$$

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, \hat{X}_c 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 $X_{y,a,c'}$ denotes the proportion of individuals of a given age that move from cell c' to c (age 0s 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)

$$R_y = \frac{4bR_0SB_y}{SB_0(1-b) + SB_y(5b-1)},$$

where b denotes steepness (the proportion of unfished recruitment produced at 20% of the unfished spawning stock biomass), 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 \hat{X}_c . Note that the recruitment spatial distribution is independent of year (and thus density).

Operating model parameterization

The red snapper stock assessment (SEDAR, 2018) spans the entire Gulf of Mexico contained within the United States 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 m in depth. 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.

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 $\sim 64\%$) and eastern GOM (mean $\sim 36\%$), 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 to 70 m in Florida waters by the total spatial cells with depths from 10 to 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 $+2$ SD of depth preference of an age-0 red snapper (see movement section). Unfished recruitment of red snapper for Florida (and thus the spatial model) was then calculated

culated 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.

Parameterizing movement

The movement matrix is 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 modelling on preference-type movement from the spatially explicit stock assessment platform/programme 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, and 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

$$\hat{X}_c = \frac{\hat{p}_c}{\sum_c \hat{p}_c}.$$

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

Preference functions

Spatially referenced red snapper catch at age data was compiled from the US Gulf of Mexico reef fish bottom long-line and vertical line observer databases. This database contained captures-at-age for red snapper ages 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 s (30 arc s = 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-s 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. 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. Bottom substrate data were collected from the NOAA Gulf of Mexico Data Atlas (<https://www.ndbc.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 >66%, and subdominant classifications if the most abundant fraction is >33%, resulting in eight substrate classes.

Distance

The distance preference function, referencing the distance from one cell to another cell, was modelled as an exponential decay by Euclidean distance in km from the midpoint of the cell.

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

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 travelled in a year, corrected for time at liberty. We omitted all recaptures where fish spent <200 d at liberty, as the daily movement rates were much higher for fish that spent <200 d at liberty. Tag-recapture data on red snapper was obtained from Addis *et al.* (2013).

Density

The density preference function was modelled as an exponential decay below a density threshold (Bentley *et al.*, 2004).

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

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 ($\sum_a N_{y,a} L_a^2$, where L_a refers to the length of a fish age a). The density threshold D^* was arbitrarily set at the 75% quantile of the unfished densities (year one of the model) in each cell. The decay rate, λ_{DD} was arbitrarily set at 0.5.

Fishing

Total effort in each year was simulated by first defining the average effort as a logistic increase from year 51 (first year after 50 years of no fishing) to 75% of the fishing time series (year 80 for a 40-year fishing time series), followed by a linear decrease in effort for the final 25% of the time series (Supplementary Table 1). 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. Variabil-

ity was then added to the average effort timeseries by drawing from a normal distribution with a specified CV (set at 0.25), to obtain the total effort in each year.

Effort was assumed to originate from a port at the coastline centre point for the 23 coastal counties in Florida that border the Gulf of Mexico. 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 with variation added by drawing from a multinomial distribution.

In one formulation of the operating model, the amount of fishing effort each spatial cell received in each year from a given port was modelled 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 modelled as a negative exponential decay (with one parameter λ_{FD}). The exploitable biomass profit function was modelled 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

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

Where $P(E_{p,c})$ 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).

In an alternative formulation of the operating model, effort can be distributed randomly over space. The random effort distribution OM formulation proves particularly useful in simulation testing estimation models which are largely correctly specified in their processes. Variation in effort allocation for each OM formulation was included using the multinomial distribution with probabilities calculated from either the gravity model or the random distribution and sample size as the total amount of effort originating from a port in a given year. A baseline of at least 1 unit of effort was assumed for all cells during active fishing years.

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) and a logistic fishery selectivity, which modelled the selectivity of fish at age within a cell (cs_a). 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).

$$F_{y,a,c} = cq \times cs_a \times E_{y,c}$$

To create dome-shaped selectivity within the OM where effort is randomly distributed over space, we changed the fishery contact selectivity solely within this OM formulation to be dome-shaped according to a double-normal functional form (Methot and Wetzel, 2013). We provide mathematical detail of the double-normal in [Appendix B](#).

Process variation

Process variation was included in recruitment using a lognormal distribution (log scale standard deviation set at 0.3), total effort time series using a normal distribution (with CV), 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.

Sampling model

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 can be thought of as analogous to sampling the catch of a subset of trips.

The SM sampled a pre-specified percentage of the total effort in the fishery each year (identified in main text). 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 (specified in the main text). 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. This generated 100 age composition datasets 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 ([Figure 2](#)).

Fishery independent survey age composition

Fishery independent surveys were simulated by randomly sampling spatial cells in the matrix at the start of the year. Within a year, a pre specified number of spatial cells were to be sampled (50). 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 ($cs_a \times N_{y,a,c}$) 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. The same contact selectivity was used for the survey as for the fishery (simple logistic). 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.

Fishery independent index

A fishery-independent index of abundance 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).

Appendix B

Double-normal selectivity

The double-normal selectivity function was parameterized with six parameters ($B1$ – $B6$) as in Methot and Wetzel (2013)

$$cs_a = asc_a \times (1 - j1_a) + j1_a \times [(1 - j2_a) + j2_a \times dsc_a].$$

Where $j1$, $j2$, asc , and dsc , can be found using

$$j1_a = \left[1 + e^{\left(-20 \frac{a-B1}{1+|a-B1|} \right)} \right]^{-1}, \quad j2_a = \left[1 + e^{\left(-20 \frac{a-peak2}{1+|a-peak2|} \right)} \right]^{-1},$$

$$asc_a = (1 - e^{-B5})^{-1} + \left[1 - (1 + e^{-B5})^{-1} \right] \times \frac{e^{\left[\frac{-(a-B1)^2}{e^{B3}} \right]} - e^{\left[\frac{-(\min(a)-B1)^2}{e^{B3}} \right]}}{1 - e^{\left[\frac{-(\min(a)-B1)^2}{e^{B3}} \right]}}$$

$$dsc_a = 1 + \left[(1 + e^{-B6})^{-1} - 1 \right] \frac{e^{\left[\frac{-(a-peak2)^2}{e^{B4}} \right]} - 1}{e^{\left[\frac{-(\max(a)-peak2)^2}{e^{B4}} \right]} - 1},$$

$peak2$ is calculated using

$$peak2 = B1 + 1 + \left[\frac{0.99 \times \max(a) - B1 - 1}{1 + e^{-B2}} \right].$$

The parameters specified used within the operating model were $B1 = 2.667$, $B2 = -15.885$, $B3 = 0.4$, $B4 = 1.372$, $B5 = -4.010$, and $B6 = 0.375$ (Supplementary Table 1). These parameters were chosen to resemble the shape of the realized selectivity, which emerged from the gravity model formulation of the OM (Figure 1).

Appendix C

For the logistic-normal negative log-likelihood [Table 3, Equation (3.8)], $V_y = K C_y K^T$, where K is a matrix with dimensions $[(b-1), b]$ formed by adding a vector of -1 s to the right side of an identity matrix with dimensions $[b-1, b-1]$, and x is a matrix where each row depicts a year and contains a vector of length $(b-1)$, filled according to

$$x_{y,a} = \log \left(\frac{P_{a,y}}{P_{b,y}} \right) - \log \left(\frac{\hat{P}_{a,y}}{\hat{P}_{b,y}} \right) \text{ for } a \text{ in } 0, 1, 2, \dots, b-1.$$

Handling editor: Ernesto Jardim