National Marine Fisheries Service NOAA

Fishery Bulletin
established in 1881

Spencer F. Baird
First U.S. Commissioner of Fisheries and founder of *Fishery Bulletin*

A comparison of 4 primary age-structured stock assessment models used in the United States

Bai Li (contact author)[1]
Kyle W. Shertzer[2]
Patrick D. Lynch[3]
James N. Ianelli[4]
Christopher M. Legault[5]
Erik H. Williams[2]
Richard D. Methot Jr.[6]
Elizabeth N. Brooks[5]
Jonathan J. Deroba[5]
Aaron M. Berger[7]
Skyler R. Sagarese[8]
Jon K. T. Brodziak[9]
Ian G. Taylor[6]
Melissa A. Karp[10]
Chantel R. Wetzel[6]
Matthew Supernaw[11]

Email address for contact author: bai.li@noaa.gov

[1] National Research Council
1315 East-West Highway, Building SSMC3
Silver Spring, Maryland 20910

[2] Southeast Fisheries Science Center
National Marine Fisheries Service, NOAA
101 Pivers Island Road
Beaufort, North Carolina 28516-9722

[3] Office of Science and Technology
National Marine Fisheries Service, NOAA
1315 East-West Highway, Building SSMC3
Silver Spring, Maryland 20910-3282

[4] Alaska Fisheries Science Center
National Marine Fisheries Service, NOAA
7600 Sand Point Way NE, Building 4
Seattle, Washington 98115-6349

[5] Northeast Fisheries Science Center
National Marine Fisheries Service, NOAA
166 Water Street
Woods Hole, Massachusetts 02543

[6] Northwest Fisheries Science Center
National Marine Fisheries Service, NOAA
2725 Montlake Boulevard East
Seattle, Washington 98112

[7] Northwest Fisheries Science Center
National Marine Fisheries Service, NOAA
2032 SE OSU Drive
Newport, Oregon 97365-5275

[8] Southeast Fisheries Science Center
National Marine Fisheries Service, NOAA
75 Virginia Beach Drive, Building 1
Miami, Florida 33149-1003

[9] Pacific Islands Fisheries Science Center
National Marine Fisheries Service, NOAA
1845 Wasp Boulevard, Building 176
Honolulu, Hawaii 96818

[10] ECS Federal, LLC
Office of Science and Technology
1315 East-West Highway, Building SSMC3
Silver Spring, Maryland 20910-3282

[11] Office of Science and Technology
National Marine Fisheries Service, NOAA
263 13th Avenue South
St. Petersburg, Florida 33701

Abstract—The National Marine Fisheries Service conducts fishery stock assessments to provide the best scientific information available for the U.S. regional fishery management councils. The assessment models applied in the United States are often region specific, although the models share similar mathematical and statistical attributes. However, comprehensive comparison studies identifying similarities and differences among these assessment models remain scarce. We developed a multi-model comparison framework to evaluate the reliability of 4 age-structured assessment models that are commonly used in the United States: the Assessment Model for Alaska, the Age Structured Assessment Program, the Beaufort Assessment Model, and Stock Synthesis. When applied to simulated data, all 4 models produced reliable estimates of assessment quantities of interest, such as fishing mortality, spawning biomass, recruitment, and biological reference points. Although there were differences among models in the calculation of the initial population numbers at age and in the bias adjustment of recruitment, their effects on model outputs were minor when estimation models were configured similarly. In addition, we provide guidelines for converting unfished recruitment and steepness between 2 methods of bias adjustment. We recommend that next-generation stock assessment models include recruitment bias adjustment and that more research be conducted to provide guidelines for which methods might be preferred under which situations.

Manuscript submitted 29 October 2020.
Manuscript accepted 12 July 2021.
Fish. Bull. 119:149–167 (2021).
Online publication date: 30 August 2021.
doi: 10.7755/FB.119.2-3.5

Fishery stock assessment models have been widely used by scientific and management communities to evaluate fish population dynamics and provide estimates of stock abundance and fishing mortality rate (F) (Hilborn and Walter, 1992; Quinn and Deriso, 1999; Maunder and Punt, 2013; Lynch et al., 2018). Over time, assessment models have become more comprehensive and more efficient to facilitate the integration of data from diverse sources and to use increasing computational power. Concurrently, the complexity of stock assessments has increased considerably, creating challenges for both analysts and reviewers of stock assessments. Given the range of stocks requiring assessments, using generic stock assessment models that are not

stock or region specific is common practice (NRC, 1998; Dichmont et al., 2016). Indeed, many of the age-based models used for stock assessments conducted in the United States use software developed for generic use.

In some literature reviews, the minimum data requirements, output, and projection capabilities of stock assessment models in the United States have been compared to facilitate model choice (NRC, 1998; Dichmont et al.[1]). In addition, simulation-based research has been designed to evaluate the performance of various models and their ability to meet the needs of fishery managers (Smith et al., 1993; Sampson and Yin, 1998; Cadrin and Dickey-Collas, 2015; Deroba et al., 2015). However, comparison of assessment models through both side-by-side code comparison and simulation tests remains scarce.

Four age-structured stock assessment models are used most commonly in the United States. We refer to them as models, although in reality they are software packages that may be configured to represent various forms of mathematical models. Three of the models, the Age Structured Assessment Program (ASAP) (Legault and Restrepo, 1999), the Beaufort Assessment Model (BAM) (Williams and Shertzer, 2015), and Stock Synthesis (SS) (Methot and Wetzel, 2013), are frequently used for assessments on the East Coast. On the West Coast of the continental United States, recent assessments are primarily conducted by using SS. The development trajectory for the Assessment Model for Alaska (AMAK) (AFSC, 2015) is similar to that for the ASAP. Alaska fishery scientists created simple age-structured statistical models by using a program that implements automatic differentiation, AD Model Builder (e.g., Ianelli and Fournier, 1998; Fournier et al., 2012), but they mostly tailored them for the individual stock characteristics and types of data available. From these bespoke models, a more general model, the AMAK, was developed and applied to a number of stocks, such as walleye pollock (*Gadus chalcogrammus*) in the Aleutian Islands region (Barbeaux et al., 2019). Although there are additional models and diagnostic tools that can be used to assess fish stocks (Dichmont et al.[1]), these 4 age-structured assessment models are the main approaches applied in the United States for "data-rich" stock assessments, and they share similar conceptual, mathematical, and statistical frameworks.

The varied development and preference among regions for different assessment models may be attributed to special requirements or features of a stock assessment model given the availability of observed data (data collection programs vary regionally) and length of time of commercial fishery operations (periods of hundreds of years on the East Coast versus periods that began roughly after World War II in Alaska or more recently on the West Coast) that create different states of initial depletion. Additional reasons for different software and modeling approaches include inertia and continuity with past practices (or application to similar stocks), region-specific training, and the presence of local expertise (Cadrin and Dickey-Collas, 2015). The availability of different assessment approaches may provide flexibility, but it also requires testing to determine how different assumptions affect results. A first step is to test whether various models of a given type produce similar estimates when configured similarly without introducing misspecifications to the models. Then identifying sources of any differences can inform and improve assumptions used in actual assessments (e.g., NRC, 1998).

Simulation testing provides a means to evaluate the accuracy of individual assessment models because we know the correct values used to generate the data. An *operating model* (OM) is configured to reflect hypotheses about true stock dynamics and is the basis for generating age-structured stock assessment inputs for each assessment model (which are referred to as *estimation models* [EMs]). The OM-EM framework to fit EMs to simulated data (with errors) has been used previously to assess the ability of assessment models to estimate stock conditions (Wetzel and Punt, 2011; Henríquez et al., 2016). Deroba et al. (2015) conducted both self-tests and cross-tests from a simulation-estimation framework to compare the robustness of assessment models to error. A self-test fits an assessment model to the simulated data generated from the same assessment model, and a cross-test fits an assessment model to data generated from a different model (Chang et al., 2015; Deroba et al., 2015). They highlighted that the lack of robustness in self-tests may indicate bias and that a lack of robustness in cross-tests may indicate differences in structural assumptions between assessment models. To avoid the bias introduced during the cross-test process, we attempted to develop an OM based on common features of the 4 EMs.

The aim of this study was to improve our understanding of both the similarities and differences among 4 primary age-structured stock assessment models used in the United States. To our knowledge, this evaluation is the first in which a comprehensive comparison of source codes and a simulation-estimation analysis of these models has been conducted. This study specifically addressed the following 4 primary questions: What are the key features and source code that need to be examined before developing an OM for comparing multiple EMs? Do the EMs give similar and accurate estimates under a range of cases? What are the sources of differences in estimates, if there are any? What recommendations can be drawn for future model development after examining the similarities and differences of the 4 EMs in our study? Addressing these questions is critical for improving the understanding of current models and for developing next-generation stock assessment models (Punt et al., 2020).

Materials and methods

General framework

To compare assessment models, we conducted a comparison of key features in the code from the 4 EMs as well as

[1] Dichmont, C. M., A. R. Deng, A. E. Punt, and L. R. Little. 2017. Stock assessment integration: a review. Fish. Res. Dev. Corp. Rep. 2014-039, 106 p. CSIRO Publ., Hobart, Australia. [Available from website.]

OM-EM simulation tests to compare estimates to true values (Fig. 1). The code comparison process helps to verify whether the code from the 4 EMs executes the intended algorithms the same way and to identify common features from the EMs to develop an OM (Table 1, Suppl. Material, Suppl. Table 1) with only those commonalities (NRC, 1998) (Table 2). The simulation-estimation process, which helps validate the accuracy of EMs, consisted of 4 main steps: 1) developing an OM to simulate annual fish population and fishery dynamics, 2) fitting the 4 EMs to the simulated data, 3) repeating the simulation-estimation 100 times with different recruitment deviations and observation errors for each iteration, and 4) comparing estimates from the EMs with the true values from the OM (Fig. 1). This process was repeated for 13 cases (Table 3). Comparisons were made among the 4 EMs within each case and across cases.

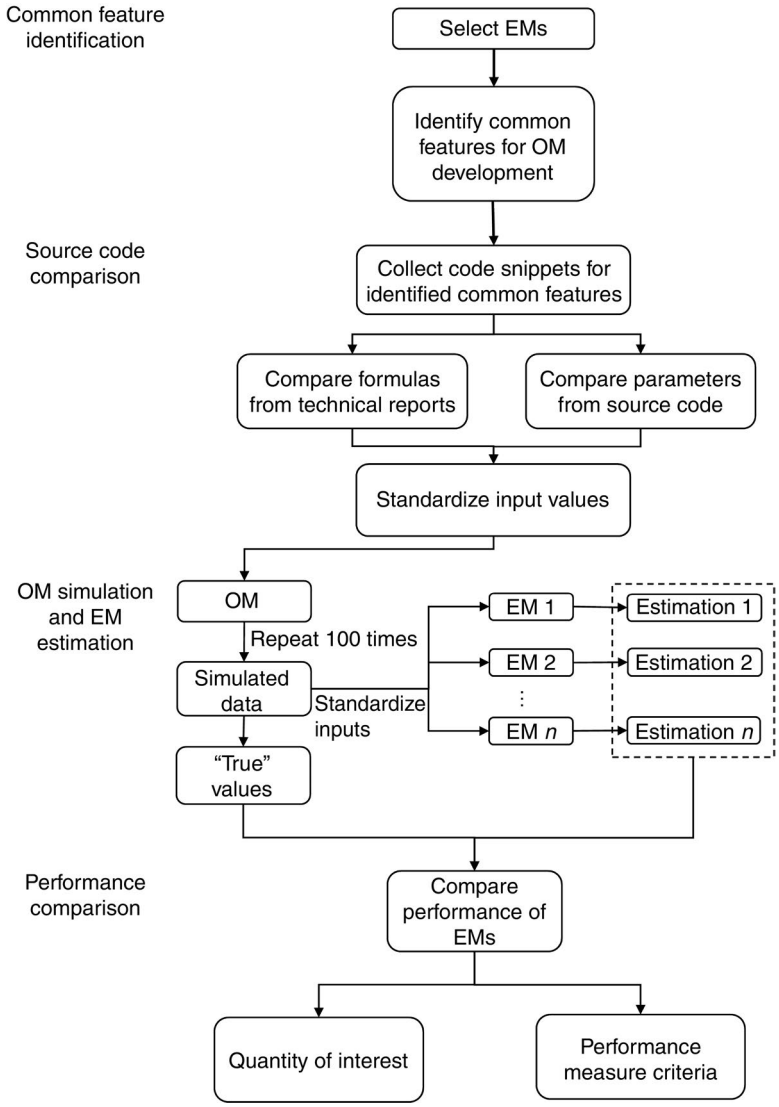

Figure 1

Flow diagram of the processes used to compare 4 age-structured stock assessment models used in the United States. Steps 1 and 2, which compose the code comparison process, involve identification of common features and source code comparison of the estimation models (EMs). Steps 3 and 4, which compose the simulation-estimation process, involve operating model (OM) development, estimation with EMs, and comparison of performance between the EMs. Common features, input standardization, quantities of interest, and performance measures are described in the "Materials and methods" section and in Table 1. The "true" values include actual values used as inputs for development of the OM and simulated true values for quantities of interest.

Operating model and comparison cases

The OM developed in this study was an age-structured model, with parameter values describing life history obtained from Siegfried et al. (2016). That study simulated a population on the basis of an amalgam of life history traits common to species found in waters of the Atlantic Ocean off the southeastern United States. In our study, the OM population was simulated with an annual time step over 30 years and a maximum age of 12 years (Table 1, Suppl. Material, Suppl. Table 1). The OM null case (case 0) had one fishing fleet and one survey, with fully selected F linearly increasing with time. A time-invariant logistic selectivity function was used for both the fishing fleet and the survey in the null case. Fishery landings and survey abundance index data were simulated yearly with observation error from year 1 to year 30. The annual sample size was 200 samples for age composition data from both the fishery and survey. In all cases except case 12, the initial equilibrium recruitment was lowered from the unfished recruitment level as determined by an initial equilibrium F (spawning biomass per recruit based on F [ϕ_F] is less than unfished spawning biomass per recruit [ϕ_0], as in equation 3.4 provided in Supplementary Table 2). The addition of case 12, for which the initial condition was equal to the unfished equilibrium population (ϕ_F=ϕ_0), as in equation 3.4 provided in Supplementary Table 2, allowed a comparison of methods for simulating the initial population. Details of parameter definitions and equations used to describe the OM under case 0 are presented in Table 1 and in the Supplementary Material (the code for creating the OM and comparing the EM results is available from website).

Eleven additional cases were explored to investigate the effects of recruitment variability, process error in F, patterns in

Table 1

Description and values for index variables, structural parameters, state variables, derived variables, and stochastic deviation used in the operating model to generate simulated data that served as inputs for the 4 age-structured stock assessment models evaluated in this study. Models were tested under different cases, including the null case (case 0), which has one fishing fleet and one survey, with fully selected fishing mortality rate (F) increasing linearly with time. Also provided are the value or expression in case 0 and indication of whether a parameter is estimated in the estimation model. The parameter natural mortality rate at age is assumed to be constant across years and age classes.

Symbol	Description	Case 0 value or expression	Estimated
Index variables			
y	Years	{1,2,. . .,Y} and Y=30	
a	Ages	{1,2,. . .,A} and A=12+	
Structural parameters			
l_∞	Asymptotic average length (in millimeters)	800	
K	Growth coefficient (per year)	0.18	
t_0	Age at mean length of zero (in years)	−1.36	
θ_1	Length–weight coefficient	2.50×10^{-8}	
θ_2	Length–weight exponent	3	
θ_3	Slope of maturity ogive	3	
a_{50}	Age at 50% maturity (in years)	2.25	
M_a	Natural mortality rate at age (per year)	0.2	
r_a	Proportion of females at each age	0.5	
$R0$	Median-unbiased unfished recruitment (in number)	1 million	Yes
h	Median-unbiased steepness	0.75	
x_1	Fishery selectivity slope	1	Yes
x_2	Fishery selectivity age at 50% selection (in years)	2	Yes
x_3	Survey selectivity slope	2	Yes
x_4	Survey selectivity age at 50% selection (in years)	1.5	Yes
f_y	Shape of the fully selected F in year y	Linear increase with f_1=0.01 and f_Y=0.40	Yes
φ_F	Annual sample size for age composition of fishery landings	200	
φ_I	Annual sample size for age composition of survey	200	
State variables			
R_y	Annual recruitment in year y (in number)		
SSB_y	Spawning biomass in year y (in metric tons of female biomass)		
$N_{a,y}$	Abundance at age a in year y (in number)		
A_y	Abundance in year y (in number)		
B_y	Biomass in year y (in metric tons)		
$L_{a,y}$	Landings at age a in year y (in number)		

(*Continued on next page*)

F, selectivity patterns, number of surveys, and bias adjustment of recruitment on performance of the EMs (Table 3). Because our goal was to demonstrate that the 4 EMs are similar at their core, we started with simple cases to compare the performance of the EMs before adding in additional complexity. It would have been harder to interpret the differences found in the results if more complicated cases had been included in the simulation. We did not implement these cases to evaluate how model misspecifications introduced in the EMs could affect the model estimates; therefore, we made the correct assumptions about parameters (e.g., fixing them at correct values) and governing processes (e.g., stock–recruit relationship and selectivity pattern).

Recruitment variability level In case 0, the standard deviation of annual recruitment variability (σ_R) in log space was 0.2. Two additional levels were explored to examine if higher recruitment variability with σ_R of 0.4 (case 1) and 0.6 (case 2) affected the performance of the EMs.

Process error in fishing mortality A comparison between case 3 and case 0 addressed the question of whether additional process error in F affected the performance of the EMs. In all cases, the standard deviation of the log fully selected F was used for generating annual deviations in fully selected F. In case 0, the same set of annual deviations was used for each iteration. In case 3, we randomly generated stochastic sets of annual deviations per iteration. Case 0 did not start with stochastic sets of annual deviations per iteration because it can be difficult to interpret differences in results if the process error in F is confounded with other settings or assumptions.

Table 1 (*continued*)

Symbol	Description	Case 0 value or expression	Estimated
L_y	Landings in year y (in metric tons)		
$I_{a,y}$	Survey abundance at age a in year y (in number)		
I_y	Survey abundance (sum across ages) in year y (in number)		
L'_y	Observed landings in year y (in metric tons) with noise		
$P_{L_{a,y}}$	Proportion at age a in year y for fishery landings		
$P'_{L_{a,y}}$	Observed proportion at age a in year y for fishery landings		
I'_y	Observed survey abundance in year y with noise (in number)		
$P_{I_{a,y}}$	Proportion at age a in year y for survey		
$P'_{I_{a,y}}$	Observed proportion at age a in year y for survey		
Derived variables			
l_a	Length at age a (in millimeters)		
w_a	Weight at age a (in metric tons)		
m_a	Proportion that reached maturity at age a		
ϕ_0	Unfished spawning biomass per recruit (in metric tons)		
$Z_{a,y}$	Total mortality rate at age a in year y		
$Fmult_y$	Fully selected F in year y (year^{-1})		
S_{F_a}	Fishery selectivity at age a		
S_{I_a}	Survey selectivity at age a		
Φ_a	Number of spawners per recruit at age		
ϕ_F	Spawning biomass per recruit given F (in metric tons)		
R_{eq}	Equilibrium recruitment (in number)		
q	Catchability coefficient for survey	3.46×10^{-7}	Yes
Stochastic deviation: process error			
σ_R	Standard deviation of log recruitment	0.2	
$Rdev_y$	Recruitment deviations in year y	$Rdev_y \sim N(0, {\sigma_R}^2)$	Yes
σ_F	Standard deviation of log fully selected F	0.2	
$fdev_y$	Fully selected F deviations in year y	$fdev_y \sim N(0, {\sigma_F}^2)$	Yes
Stochastic deviation: observation error			
CV_L	Coefficient of variation of fishery landings	0.05	
ε_{1_y}	Landings deviations in year y	$\varepsilon_{1_y} \sim N\left(0, \log\left(1 + CV_L^2\right)\right)$	
CV_I	Coefficient of variation for survey	0.2	
ε_{2_y}	Survey abundance deviations in year y	$\varepsilon_{2_y} \sim N\left(0, \log\left(1 + CV_I^2\right)\right)$	

Fishing mortality patterns In case 0, the fully selected F increased over time from a relatively low value (0.01) to a higher rate (F_{high}) (Fig. 2). Two additional trends in F were investigated on the basis of methods from Johnson et al. (2015) and Ono et al. (2015). In this study, the F either increased from a low value (0.01) to F_{high} during the first 24 years and then decreased to a lower rate (F_{low}; case 4; Fig. 2) or that F remained constant over time across 3 levels: F_{low}, the F that corresponds to maximum sustainable yield (F_{MSY}), or F_{high} (cases 5–7; Fig. 2). Values of F_{low} and F_{high} corresponded to 80% of maximum sustainable yield (MSY). Comparing cases 4–7 with case 0 allowed an evaluation of whether different fishing patterns affected the magnitude of error in estimating parameters of interest.

Selectivity patterns A comparison between case 8 and case 0 was used to examine the influence on assessment performance of using double-logistic selectivity for the fishery and the survey instead of simple logistic selectivity (case 0).

Multiple surveys Comparing case 9 (which includes 2 surveys with the same level of observation errors) with case 0 (which includes 1 survey) allowed an evaluation of whether use of an additional survey reduced error in estimating quantities of interest.

Bias adjustment of recruitment For case 10, the median-unbiased spawner-recruit parameters were used, whereas for case 11 the mean-unbiased spawner-recruit parameters were used to conduct a bias adjustment in the OM. The arithmetic mean curve of recruitment that is associated with the mean-unbiased spawner-recruit parameters is higher than the geometric mean curve of recruitment

Table 2

Comparison of features between the 4 age-structured estimation models (EMs) evaluated in this study: the Assessment Model for Alaska (AMAK), the Age Structured Assessment Program (ASAP), the Beaufort Assessment Model (BAM), and Stock Synthesis (SS). The letter *Y* indicates that the feature is implemented in the EM. *F*=fishing mortality rate.

Feature	AMAK	ASAP	BAM	SS
Age modeled	1+	1+	1+	0+/1+
Timing of spawning	Real month	Fraction	Fraction	Real month
Timing of survey	Real month	Real month	Fraction	Real month
Survey index unit	Biomass/number	Biomass/number	Biomass/number	Biomass/number
Spawner-recruit model				
Standard Beverton–Holt	Y	Y	Y	Y
Ricker	Y		Y	Y
Average recruitment	Y	Y	Y	Y
Bias adjustment of recruitment			Y	Y
Types of selectivity available				
Free parameter approach	Bound	Random walk	Logit	Random walk/logit
Simple-logistic function	Y	Y	Y	Y
Double-logistic function	Y (3 parameters)	Y (4 parameters)	Y (4 parameters)	Y (4 parameters)
Logistic-exponential function			Y	Y
Joint-logistic function			Y	Y
Double-Gaussian function			Y	Y
F in terminal year	Last year	Last year	Flexible	Last year
Definition of *F*	Flexible	Apical *F*	Apical *F*	Flexible
Likelihoods available				
Landings, lognormal	Y	Y	Y	Y
Survey index, lognormal	Y	Y	Y	Y
Age composition, standard multinomial	Y	Y	Y	Y
Age composition, Dirichlet multinomial			Y	Y
Priors				
None	Y	Y	Y	Y
Lognormal	Y	Y	Y	Y
Beta			Y	Y
Normal			Y	Y
Reference or website (see for details of other features)	website	website	Williams and Shertzer (2015)	website

that is associated with median-unbiased spanwer-recruit parameters because of lognormal deviation in recruitment residuals (Methot and Taylor, 2011).

Bias adjustment is handled differently in the 4 EMs. In the BAM, a bias adjustment is applied when median-unbiased parameters are used to compute equilibrium recruitment for the spawner-recruit model (Suppl. Table 3) (Williams and Shertzer, 2015). In contrast, in SS, mean-unbiased parameters are used for the spawner-recruit model, and then a bias adjustment is applied when computing annual recruitment (Suppl. Table 3) (Methot and Taylor, 2011). In the AMAK and ASAP, bias adjustment is not included as part of the internal machinery. Details of differences between the EMs are documented in the "Spawner-recruit parameters in bias adjustment of recruitment" subsection of the "Results" section and in Supplementary Table 3. For cases 10 and 11, σ_R was set to 0.6 to make the differences in estimates noticeable, if any were present. We also adjusted the estimates of MSY-based reference points from the AMAK and ASAP with the BAM bias adjustment method in case 10 to make estimates comparable among all 4 EMs (Suppl. Figs. 1 and 2). We adjusted the estimates of unfished recruitment ($R0$) and steepness (h) to mean-unbiased values and then adjusted MSY-based reference points from the AMAK and ASAP in case 11 (Suppl. Figs. 1 and 2). The estimates from cases 10 and 11 were compared with the estimates from case 2 to quantify the effect of bias adjustment methods on EM performance.

Estimation models

Four EMs were evaluated in this study. The AMAK was compiled with AD Model Builder, vers. 12.1, from source code available on GitHub (website, accessed November 2019). We used the executable file of the ASAP, vers. 3.0.16, available on the National Marine Fisheries Service Integrated Toolbox website (accessed December 2019). For the BAM, the source code is available from the authors or in the appendix of a NOAA publication (Williams and Shertzer, 2015) and was compiled by using AD Model Builder. We used the executable file of SS, vers. 3.30.15,

Table 3

Settings for recruitment variability (σ_R), deviations in fishing mortality rates (F), patterns of F, selectivity patterns, and recruitment bias adjustment in the operating model (OM) that was used to evaluate 4 age-structured stock assessment models under different cases. A dash denotes that the value or information is the same as that given for the null case (case 0). For case 0, the OM creates the same set of F deviations for a given iteration. For case 3, the OM randomly generates a unique set of F deviations per iteration. F_{low}=low F value; F_{MSY}=F that corresponds to maximum sustainable yield; F_{high}=high F value; ϕ_F=spawning biomass per recruit based on F; ϕ_0=unfished spawning biomass per recruit; *R0*=unfished recruitment; and h=steepness.

		Details						
Setting	Case	σ_R	F deviations	F patterns	Fishery and survey selectivity	No. of surveys	Bias adjustment	Initial condition
Null case	0	0.2	Same iteration	Increase	Simple-logistic	1	No	$\phi_F \neq \phi_0$
Recruitment variability	1	0.4	–	–	–	–	–	–
	2	0.6	–	–	–	–	–	–
Stochastic F	3	–	Stochastic iteration	–	–	–	–	–
F patterns	4	–	–	Roller coaster	–	–	–	–
	5	–	–	Constant F_{low}	–	–	–	–
	6	–	–	Constant F_{MSY}	–	–	–	–
	7	–	–	Constant F_{high}	–	–	–	–
Double-logistic selectivity	8	–	–	–	Double-logistic	–	–	–
Multiple surveys	9	–	–	–	–	2	–	–
Recruitment bias adjustment	10	0.6	–	–	–	–	Yes (median-unbiased *R0* and h)	–
	11	0.6	–	–	–	–	Yes (mean-unbiased *R0* and h)	–
Initial condition	12	–	–	–	–	–	–	$\phi_F = \phi_0$

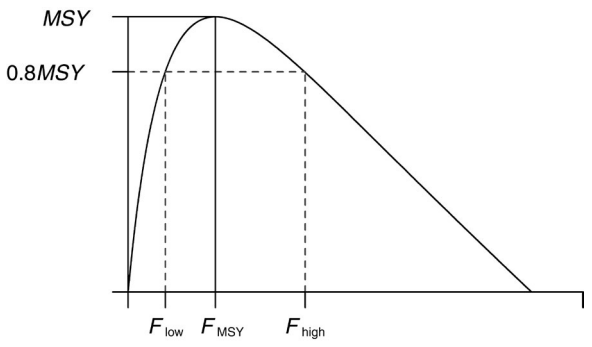

Figure 2

The curve of the relationship of yield and fishing mortality rate (F) and the definitions of the low F value (F_{low}) and the high F value (F_{high}) used in creation of various patterns of F in the operating model. The horizontal black line indicates maximum sustainable yield (MSY), and the dashed gray lines indicate 0.8*MSY*, which corresponds to F_{low} and F_{high}. The vertical black line indicates the F that corresponds to MSY (F_{MSY}).

available from the NOAA Virtual Laboratory (website, accessed November 2019).

The 4 EMs estimated time series of spawning stock biomass (SSB), recruitment, F, abundance, biomass, and landings in weight. Annual age composition of landings from the fleet and from surveys was also estimated. The annual F values used in the comparison were apical F values. The common output of biological reference points used in the comparison project included MSY, F_{MSY}, and SSB at MSY (SSB_{MSY}). The relative F (calculated as F/F_{MSY}) and relative SSB (calculated as SSB/SSB_{MSY}) for MSY-based reference points were also examined.

Performance measures

For each of the cases, performance of the EMs was evaluated by comparing estimated values against the true values from the OM. The quantities of interest included *R0*, F in the "terminal" year ($F_{terminal}$), $SSB_{terminal}$, MSY, F_{MSY}, SSB_{MSY}, relative F, and relative SSB. To evaluate performance related to a management application (e.g., stock status determination), F_{limit} was set to F_{MSY} and SSB_{limit} was set to half of SSB_{MSY} (Federal Register, 1998; Restrepo

et al., 1998; Gabriel and Mace, 1999; Methot et al., 2014). The estimated values of F/F_{limit} and SSB/SSB_{limit} from the EMs were compared with the true values from the OM to verify whether the estimated stock status (i.e., overfished or overfishing) matched the true status.

The bias and variability of the bias of the EMs were determined by calculating relative error (RE) and median absolute relative error (MARE) for key parameters. The RE and MARE for each EM within a case were calculated as follows:

$$RE_{i,j,t} = (E_{i,j,t} - T_{i,j,t}) / T_{i,j,t} \text{ and} \quad (1)$$

$$MARE_{i,t} = median(|RE_{i,j,t}|), \quad (2)$$

where E = the estimated quantity of interest;
T = the true value from the OM;
i = the quantity of interest;
j = the iteration number; and
t = the year, if applicable.

For evaluating performance related to making stock status determinations (i.e., using F/F_{limit} and SSB/SSB_{limit}), the accuracy of each model under a case equals the number of correctly identified positives and negatives divided by the total number of iterations.

Results

Code comparison process

Identification of common features The structure of the OM and cases were defined on the basis of the similarities and differences found among the 4 EMs (Table 2). On the basis of the comparisons, we found that the common available spawner-recruit model among the EMs is the compensatory Beverton–Holt spawner-recruit model (Beverton and Holt, 1957) with lognormally distributed recruitment deviations. For selectivity patterns, all of the EMs have both a simple-logistic function and a double-logistic function. For the double-logistic function, the ASAP, BAM, and SS each require 4 parameters, and the AMAK requires 3 parameters. Bias adjustment of estimated mean recruitments is implemented in only the BAM and SS. For some features (e.g., available selectivity patterns), some of the EMs have more options than those listed in Table 2. We limited the options to features available in at least 2 EMs. All available options can be found in the technical manual for each EM. Differences among other features have been summarized in Dichmont et al. (2016), Dichmont et al.[1], and Punt et al. (2020).

Basic settings to ensure similar configurations Results of the comparison of formulas used in source code for key features indicate that analysts can manually adjust some basic settings to ensure that all 4 EMs are configured similarly and to ensure that a comparison study is effective. For example, in the AMAK, ASAP, and BAM, the population starts at age 1, but in SS the population routinely starts at age 0 and can be configured to start at older ages, as was done for this study. Survey index units, biomass or number of fish, can be used as input for the ASAP, BAM, and SS. In the AMAK, the default unit is biomass, but numbers could be used by setting all entries in the weight-at-age matrix to the value of 1.

Selectivity-at-age outputs can be used directly for comparison, but estimated selectivity parameters need to be further converted before being compared because they are modeled differently in EMs. The simple logistic selectivity in the AMAK and BAM share the same formula as the OM (i.e., equation 5.1 from Supplementary Table 1). In the ASAP and SS, simple-logistic selectivity is calculated as follows:

$$S_{F_a} = \frac{1}{1 + e^{-(a-\mu_1)/\mu_2}} \text{and} \quad (3)$$

$$S_{F_a} = \frac{1}{1 + e^{-\ln(19)(a-\nu_1)/\nu_2}}, \text{ respectively,} \quad (4)$$

where S_{F_a} = fishery selectivity at age;
μ_1 = age at 50% selection parameter (equal to x_2 from equation 5.1 in Supplementary Table 1, where x_2 is age at 50% selection parameter);
μ_2 = slope parameter (equal to $1/x_1$ from equation 5.1 in Supplementary Table 1);
a = ages;
ν_1 = age at 50% selection for fishery parameter (equal to x_2 from equation 5.1 in Supplementary Table 1); and
ν_2 = slope parameter (equal to $\ln(19)/x_1$ from equation 5.1 in Supplementary Table 1).

Similarly, the BAM and SS share the same double-logistic selectivity formula (Equation 5), and the formula from the AMAK (Equation 6) and ASAP (Equation 7) are different. Consequently, parameter values cannot be directly compared between models. However, the resultant curves and selectivity at age can be compared with the following equations:

$$S'_{F_a} = \frac{1}{1 + e^{-x_1(a-x_2)}} \left(1 - \frac{1}{1 + e^{-\beta_1(a-\beta_2)}}\right) \text{and} \quad (5)$$

$$S_{F_a} = S'_{F_a} / \max(S'_{F_a});$$

$$\gamma_1 = p_1 + p_2, \quad (6)$$

$$\gamma_2 = p_1 + y_1 + p_3, \text{ and}$$

$$S_{F_a} = \frac{1}{1 + e^{\frac{-\ln(19)(a-\gamma_1)}{p_1}}} \left(1 - \frac{1}{1 + e^{\frac{-\ln(19)(a-\gamma_2)}{p_3}}}\right) / 0.95^2; \text{ and}$$

$$S'_{F_a} = \frac{1}{1 + e^{-\frac{a-\mu_1}{\mu_2}}} \left(\frac{1}{1 + e^{\frac{a-\mu_3}{\mu_4}}}\right) \text{and} \quad (7)$$

$$S_{F_a} = S'_{F_a} / \max(S'_{F_a}),$$

where S'_{F_a} = fishery selectivity at age before rescaling to ensure that it peaks at 1;

x_1 = first slope of the double-logistic selectivity function;
x_2 = first inflection point of the double-logistic selectivity function;
β_1 = second slope of the double-logistic selectivity function;
β_2 = second inflection point of the double-logistic selectivity function;
γ_1 = first inflection point of the selectivity function;
γ_2 = second inflection point of the selectivity function;
p_1 = distance between first inflection point and age at 95% selection;
p_2 = parameter to be used with p_1 to get first inflection point of the selectivity function;
p_3 = second slope of the selectivity function;
μ_1 = first inflection point of the double-logistic selectivity function;
μ_2 = first slope of the double-logistic selectivity function;
μ_3 = second inflection point of the double-logistic selectivity function; and
μ_4 = second slope of the double-logistic selectivity function.

Spawner-recruit parameters in bias adjustment of recruitment In both the models in which a bias adjustment to recruitment was applied, the BAM and SS, the mean-unbiased relationship is used for estimating MSY-based reference points. The F_{MSY} from the OM was based on searching for the F that provides maximum equilibrium catch. The calculation started with computing equilibrium recruitment at F and involved the same equations that were used to calculate the initial equilibrium conditions (equation 3.4 in Supplementary Table 1). That step of calculation involved $R0$ (Suppl. Table 3), and the use of $R0$ in the calculation results in differences in estimated MSY-based reference points depending on how bias adjustment of recruitment was addressed in each EM. The median-unbiased spawner-recruit parameters from the BAM correspond to the geometric mean curve of recruitment, and the mean-unbiased parameters from SS correspond to the arithmetic mean curve of recruitment (Table 4, Suppl. Table 3).

For direct comparisons, we derived a function to convert combinations of median-unbiased $R0$ and h to mean-unbiased values, or to convert mean-unbiased values to median-unbiased values, when the Beverton–Holt spawner-recruit model was used. The derivation of the median-unbiased approach starts with calculating the median-unbiased $R0$ by using the Beverton–Holt model:

$$R0 = \frac{0.8R0hS0}{0.2\phi_0 R0(1-h)+(h-0.2)S0}, \tag{8}$$

where $S0$ = unfished spawning biomass.

Given that the bias adjustment component b equals $e^{\sigma_R^2/2}$, the mean-unbiased $R0'$ that corresponds to the arithmetic mean curve of recruitment is determined with this equation:

$$R0' = \frac{b0.8R0hS0'}{0.2\phi_0 R0(1-h)+(h-0.2)S0'}, \tag{9}$$

where $S0'$ = the mean-unbiased unfished SSB.

Then, following the derivations below, the bias-adjusted mean-unbiased $S0'$ can be obtained:

$$0.2\phi_0 R0(1-h)+(h-0.2)S0' = \frac{b0.8R0hS0'}{R0'} = b0.8R0h\phi_0, \tag{10}$$

$$0.2R0(1-h)+(h-0.2)\frac{S0'}{\phi_0} = b0.8R0h,$$

$$(h-0.2)\frac{S0'}{\phi_0} = b0.8R0h - 0.2R0(1-h), \text{ and}$$

$$S0' = \frac{b0.8R0h\phi_0 - 0.2R0(1-h)\phi_0}{h-0.2}.$$

When converting median-unbiased values to mean-unbiased values, the bias adjustment component b equals $e^{\sigma_R^2/2}$. Inputs of the conversion function include median-unbiased $R0$, h, and ϕ_0. The outputs of the conversion include adjusted unfished SSB ($S0_b$), adjusted unfished recruitment ($R0_b$), and adjusted steepness (h_b) in mean-unbiased levels:

$$S0_b = \frac{b0.8R0h\phi_0 - 0.2\phi_0 R0(1-h)}{h-0.2}, \tag{11}$$

$$R0_b = \frac{S0_b}{\phi_0}, \tag{12}$$

$$R_{new} = \frac{b0.8R0h0.2S0_b}{0.2\phi_0 R0(1-h)+(h-0.2)0.2S0_b}, \text{ and} \tag{13}$$

$$h_b = \frac{R_{new}}{R0_b}, \tag{14}$$

where R_{new} = recruitment when the SSB is 20% of its unfished level.

When converting mean-unbiased values to median-unbiased values, b equals $e^{-\sigma_R^2/2}$, and the inputs from the equation (Equation 11) need to be mean-unbiased values (the outputs are median-unbiased $S0_b$, $R0_b$, and h_b). Without these conversions, the difference between mean- and median-unbiased $R0$ and the difference between mean- and median-unbiased h gradually increased when the true median-unbiased h was reduced and when σ_R was increased (Fig. 3).

Simulation-estimation process

Null case (case 0) Model parameters of $R0$ and catchability (q) were accurately estimated in all EMs with low bias.

Table 4

Methods for bias adjustment of recruitment from 2 of the age-structured stock assessment models evaluated in this study: the Beaufort Assessment Model (BAM) and Stock Synthesis (SS). MSY=maximum sustainable yield; *R0*=unfished recruitment.

Parameter	BAM method	SS method
Unfished recruitment	Median-unbiased *R0* as input Median-unbiased *R0* and mean-unbiased *R0* as output	Mean-unbiased *R0* as input and output
Unfished biomass	Based on mean-unbiased *R0*	Based on mean-unbiased *R0*
Equilibrium recruitment	Mean-unbiased	Mean-unbiased
Spawner-recruit parameters	Median-unbiased	Mean-unbiased
MSY-based reference points	Mean-unbiased	Mean-unbiased

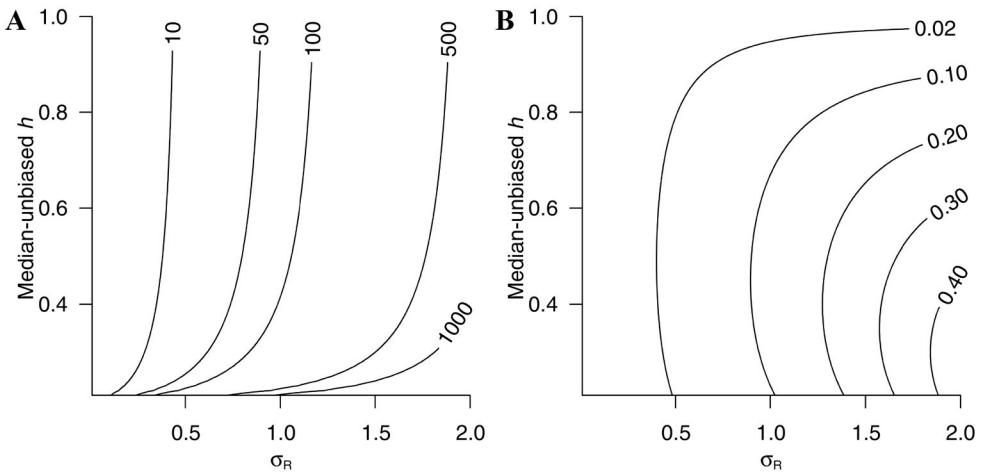

Figure 3

(**A**) Relative difference (%) in unfished recruitment (*R0*) and (**B**) difference in steepness (h) over possible combinations of median-unbiased h and standard deviation of log recruitment (σ_R), both indicated by the contour lines. The differences were determined by using the spawner-recruit parameter conversion function. Relative difference in unfished recruitment is defined as 100(mean-unbiased *R0*–median-unbiased *R0*)/median-unbiased *R0*. Relative difference in steepness is defined as mean-unbiased h–median-unbiased h.

The MAREs of *R0* and q were below 10% for all EMs (Table 5). The median *R0* over 100 iterations from the AMAK was generally lower compared with the estimates from the other EMs (cases 0–9), indicating that the AMAK initializes the population differently compared with the OM (Fig. 4). For case 12, in which the initial condition was simulated for an unfished equilibrium population, the estimated *R0* from the AMAK was similar to the estimates from the other EMs (Fig. 4).

Under case 0, the estimated median selectivity at age over 100 iterations from all EMs has almost identical patterns compared with the selectivity curves based on true values from fishery and survey sources (Suppl. Fig. 3). The MAREs of MSY, F_{MSY}, and SSB_{MSY} were below 10% (Table 6), and the REs indicate similar variability among all EMs (Fig. 5). The AMAK produced relatively lower MSY and SSB_{MSY}, along with a lower estimate of *R0*, compared with the true values and estimates from the other stock assessment models. The REs for SSB, R, F, relative SSB, and relative F centered around zero over time and had similar variability patterns (Fig. 6, Suppl. Figs. 4–8). The accuracy of stock status determination was 100% for overfished status determination, and the accuracy of the overfishing status was 100% in most but not all years (Fig. 7).

Recruitment variability level The 4 EMs accurately estimated model parameters, but the range of RE increased when σ_R increased. The MARE in key parameters increased when σ_R increased from the null case value of 0.2 (Tables 5 and 6).

Table 5

Median absolute relative error (%) for the model parameters unfished recruitment (*R0*) and catchability of survey abundance index (q) from each simulated case used to compare the 4 age-structured stock assessment models used most commonly in the United States: the Assessment Model for Alaska (AMAK), the Age Structured Assessment Program (ASAP), the Beaufort Assessment Model (BAM), and Stock Synthesis (SS).

	R0				*q*			
Case	AMAK	ASAP	BAM	SS	AMAK	ASAP	BAM	SS
Case 0	3.05	3.87	3.86	3.88	1.60	1.67	1.68	1.63
Case 1	5.69	5.72	5.68	5.80	2.27	2.26	2.24	2.19
Case 2	8.10	8.63	8.60	8.79	2.14	2.14	2.14	2.14
Case 3	2.81	2.68	2.67	2.72	1.92	1.99	1.98	1.93
Case 4	3.25	3.20	3.18	3.18	1.77	1.87	1.87	1.83
Case 5	4.45	4.30	4.28	4.20	4.09	3.88	3.85	4.11
Case 6	3.40	3.73	3.71	3.57	1.99	1.99	2.02	2.05
Case 7	3.97	3.53	3.17	3.12	1.96	2.15	2.04	2.11
Case 8	4.15	3.94	3.84	3.72	3.56	3.44	3.88	3.42
Case 9	3.15	2.97	2.95	2.98	3.17	1.79	1.79	1.80 (survey1)
Case 9	3.15	2.97	2.95	2.98	3.28	1.61	1.62	1.73 (survey2)
Case 10	7.83	9.08	9.26	9.28	1.86	1.95	1.92	1.88
Case 11	7.13	8.77	8.74	8.81	2.06	2.04	2.06	2.01
Case 12	3.04	2.96	2.95	2.67	2.20	2.19	2.22	2.14

The range of RE in MSY-based reference points became wider when recruitment variability increased (Fig. 5). Furthermore, the increased variability in SSB_{MSY} induced a wider range of RE in relative SSB (Table 6, Fig. 6). The accuracy of determining overfishing status for all EMs was not greatly affected by changes in recruitment variability, and the same trends were retained over time compared with the accuracy trends for the null case (Fig. 7, Suppl. Figs. 7 and 8). The accuracy of determining overfished status was 100% over time for all EMs.

Process error in fishing mortality There were no considerable differences found in key parameter estimates when examining the effect of process error in F (case 3 versus case 0 in Figures 5 and 6). For all 4 EMs, RE patterns and estimates of key parameters were almost identical. However, the accuracy of the overfishing determination from case 3 fluctuated at levels under 100% for a longer period than that from case 0, although the true F values were not always close to the true F_{MSY} (Fig. 7). This result indicates that incorporating stochastic sets of annual deviations in F induces sensitivity in the determination of overfishing, but the overall accuracy of the overfishing determination was still high with values often exceeding 94%. The accuracy of determining an overfished status was still 100% over time for all EMs.

Fishing mortality patterns The EMs produced low MARE in estimates of key parameters when patterns of F were different. The variability of RE in all 4 EMs was consistently higher than in case 0 when the F pattern was a constant F_{low}, indicating that the estimates had greater bias when there was not much contrast between initial F and the later period of constant F (case 5 versus case 0 in Table 6 and Figures 5 and 6). When there was a fair amount of contrast in F across time, the MARE among the 4 models remained similar in trends and magnitude. Although accuracy in determining an overfished status was 100% for all EMs, the accuracy in determining overfishing status was consistently lower compared with that for other cases when F fluctuated around F_{MSY} (case 6 in Figure 7).

Selectivity patterns When both the fishing fleet and survey had double-logistic selectivity, the estimated median selectivity at age over 100 iterations from all 4 EMs was close to the true selectivity at age. The estimates of key parameters from all EMs were accurate compared with the true values from the OM (Table 5). The range of RE in key estimates became wider when the fishery and survey had double-logistic selectivity (case 8 versus case 0 in Table 6, Figure 6, and Supplementary Figures 3–8), especially in the early years. For years when fishing was not near F_{MSY}, the overfishing status determination was 100% accurate, and the overfished status determination was 100% accurate over all years.

Multiple surveys The differences in key parameter estimates between case 9 and case 0 were not considerable (Fig. 5). However, the range of RE in SSB, recruitment, F, relative SSB, and relative F became narrower after an additional survey index with the same level of observation error

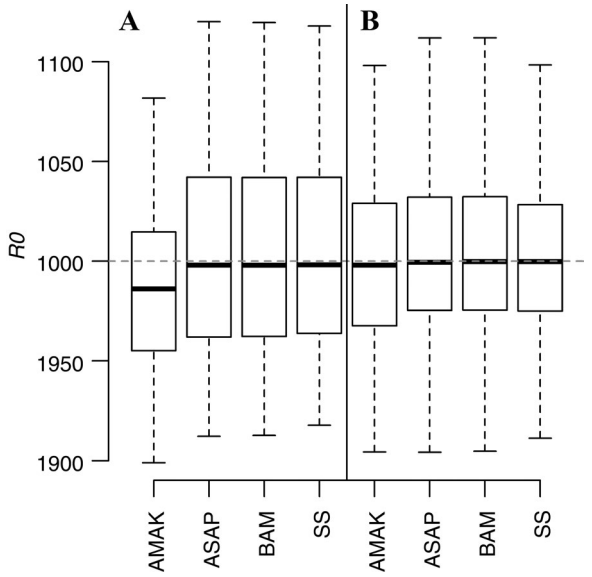

Figure 4

Box plots of true and estimated unfished recruitment (*R0*) from 4 age-structured estimation models under (**A**) case 0 and (**B**) case 12. The models, evaluated in this study for use in stock assessments, include the Assessment Model for Alaska (AMAK), the Age Structured Assessment Program (ASAP), the Beaufort Assessment Model (BAM), and Stock Synthesis (SS). The horizontal gray dashed line represents the true *R0* under the 2 cases. For case 0, which is the null case, initial equilibrium recruitment was lowered from the unfished recruitment level as determined by an initial equilibrium fishing mortality rate. For case 12, the initial condition was equal to the unfished equilibrium population. The upper and lower parts of each box represent the first and third quartiles (the 25th and 75th percentiles), and the thick horizontal line is the median. The whiskers extending above and below the box correspond to 1.5 times the interquartile range.

was included in case 9 (Table 6, Fig. 6, Suppl. Figs. 4–8). Results were as expected, given that there was no reduction in bias but an increase in precision. The accuracy of determining overfishing and overfished status shared similar trends compared with case 0 (Fig. 7).

Bias adjustment of recruitment The accuracy of parameter estimates was high if, in the EMs, the conversion function was used before estimation when a bias adjustment of recruitment was incorporated in the OM. Median relative errors were close to zero for MSY-based reference points when the conversion function was used in the BAM and SS (Fig. 5). With ad hoc adjustment (i.e., after estimation) in the AMAK and ASAP, the median relative errors in MSY-based reference points were reduced (Fig. 5, Suppl. Figs. 1 and 2). Estimated SSB, recruitment, and *F* remained highly accurate over time (Fig. 6). The trends in accuracy of stock status determination were similar to the trends from case 0 (Fig. 7).

Discussion

Similarity of estimates from the 4 models

In this study, the 4 stock assessment models, or EMs, produced similar estimates, an outcome that can be attributed to the fact that the models share similar mathematical and statistical attributes. Prior to our study, this supposition was expected to be true but was unverified. Under cases that are associated with different recruitment variability levels, process error in *F*, diverse patterns of *F*, various selectivity shapes, and multiple surveys, the median relative errors in key parameters remained low, and the variability of the REs were similar among the EMs. Of the 5 cases described here, the level of recruitment variability caused the most change in RE patterns. The range of REs in all 4 EMs became wider when recruitment variability increased and remained stable over other cases. Furthermore, the temporal trend in the accuracy of overfishing and overfished status determination was the same among the 4 EMs.

These findings indicate that the 4 EMs produce similar estimates when the same data are analyzed and the EMs are configured similarly. Nevertheless, the results would differ if different options of features are used for different EMs, depending on the stock-specific data and issues that assessment analysts must face. In practice, stock assessment analysts may make different configuration choices given the same data and the same model. We encourage analysts to clearly document the assumptions made in an assessment, especially when the analysis involves comparisons among multiple models. In addition, model misspecification may result from different assumptions about parameters, governing processes, and statistical properties that can have a substantial effect on stock assessment results and subsequent management advice (Piner et al., 2011; Maunder and Piner, 2015). More simulation-estimation studies could be done to quantify the effect of model misspecification on model estimates.

The fundamental differences in mathematical and statistical attributes found in this study could also serve as a starting point for diagnostics that can be used to identify the source of variation. In addition to confirming similar estimates, we identified that different approaches of computing initial numbers at age induced differences in estimates, especially those associated with *R0* and MSY-based reference points. Estimates among the 4 EMs also differed if bias adjustment of recruitment was not addressed carefully. The effects of the initial numbers at age setup and recruitment bias adjustment on EM performance are discussed in detail in subsequent sections. We also noticed that determining the overfishing status was not 100% accurate across time because, in the binary classification applied in our study, the overfishing determination was based on the maximum likelihood estimate. Use of the estimated model uncertainty interval may better capture the true overfishing determination. Aggregating this binary determination over years from one EM and using

Table 6

Median absolute relative error (%) in maximum sustainable yield (MSY), fishing mortality rate that corresponds to MSY (F_{MSY}), spawning stock biomass at MSY (SSB_{MSY}), SSB, recruitment (R), fishing mortality rate (F), relative SSB, and relative F for each case used to compare the 4 age-structured stock assessment models evaluated in this study: the Assessment Model for Alaska (AMAK), the Age Structured Assessment Program (ASAP), the Beaufort Assessment Model (BAM), and Stock Synthesis (SS).

	MSY				F_{MSY}				SSB_{MSY}			
Case	AMAK	ASAP	BAM	SS	AMAK	ASAP	BAM	SS	AMAK	ASAP	BAM	SS
Case 0	3.25	4.10	4.07	3.84	1.04	1.04	1.04	1.04	3.13	3.91	3.95	3.92
Case 1	5.47	5.83	5.80	5.77	1.04	1.04	1.04	1.04	5.69	5.68	5.62	5.67
Case 2	7.90	9.02	8.97	9.08	1.04	1.04	1.04	1.04	8.17	8.48	8.42	8.78
Case 3	3.06	2.76	2.74	2.81	1.04	1.04	1.04	1.04	3.03	2.67	2.76	2.65
Case 4	3.41	3.17	3.12	3.20	1.04	1.04	1.04	1.04	3.33	3.28	3.16	3.30
Case 5	4.57	4.43	4.42	4.09	1.04	1.04	1.04	1.04	4.63	4.28	4.37	4.08
Case 6	3.13	3.39	3.36	3.44	1.04	1.04	1.04	1.04	3.32	3.66	3.63	3.51
Case 7	4.06	3.55	3.40	3.28	1.04	1.04	1.04	1.04	4.08	3.32	3.26	3.10
Case 8	4.07	4.02	3.98	3.62	1.22	1.22	1.22	1.22	3.96	3.78	3.79	3.75
Case 9	3.22	3.24	3.23	3.24	1.04	1.04	1.04	1.04	3.14	3.00	2.88	2.92
Case 10	7.68	9.70	9.75	9.50	0.98	0.98	0.98	0.98	7.99	9.52	9.50	9.27
Case 11	7.19	9.10	9.12	9.17	0.98	0.98	0.98	0.98	6.83	8.55	8.56	8.77
Case 12	3.09	3.01	2.97	2.57	1.04	1.04	1.04	1.04	3.13	2.92	3.05	2.59

	SSB				R				F			
Case	AMAK	ASAP	BAM	SS	AMAK	ASAP	BAM	SS	AMAK	ASAP	BAM	SS
Case 0	2.11	2.11	2.12	2.07	3.82	3.80	3.80	3.81	2.74	2.67	2.67	2.63
Case 1	1.93	2.03	2.03	2.00	3.87	3.94	3.93	3.92	2.46	2.50	2.49	2.49
Case 2	2.08	2.07	2.07	2.06	4.17	4.17	4.17	4.17	2.58	2.63	2.63	2.58
Case 3	2.01	2.09	2.09	2.07	3.72	3.75	3.74	3.75	2.42	2.44	2.45	2.40
Case 4	1.75	1.77	1.76	1.76	3.83	3.82	3.82	3.80	2.34	2.35	2.35	2.27
Case 5	4.00	3.92	3.94	3.95	4.78	4.79	4.77	4.71	4.21	4.23	4.25	4.22
Case 6	2.07	2.13	2.13	2.13	3.87	3.90	3.92	3.91	2.49	2.61	2.62	2.60
Case 7	1.73	1.98	1.79	1.77	3.72	3.85	3.64	3.62	2.37	2.72	2.50	2.48
Case 8	3.62	4.93	5.95	4.18	4.15	4.20	4.25	4.15	3.45	4.66	4.92	4.04
Case 9	1.68	1.53	1.54	1.51	3.19	3.20	3.21	3.19	2.01	1.87	1.87	1.84
Case 10	2.01	2.04	2.04	2.03	4.25	4.26	4.23	4.27	2.65	2.68	2.64	2.64
Case 11	2.02	2.02	2.02	2.00	4.17	4.13	4.13	4.15	2.59	2.58	2.58	2.55
Case 12	2.01	2.05	2.05	1.95	3.81	3.79	3.79	3.86	2.61	2.61	2.63	2.47

	Relative SSB				Relative F			
Case	AMAK	ASAP	BAM	SS	AMAK	ASAP	BAM	SS
Case 0	3.73	4.29	4.31	4.30	2.06	2.10	2.03	2.01
Case 1	5.99	6.39	6.39	6.51	2.06	2.04	2.01	2.02
Case 2	8.19	9.83	9.75	9.95	2.16	2.22	2.12	2.17
Case 3	3.67	3.28	3.25	3.19	2.08	2.17	2.14	2.13
Case 4	3.80	3.80	3.80	3.69	1.85	1.90	1.86	1.87
Case 5	3.40	3.24	3.20	3.18	3.98	4.00	3.92	3.96
Case 6	3.68	4.44	4.45	4.31	2.21	2.38	2.26	2.27
Case 7	4.44	4.50	4.14	4.13	1.96	2.26	2.00	1.98
Case 8	3.98	5.50	5.97	4.99	3.29	4.40	4.96	3.81
Case 9	3.43	3.46	3.45	3.46	1.76	1.74	1.68	1.69
Case 10	8.25	9.73	9.71	10.14	2.06	2.06	2.06	2.09
Case 11	7.83	9.43	9.41	9.77	2.08	2.06	2.07	2.09
Case 12	3.49	3.51	3.52	3.04	2.13	2.08	2.08	2.01

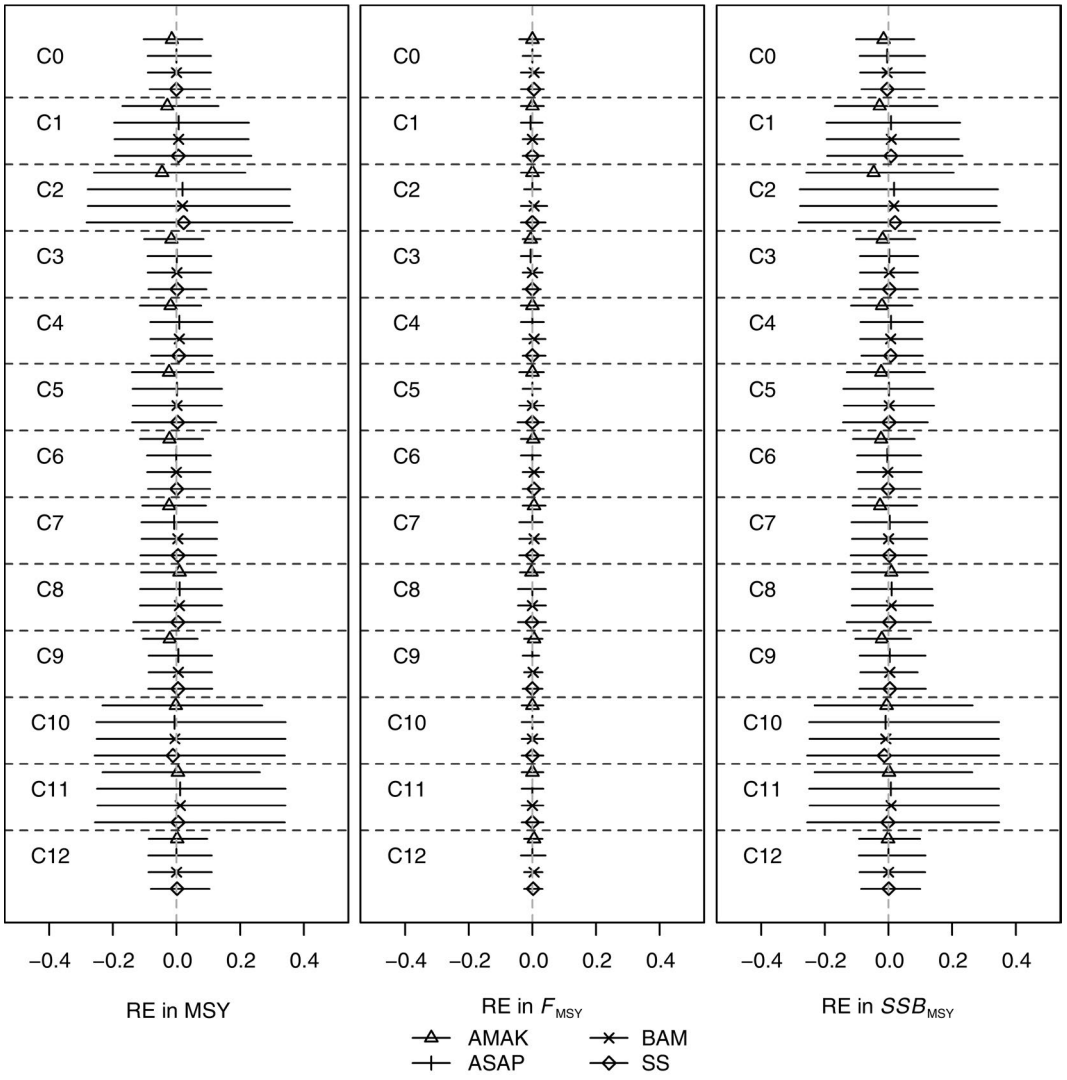

Figure 5

Relative error (RE) in maximum sustainable yield (MSY), fishing mortality rate that corresponds to MSY (F_{MSY}), and spawning stock biomass at MSY (SSB_{MSY}) for 4 age-structured estimation models under cases 0–12 (C0–C12). The models, evaluated in this study for use in stock assessments, include the Assessment Model for Alaska (AMAK), the Age Structured Assessment Program (ASAP), the Beaufort Assessment Model (BAM), and Stock Synthesis (SS). The bar is equivalent to the whisker of a box plot. The left bar represents the smallest value larger than 1.5 times the interquartile range below the first quartile. The right bar represents the largest value smaller than 1.5 times the interquartile range above the third quartile. The symbols in the center represent median RE.

receiver operating characteristic curves may also help summarize the overall degree of agreement (e.g., true positive and false negative classifications, accuracy rate, error rate, and sensitivity) between estimates and true status (Connors and Cooper, 2014; Cortés and Brooks, 2018).

Examination of the value of the code comparison process

This study is unique among other model comparison studies because a code comparison process was included before the simulation-estimation process. Identifying common features among the EMs and comparing source code can reduce contamination by parameter misspecification or analyst effect if the parameters are fixed in a comparison study. Although we estimated selectivity in this comparison study, during the early stage of the study, we fixed the selectivity at the true values. It is important to carefully apply fixed values because, in the 4 models, simple logistic selectivity was defined with parameters that had the same name (e.g., slope) but different interpretations. Otherwise,

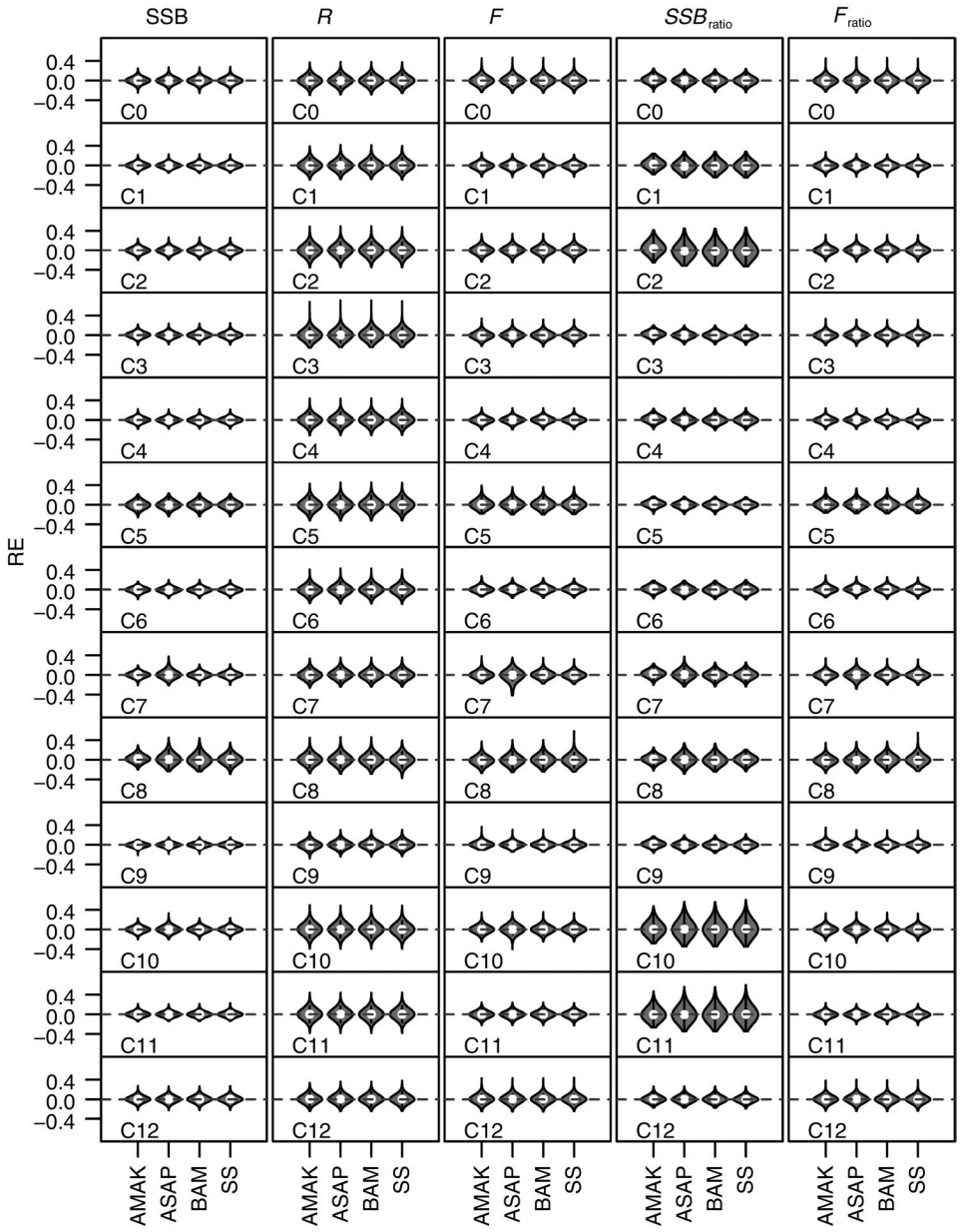

Figure 6

Violin plot of relative error (RE) across years and iterations for spawning stock biomass (SSB), recruitment (R), fishing mortality rate (F), SSB_{ratio} (ratio of SSB to SSB at maximum sustainable yield [MSY]), and F_{ratio} (ratio of F to F at MSY) for each of 4 estimation models under cases 0–12 (C0–12). The models, evaluated in this study for use in stock assessments, include the Assessment Model for Alaska (AMAK), the Age Structured Assessment Program (ASAP), the Beaufort Assessment Model (BAM), and Stock Synthesis (SS).

differences in estimates might have been forced unwittingly, because of parameter misspecification.

In addition to differences in selectivity parameterization, we also identified differences in how age is modeled in the EMs. If ages modeled in SS start with age 0 (the default) and not age 1, a mismatch between SS and the other EMs would have been induced. In this case, the mismatch would manifest only in scaling of recruitment to account for natural mortality at age 0. Therefore, identification of common features and comparison of source codes are particularly important in cross-testing a set of assessment models. The comparison framework developed (Fig. 1) and the approach

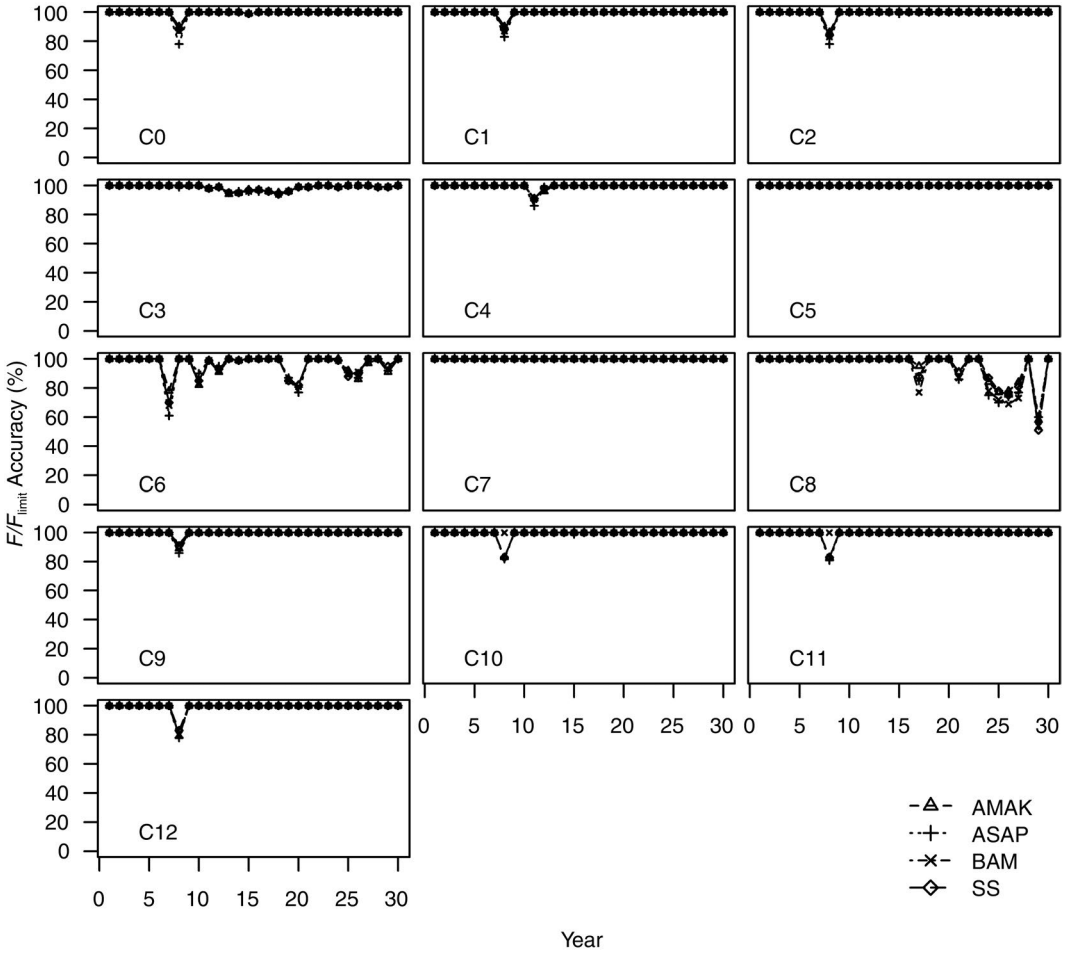

Figure 7

Accuracy (%) of determining overfishing status over time for each of 4 estimation models under cases 0–12 (C0–C12). Overfishing status is determined by dividing fishing mortality rate (F) by F_{limit}, which was set to the F that corresponds to maximum sustainable yield. The models, evaluated in this study for use in stock assessments, include the Assessment Model for Alaska (AMAK), the Age Structured Assessment Program (ASAP), the Beaufort Assessment Model (BAM), and Stock Synthesis (SS).

to identification of common features used (Table 2) in this study could be applied to multi-model comparisons in other studies. They might also prove useful in ensemble modeling, which is now gaining traction in fisheries science as a means to combine the estimates from multiple stock assessment methods (Brodziak and Legault, 2005; Brodziak and Piner, 2010). Ensemble modeling loosens the assumptions associated with selecting a single "best" assessment model (Rosenberg et al., 2018). Stewart and Martell (2015) proved that ensemble modeling benefits from guidelines for developing sets of candidate models. The steps developed in this study to identify common features across alternative models may facilitate selection of plausible models and identification of the sources of differences among estimates before constructing an ensemble.

Examination of similarity in initial numbers at age

The approach used in each EM to estimate the initial numbers at age depends on 3 factors: the selected initial year; the level of fishing, if any, that typically occurred prior to that initial year; and the availability of age data beginning with the initial year, such that the initial non-equilibrium age composition can be estimated. In principle, if informative age data are available across years, the EM results should not be dependent on the choice of the initial year because the estimated age composition in the initial year could have been alternatively estimated by starting the model at an earlier year and estimating the age composition for that year as projected from earlier recruitments. Also, when age composition data from the fleet and survey are available in the first year, the numbers at age for the

population can be estimated directly as parameters. In all 4 EMs, this feature of estimating the initial numbers at age is implemented by estimating the initial age composition as deviations from an equilibrium age composition. The level of recruitment that anchors the equilibrium age composition is calculated by using the spawner–recruit relationship. In other words, the initial equilibrium recruitment is lowered from the unfished recruitment level by an initial equilibrium F. However, in the AMAK, the initial equilibrium stock size is treated independently of historical fishing F because it typically is used in situations where this value would be negligible.

In this study, results from the comparison between case 0 and case 12 indicate that the AMAK, which does not specify the initial F, will scale the $R0$ downwards to match the numbers at age 1 in the initial year and, consequently, the EM will produce lower MSY and SSB_{MSY} compared with the true values. The magnitude of RE will increase when initial F becomes higher.

This study was not designed to compare the performance of EMs under different initial population conditions, and the initial F and recruitment variability were fixed at low levels in cases 0 and 12. The initial numbers at age in year 1 may be quite different from those for the unfished equilibrium populations, especially when fishing occurred many years prior to the first year of data or with large variability in recruitment. When fishery or survey composition data are available near the start of commercial fishing, the estimates of unfished condition or stock status may accurately reflect the true conditions when initial equilibrium stock size is treated independently of historical fishing F in an EM. In addition to having this configuration option, the ASAP, BAM, and SS allow the initial numbers at age to be controlled through different processes (Legault and Restrepo, 1999; Methot and Wetzel, 2013; Williams and Shertzer, 2015). Future work on assessment model development should consider which options are most accurate and efficient for computing initial numbers at age.

Spawner-recruit parameters: median-unbiased or mean-unbiased

The results from cases 10 and 11 have clear implications for bias adjustment of recruitment. First, we identified the fundamental differences between 2 bias adjustment methods (Table 4, Suppl. Table 3). In the BAM, median-unbiased spawner-recruit parameters are used, and in SS mean-unbiased parameters are used. Therefore, the inputs and outputs of $R0$ and h from the 2 models are comparable only after conversion, for example, by using the function introduced here (Equations 11–14). These findings highlight the importance of clarifying in assessment reports and meta-analyses whether estimates of spawner-recruit parameters correspond to geometric mean or arithmetic mean curves of recruitment and the significance of the need for developing functions for conversion of mean-unbiased parameters to median-unbiased parameters (and vice versa) for other spawner-recruit models (e.g., Ricker spawner-recruit model; Hilborn, 1985). In various studies, bias adjustment of recruitment has been implemented differently, but no study has clearly demonstrated the strengths and weaknesses of different bias adjustment methods (Walters, 1990; Chen, 2004; Yin and Sampson, 2004; Methot and Taylor, 2011; Subbey et al., 2014). We recommend further work on bias adjustment to derive conversion functions for other spawner-recruit models and to provide clear guidance on which estimation process (mean-unbiased or median-unbiased) might be preferred under different situations.

Second, we established that ad hoc bias adjustment of recruitment can be implemented in EMs that do not have the bias adjustment feature (Suppl. Fig. 2). The ad hoc adjustment affects recruitment and fishery management parameters, such as MSY-based reference points. In this study, we found that the AMAK and ASAP produced estimates of $R0$ and MSY-based reference points that are similar to the true values, if the estimates from those models were adjusted from a median-unbiased relationship to a mean-unbiased relationship.

Limits and future research

In addition to the specific recommendations coming from the issues found in this study, we think the comparison design could be extended to address other specific needs, such as quantifying the value of estimation of time-varying parameters as random effects (e.g., numbers at age, selectivity, and F), estimation of spawner-recruit parameters, data weighting, spatial structure, and other attributes to the performance of EMs. Growth was assumed to be known in this study because not all EMs have the capability to estimate growth. We can further compare EMs with more complicated cases, such as those that involve estimating growth within the assessments or using lower quality weight-at-age data. It would also be useful to conduct comparisons across life history patterns (e.g., patterns of long-lived versus short-lived species or patterns of demersal versus pelagic species), but further work on development of more complex OMs for simulation testing would be required. Punt et al. (2020) outlined essential features that should be considered for the next-generation stock assessment model, and they highlighted the importance of simulation testing in evaluation of estimation performance. Continued development of the OM used in this study through addition of essential features would result in an OM that can serve as an independent test bed to validate existing models as well as next-generation stock assessment models.

The comparison framework used in our study focused on age-structured models. Other age-structured stock assessment models that were not included in this study can be evaluated by using the comparison framework and creating connection files that automatically write input files, run the model, and save standard outputs. In addition, the comparison framework can be further applied to include other types of stock assessment models (e.g., surplus production, length-based, and catch-only models).

Having more comparison practices that involve both next-generation stock assessment models and existing models (inside and outside the United States) will enhance the communication among model developers and users, facilitate the interpretation of comparison results among models, and improve future assessments.

Conclusions

This study was designed to verify if the assessment models developed in different regions of the United States can produce similar estimates when given the same input data and configured similarly. However, it had a secondary objective of informing development of next-generation models (Punt et al., 2020). It is clear that all 4 models tested in this study provide similar and accurate estimates of quantities of interest under the tested cases. This outcome was expected given that the 4 EMs share similar mathematical and statistical attributes and that the simulated data were very informative. Nevertheless, it was expected also because we carefully evaluated the conversions among models to ensure that model configurations were similar to each other and model outputs were comparable. For future model comparison work or ensemble work, we recommend comparison of key features in source code before any multi-model analysis is done in order to identify differences in parameterization that could be misleading when results are compared (e.g., selectivity function parameters). We also recommend minimizing the variations of parameterizations for the same feature during development of next-generation stock assessment models. Standardized inputs and outputs for common parameters would allow easy comparisons of results from different models.

In this study, we have identified the sources of slight differences among model estimates under different cases. The differences are associated with computation of initial numbers at age and bias adjustment of recruitment. Improved insights on these key differences should help the development of next-generation stock assessment models. Key potential areas for future improvements include better clarification of terminology used in assessment reports, use of the conversion function developed in this study to convert between median-unbiased and mean-unbiased spawner-recruit parameters in stock assessment, use of the conversion function in other meta-analyses to ensure the inputs of meta-analysis are comparable, and development of guidance on which bias adjustment method is preferable under which situations.

Acknowledgments

This research was performed while the senior author held a National Research Council Research Associateship award at the Office of Science and Technology, National Marine Fisheries Service. We thank K. Doering, C. Stawitz, E. Dick, M. Masi, K. Johnson, and C. Bassin for their comments and suggestions for improving the development of the operating model in this study and this paper.

Literature cited

AFSC (Alaska Fisheries Science Center).
2015. Assessment Model for Alaska description of GUI and instructions, 50 p. [Available from website.]

Barbeaux, S. J., J. Ianelli, and W. Palsson.
2019. Assessment of the pollock stock in the Aleutian Islands, 3 p. *In* Stock assessment and fishery evaluation report for the groundfish resources of the Bering Sea/Aleutian Islands regions. North Pac. Fish. Manage. Counc., Anchorage, AK. [Available from website.]

Beverton, R. J. H., and S. J. Holt.
1957. On the dynamics of exploited fish populations. Fish. Investig. (G.B. Minist. Agric. Fish. Food), Ser. 2, vol. 19, 533 p. HMSO, London.

Brodziak, J., and C. M. Legault.
2005. Model averaging to estimate rebuilding targets for overfished stocks. Can. J. Fish. Aquat. Sci. 62:544–562. Crossref

Brodziak, J., and K. Piner.
2010. Model averaging and probable status of North Pacific striped marlin, *Tetrapturus audax*. Can. J. Fish. Aquat. Sci. 67:793–805. Crossref

Cadrin, S. X., and M. Dickey-Collas.
2015. Stock assessment methods for sustainable fisheries. ICES J. Mar. Sci. 72:1–6. Crossref

Chang, Y.-J., J. Brodziak, J. O'Malley, H.-H. Lee, G. DiNardo, and C.-L. Sun.
2015. Model selection and multi-model inference for Bayesian surplus production models: a case study for Pacific blue and striped marlin. Fish. Res. 166:129–139. Crossref

Chen, D. G.
2004. Bias and bias correction in fish recruitment prediction. North Am. J. Fish. Manage. 24:724–730. Crossref

Connors, B. M., and A. B. Cooper.
2014. Determining decision thresholds and evaluating indicators when conservation status is measured as a continuum. Conserv. Biol. 28:1626–1635. Crossref

Cortés, E., and E. N. Brooks.
2018. Stock status and reference points for sharks using data-limited methods and life history. Fish Fish. 19:1110–1129. Crossref

Deroba, J. J., D. S. Butterworth, R. D. Methot Jr., J. A. A. De Oliveira, C. Fernandez, A. Nielsen, S. X. Cadrin, M. Dickey-Collas, C. M. Legault, J. Ianelli, et al.
2015. Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods. ICES J. Mar. Sci. 72:19–30. Crossref

Dichmont, C. M., R. A. Deng, A. E. Punt, J. Brodziak, Y.-J. Chang, J. M. Cope, J. N. Ianelli, C. M. Legault, R. D. Methot Jr., C. E. Porch, et al.
2016. A review of stock assessment packages in the United States. Fish. Res. 183:447–460. Crossref

Federal Register.
1998. Magnuson-Stevens Act provisions; national standard guidelines. Fed. Regist. 63:24212–24237.

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

Gabriel, W. L., and P. M. Mace.
1999. A review of biological reference points in the context of the precautionary approach. *In* Proceedings of the 5th annual NMFS National Stock Assessment Workshop, p. 34–45. NOAA Tech. Memo. NMFS-F/SPO-40.

Henríquez, V., R. Licandeo, L. A. Cubillos, and S. P. Cox.
2016. Interactions between ageing error and selectivity in statistical catch-at-age models: simulations and implications for assessment of the Chilean Patagonian toothfish fishery. ICES J. Mar. Sci. 73:1074–1090. Crossref

Hilborn, R.
1985. Simplified calculation of optimum spawning stock size from Ricker's stock recruitment curve. Can. J. Fish. Aquat. Sci. 42:1833–1834. Crossref

Hilborn, R., and C. J. Walters.
1992. Quantitative fisheries stock assessment, choice, dynamics and uncertainty, 570 p. Routledge, Chapman and Hall, London.

Ianelli, J. N., and D. A. Fournier.
1998. Alternative age-structured analyses of the NRC simulated stock assessment data. *In* Analyses of simulated data sets in support of the NRC study on stock assessment methods (V. R. Restrepo, ed.), p. 81–96. NOAA Tech. Memo. NMFS-F/SPO-30.

Johnson, K. F., C. C. Monnahan, C. R. McGilliard, K. A. Vert-pre, S. C. Anderson, C. J. Cunningham, F. Hurtado-Ferro, R. R. Licandeo, M. L. Muradian, K. Ono, et al.
2015. Time-varying natural mortality in fisheries stock assessment models: identifying a default approach. ICES J. Mar. Sci. 72:137–150. Crossref

Legault, C. M., and V. R. Restrepo.
1999. A flexible forward age-structured assessment program. ICCAT Collect. Vol. Sci. Pap. 49(2):246–253. SCRS/98/058. [Available from website.]

Lynch, P. D., R. D. Methot, and J. S. Link (eds.).
2018. Implementing a next generation stock assessment enterprise. An update to the NOAA Fisheries stock assessment improvement plan. NOAA Tech. Memo. NMFS-F/SPO-183, 127 p.

Maunder, M. N., and A. E. Punt.
2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142:61–74. Crossref

Maunder, M. N., and K. R. Piner.
2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. 72:7–18. Crossref

Methot, R. D., Jr., and I. G. Taylor.
2011. Adjusting for bias due to variability of estimated recruitments in fishery assessment models. Can. J. Fish. Aquat. Sci. 68:1744–1760. Crossref

Methot, R. D., Jr., and C. R. Wetzel.
2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142:86–99. Crossref

Methot, R. D., Jr., G. R. Tromble, D. M. Lambert, and K. E. Greene.
2014. Implementing a science-based system for preventing overfishing and guiding sustainable fisheries in the United States. ICES J. Mar. Sci. 71:183–194. Crossref

NRC (National Research Council).
1998. Improving fish stock assessments, 188 p. National Academies Press, Washington, D.C.

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

Piner, K. R., H.-H. Lee, M. N. Maunder, and R. D. Methot.
2011. A simulation-based method to determine model misspecification: examples using natural mortality and population dynamics models. Mar. Coast. Fish. 3:336–343. Crossref

Punt, A. E., A. Dunn, B. Þ. Elvarsson, J. Hampton, S. D. Hoyle, M. N. Maunder, R. D. Methot, and A. Nielsen.
2020. Essential features of the next-generation integrated fisheries stock assessment package: a perspective. Fish. Res. 229:105617. Crossref

Quinn, T. J., II, and R. B. Deriso.
1999. Quantitative fish dynamics, 540 p. Oxford Univ. Press, Oxford, UK.

Restrepo, V. R., G. G. Thompson, P. M. Mace, W. L. Gabriel, L. L. Low, A. D. MacCall, R. D. Methot, J. E. Powers, B. L. Taylor, P. R. Wade, et al.
1998. Technical guidance on the use of precautionary approaches to implementing National Standard 1 of the Magnuson-Stevens Fishery Conservation and Management Act. NOAA Tech Memo NMFS-F/SPO-31, 48 p.

Rosenberg, A. A., K. M. Kleisner, J. Afflerbach, S. C. Anderson, M. Dickey-Collas, A. B. Cooper, M. J. Fogarty, E. A. Fulton, N. L. Gutiérrez, K. J. W. Hyde, et al.
2018. Applying a new ensemble approach to estimating stock status of marine fisheries around the world. Conserv. Lett. 11:e12363. Crossref

Sampson, D. B., and Y. Yin.
1998. A Monte Carlo evaluation of the Stock Synthesis assessment program. *In* Fishery stock assessment models. Alaska Sea Grant Rep. 98-01 (D. Sampson and Y. Yin, eds.), p. 315–338. Alaska Sea Grant College Prog., Univ. Alaska Fairbanks, Fairbanks, AK.

Siegfried, K. I., E. H. Williams, K. W. Shertzer, and L. G. Coggins.
2016. Improving stock assessments through data prioritization. Can. J. Fish. Aquat. Sci. 73:1703–1711. Crossref

Smith, S. J., J. J. Hunt, and D. Rivard (eds.).
1993. Risk evaluation and biological reference points for fisheries management. Can. Spec. Publ. Fish. Aquat. Sci. 120, 442 p.

Stewart, I. J., and S. J. D. Martell.
2015. Reconciling stock assessment paradigms to better inform fisheries management. ICES J. Mar. Sci. 72:2187–2196. Crossref

Subbey, S., J. A. Devine, U. Schaarschmidt, and R. D. M. Nash.
2014. Modelling and forecasting stock–recruitment: current and future perspectives. ICES J. Mar. Sci. 71:2307–2322. Crossref

Walters, C. J.
1990. A partial bias correction factor for stock–recruitment parameter estimation in the presence of autocorrelated environmental effects. Can. J. Fish. Aquat. Sci. 47:516–519. Crossref

Wetzel, C. R., and A. E. Punt.
2011. Performance of a fisheries catch-at-age model (Stock Synthesis) in data-limited situations. Mar. Freshw. Res. 62:927–936. Crossref

Williams, E. H., and K. W. Shertzer.
2015. Technical documentation of the Beaufort Assessment Model (BAM). NOAA Tech. Memo. NMFS-SEFSC-671, 43 p.

Yin, Y., and D. B. Sampson.
2004. Bias and precision of estimates from an age-structured stock assessment program in relation to stock and data characteristics. North Am. J. Fish. Manage. 24:865–879. Crossref