ICES Journal of Marine Science, 2024, Vol. 81, Issue 6, 1053–1072
https://doi.org/10.1093/icesjms/fsae067
Received: 8 November 2023; **revised:** 16 April 2024; **accepted:** 17 April 2024
Advance access publication date: 28 May 2024
Original Article

Simulation testing performance of ensemble models when catch data are underreported

Elizabeth N. Brooks [1,*] **and Jon K.T. Brodziak** [2]

[1]Northeast Fisheries Science Center, 166 Water Street, Woods Hole, MA 02543, United States
[2]Pacific Islands Fisheries Science Center, 1845 Wasp Blvd., Honolulu, HI 96818, United States

*Corresponding author. Northeast Fisheries Science Center, 166 Water Street, Woods Hole, MA 02543, United States. E-mail: Liz.Brooks@noaa.gov

Abstract

Ensemble model use in stock assessment is increasing, yet guidance on construction and an evaluation of performance relative to single models is lacking. Ensemble models can characterize structural uncertainty and avoid the conundrum of selecting a "best" assessment model when alternative models explain observed data equally well. Through simulation, we explore the importance of identifying candidate models for both assessment and short-term forecasts and the consequences of different ensemble weighting methods on estimated quantities. Ensemble performance exceeded a single best model only when the set of candidate models spanned the true model configuration. Accuracy and precision depended on the model weighting scheme, and varied between two case studies investigating the impact of catch accuracy. Information theoretic weighting methods performed well in the case study with accurate catch, while equal weighting performed best when catch was underreported. In both cases, equal weighting produced multimodality. Ensuring that an ensemble spans the true state of nature will be challenging, but we observed that a change in sign of Mohn's rho across candidate models coincided with the true OM being bounded. Further development of protocols to select an objective and balanced set of candidate models, and diagnostics to assess adequacy of candidates are recommended.

Keywords: ensemble model; stock assessment; model weights; cross-validation

Introduction

Stock assessments are important for making informed management decisions to support the sustainable harvest of marine fishery resources. The quantitative methods for stock assessment have evolved from simple biomass dynamics and virtual population analyses to integrated likelihood-based estimation models (EMs) that include multiple hierarchies, such as population age, gender, size, or spatial structure. More recently, ensemble modeling approaches are being considered with the aim of better representing uncertainty about the fishing patterns and population dynamics of the fishery system. Regardless of the complexity, conducting a stock assessment involves several interrelated steps (Table 1). Whether pursuing a single model or an ensemble model to provide management advice, a crucial step in the assessment process is the identification of model structures and plausible model configurations. These are often hypothesis-driven or based on experience with similar fisheries or stocks, and will vary depending on the degree of aggregation, the length of time series, and the availability and quality of the catch, population life history, and abundance data.

Under the prevailing single-model paradigm for stock assessment, plausible models and sensitivity analyses are compared, and a "best model" is selected based on model diagnostics. The characterization of processes in the selected model is considered adequate for making accurate predictions for management advice. Analytical techniques for decision analysis and risk management are well established for single assessment models (e.g. Smith et al. 1993, Patterson et al. 2001, Privitera-Johnson and Punt 2020), and most of the scientific advice provided to fisheries managers has been framed under the expectation of having correctly identified one model to describe the system, or so-called target-directed modeling (Weisberg 2013). If system dynamics are well specified and approximately stationary, then this optimism may be well founded. In reality, there is ambiguity about which model is the best. Model diagnostics may not provide a clear basis for selecting one model over others, and reporting or emphasizing the single outcome will underestimate uncertainty and produce overconfidence in assessment results used for management advice. One aspect of underestimated uncertainty is due to the single-model assessment approach accounting only for parametric, not structural, uncertainty, which is often the key impediment to understanding and predicting the dynamics of complex fishery systems (e.g. Hilborn and Stearns 1982, Brodziak et al. 2008).

When there are distinct uncertainties about the dynamics of the fish stock or fishery system, or when a given stock assessment is contested with two or more competing assessment models (e.g. Starr et al. 1998), then a multi-model inference approach may be preferred (e.g. Burnham and Andersen 2002; Jardim et al. 2020). In the ensemble paradigm, the goal is to build a set of plausible assessment models that accurately characterize and bound the true but unknown dynamics and avoid redundancy and favoritism (Chamberlin 1965, Jardim et al. 2020). This can be approached by defining axes of uncertainty that correspond to alternative hypotheses about major system processes, and then constructing models with parameters and structures that span those uncertainties in model space. For example, one dimension of uncertainty could be the functional form of fishery selectivity (e.g. a parametric form such as a logistic function, or a non-parametric smoother), or it could be

Table 1. Steps and associated decisions for stock assessment along with some examples.

Ensemble process	Decisions to be made	Examples of decision points
Data preparation	• Stock definition • Spatiotemporal representation • Data aggregation • Alternative data sets*	• Which indices to include? • Use 5 or 7 regions for spatial model? • Use pooled-sex or two-sex model? • Use alternative natural mortality rates?
Model specification	• Matching data availability • Model structure* • *Ensemble members*	• Choose distribution for initial equilibrium catch? • Choose among submodel structures? • Choose plausible models or cut implausible ones?
Parameter estimation	• Error distributions* • Estimation framework* • *Ensemble weights*	• Use symmetric or skewed observed errors? • Use frequentist or random effects framework? • Use equal or tactical or Bayesian weights?
Model selection	• Model diagnostics • *Multimodel inference*	• Use diagnostics for indices or size compositions? • Use averaged results for mean and precision?
Management advice	• Probabilistic status determination • Characterization of uncertainty • Risk analysis	• Combine status estimates across models? • Combine quantities of interest (QOIs) • Combine to form one ensemble or use clustering?

*Indicates a decision that complicates some diagnostic metrics or ability to weight results in an ensemble modeling framework; italics emphasize decisions that are ensemble-specific and differ from a single model assessment approach.

the type of process error about the expected stock-recruitment relationship, modeled with or without autocorrelated errors.

Application of ensemble models for stock assessments has seen increased use in the last decade, especially for highly migratory species in international fisheries (Supplementary Table S1). Early examples include southwest Pacific swordfish (*Xiphias gladius,* WCPFC 2011), south Pacific albacore tuna (*Thunnus alalunga*, WCPFC 2012), and Pacific Halibut (*Hippoglossus stenolepis*, Stewart and Martell 2014), where management quantities of interest were reported as medians from the ensemble distributions that weighted individual assessment models equally. Common uncertainties considered in these ensemble examples include life history parameters (steepness, recruitment, natural mortality, and growth) and fishery processes (selectivity or size composition weighting) that reflect hypotheses about credible states of nature, similar to the approach recommended in Millar et al. (2015). These species-specific applications differed in whether each axis was explored factorially, parametrically with a prior, or combined with other axes to reduce the overall dimension (Supplementary Table S1).

Ensemble construction requires a weighting approach to combine the individual model results (Levins 1966, Burnham and Anderson 2002, Jardim et al. 2020). Equal weighting has been the *de facto* method, although unequal model weighting has been used in a few recent assessments (Supplementary Table S1). Dormann et al. (2018) discuss four general approaches to setting weights based on Bayesian modeling, information-theoretic criteria, tactical approaches as well as expert opinion. Bayesian model weighting uses the Bayes factor, or ratio of marginal likelihoods, to calculate the posterior model weights. Information-theoretic weights are based on some approximation of the expected distance between the real system and the approximating model under a fixed likelihood structure, e.g. Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) (Burnham and Anderson 2002). Tactical model weights are set based on the goal of maximizing the predictive accuracy of the ensemble. Expert judgment is a more subjective approach (e.g. Morgan and Henrion 1990) and may lead to equal model weightings when there is a perceived absence of relevant evidence or knowledge. While model averaging can improve predictive accuracy or precision and generally provides a better representation of uncertainty, performance depends on the relative biases, model weights, and associated covariances among model results (e.g. Dormann et al. 2018). Treating model weights as random variables and estimating the weights from data will affect the uncertainty of the resulting ensemble model averages of quantities of interest (QOI) and can lead to suboptimal weighting results (Nguefack-Tsague 2014, Claeskens et al. 2016, Dormann et al. 2018). Thus, there are many analytical aspects of setting model weights that warrant further investigation.

In this paper, we use simulated data to explore the performance of ensembles formed from different subsets of candidate models across different model weighting schemes for two contrasting situations: when catch data are accurate or when catch data are underreported. These two case studies allow us to evaluate consequences of steps in ensemble construction in a best case (accurate catch) while also demonstrating how a more realistic case (underreported catch) alters those results. Ensemble performance relative to the single best model paradigm is also characterized so that we can address under what conditions an ensemble is "better" than a single model. Our simulated assessment process has three steps: estimate current abundance, estimate reference points and stock status, and make projections for future catch and stock conditions. We demonstrate how ensemble modeling can be applied in each of these steps, including how model weights can be updated as a regular part of the assessment update process.

Methods

Simulation

A typical groundfish life history is simulated from an operating model (OM) that follows standard age-structured population dynamics (Table 2, Appendix A). The simulated stock is fished for 75 years from unexploited conditions, and the data generated for years 30–75 (i.e. the last 46 years) serve as input to a statistical catch at age model (ASAP, Legault and Re-

Table 2. Parameter values used in OM for generating simulated data.

Model parameters	OM value	EM value
Years of simulated data	1–75	46–75
Plus group	30	10
Steepness (*h*)	0.65	est
Virgin recruitment (*R0*)	1.0E + 07	est
σ_R	0.5	1
Spawning time	Jan 1	Jan 1
Fishery selectivity (a50, slope)	3.5, 0.75	est
Index-1 selectivity (a50, slope)	1,5, 0.7	est
Index-2 selectivity (a50, slope)	3.5, 0.5	est
Catchability (q1, q2)	1.0e-4, 1.5e-5	est
Natural mortality (*M*)	0.275	fixed values
Maturity ogive (a50, slope)	1.5, 1.0	1.5, 1.1
Weight at length (a, b)	6.63E-6, 3.1	6.63E-6, 3.2
t0	−0.05	NA
K	0.27	NA
Linfinity	106	NA
Aggregate catch CV	0.075	0.075
Aggregate index CV	0.2	0.2
Effective sample size (catch age comp)	115	100
Effective sample size (index age comp)	65	50

In the EM column, "est" is for parameters that were estimated in the assessment models, "fixed values" refers to natural mortality (M) being fixed at a range of different values in candidate EMs, while "NA" is a parameter that was only needed for the OM data generation. Observation error on aggregate data was assumed to be lognormal, while age composition was assumed to be multinomial (matches likelihood assumptions in ASAP).

strepo 1998, https://noaa-fisheries-integrated-toolbox.github.io/ASAP) for assessment and projection (AGEPRO, Brodziak et al. 1998, https://noaa-fisheries-integrated-toolbox.github.io/AGEPRO). From these known simulated stock dynamics, two case studies were generated: Case 1, where aggregate catch is reported accurately and Case 2, where aggregate catch for the last 23 years is rescaled by the random multiplier u_y ~Uniform [0.4, 0.7], which reduces the reported catch used in the assessment by 45% on average. All other aspects of the operating models for the two case studies were the same (Table 2, Supplementary Fig. S1).

Conducting an ensemble assessment requires the same three analytical steps for the assessment process as the individual model approach, which are vectorized and conducted for each model comprising the ensemble: (i) fit the assessment models; (ii) calculate reference points and stock status for each model; and (iii) make short-term forecasts from each model for catch advice. For the first step, two key decisions are required: what candidate assessment models should be included in the ensemble, and how should model weights be assigned in order to combine the model results? A related question is the process of how those model weights evolve over time as assessments are updated with additional years of data. For the second step, it is important that the ensemble stock status result is consistent with the set of individual model results. To achieve this, stock status is calculated for each candidate model with respect to its estimated reference points, and then the set of stock status estimates is combined with the same model-specific weights assigned in the first step (Brodziak and Piner 2010). Averaging the individual stock status estimates, rather than dividing the average of model estimates by the average of reference points, avoids the pitfall of averaging across reference points that were derived from different model structures and parameterizations. For the third step, we emphasize that uncertainty in the forecasts derives not just from the suite of candidate assessment models (step 1), but potentially includes a distinct set of forecast models nested within each candidate assessment model. To illustrate these three steps of the ensemble framework, including the topic of evolving model weights as assessments are updated in the future, we conduct three cycles of the assessment process. The first assessment cycle omits the last 9 years of data, calculates status, and then projects for 3 years. For the second and third assessment cycles, three additional years of simulated data are added, the models are fit, and projections are made (Supplementary Fig. S2). Detailed diagnostics were not examined, but all models met convergence criteria (maximum gradient smaller than 10E-3, the hessian was positive definite), correlations between estimated parameters were checked to rule out confounding, and CVs of estimated parameters were neither too large (indicating inestimability) nor too close to zero (indicating possible boundary solution) (Carvalho et al. 2021, Brooks and Legault 2022). Mohn's rho (Mohn 1999) was calculated for each individual model in each assessment cycle by removing 7 years of data.

Ensemble candidate model specification

We used a factorial approach to construct ensembles of assessment models, focusing on four primary axes of uncertainty that have been addressed most frequently in contemporary ensemble applications (Supplementary Table S1). These included: the assumed rate of natural mortality (*M*, a constant with six levels of age- and time-invariant *M*), the shape of selectivity for the fishery and two survey indices (all were flat-topped or all were domed, two levels), and the functional form of the stock-recruit relationship (SRR: Beverton–Holt, Ricker, or mean recruitment with deviations, three levels). Treating the selectivity shape for both indices and the fishery as a single factor was done to limit the number of models considered to 36 (Table 3). Initially, the *M* values were chosen to be symmetric about the true value (0.275), but models with the lowest value of *M* (0.05) did not converge, so we increased *M* until we got convergence in all models (0.15). None of the models in the ensemble exactly matched the true operating model (OM) structure, but the true OM specifications lie between models M_3 and M_4.

Table 3. Model structure for the 36 candidate models in the ensemble.

	Natural mortality						Fishery and index selectivity		SR relationship		
Model number	0.15	0.2	0.25	0.3	0.35	0.5	All logistic	All dome	Bev–Holt	Ricker	No SR
1	×						×		×		
2		×					×		×		
3			×				×		×		
4				×			×		×		
5					×		×		×		
6						×	×		×		
7	×							×	×		
8		×						×	×		
9			×					×	×		
10				×				×	×		
11					×			×	×		
12						×		×	×		
13	×						×			×	
14		×					×			×	
15			×				×			×	
16				×			×			×	
17					×		×			×	
18						×	×			×	
19	×							×		×	
20		×						×		×	
21			×					×		×	
22				×				×		×	
23					×			×		×	
24						×		×		×	
25	×						×				×
26		×					×				×
27			×				×				×
28				×			×				×
29					×		×				×
30						×	×				×
31	×							×			×
32		×						×			×
33			×					×			×
34				×				×			×
35					×			×			×
36						×		×			×

Natural mortality was constant across all ages and years, there was a single fishery selectivity block, two fishery-independent indices of abundance, and the "Mean" SR modeled mean recruitment with annual deviations. True OM specifications lie between models M_3 and M_4.

Simulated data from the OM were formatted as input to the assessment model ASAP (Legault and Restrepo 1998). For each of the 36 candidate models for both case studies, the data were fit with no post-hoc tuning or data re-weighting. Different subsets of these 36 models (within a case study) were then used to define a model ensemble. For example, let M_m be the mth of the 36 model structures, and E [i:j] denote the ensemble constructed from the subset of models E [i:j] = {M_i, M_{i+1}, …, M_j} indexed for a given case study. We summarize results below for four ensembles: E [1:6], E [1:12], E [1:24], and E [1:36]. Of these, ensemble E [1:6] includes uncertainty in natural mortality only, E [1:12] accounts for uncertainty in natural mortality as well as fishery and index selectivity, and ensembles E [1:24] and E [1:36] incorporate additional uncertainty in the stock-recruit relationship. All four of these ensembles span the true OM, which would be the goal in practice but is only possible to know with certainty in a simulation. To evaluate outcomes when the true model specifications are not contained within the model set, we also consider two small ensembles for each case study where all candidate models are either "below" the true value of natural mortality in the OM (E [1:3]) or "above" the true value (E [4:6]).

It has been suggested that ensemble modeling may perform better than the single best model approach for fishery applications (Karp et al. 2018), but we are unaware of this being specifically tested in an integrated stock assessment modeling context (e.g. Dormann et al. 2018). Therefore, for each of the six ensembles (E [1:6], E [1:12], E [1:24], E [1:36], E [1:3], and E [4:6]), we evaluate ensemble performance relative to the best individual model comprising the ensemble in each case study, where "best" is identified by having the lowest AIC among the individual models in the ensemble. Relative errors for SSB, the fully selected F ("Fmult" in tables and figures), SSB/SSB [F40%], and F/F [40%] are summarized for ensembles across all years, as well as the last 15 years to characterize recent performance. The reference point of SPR40% corresponded to the true OM F_{MSY}, and was used instead of MSY because 12 of the 36 models (M_{25}–M_{36}) did not have a

stock-recruit relationship. SSB [F40%] was calculated for all models as the product of spawning biomass per recruit when fishing at F40% scaled by the mean recruitment for the time series of estimated recruitment in a given assessment (dropping the last 2 years due to greater uncertainty in terminal year estimates).

For both case studies, we evaluated the forecast skill of two forecast models (F_1 and F_2) that project dynamics 3 years from the terminal year of each assessment model, M_m. Both forecast models address short-term uncertainty about future recruitment in different ways. The first forecast model (F_1) generates random samples from the empirical cumulative distribution function (ecdf) of assessment estimates of recruitment for all but the last 2 years of the assessment time period. This reflects the hypothesis that recruitment has fluctuated without trend about its mean during the assessment years and will continue to do so in the near-term future. The second forecast model (F_2) generates random samples from the ecdf of the most recent 10 years, which reflects the hypothesis that the recent empirical distribution of recruitment will continue in the near-term future. After each candidate assessment model was fitted to the case study data, Markov Chain Monte Carlo (MCMC) simulation was performed with a thinning rate of 50 and burn-in of 700, with 1000 draws saved. Forecasts are then made from the MCMC estimates of numbers at age in the final assessment year using AGEPRO (Brodziak et al. 1998).

Weighting factors for assessment models

We evaluated a dozen different weighting methods for constructing an ensemble mode, which can be grouped into four categories: information theoretic, L2 risk (i.e. Euclidean distance), predictive accuracy, and uninformative. For the information theoretic weights, we include six different metrics: Akaike's Information Criteria (AIC) is calculated from the full model likelihood; the AIC contribution is associated with either aggregate catch (AIC.Catch), aggregate indices (AIC.Index), age composition of catch (AIC.C.comp), or age composition of indices (AIC.I.comp), and Bayesian Information Criteria (BIC; Schwarz 1978). Weighting methods based on the fit to a single data component are motivated by the Focus Information Criterion (FIC; Claeskens and Hjort 2003, 2008, Claeskens 2016). L2 risk is calculated as the mean squared error (MSE) of the predicted values for the same four data components as FIC. Predictive accuracy was calculated via cross-validation by removing the last 2 years of data (aggregate and composition) from both surveys and calculating the MSE between the model predicted index for the missing years and the "observed" index values (i.e. the simulated indices with observation error) in those same years. Because a lower MSE score is better, in order to use this as a weight, we take the inverse of each model's MSE_m score, i.e. $MSE'_m = {}^{1}\!/\!_{MSE_m}$. Finally, the uninformative weighting scheme assigns equal weights of $1/n$, to each of n models in the ensemble. Given an ensemble E [i:j], model-specific weights for the ensemble members are denoted as $w_{m,k}$ [i:j] for model M_m, where $k = 1, 2, \ldots, 12$ refers to the different methods of calculating weights. For a given weighting method k and ensemble E [i:j], we normalized the model weights to sum to unity, $\sum_{m=i}^{j} w_{m,k}[i:j] = 1$, in each case study.

Ensemble construction

Ensemble point estimates ($\Psi_k[i:j]$) and ensemble distributions ($\Delta_k[i:j]$) based on weighting method k are summarized for the following QOIs: spawning stock biomass SSB, fishing mortality F, and stock status (SSB/SSB [F40%], F/F [40%]). For each case study, ensemble E [i:j] with model weights $w_{m,k}$ [i:j], the point estimates of QOI are calculated as

$$\Psi_{\mathrm{k}}\,[\mathrm{i}:\mathrm{j}] \;= \sum_{m=i}^{j} w_{m,k}\,[i:j]\,QOI_m, \tag{1}$$

Calculating distributions for ensemble QOIs ($\Delta_k[i:j]$) is accomplished by sampling from each model's posterior distribution of time series estimates. Given weighting method k, a sample size $n_{m,k}$ [i:j] $= w_{m,k}$ [i:j] $*$D is drawn with replacement from the 1000 saved MCMC draws for model M_m, where D is the total number of draws over the whole ensemble. We set D = 10 000 in order to achieve smooth estimates of ensemble distributions, meaning that a model with $w_{m,k}$ [i:j] = 0.1 would have 1000 random samples with replacements drawn from its saved MCMC simulations. The ensemble distributions are then constructed by combining year-specific samples across all models in the ensemble to achieve D total estimated values in each year.

Weighting factors for forecast models

Weighting factors for forecast models differ from those considered for assessment models because there are no observations being fit. We explored a process of weighting projections as follows, similar in approach to Brooks and Legault's (2016) retrospective forecasting. First, an assessment model M_m is fit to data through year y_t and 3-year projections are made for years y_{t+1}, y_{t+2}, and y_{t+3}, producing estimates of QOI (catch, SSB, and recruitment). Then, when M_m is updated with data through year y_{t+3}, the projected QOI from the previous assessment cycle ($proj_{m,Q,y}$) is compared with estimates from the updated assessment ($est_{m,Q,y}$). We can calculate the performance skill (Ω) of each forecast model F_f from M_m, for a given QOI Q, in a given projection year y, over all 1000 MCMC iterations, i, as

$$\Omega_{m,f,Q,y} = \frac{1000}{\sum_{i=1}^{1000}\left(proj_{m,Q,y,i} - est_{m,Q,y}\right)^2}, \tag{2}$$

which is the inverse of the MSE of the projected value for year y (higher skill implies lower MSE). If all years have equal importance in management advice, then the skill across all years is just the sum of $\Omega_{m,f,Q,y}$. An alternative would be to apply an economic discounting factor, d_y, where skill in later years is less valuable than skill in earlier years of a projection. Generically, then, the model-specific forecast skill across all years is

$$\Omega_{m,f,Q} = \sum_{y} d_y \Omega_{m,f,Q,y}. \tag{3}$$

We set the discounting factor to unity in the simulation tests presented here. In a similar vein, the overall skill of a forecast model depends on the relative weight assigned to quantity-specific skill, $\tilde{w}_Q$, across all quantities being evaluated (catch, SSB, and recruitment in our case)

$$\Omega_{m,f} = \sum_{Q} \tilde{w}_Q \Omega_{m,f,Q}. \tag{4}$$

We use a scale of 0–1 for the quantity-specific relative weights ($\tilde{w}_Q$), but the scale is irrelevant since the model skill is scaled to sum to 1 when calculating forecast model weights $w_{m,f}$

$$w_{m,f} = \frac{\Omega_{m,f}}{\sum_f \Omega_{m,f}}. \quad (5)$$

Applying the forecast model weights to the projected QOI_m within a given assessment model M_m produces an ensemble projection P_m [i:j]. In simple terms, this means that the performance of each forecast model in the previous projection cycle is used to weight the forecast models in the current projection cycle. In the case of the first assessment cycle and the first set of projections, equal weights were applied. We can then append these P_m [i:j] to the end of the time series of each model-specific QOI_m, and then use the weights calculated for the assessment models ($w_{m,k}$ [i:j]) to get a seamless ensemble time series reflecting the ensemble of assessment estimates as well as the ensemble forecasts.

Results

Performance of individual models

The process by which a single best model is identified can vary widely by management organization, and may depend on a variety of factors, including assessment team and review panel composition (Ralston et al. 2011). For our case studies, the best model was identified as the model having the lowest AIC, comprised of the sum of negative log-likelihoods from fitting to observed data components plus a parameter penalty (Helu et al. 2000, Maunder and Punt 2013). The best-performing model may change through time, and across assessment updates, as the specified model structure becomes more or less able to match patterns in the data. Initial performance for M_1–M_6 for Cases 1 and 2 is quite close in the early model years, but after ~10 years, the poor performing models are easily identifiable while the better models show similar fits to the data (Fig. 1). With no catch misreporting (Case 1), the single best model across the three assessment updates was consistently M_4 but model 3 was very close in fit. In Case 2, where catch was misspecified, M_5 was the best in the first assessment but M_6 had a better fit in the second and third assessments. In both cases, multiple models fit the data similarly but the uncertainty in abundance, stock status, and catch advice across the multiple models is not conveyed to managers if results are only reported from the best model.

Inspecting the fits to the individual data components, the fit to the aggregate indices had the largest scale for difference in negative log-likelihood (NLL) among models (Fig. 2). Case 2 had underreported catch for the final 23 model years, and models with higher rates of natural mortality (compared to the true M) performed better. As additional years of data (with misreporting) are added for the second and third assessments, the improved fit of M_6 ($M = 0.5$) compared to M_5 ($M = 0.35$) is not surprising because a larger M compensates for the missing catch. While these models are compared with respect to NLL, they all have the same number of parameters, so conclusions hold for model-specific AIC scores as well. Furthermore, the model providing the best fit to individual data components (NLL) differs by data component, highlighting the fact that tuning data weights can influence which model has the lowest total NLL and AIC.

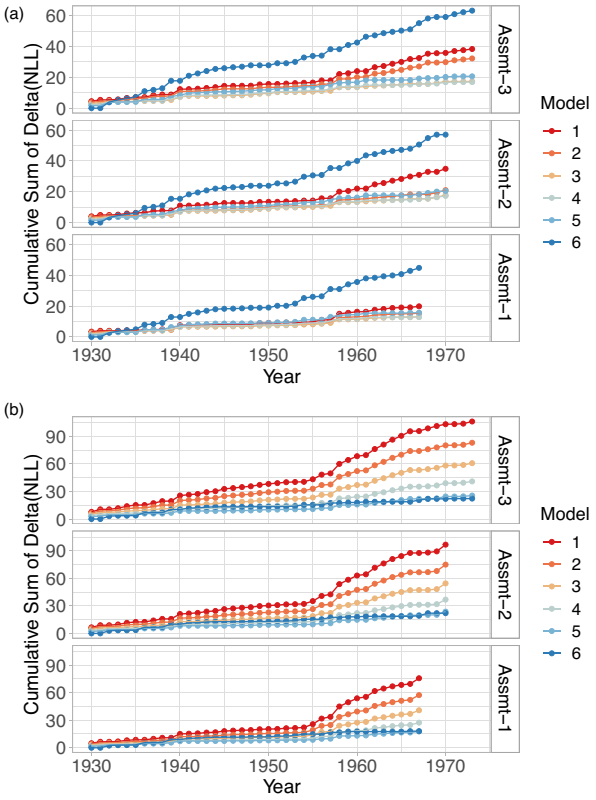

Figure 1. Plots of "model flow" showing cumulative negative log likelihood (NLL) differences for Case 1 (a) and Case 2 (b) for candidate models 1–6. Each panel is an assessment, with three assessments per case study. Three years of data are added for the second and third assessments.

Considering a larger suite of individual model fits (M_1–M_{12}, M_1–M_{24}, or M_1–M_{36}), the pattern described for Case 1 with M_1–M_6 remains, i.e. M_3–M_4, and additionally, M_{15}–M_{16}, had the lowest AIC scores. Notable is that each of the consecutive model number pairs bounds the true M and true selectivity function in the OM. For Case 2, when additional models are considered—then the lowest AIC scores were for M_{11}–M_{23}—where the higher M and doming in the fishery, fishery independent surveys, and the SRR (via a Ricker function) absorbed some of the misspecification from underreported catch (Supplementary Fig. S3). The best individual models identified for Case 1 closely followed the true OM trajectories for SSB and F, whereas the best models for Case 2 did not (Fig. 3).

Reference points estimated for each of the 36 individual models separated into 6 clusters, with the fixed M value driving the greatest difference (Fig. 4). Within the M-clusters, additional variability is driven by the functional form of selectivity for the fishery and surveys and the population scale ($SSB_{F40\%}$ is scaled by individual models' estimates of mean recruitment). Lastly, the $F_{40\%}$ reference points were slightly lower for Case 2 and $SSB_{F40\%}$ slightly higher (Fig. 4a). True stock status in the final model year was bounded by the models with fixed M closest to the true M of 0.275 (Case 1), while for Case 2, the bias in estimated F due to underreported catch produced very inaccurate overfishing status. Nearly all Case 2 models (33 out of 36) estimated F to be less than the reference point, when in fact the true F in the last 10 years of the OM was exactly equal

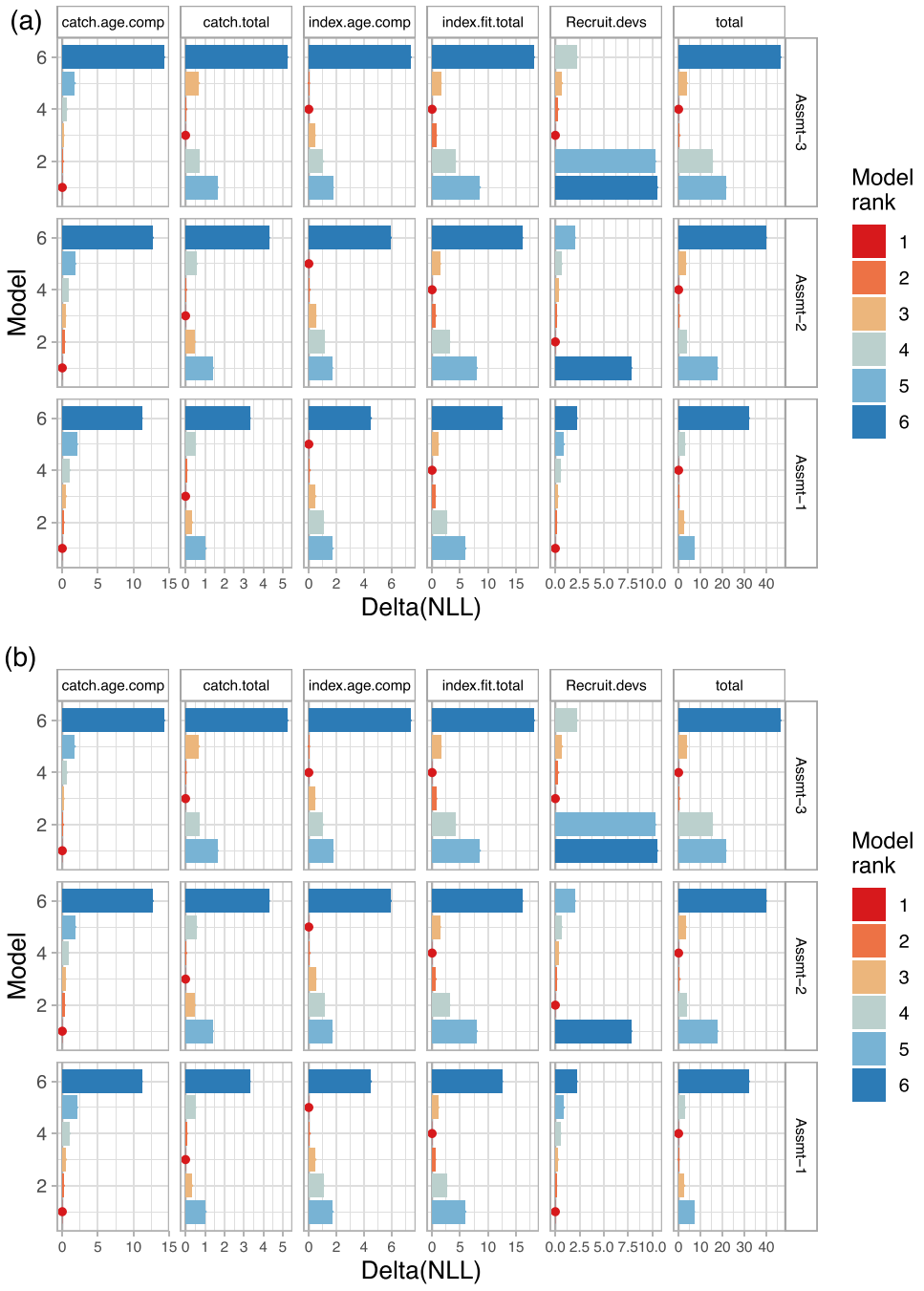

Figure 2. Differences in negative log likelihoods for data components, and for the total negative log likelihood (last column) for Cases 1 (a) and 2 (b). Each row facet is an assessment, with three assessments per case study, where three years of data are added for the second and third assessments. A filled circle indicates the model with the lowest score for that data component of a given assessment.

to the reference point, i.e. the true *F* status specified in the OM was exactly 1 (Fig. 4b).

Model weighting methods and ensemble performance

The alternative model weighting metrics generated three broad patterns that produced different estimates of central tendency for each ensemble. Information theoretic weighting methods distributed the weight across a small subset of the models, while MSE-based weights and Equal Weight allocated weight more broadly across ensemble members. MSE weights are derived from an L2, or squared, loss function, whereas the information theoretic weights are derived by exponentiating the differences in AIC, and this difference in scaling likely explains why AIC weights emphasize very few models compared to MSE weights. Cross-validation weights produced results that were intermediate to these two broad patterns (Fig. 5). For a given case study ensemble, the information theoretic weights varied the most between assessment updates, but this variability was strongest for Case 2, where the model with lowest AIC changed with just 3 years of additional data. For a given weighting metric, the weight assigned to a given model also varied depending on the compo-

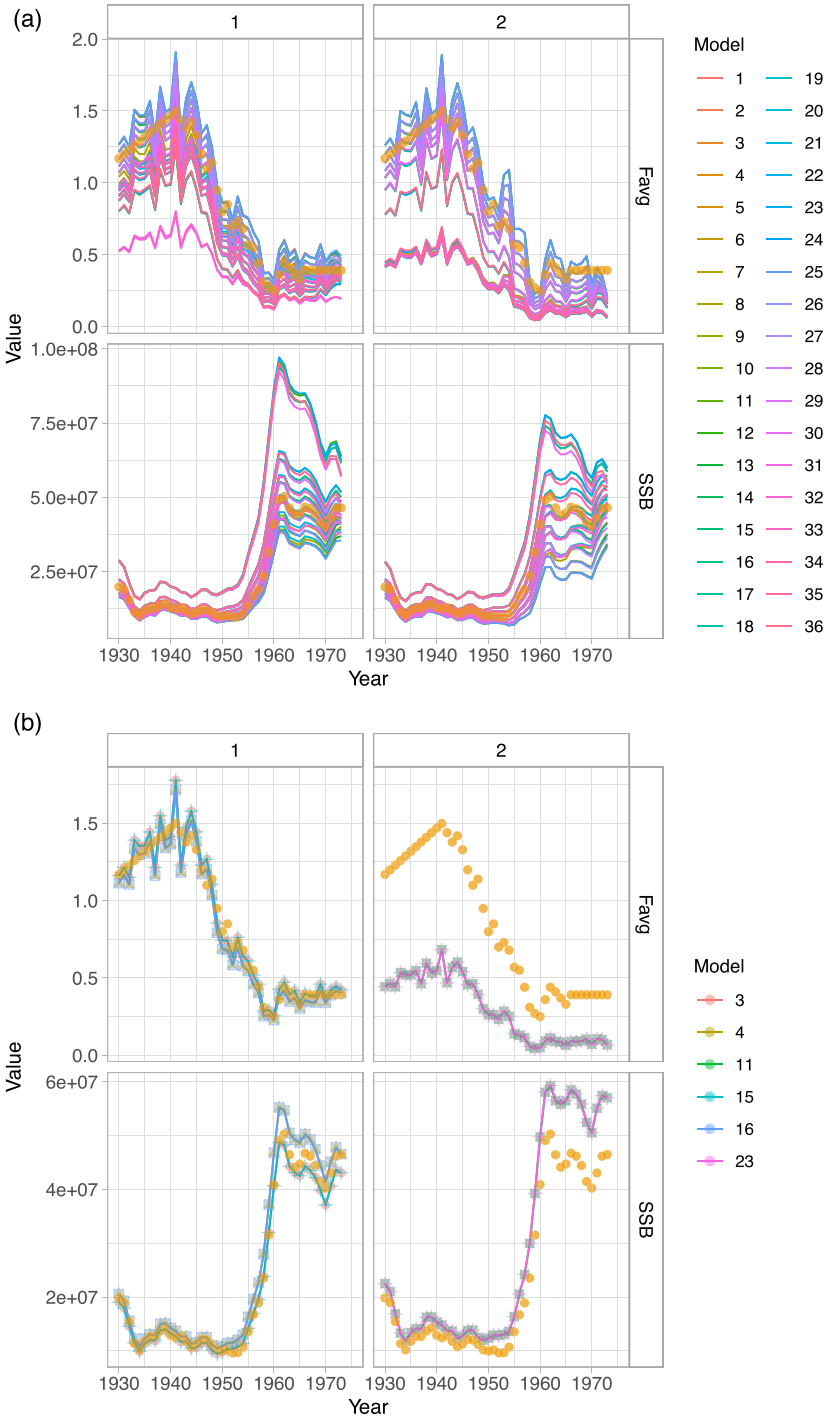

Figure 3. (a) Individual model estimates of spawning stock biomass (SSB) and average fishing mortality (Favg) on ages 9–10. (b) Best individual model estimates of spawning stock biomass (SSB) and average fishing mortality on ages 9–10 (Favg). The true OM values are indicated by filled circles.

sition of the ensemble (Supplementary Fig. S4), although the weights from E [1:6] were generally just parsed in blocks of 12 that aligned with the blocking of factors in our EM design (Table 3).

While the true state of nature is never known, the ensemble approach aims to encompass it in QOI distributions by integrating across parametric and structural uncertainties, which is achieved by selecting a model weighting method. Given the widely different patterns in weight distribution and variability of weight allocation depending on ensemble members, we provide high-level summaries of ensemble performance for all weighting schemes, and detailed summaries for a subset of weighting schemes. Specifically, for ensemble distributions of QOI, we summarize median absolute relative error (MARE) to characterize bias relative to the true OM quantity, median absolute deviation (MAD) to characterize the spread of ensemble distributions, and plot relative error (RE) to illustrate patterns of over- and underestimation. We used the knowledge that the true OM model specifications are bounded by M_3–M_4, to help select four weighting schemes for more de-

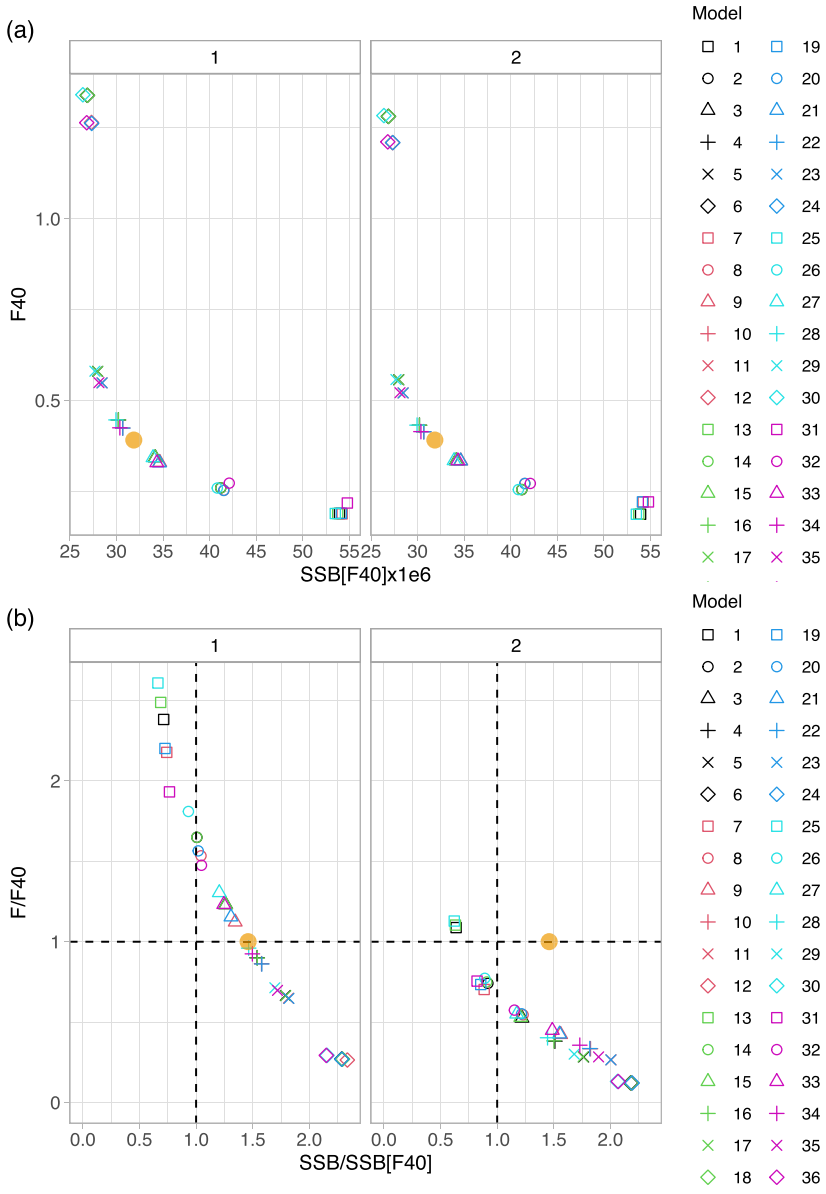

Figure 4. Estimated reference points (a) and stock status (b) for all 36 individual models for Cases 1 (left panels) and 2 (right panels). Color indicates blocking in model specifications (see Table 3), while symbols indicate a fixed value of natural mortality (M) ranging from 0.15 (open square) to 0.5 (open diamond). The solid circle is the true OM reference point (a) or true OM stock status (b).

tailed analysis. For Case 1, weights based on the AIC.total, AIC.Index, and BIC consistently identified these models as having the highest weights (Fig. 5). We therefore focus on AIC.total and AIC.Index (given their ability to identify the true OM structure), and consider cross-validation and equal weights in order to contrast the information theoretic scores with predictive skill and uninformed choice (which is often a default).

Comparing estimates of ensemble QOIs between Cases 1 and 2 highlights that the "best" weighting method depends on the veracity of the data used in the model. Nearly all weighting metrics for Case 1 were median unbiased, but information theoretic-based weights (BIC, AIC total, and AIC.Index) had the lowest MARE and RE (median across all years) for SSB and Fmult, while MSE-based and equal weight had the largest; cross validation was intermediate. However, with catch underreporting (Case 2), the metrics that tended to spread the weighting factor across more models (equal weight and MSE) and AIC based on the catch composition had the lowest MARE and the interquartile range of RE contained the true value, while all of the other AIC-based weights and BIC were biased by 40% or more (Fig. 6a). The same patterns in RE by case and QOI held for stock status distributions, but the magnitude of bias was greater (Supplementary Fig. S5). In the last 15 model years, bias for the four weight metrics was negligible for Case 1 but was substantial for AIC, AIC.Index, and cross validation in Case 2, and even the equal weight method was negatively biased (Fig. 6b). Among the individual models comprising each ensemble, the MARE and RE for the single best model (as identified by AIC), always had one of the lowest MARE and RE for Case 1 (Fg. 6a). In Case 2, the MARE for the single best model ranked inconsistently between SSB and F quantities, between the full vs the last 15 model years, and the ensemble composition.

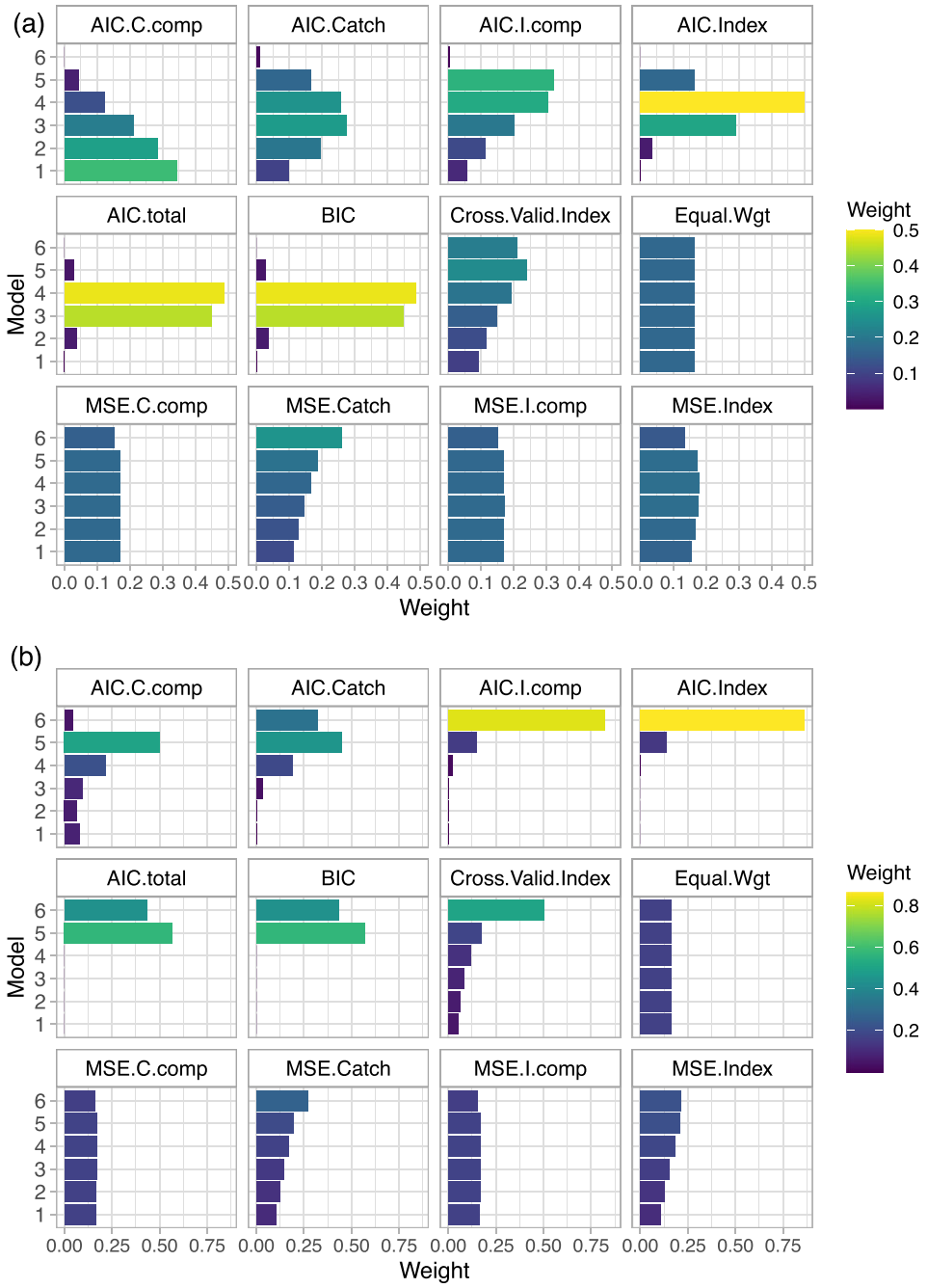

Figure 5. Weights assigned to an ensemble comprised of models 1–6 for Cases 1 (a) and 2 (b) for the first assessment. The true OM model structure is midway between M_3 and M_4.

With respect to precision of the QOI distributions, the single best model nearly always had the lowest MAD, and this finding held for both Cases 1 and 2. This is expected given that the correlations among the individual model estimates of SSB and F were nearly 1; therefore, the expected variances for an ensemble would necessarily be greater than for an individual model. Similarly, the weighting schemes that spread weight among more models had larger MAD (e.g. equal weight and MSE-based weights), and often produced multimodal distributions (Fig. 7).

A pair of 3-model ensembles that were intentionally constructed to not span the true OM were also examined (E [1:3] and E [4:6]), because in practice one does not know the true model specifications. For Case 1, the patterns in relative bias are as expected, with mostly negative bias in the ensembles for E [1:3] (because all *M* values are below the true), and positive bias for E [4:6] (all *M* values are above the true) (Supplementary Fig. S5). For Case 2, directional bias was similar for E [1:3] and E [4:6], but the magnitude of bias was greater and precision was lower. In both case studies, bias and precision of the best individual model comprising these 3-model ensembles were better than for the ensembles.

Retrospective performance of ensembles vs individual models

For both case studies, the magnitude of Mohn's rho for individual models was smallest for models with the lowest AIC, and was smaller in general for Case 1 due to the underreported

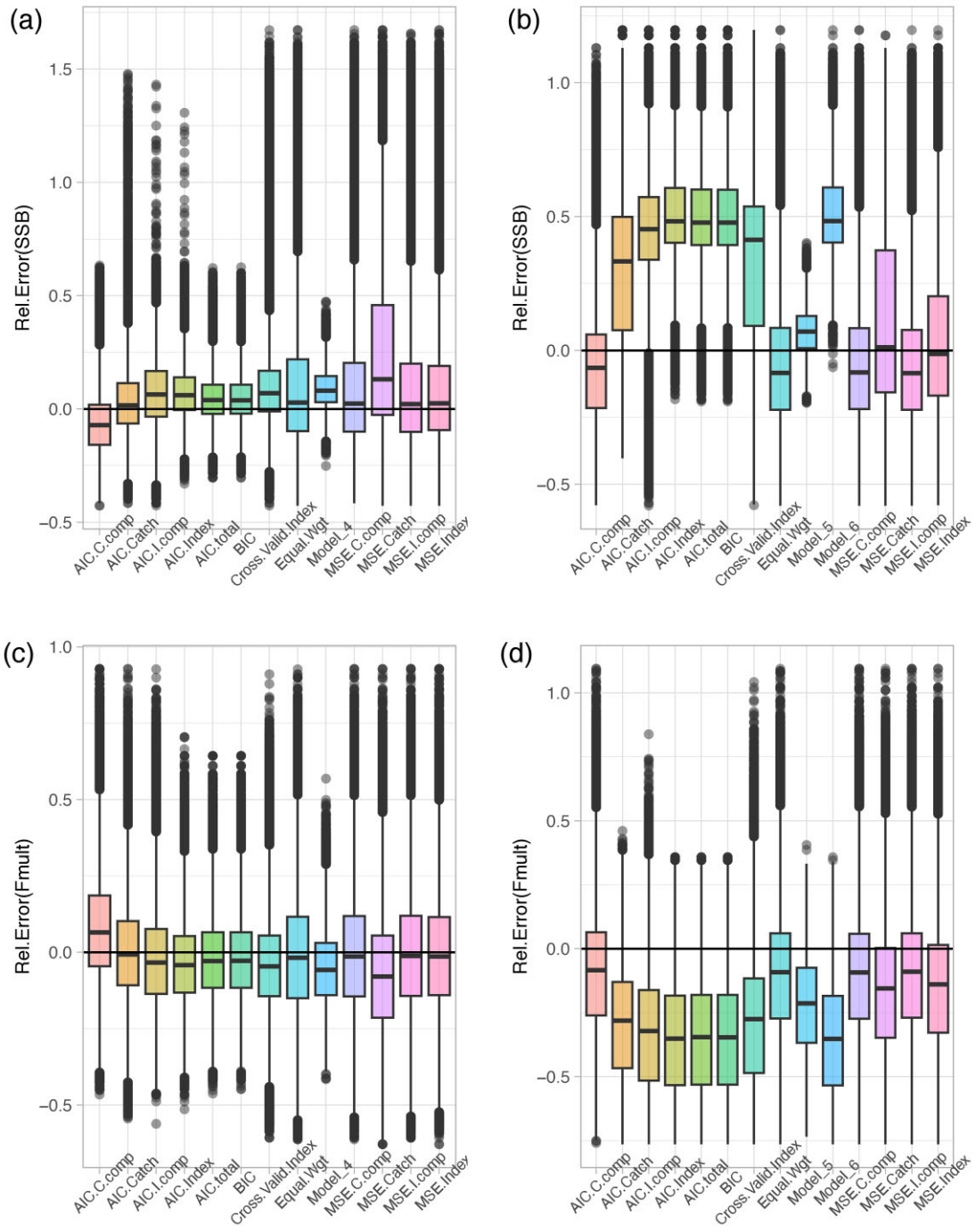

Figure 6. (a) Relative error (RE) in ensemble estimates of SSB (a, b) and Fmult (c, d) from weighting M_1–M_6 by the indicated weighting method. RE was aggregated over all years for the third assessment for Case 1 (left column) and Case 2 (right column). (b) Relative error (RE) in ensemble estimates of SSB (a, b) and Fmult (c, d) from weighting M_1–M_6 by the indicated weighting method as well as the best performing individual model(s) as identified by AIC. RE is shown for the last 15 years for the third assessment cycle for Case 1 (left column) and Case 2 (right column).

catch in Case 2 (Fig. 8a). Consistent with the AIC result for individual models, the retrospective pattern for ensembles was smallest when AIC was the weighting metric. In general, regardless of the weighting metric, Mohn's rho for ensembles was small due to combining bidirectional errors. The exception was for E [1:3], which was biased for all weighting metrics because all three models in that ensemble had the same sign for Mohn's rho. We also observed a change in sign for Mohn's rho across individual models spanning the true OM; however, the distance away from those OM specifications (in model space) differed between cases.

The lowest retrospective patterns for Case 1 corresponded to the individual models closest to the true OM structure and for ensembles that were weighted by information theoretic methods, which is consistent with the MARE, RE, and MAD results. This suggests that if the data are reliable, then standard model selection criteria such as AIC will select the models with the most appropriate structure and acceptable diagnos-

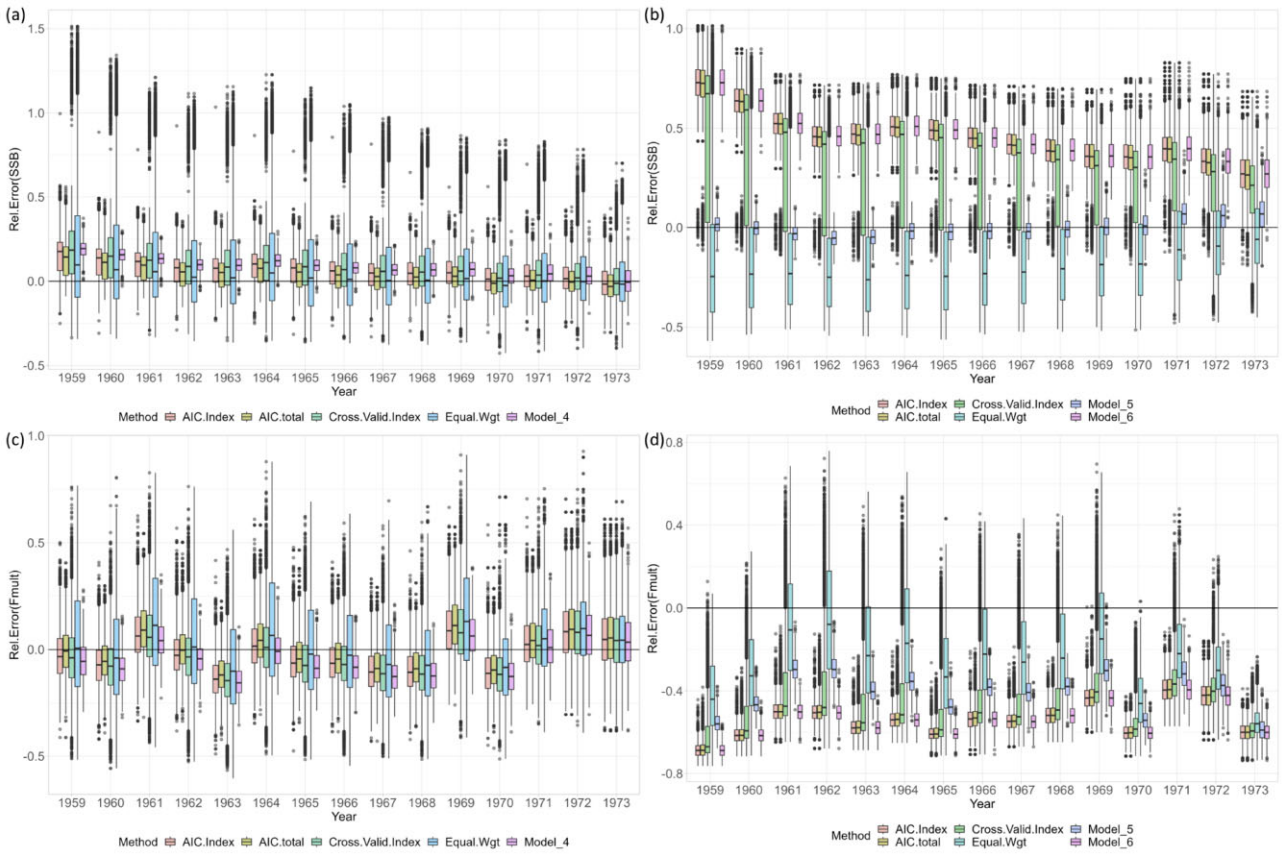

Figure 6 —*continued*

tics. However, for Case 2, MARE and RE were best for weighting methods that distributed weight across more models but these same methods had worse MAD and larger Mohn's rho. Therefore, simultaneously achieving accuracy, precision, and a small retrospective pattern will be challenging if one suspects problems with the underlying data.

Performance of forecasts

The two forecast models reflected different hypotheses about future recruitment, and the resulting forecast weights (derived from comparing previously forecasted QOI with estimates of the same QOI in the updated assessment) depended on the QOI being evaluated. When recruitment was the QOI for calculating forecast weights, this produced the largest difference in weights between the two forecast models (Supplementary Fig. S6), followed by catch and SSB. The small difference in forecast weights when catch or SSB was the QOI is due to the short length of projections (3 years) and smaller relative contribution of new recruits to those quantities in the near term (low selectivity to the fishery, low proportion mature, and low weights at ages 1–3). Exploring additional forecast models that specify different selectivities or weights at age to address uncertainty about those quantities in the short term might be reasonable and could result in greater differences in the forecast weights derived from catch and SSB forecast skill.

The ability of the two forecast models to predict recruitment differed between the assessment updates, most likely reflecting variability in the recruitment process (Supplementary Fig. S1). The forecast weights assigned to forecast model F1 (using the full time series of recruits) vs F2 (using the recent time series) ranged from ~F1:F2 = 65:35 to 50:50, depending on the assessment model and the case study. The two forecasts per model M_m are first combined based on forecast skill yielding m trajectories. The full ensembles (of length assessment + forecast years) were then formed by applying the assessment model weights ($w_{m,k}$) for all k weighting methods to those m trajectories. Consequently, the weighting method producing the best ensemble forecast was case-specific—for Case 1, information theoretic weights were accurate and more precise than equal weight methods, but Case 2, ensemble forecasts only included the true OM if the weighting method spread weights across many models (such as equal weight; Fig. 9).

Discussion

Applications of ensemble modeling in stock assessment have been increasing, yet controlled evaluations of the impact of decisions needed to implement this approach are lacking. The claim that ensembles will outperform individual models has not been investigated to determine under what conditions this may be true. As a first step to address this need, we developed a simulation study to explore the construction of ensembles in all aspects of integrated stock assessment models, including reference points, stock status, and short-term forecasts from two candidate forecast models. By doing so with simulated case studies, we were able to evaluate aspects of the ensemble process such as candidate model composition, weighting metric, and how model weights evolve with new years of data. Comparing estimates of assessment QOI between ensembles

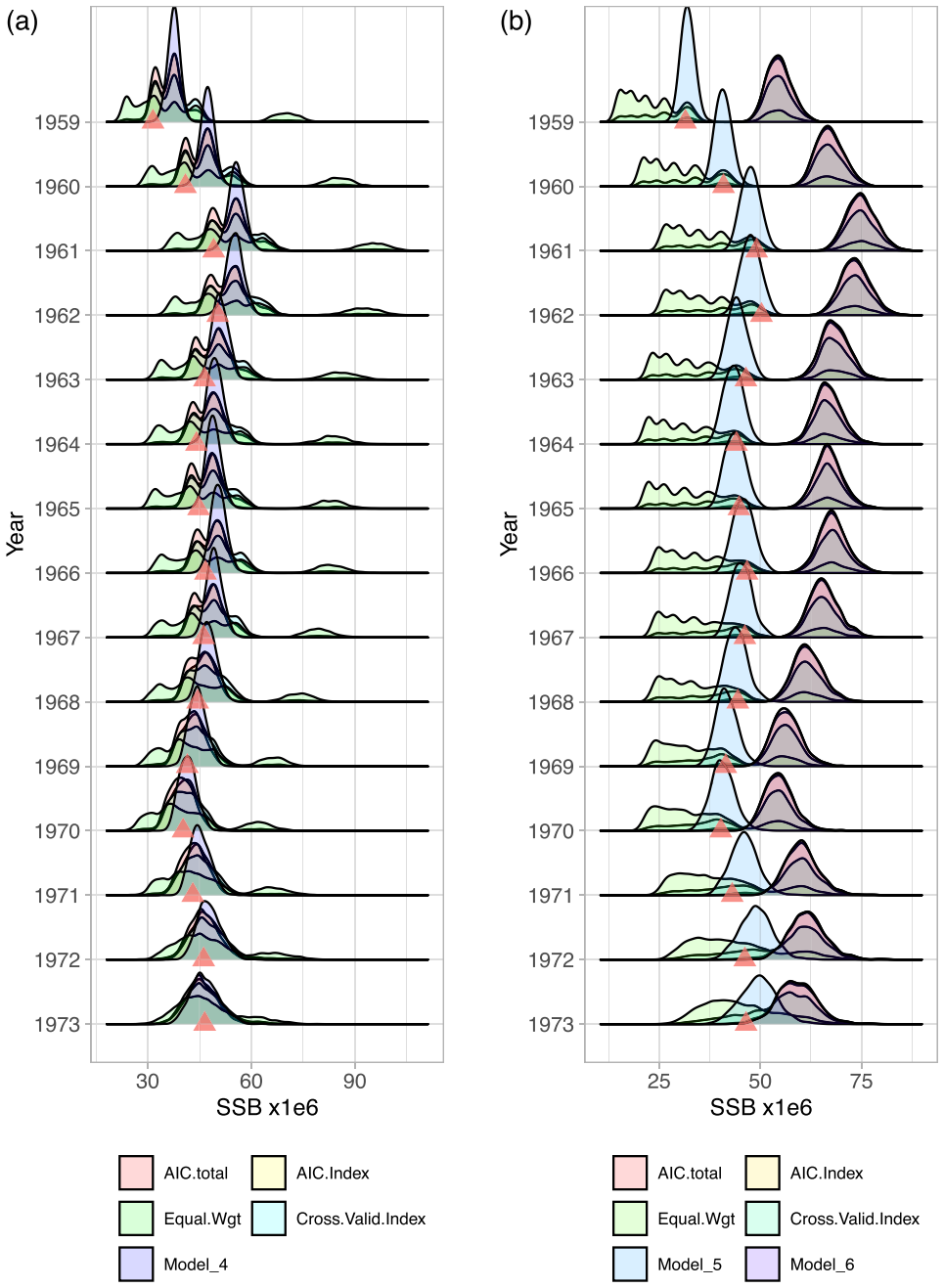

Figure 7. Ensemble distribution of SSB for the last 15 model years from M_1 to M_6 for Cases 1 (a) and 2 (b) for the third assessment cycle. Ensemble distributions are plotted for four model weights in addition to the individual model(s) identified as best by model AIC. The solid triangle under each annual distribution indicates the true OM SSB in that year.

and the single "best" model, we found that ensembles generally had lower MARE but this depended on the chosen weighting metric and on the ensemble spanning the true OM model structure.

By design, none of our EMs matched the OM, but we knew whether or not an ensemble spanned the correct configuration. In real stock assessments, ensuring that the true state of nature is bound by the models comprising the ensemble will be a challenge. In our case studies, we found that the sign of Mohn's rho changed across the candidate models for a given ensemble QOI (SSB, Fmult) whenever the ensemble spanned the true OM. We also found that the "best" individual model (lowest AIC) varied between assessment updates when there was misspecification (Case 2), indicating lack of robustness in the best model structure. These two observations may lend insight as to the sufficiency of the candidate model set (when a change in sign for Mohn's rho is observed) and potentially the mechanism i.e. driving misspecification (when the "best" individual model changes over time). We recommend checking if these observations hold in future ensemble model simulation studies for other (or multiple) sources of misspecification.

The choice of weighting metric affected ensemble bias and precision. AIC was able to identify the individual models that most closely matched the OM for Case 1 (no underreported

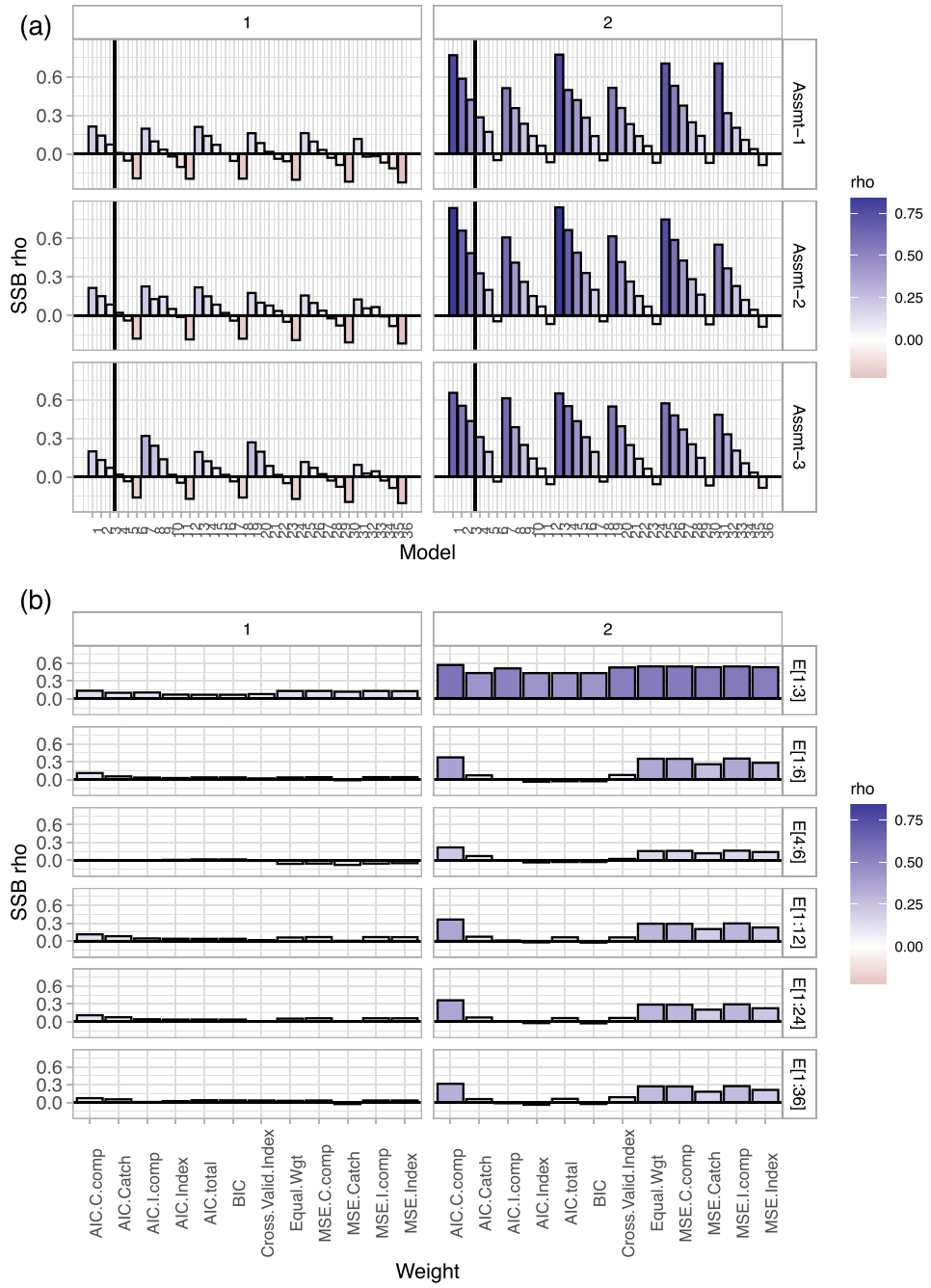

Figure 8. (a) Mohn's rho for SSB for each individual model (a) and for ensembles (b) composed of M_1–M_3, M_1–M_6, M_4–M_6, M_1–M_{12}, M_1–M_{24}, and M_1–M_{36} for all 12 weighting metrics. The solid vertical line in (a) shows location of true OM specifications. (b) Mohn's rho for *F* for each individual model (a) and for ensembles (b) composed of of M_1–M_3, M_1–M_6, M_4–M_6, M_1–M_{12}, M_1–M_{24}, and M_1–M_{36} for all 12 weighting metrics. .

catch). Using AIC weights led to generally unbiased ensembles because the models receiving the most weight spanned the true OM. Ensembles based on equal weights for Case 1 were also generally unbiased, but the disadvantage was far less precision and multimodality in the QOI distributions. Due to the misspecification of catch in Case 2, the AIC was lowest for individual model structures that accommodated this data misspecification (higher *M*, doming in selectivity, and the SRR). Consequently, ensembles formed with AIC weights were biased, while weighting metrics that allocated model weights more broadly had less bias (equal weight and also the MSE weighting methods). But as with Case 1, these metrics produced ensembles with less precision and multimodality. Model weights based on cross validation never outperformed AIC in Case 1 and performed worse than equal weight and MSE weights in Case 2. This is because in both cases, the cross-validation scores were similar to the AIC.Index and AIC.I.comp scores, as might be expected given those were the data components dropped for skill evaluation (Fig. 5).

There is probably no single approach to setting model weights that optimizes predictive accuracy within a given statistical estimation framework (e.g. Wolpert and Macready 1997) because there are too many ways for individual assessment data-model configurations to differ from the true state

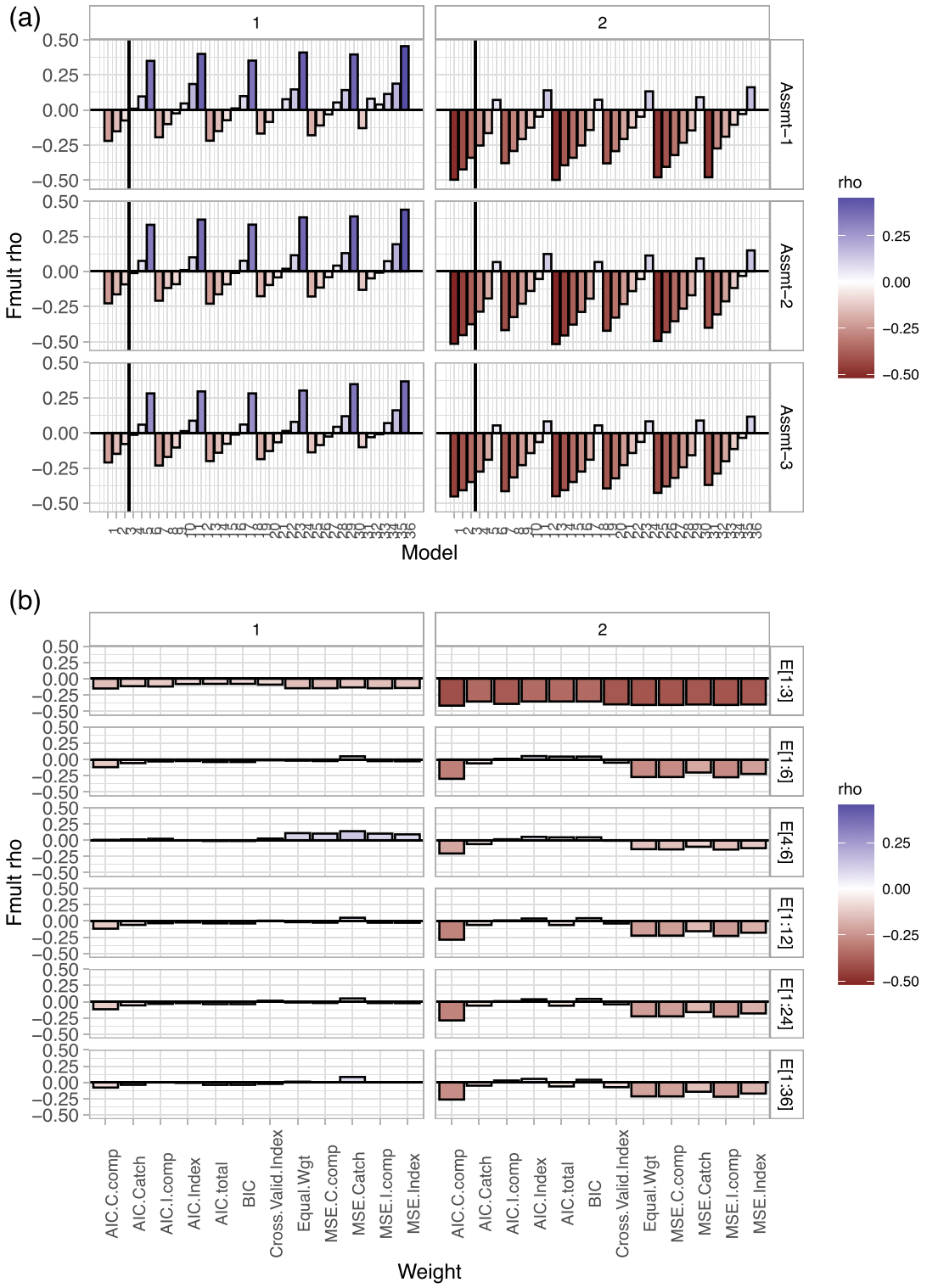

Figure 8 —*continued*

of nature. Of our two case studies, Case 1 is the ideal, and the traditional model selection tool (AIC) identified the most appropriate individual models and produced the optimal ensemble (low bias, high precision). However, we want to emphasize the findings of Case 2, where equal model weights performed best because we expect that mismatches between assumptions about the data and structural choices in the models will be more common in real assessments. Accepting that some misspecification is unavoidable, AIC is probably not the ideal weighting metric for ensembles but may serve as a useful diagnostic. As a default, assigning equal weights could be expected to have less bias, but careful consideration of candidate models for inclusion in the ensemble is needed because disparate MCMC distributions for QOI from each model are

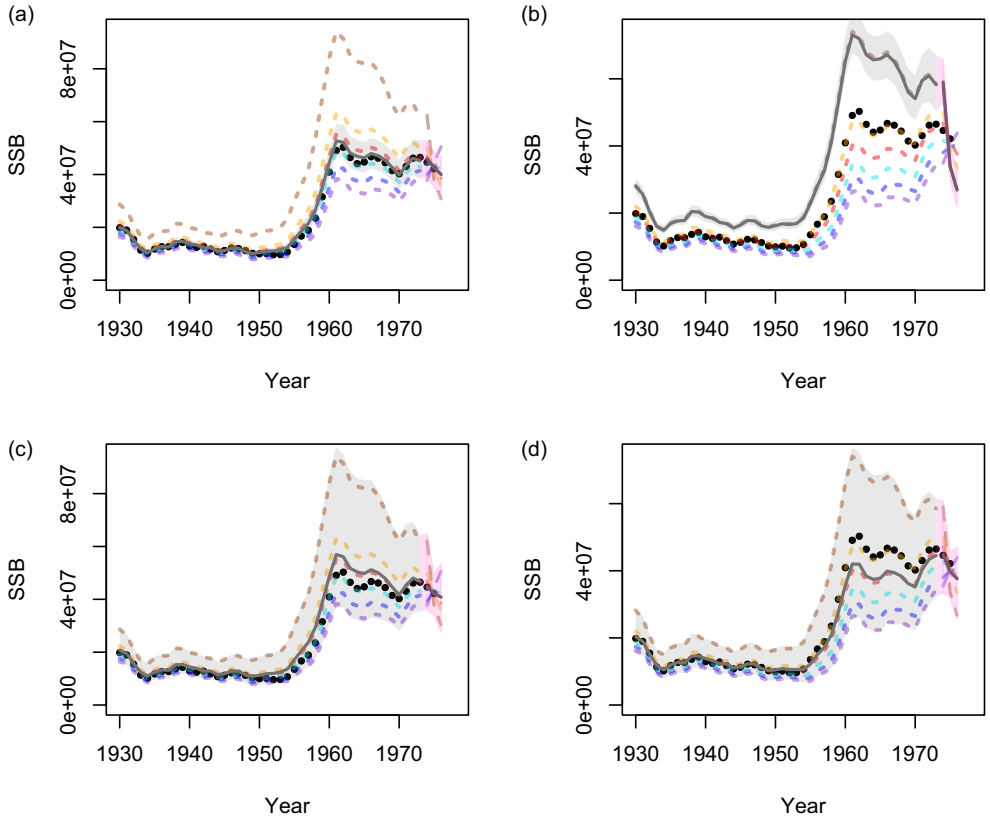

Figure 9. Ensemble distribution of SSB for Case 1 (left) and Case 2 (right) for M_1–M_6. The shaded polygon through 1973 is the 95% CI of the ensemble assessment weighted by AIC.total (a, b) or equal weights (c, d). The polygon in the final three years of the plotted trajectory is the 95% CI of the ensemble forecast, where forecast models within each *m* assessment were first weighted based on forecast skill for predicting recruitment, and then those weighted forecasts were combined with the ensemble assessment weights for each assessment model. Solid line is the ensemble median, filled circles are the true OM values, and dashed lines are individual model estimates of SSB (short dash = assessment, long dash = composite of the two forecast models per assessment)..

being combined as if they are all equally plausible, and this is what produces the multimodality. The distance between the modes in the ensemble QOI distribution depends on how far apart the models are in parameter space (such as levels of *M* in our model set), although one approach to minimize those gaps between modes would be to construct the ensemble from individual models with finer parametric or structural resolution. Multimodality can complicate risk characterization in management advice, particularly when the modes have similar probability density or when the median falls between modes where there is very little probability density (Fig. 7).

Although AIC may not serve as a default basis for setting model weights, we suggest that there is still value in calculating it for individual models. The use of AIC has a statistical foundation, is transparent, is repeatable, and has performed well in simulation studies (Helu et al. 2000, Maunder and Punt 2013). Our finding that the best model (by AIC) varied with only 3 years of additional data suggests that examining model selection across consecutive assessments, and in retrospective peels, should be a standard diagnostic for characterizing whether structural uncertainty is present (e.g. Fig. 3 in Brooks 2023). Instability in model selection between updates or across years in retrospective peels could indicate a change in an underlying process, i.e. not captured in the model structure (a non-stationary process) or misspecification with respect to what is assumed about data accuracy or suitability. But AIC can only be compared between models if the data (and the data component weights or variances) are consistent, the same likelihood functions are used, and the models are of the same complexity (all age-structured, as opposed to modeling across both age-structured and production models, e.g.). In practice, variance parameters are usually tuned to achieve acceptable diagnostics, to downweight contradictory or less reliable data sources (Maunder and Punt 2013), and because the appropriate sample size is difficult to specify due to high correlation in age or length samples (Pennington and Volstad 1994; Francis 2011). To preserve the utility of AIC as a diagnostic tool for ensembles, we suggest that any tuning to data weighs be done once, on a consensus base model, with no further tuning on the other candidate models, in order to emphasize the underlying probability models as the informational basis for estimating and comparing assessment model parameters and QOI.

In the status quo assessment process, a single best model is selected during what is typically referred to as a benchmark assessment, and then that model is used in all subsequent assessment updates until a new benchmark is held. Unrepresented uncertainty associated with selecting a single model at a benchmark is compounded in updates if the true dynamics evolve away from that model (e.g. Case 2 in Fig. 1). The ensemble approach, in principle solves, this status quo assessment issue if model weights are recalculated at each update. However, if one defaults to equal weights, then model weights would only change if the number of candidate mod-

els changed. This could happen if criteria for determining candidate models are evaluated at each update, leading to either dropping formerly included models or including models that previously did not meet the ensemble membership criteria. As knowledge accumulation and performance feedback stimulate new hypothesis generation, we expect there to be interest in evaluating new model structures when there is an opportunity to do so. Reevaluating ensemble model composition as part of the assessment update should be appropriate, with the reminder that expectations of low bias result from bidirectional prediction errors, and admitting or excluding different models could influence the balance of those errors. We caution, however, that asymmetries in risk associated with model configurations could provide incentives for "subjective influence" (Hordyk et al. 2019). For this reason, we suggest that the models be identified based on the merits of their hypotheses, i.e. before viewing full assessment results.

To maintain a balance in model composition, a two-step approach that first weights candidate models according to hypotheses and associated process models about states of nature (perhaps with the aid of *post hoc* clustering analysis) and then weights models within those categories by model skill or equal weights could be explored (see e.g. Burnham and Anderson 2002, Jardim et al. 2020). Algorithmic approaches to regularization may be useful for model selection, ensemble averaging, and variance stabilization in fisheries applications. Importance sampling has also been applied to select more probable ensemble members for averaging (Ducharme-Barth and Vincent 2022). However, if all individual models produce estimates of Mohn's rho that all have the same sign (e.g. ensemble E [1:3]), we suggest either revisiting the range of current axes of uncertainty or considering additional axes. We recommend further testing of these approaches to identify candidate models through simulation to evaluate if they can alleviate overemphasis of a subset of configurations that arise either inadvertently or unscrupulously.

Retrospective patterns in ensembles were very small, even when the retrospective patterns of some of the individual models were large, due to the canceling effect of bidirectional error. An alternative ensemble approach was proposed by Legault (2020), where ensembles formed by retro-adjusting the terminal year estimate of individual models (the "Rho" approach) were compared with ensembles formed from individual models that were modified (structurally or parametrically across different years) until the retrospective pattern was removed (the "Rose" approach). For our case studies, the retrospective adjustment of terminal year estimates improved accuracy of the F estimates in Case 2 (which were biased due to unreported catch), but for SSB in Case 2 and for SSB and F in Case 1, the rho-adjustment pushed the individual model estimates further from the true value (Supplementary Fig. S8). Similarly, combining the individual rho-adjusted models into an ensemble generally did not improve accuracy except for F in Case 2 (Supplementary Fig. S9). However, assuming the retrospective cause continues into the future (which it does across our assessment updates), then making a rho-adjustment prior to projections improved the catch advice for Case 2, making it closer to what the underreported catch would be (Supplementary Fig. S10). We did not attempt the Rose approach, but the nature of retrospective patterns is that their magnitude (and occasionally their direction) varies as data are added (Brooks and Legault 2016), which would necessitate repeating the time-consuming search for model modifications each time the assessment is updated. Given the importance of bounding the true but unknown model structure, it may be more efficient to ensemble across individual models that produce both positive and negative retrospective patterns to ensure the cancelation of prediction error. We recommend more simulated case studies to evaluate different misspecifications, and to develop appropriate guidelines for dealing with retrospective patterns in ensembles.

The EMs fitted to these case studies were traditional SCAAs, and the OM model was simple in that selectivity in the fishery, indices, and biological parameters were constant over the whole time series. This is an appropriate starting point for understanding the behavior of ensembles with respect to candidate model composition and weighting metrics. However, real data are typically heterogeneous and exhibit much more variability. Future ensembles could consider state space assessment models as candidates, where some of the fixed model structures in our EMs could instead be modeled to vary annually via random effects (e.g. annually varying selectivity, or deviations in natural mortality; Nielsen and Berg 2014, Stock and Miller 2021). Treating model uncertainty via different process error structures may help limit the number of candidate models in an ensemble of state space models relative to specifying multiple models with different fixed selectivity blocking and fixed parameter intervals in the traditional SCAA—testing this through simulation studies is recommended.

The importance of addressing uncertainty is widely recognized but its treatment tends to depend on the scientific field (e.g. Simmonds et al. 2022). Stock assessment scientists often impose structural and parametric assumptions to reduce dimensionality and make the models tractable. These assumptions can be consequential, but the uncertainty associated with alternative assumptions is not formally accounted for and risk is therefore mischaracterized in the resulting advice. Reliance on a single model that only reflects parametric and not structural uncertainties is likely insufficient for fisheries assessment and management, and accounting for model-based uncertainties is needed (Peterman 2004, Jardim et al. 2020, Kell et al. 2021, Ducharme-Barth and Vincent 2022). This is acknowledged to some extent in the *ad hoc* approach of inflating uncertainty around assessment estimates depending on tiers of perceived quality and complexity of the model (e.g. Ralston et al. 2011, MAFMC 2019). These *ad hoc* approaches to account for model-based uncertainty could be dealt with more directly through ensembles, where the inclusion of multiple candidate models explicitly captures uncertainty.

Schnute and Richards (2001) advise that scientists remain cognizant of their imposed model constraints and consequent limitations to advice, and maintain skepticism of the chosen model by actively seeking alternative explanations and implementing robust strategies. Monte Carlo simulations across uncertain parameters, Bayesian approaches to integrate alternative hypotheses, decision tables, management procedures (MP), or management strategy evaluation (MSE) have all been developed to quantify management risk associated with model uncertainty (Restrepto et al. 1992, Punt and Hilborn 1997, Butterworth 2007, Punt et al. 2016, Hordyk et al. 2019). Ensemble models are another tool that can be used to address robustness by aggregating results of alternative models into QOI distributions that inform management advice. In our experience, managers are more comfortable making decisions from one model rather than many, i.e. "Give me one set of numbers,

please!" Thus, adapting the management advice process to an ensemble context will take careful dialogue and demonstration for successful implementation.

Similar to the other approaches to risk management, practical implementation of ensemble models comes with additional costs such as increased computing time, possible intractability to examine detailed diagnostics (for large ensembles), and issues related to objective candidate model selection (Jardim et al. 2020, Legault 2020). However, as shown in our simulation results for Case 2, biased fisheries catch data will lead to inaccurate predictions under some model weighting schemes if the bias in the observed data is not modeled or the retrospective pattern has the same sign for all individual models. Thus, ensemble modeling will not provide a panacea for unrepresentative data. Nonetheless, relying on single model assessments in situations where structural uncertainty is important can oversimplify the system dynamics and give a false sense of predictive accuracy and uncertainty representation for decision makers and stakeholders. Further investigations of the benefits and trade-offs of applying an ensemble approach (perhaps in concert with MP) are needed to position the assessment process to handle future challenges posed by changes in the environment as well as in the data we sample to track those changes (e.g. Brodziak and Link 2002).

Acknowledgments

We gratefully acknowledge the NOAA Office of Science and Technology for the support that allowed NOAA personnel to travel to ICES working group meetings (WGRFE) and collaborations through the European Commission Joint Research Centre Exploratory Research Program, via funded proposals through the Stock Assessment Analytical Methods program. We also acknowledge fruitful discussions following early presentation of this work at the 13th National Stock Assessment Workshop (2018) and during virtual meetings of the ICES WKENSEMBLE. We thank Chris Legault, the editor, and two anonymous reviewers for comments that improved the manuscript.

Author contributions

E.B.: conceptualization; formal analysis; methodology; software; visualization; writing, review, and editing and J.B.: conceptualization; methodology; writing, review, and editing

Supplementary data

Supplementary data is available at *ICES Journal of Marine Science* online.

Conflict of interest: The authors have no conflict of interest to declare.

Data availability

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

Appendix A. The simulation model

Population dynamics for numbers at age including recruitment ($a = 1$), survival of true age classes ($1 < a < A$), and survival for the plus group ($a = A$).

$$N_{a,y+1} = \begin{cases} 4hR_0SSB/[R_0\varphi_0(1-h) \\ \quad + (5h-1)SSB] & a = 1 \\ N_{a-1,\,y}\exp(-\mathrm{M}_{\mathrm{a-1,y}}, -\mathrm{F}_{\mathrm{a-1,y}}) & 1 < a < A \\ N_{a-1,\,y}\exp(-M_{a-1,y} - F_{a-1,y}) & \\ \quad + N_{A,y}\exp(-M_{a,y} - F_{a,y}) & a = A \end{cases} \quad \text{(A1)}$$

where a is age, y is year, $M_{a,y}$ and $F_{a,y}$ are natural and fishing mortality at age in year y, and A is the plus group. Recruitment occurs at age 1 following the Beverton–Holt stock-recruit function parameterized in terms of spawners (SSB), steepness (h), unexploited recruitment (R_0), and unexploited spawners per recruit (φ_0). Recruitment was assumed to have a lognormal error with $\sigma_{\mathrm{R}} = 0.5$. Natural mortality ($M$) was time and age invariant, while maturity and weight at age were time invariant (Table 2).

Aggregate data for fisheries and indices were generated as follows:

$$C_y = \sum_{a=1}^{A} N_{y,a}(1 - \exp\left(-M - F_y s_a\right)\frac{F_y s_a}{M + F_y s_a}, \quad \text{(A2)}$$

$$I_{y,\,i} = \sum_{a=1}^{A} N_{y,a}\exp\left(\Delta_i\left[-M - F_y s_a\right]\right)\pi_{a,i}q_i, \quad \text{(A3)}$$

where catch in year y, C_y, is calculated from annual fishing mortality (F_y) and time-invariant fishery selectivity at age (s_a). Two survey abundance indices were simulated, where the annual aggregate index in numbers ($I_{y,i}$) was calculated from index-specific selectivity at age ($\pi_{a,i}$), index-specific catchability (q_i), and numbers at age decremented for time elapsed prior to the start of the survey (Δ_i, where indices were observed in months 1 and 6). Lognormal error was added to create "observed" aggregate catch and aggregate indices.

Age composition for catch and index are defined inside the summation for Appendices (A2) and (A3), and multinomial observation error was added to these to create "observed" time series of composition data.

References

Brodziak J, Cadrin SX, Legault CM *et al.* Goals and strategies for rebuilding New England groundfish stocks. *Fish Res* 2008;**94**:355–66. https://doi.org/10.1016/j.fishres.2008.03.008

Brodziak J, Legault CM. Model averaging to estimate rebuilding targets for overfished stocks. *Can J Fish Aquat Sci* 2005;**62**:544–62. https://doi.org/10.1139/f04-199

Brodziak J, Link J. Ecosystem-based fishery management: what is it and how can we do it? *Bull Mar Sci* 2002;**70**:589–611.

Brodziak J, Piner K. Model averaging and probable status of North Pacific striped marlin, *Tetrapturus audax*. *Can J Fish Aquat Sci* 2010;**67**:793–805. https://doi.org/10.1139/F10-029

Brodziak J, Rago P, Conser R. A general approach for making short-term stochastic projections from an age-structured fisheries assessment model. In: F Funk, T Quinn, J Heifetz *et al.* (eds), *Proceedings of the International Symposium on Fishery Stock Assessment Models for the 21st Century*. Fairbanks, AK: Alaska Sea Grant College Program, Univ. of Alaska, 1998. https://doi.org/10.4027/fsam.1998.52

Brooks EN, Legault CM. ASAPplots: creates standard plots and PDFs for ASAP3. R package version 0.2.18. 2022. https://rdrr.io/github/cmlegault/ASAPplots/ (17 May 2024, date last accessed).

Brooks EN, Legault CM. Retrospective forecasting—evaluating performance of stock projections for New England groundfish stocks. *Can J Fish Aquat Sci* 2016;**73**:935–50. https://doi.org/10.1139/cjfas-2015-0163

Brooks EN. Pragmatic approaches to modeling recruitment in fisheries stock assessment. *Fish Res* 2023;**270**:106896. https://doi.org/10.1016/j.fishres.2023.106896

Burnham K, Anderson D. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. 2nd edn. New York, NY: Springer, 2002, 488

Butterworth DS. Why a management procedure approach? Some positives and negatives. *ICES J Mar Sci* 2007;**64**:613–7. https://doi.org/10.1093/icesjms/fsm003

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

Chamberlin TC. The method of multiple working hypotheses. *Science* 1965;**148**:754–9. https://doi.org/10.1126/science.148.3671.754

Claeskens G, Hjort N. *Model Selection and Model Averaging*. Cambridge: Cambridge University Press, 2008, 312.

Claeskens G, Hjort N. The focused information criterion. *J Am Stat Assoc* 2003;**98**:900–16. https://doi.org/10.1198/016214503000000819

Claeskens G. Statistical model choice. *Ann Rev Statist Appl* 2016;**3**:233–56. https://doi.org/10.1146/annurev-statistics-041715-033413

Dormann CF, Calabrese JM, Guillera-Arroita G *et al.* Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecol Monogr* 2018;**88**:485–504. https://doi.org/10.1002/ecm.1309

Ducharme-Barth N, Vincent T. Focusing on the front end: a framework for incorporating uncertainty in biological parameters in model ensembles of integrated stock assessments. *Fish Res* 2022;**255**:1–15. https://doi.org/10.1016/j.fishres.2022.106452

Francis RICC. Data weighting in statistical fisheries stock assessment models. *Can J Fish Aquat Sci* 2011;**68**:1124–38. https://doi.org/10.1139/f2011-025

Helu SL, Sampson DB, Yin Y. Application of statistical model selection criteria to the Stock Synthesis assessment program. *Can J Fish Aquat Sci* 2000;**57**:1784–93. https://doi.org/10.1139/f00-137

Hilborn R, Stearns S. On inference in ecology and evolutionary biology: the problem of multiple causes. *Acta Biotheor* 1982;**31**:145–64. https://doi.org/10.1007/BF01857238

Hordyk AR, Huynh QC, Carruthers TR. Misspecification in stock assessments: common uncertainties and asymmetric risks. *Fish Fish* 2019;**20**:888–902. https://doi.org/10.1111/faf.12382

Jardim E, Azevedo M, Brodziak J *et al.* Operationalizing model ensembles for scientific advice to fisheries management. *ICES J Mar Sci* 2020;**78**:fsab010. https://doi.org/10.1093/icesjms/fsab010

Karp MA, Blackhart K, Lynch PD *et al.* (eds). *Proceedings of the 13th National Stock Assessment Workshop: Model Complexity, Model Stability, and Ensemble Modeling*. Washington, D.C.: U.S. Dept. of Commerce, NOAA. NOAA Technical Memoranda NMFS-F/SPO-189, 2018, 49.

Kell LT, Sharma R, Kitakado T *et al.* Validation of stock assessment methods: is it me or my model talking? *ICES J Mar Sci* 2021;**78**:2244–55. https://doi.org/10.1093/icesjms/fsab104

Legault CM, Restrepo V. A flexible forward age-structured assessment program. *Collect Vol Sci Pap ICCAT* 1998;**49**:246–53.

Legault CM. Rose vs. rho: a comparison of two approaches to address retrospective patterns in stock assessments. *ICES J Mar Sci* 2020;**77**:3106–030. https://doi.org/10.1093/icesjms/fsaa184

Levins R. The strategy of model building in population biology. *Am Sci* 1966;**54**:421–31.

MAFMC. 2019. Mid-Atlantic Fishery Management Council Scientific and Statistical Committee OFL CV Guidance Document. https://www.mafmc.org/s/Final_Revised-OFL-CV-guidance-document_06_19_20.pdf

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

Millar CP, Jardim E, Scott F *et al.* Model averaging to streamline the stock assessment process. *ICES J Mar Sci* 2015;**72**:93–8. https://doi.org/10.1093/icesjms/fsu043

Mohn R. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. *ICES J Mar Sci* 1999;**56**:473–88. https://doi.org/10.1006/jmsc.1999.0481

Morgan MG, Henrion M. *Uncertainty*. New York, NY: Cambridge University Press, 1990, 332.

Nguefack-Tsague G. On optimal weighting scheme in model averaging. *Am J Appl Maths Statist* 2014;**2**:150–6. https://doi.org/10.12691/ajams-2-3-9

Nielsen A, Berg C. Estimation of time-varying selectivity in stock assessments using state-space models. *Fish Res* 2014;**158**:96–101. https://doi.org/10.1016/j.fishres.2014.01.014

Patterson K *et al.* Estimating uncertainty in fish stock assessment and forecasting. *Fish Fish* 2001;**2**:125–57. https://doi.org/10.1046/j.1467-2960.2001.00042.x

Pennington M, Vølstad JH. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from trawl surveys. *Biometrics* 1994;**50**:725–32. https://doi.org/10.2307/2532786

Peterman R. Possible solutions to some challenges facing fisheries scientists and managers. *ICES J Mar Sci* 2004;**61**:1331–43. https://doi.org/10.1016/j.icesjms.2004.08.017

Privitera-Johnson K, Punt AE. A review of approaches to quantifying uncertainty in fisheries stock assessments. *Fish Res* 2020;**226**:105503. https://doi.org/10.1016/j.fishres.2020.105503

Punt AE, Butterworth DS, de Moor CL *et al.* 2016. Management strategy evaluation: best practices. *Fish and Fisheries* **17**:303–334.

Punt AE, Donovan GP. Developing management procedures that are robust to uncertainty: lessons from the International Whaling Commission. *ICES J Mar Sci* 2007;**64**:603–12. https://doi.org/10.1093/icesjms/fsm035

Punt AE, Hilborn R. Fisheries stock assessment and decision analysis: the Bayesian approach. *Rev Fish Biol Fish* 1997;**7**:35–63. https://doi.org/10.1023/A:1018419207494

Ralston S, Punt AE, Hamel OS *et al.* A meta-analytic approach to quantifying scientific uncertainty in stock assessments. *Fish Bull (US)* 2011;**109**:217–31.

Restrepo VR, Hoenig JM, Powers JE *et al.* A simple simulation approach to risk and cost analysis, with applications to swordfish and cod fisheries. *Fish Bull* 1992;**90**:736–48.

Schnute JT, Richards LJ. Use and abuse of fishery models. *Can J Fish Aquat Sci* 2001;**58**:10–7. https://doi.org/10.1139/f00-150

Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978;**6**:461–4. https://www.jstor.org/stable/2958889

Scott F, Jardim E, Millar CP *et al.* An applied framework for incorporating multiple sources of uncertainty in fisheries stock assessments. *PLoS One* 2016;**11**:e0154922. https://doi.org/10.1371/journal.pone.0154922

Simmonds EG, Adjei KP, Andersen CW *et al.* Insights into the quantification and reporting of model-related uncertainty across different disciplines. *iScience* 2022;**25**:1–16. https://doi.org/10.1016/j.isci.2022.105512

Smith SJ, Hunt JJ, Rivard D (eds). Risk evaluation and biological reference points for fisheries management. *Can Spec Publ Fish Aquat Sci* 1993;442.

Starr P, Annala JH, Hilborn R. Contested stock assessment: two case studies. *Can J Fish Aquat Sci* 1998;**55**:529–37. https://doi.org/10.1139/f97-230

Stewart IJ, Hicks AC. Interannual stability from ensemble modelling. *Can J Fish Aquat Sci* 2018;**75**:2109–13. https://doi.org/10.1139/cjfas-2018-0238

Stewart IJ, Martell SJD. Assessment of the Pacific halibut stock at the end of 2013. *IPHC Rep Ass Res Acti* 2014;**2013**:169–96.

Stewart IJ, Martell SJD. Reconciling stock assessment paradigms to better inform fisheries management. *ICES J Mar Sci* 2015;**72**:2187–96. https://doi.org/10.1093/icesjms/fsv061

Stock BC, Miller TJ. The Woods Hole Assessment Model (WHAM): a general state-space assessment framework that incorporates time- and age-varying processes via random effects and links to environmental covariates. *Fish Res* 2021;**240**:105967. https://doi.org/10.1016/j.fishres.2021.105967

WCPFC. Scientific Committee. Eighth regular session, Busan, Korea, 7-15 August 2012 : Summary report. Kolonia, Pohnpei: Western and Central Pacific Fisheries Commission, 2012, 192, Available at: https://www.wcpfc.int/meetings/8th-regular-session-scientific-committee.

WCPFC. Scientific Committee. Seventh regular session, Pohnpei, Federated States of Micronesia, 9-17 August 2011: Summary report. Kolonia, Pohnpei: Western and Central Pacific Fisheries Commission, 2011, 203, Available at: https://www.wcpfc.int/meetings/7th-regular-session-scientific-committee.

Weisberg M. *Simulation and Similarity: Using Models to Understand the World*. New York, NY: Oxford University Press, 2013, 190.

Wolpert D, Macready W. No free lunch theorems for optimization. *IEEE Trans Evol Comp* 1997;**1**:67–82.

Handling Editor: Ruben Roa-Ureta