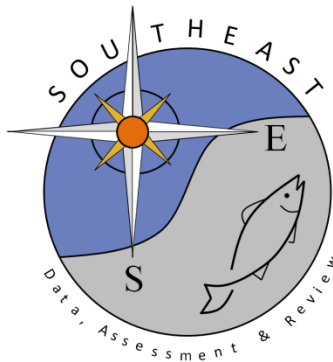


Retrospective forecasting - evaluating performance of stock projections for New England groundfish stocks

Elizabeth N. Brooks and Christopher M. Legault

SEDAR90-RD-43



This information is distributed solely for the purpose of pre-dissemination peer review. It does not represent and should not be construed to represent any agency determination or policy.

Retrospective forecasting — evaluating performance of stock projections for New England groundfish stocks

Elizabeth N. Brooks and Christopher M. Legault

Abstract: Projections are used to explore scenarios for catch advice and rebuilding and are an important tool for sustainably managing fisheries. We tested each projection specification for 12 groundfish stocks in the Northwest Atlantic to identify sources of bias and evaluate techniques for reducing bias. Projections were made from assessments using virtual population analysis (VPA) with 1–7 years of recent data removed from the full time series and were then compared with results from a VPA assessment on the full time series of data. The main source of bias in projections was the assessment model estimates of the numbers at age in the terminal model year + 1 ($N_{a,T+1}$). Recruitment was responsible for more bias in projections beyond 3 years, when population numbers begin to be dominated by cohorts that were statistically generated. Retrospective analysis was performed and several adjustment factors to reduce bias were tested. Even after adjusting for bias, the remaining bias in projections was non-negligible. The direction of bias generally resulted in projected spawning stock biomass (SSB) and catch being overestimated, and the bias in catch was nearly always larger than in SSB. Scientists need to clearly communicate the direction and magnitude of this bias, managers need to consider this additional uncertainty when specifying future catch limits, and both scientists and managers need to develop more robust control rules so that objectives are achieved.

Résumé : Les projections sont utilisées pour explorer des scénarios de recommandation de prises et de reconstitution de stock et constituent un important outil de gestion durable des pêches. Nous avons évalué les critères de projection pour 12 stocks de poissons de fond du nord-ouest de l'Atlantique afin de cerner les sources de biais et d'évaluer des techniques de réduction du biais. Les projections ont été faites à partir d'évaluations par analyse virtuelle des populations (AVP) de données récentes pour d'une à sept années, extraites de séries chronologiques complètes, et ont été comparées aux résultats d'une évaluation par AVP sur les séries chronologiques complètes. Les estimations tirées du modèle d'évaluation du nombre selon l'âge durant la dernière année + 1 modélisée ($N_{a,T+1}$) constituaient la principale source de biais dans les projections. La part du biais dans les projections attribuable au recrutement augmentait après trois ans, quand les chiffres sur les populations commençaient à être dominés par des cohortes générées statistiquement. Une analyse rétrospective a été réalisée et plusieurs facteurs d'ajustement servant à réduire le biais ont été évalués. Même après ajustement pour le biais, il restait un biais non négligeable dans les projections. La direction du biais se traduisait généralement par une surestimation de la biomasse du stock reproducteur (BSR) et des prises projetées, et le biais des prises était presque toujours plus grand que celui de la BSR. Les scientifiques doivent communiquer clairement la direction et la magnitude de ce biais, les aménagés doivent tenir compte de cette incertitude supplémentaire dans l'établissement des limites de prises futures et tant les scientifiques que les aménagés doivent élaborer des règles de contrôle plus robustes pour que les objectifs soient atteints. [Traduit par la Rédaction]

Introduction

The provision of sound catch advice is an important component to managing sustainable fisheries and is contingent upon an accurate estimate of population size from a stock assessment. Catch advice is derived by forecasting the fish population, as estimated in the assessment, several years into the future. Forecasts require assumptions about biological parameters and how the fishery will operate. In the short term, biological parameters and fishery selectivity are usually assumed to be similar to recent observed values, and for the fishery operation, either a fishing mortality rate, a catch amount, or some combination is specified for each year in the forecast. From the results of these forecasts, managers compare the impacts of different fishery scenarios on future population abundance, often trying to balance biological and socioeconomic objectives.

Risks to the fish stock and the fishery should reflect both the probability of an event and the magnitude of the consequence. To allow for an accurate evaluation of risk, forecasts should account

for uncertainties in the estimated parameters and reference points, as well as uncertainties inherent in forecasting (accuracy of biological and fishery assumptions). Risk could also consider uncertainty in the model structure (Patterson et al. 2001). In the best case scenario, forecasts are expected to produce unbiased distributions of future stock abundance that are associated with different harvest options (e.g., catch). From these distributions, probabilities of overfishing or of the stock being overfished can be calculated, which allows managers some flexibility in specifying their risk tolerance (Caddy and McGarvey 1996; Prager et al. 2003; Shertzer et al. 2008). Control rules for determining catch in federally managed US fisheries cannot exceed a 50% probability of overfishing (Code of Federal Regulations, 50 CFR Ch. VI:600.310, 2012).

While assessment models vary in their degree of complexity and the types of uncertainty that are estimated, reliability of results depends on the accuracy of model assumptions relative to the fishery and the data (Yin and Sampson 2004; Wang et al. 2009; Maunder and Piner 2015). For example, it is usually assumed that

Received 23 March 2015. Accepted 13 November 2015.

Paper handled by Associate Editor Michael Wilberg.

E.N. Brooks and C.M. Legault. Northeast Fisheries Science Center, Population Dynamics Branch, 166 Water Street, Woods Hole, MA 02543, USA.

Corresponding author: Elizabeth N. Brooks (email: Liz.Brooks@noaa.gov).

catches are reported accurately or that bias is minimal (Patterson 1998; Sampson 2011), mortality (M) is correctly specified (Mertz and Myers 1997; Clark 1999; Deroba and Schueller 2013), biological sampling for length, weight, and maturity are representative (Beverton and Holt 1957), the functional form of selectivity is appropriate (Goodyear 1996; Hulson and Hanselman 2014; Punt et al. 2014), and assumed error distributions are appropriate (Maunder 2011; Legault 2014). The consequences of model misspecification include biased parameter estimates, underestimated variance, and mischaracterization of stock status (Thompson 1994; Deroba and Schueller 2013), all of which impact management advice.

While there is no expectation that stock assessment model estimates will be 100% accurate, there is an expectation that deviations or residuals between the unknown “truth” and the model estimate will not display persistent trends of over- or underestimation. Furthermore, when the assessment is updated with additional years of data, values for previously estimated parameters are expected to vary, but not in any systematic way. When updated parameter estimates have a systematic bias compared with earlier estimates, it suggests that something may be misspecified in the model. This systematic bias is commonly referred to as retrospective bias (Sinclair et al. 1991).

Stock assessment scientists have been aware for some time of the problem of persistent trends when model estimates are updated (Rivard and Foy 1987; Sinclair et al. 1991; Mohn 1993, 1999; Evans 1996). As these previous studies have noted, the most common trend in updated estimates suggests that earlier estimates of spawning stock biomass (SSB) were overestimated and earlier estimates of fishing mortality (F) were underestimated. Many working groups have focused on, and much research has been devoted to, exploring possible mechanisms (e.g., ICES 2002; Legault 2009). Sinclair et al. (1991) describe a combination of case studies that explored alternative model configuration of the ADAPT virtual population analysis (VPA) (Gavaris 1988), as well as simulations with known misspecification, to characterize a range of possible retrospective causes. Mohn (1993) compared parameter estimates and their variance via bootstrapping, while Mohn (1999) proposed a metric to measure the retrospective bias, known as Mohn's ρ . ICES (2002) examined different data-related causes of retrospective patterns and recommended simulation testing using known causes. ICES (2007) conducted such simulation testing and found that the cause of a retrospective pattern could not be identified based on any of the explored diagnostics. This result was recently confirmed by Hurtado-Ferro et al. (2015) using a more extensive factorial design. Preliminary management strategy evaluations demonstrated that fixing the retrospective pattern, even using the wrong fix relative to the true cause, produced better performing management advice than ignoring the retrospective pattern (ICES 2008). The determination that corrections performed better are almost exclusively based on evaluating impacts on forecasted biomass.

The implications of projecting from models with retrospective patterns remains a topic of major concern in many scientific and management arenas because of its impact on catch advice (Deroba 2014). Quotas that appear to meet rebuilding requirements and to avoid overfishing may be specified based on a stock assessment, yet when that assessment is updated, the new estimates may suggest that those quotas were too high. As a result, future quotas might need to be severely reduced to correct for the previous quotas that were unintentionally set too high. When this type of quota and stock status revision is frequent, it degrades confidence that stakeholders have in assessment results — and in some cases, in the assessment models and data. These types of revisions also reduce the ability of managers to adequately manage risk, because they indicate a source of uncertainty that has not been accounted for.

This study systematically examined the performance of stock projections from VPA assessments for 12 groundfish stocks in the

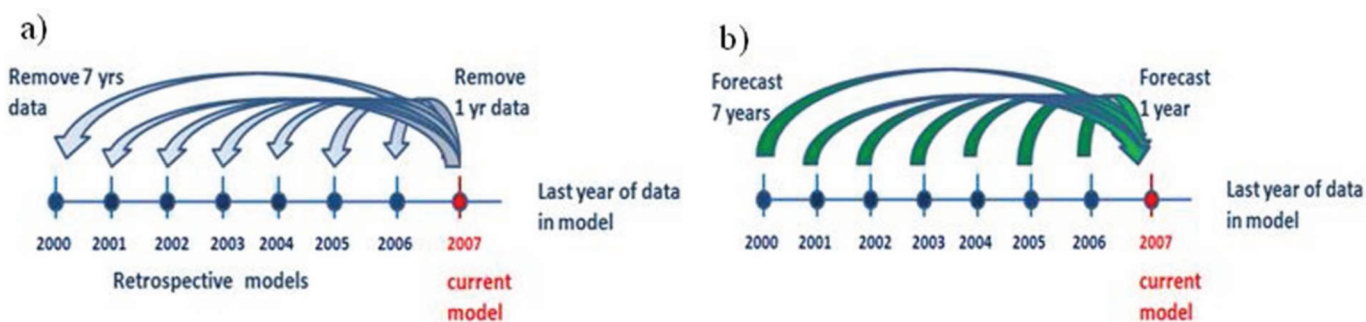
Northwest Atlantic, some of which have historically displayed retrospective patterns. Our goal was to characterize the retrospective patterns and provide robust advice for how to address them. First, we evaluated bias in forecasted catch and SSB trajectories. We then examined possible sources of bias by sequentially testing each assumption in the forecast interval to determine if the status quo assumptions are acceptable or could be improved. Performance of projections was also evaluated with respect to the length of the projection horizon in an attempt to inform on the number of years where catch advice may be reliable. Finally, we explored techniques to adjust for bias prior to making stock forecasts based on measures of retrospective bias in the stock assessments. Our analysis used truncated assessment data (1 or more years of data were removed from the full time series) and made projections through the end of the full time series so that we could compare projection results with the results when performing an assessment on the full dataset. We refer to this approach as “retrospective forecasting.”

Methods

The assessments of 12 Northwest Atlantic groundfish stocks conducted in 2008 (NEFSC 2008) were used to explore the performance of stock forecasts. All 12 stocks were assessed using VPA (Parrack 1986; Gavaris 1988; Conser and Powers 1989; NMFS Toolbox VPA v3.4; <http://nft.nefsc.noaa.gov/>). VPA is an age-structured assessment model that assumes catches at age are known without error. The solution proceeds backwards along cohorts (i.e., from the end of the time series to the beginning) and age-specific indices are used for tuning. There is no linkage between spawners and recruits in a given year; all recruits are estimated as part of the backward cohort reconstruction. Values for natural mortality at age (M_a) are fixed at an assumed rate (0.2 for all stocks in this study). Estimates of weights at age (W_a) and maturity at age (Mat_a) are fixed at values estimated from fishery catch data and research surveys and are used in calculations after the VPA solution. VPA estimates a catchability constant for each index in the model (q_i), as well as $N_{a,y}$ corresponding to 1 January for years $y = 1, 2, \dots, T, T+1$, and fishing mortality at age ($F_{a,y}$) for $y = 1, 2, \dots, T$. From these $N_{a,y}$ and $F_{a,y}$, SSB _{y} is calculated for $y = 1, 2, \dots, T$, as $SSB_y = \sum_{a=1}^A N_{a,y} W_{a,y} Mat_{a,y} \exp[(-M_{a,y} - F_{a,y})\delta]$, where a is age, A is the plus-group (containing fish aged A and older), and δ is the fraction of the year that elapses prior to spawning, implicitly assuming that mortality is evenly distributed throughout the year. Selectivity at age in a given year can be calculated by dividing $F_{a,y}$ by $\max(F_{a,y})$ in each year.

All assessments had data for catch and indices of abundance at age through the year 2007; the first year of data varied by stock and ranged from the mid-1960s to the early 1980s. Although more recent assessments exist for these stocks, the 2008 version was chosen for two reasons. First, a new research survey vessel began conducting the primary surveys used in the analyses in 2009. While an extensive calibration experiment was conducted and calibration coefficients were estimated (Miller et al. 2010), we did not want the calibration of surveys to be a misleading factor when interpreting projection results. Second, the assessment model for several of these stocks has changed — in some cases, a statistical catch at age model (ASAP; Legault and Restrepo 1999) is now used, while in other cases a model-based assessment has been completely abandoned because of severe conflict between trends implied by catch and by indices (Georges Bank yellowtail flounder (*Limanda ferruginea*) and Gulf of Maine winter flounder (*Pseudopleuronectes americanus*)). These results should be considered illustrative rather than definitive for the stocks examined as we modified the original configuration of some stock-specific models to facilitate this study (e.g., short survey indices that did not have data from at least 1990 onwards were dropped to allow all desired retrospective analyses to be conducted).

Fig. 1. Illustration of “retrospective forecasting”. (a) Retrospective models are created by removing 1, 2, ..., 7 years of data from the full time series (current model). (b) Forecasts are then made from each retrospective model to the end of year 2007.



Groundfish stock examples

Assessed stocks in the Northeast US are typically managed in three geographic areas delineated by the Gulf of Maine, Georges Bank, and southern New England. Although the exact stock boundaries vary slightly by stock, the delineation is generally defined based on a combination of species-specific differences in biological parameters and morphology, historical fishing grounds, and oceanographic features that might inhibit movement across boundaries (shelf breaks, deep channels, circulation patterns). Within these three geographic regions, the VPA-assessed stocks for this analysis include two stocks each of cod (*Gadus morhua*) and haddock (*Melanogrammus aeglefinus*) and three stocks each of yellowtail flounder and winter flounder. Two additional VPA-assessed stocks that we examined are American plaice (*Hippoglossoides platessoides*) and witch flounder (*Glyptocephalus cynoglossus*). These two stocks are not subdivided by region, although both are primarily found in the Gulf of Maine. While these 12 stocks are found in the same general region, share the same research surveys, and are exploited by similar fisheries, not all of these stocks have retrospective patterns. We believe that these similarities amount to holding some factors “constant” in our study so that the performance of projections, and sources of bias, can be isolated.

Retrospective forecasting

To evaluate the accuracy of projections, one needs to have a defined “true” value with which to compare. A common approach that allows accuracy to be measured is to simulate data with known characteristics and then to test models that ignore or mis-specify some aspect of that data. However, simulated data are rarely as noisy as real data, and our single observation of reality is a far cry from factorially designed experiments. Our approach was to use the result of the 2008 VPAs (with data through 2007, modified from actual assessments as noted above) for each stock as the “true” value for the abundance at age (N_a), SSB, and F on each stock. We then created seven “retrospective” assessment models from each original VPA by removing 1, 2, ..., 7 years of data from the end of the time series (hereinafter “retro models”). The model configuration and assumptions for all retro models are identical to the configuration of the 2008 VPAs. Next, for each of these retro models (with data through 2000, 2001, ..., 2006), the VPA was conducted and a nonparametric bootstrap was implemented as a means of characterizing parameter distributions and parameter uncertainty. The uncertainty measured by the bootstrap will likely be an underestimate of true uncertainty, as it is conditional on the assumed model structure and only reflects variability in survey indices. For a given assessment, a bootstrap dataset consisted of the original catch at age (unmodified) and bootstrapped indices, which were generated by normalizing residuals from the original indices, resampling with replacement from the set of all standardized index residuals, rescaling residuals by index-specific standard deviations, and then adding the scaled residuals to the

original index values. In the present analysis, we performed 1000 bootstraps for all seven retro models for each stock.

Performing the VPA on the bootstrapped datasets produced a distribution of estimated assessment parameters through the terminal year (T) of data as well as an estimate of numbers at age at the start of year $T + 1$ ($N_{a,T+1}$). These estimates of $N_{a,T+1}$ served as the starting point in forecasts, which were made through the end of year 2007 (Fig. 1). For example, a VPA and bootstrapping is performed on the retro model with data through year 2000, producing an estimate of N_a at the start of year 2001; from those $N_{a,2001}$, a 7-year forecast is made through the end of year 2007.

The basic specifications needed to make an age-structured stock forecast are (i) starting numbers at age; (ii) weights at age; (iii) maturity at age; (iv) fishery selectivity at age; and (v) a function that predicts recruitment. A natural mortality rate must also be specified, and for this exercise, the constant value of 0.2 assumed in each stock assessment was used. In addition, one needs to specify either a fishing mortality rate or a catch amount that is implemented in each projection year. We implemented an approach similar to that taken for New England groundfish assessments when specifying items (ii)–(iv); namely, a mean of recent years (we used 5 years) was calculated for weights, maturity, and fishery selectivity, and those age-specific values were assumed to be constant during the projection horizon.

With respect to projecting recruitment, a previous study examined estimability of stock–recruit relationships for these stocks and concluded that they were not identifiable for reasons that included high variability in recruitment, low contrast in spawning biomass, and boundary solutions for some of the stock–recruit parameters (NEFSC 2008). Therefore, instead of predicting future recruitments from a stock–recruit curve, an alternative approach was implemented. This simple alternative involves sampling from the cumulative distribution function (cdf) of recruitment estimates from the entire time series for each retro model, excluding the most recent 2 years. We chose to exclude the final two recruitment estimates from the cdf either because of high uncertainty or because the estimate was already derived from earlier recruitments (e.g., the final estimate in year $T + 1$ was sometimes specified to be the geometric mean of the most recent 10 years). Given the number of recruitment estimates, the median is rather robust, and exclusion of the two most recent estimates did not have much influence; nevertheless, we dropped those two values for the technical reasons we identified above. The cdf is sampled as

$$(1) \quad R = R_S + (T - 1)(R_{S+1} - R_S) \left(U - \frac{S - 1}{T - 1} \right)$$

where R_S denotes the S element in the ordered set of T recruitment values, U is a uniform random number on the interval (0,1), and S is selected as the largest counter with $U < S/(T - 1)$. Effectively, a random number is drawn that allows linear interpolation

Table 1. Description of cases examined and summary of the main findings.

Case	Description	Objective	Result
1. Null case	Project with all true values	Demonstrate whether approach works	Projected values from all retro models match full VPA — approach works (Figs. 4a–4b)
2. Status quo case	Project with true F , use current protocol for all other projection values	Characterize performance of current approach	Very large bias for many stocks; negligible bias for few stocks (Tables 2, 3; Figs. 4c–4f)
3. Bootstrap bias correction case	Adjust $N_{a,T+1}$ with correction for difference between bootstrap mean and VPA point estimate	Determine if bootstrap bias is causing large bias in projections	Bias from the bootstraps is negligible (Figs. 3, 6)
4–8. One true cases	Fix one of projection factors at true value ($N_{a,T+1}$, recruitment, age-specific maturity, selectivity, or weights at age)	Determine which factor causes most bias in projections	$N_{a,T+1}$ caused most bias in years 1–3 or 4; recruitment generally responsible for most bias in projections ≥ 4 years (Tables 4, 5)
9–12. Two true cases	Fix true $N_{a,T+1}$ and fix one of remaining factors at true value	Determine which factor causes most bias, if initial conditions ($N_{a,T+1}$) could be specified without bias	Weights at age generally caused most bias in years 1–3 or 4, recruitment responsible for more bias in projections ≥ 4 years
13–19. NAA bias adjustment cases	Adjust $N_{a,T+1}$ by measure of age-specific bias (ρ_{age}) calculated from 1, 2, ..., 7 years of recent model estimates	Determine if past measures of bias at age can be used to correct for current bias; evaluate how many years to use to produce best accuracy	Bias typically reduced by ρ adjustment; ρ based on 5–7 years was best adjustment
20–26. SSB bias adjustment cases	Adjust $N_{a,T+1}$ by measure of SSB bias (ρ_{SSB}) calculated from 1, 2, ..., 7 years of recent model estimates	Determine if past measures of bias in SSB can be used to correct for current bias; evaluate how many years to use to produce best accuracy	Bias typically reduced by ρ adjustment; ρ based on 5–7 years was best adjustment; ρ_{SSB} generally performed as well or better than ρ_{age} (Tables 2, 3; Figs. 3, 6)

Note: In cases 3–26, the description indicates an additional projection specification, and all other settings for those cases are as for Case 2. “Current” projection protocol is described in the Methods section “Retrospective forecasting”. NAA refers to numbers at age; $N_{a,T+1}$ are the VPA estimates of numbers at age in the terminal model year $T + 1$; SSB is spawning stock biomass; ρ refers to Mohn’s ρ (eq. 6).

between the set of ordered recruitments. Instead of simply resampling with replacement from the time series of recruitment estimates, this approach produces a smoother distribution of recruitment values while still respecting the minimum and maximum from the estimated time series.

Finally, with respect to fishery specifications, all projections were performed using the 2008 VPA estimates of F_y in 2001–2007 (the “true” F). An alternative to using the true F would have been to use the true observed catch in the projections. We explored this option as well. However, when specifying an absolute amount of catch rather than a rate of fishing, there is the potential for a catch amount to exceed the population biomass if abundance is severely underestimated. When this happens, that projection is terminated and that iteration is not included in the summary of results, meaning that the number of complete iterations in each projection is not necessarily the same; moreover, summarizing only the completed iterations gives an optimistic view of performance. For this reason, we present results for the true F projections below. We note, however, that using true catch produced conclusions that were qualitatively the same with respect to direction of bias and factors most responsible for that bias.

The point of using the true fishery specification in forecasts is so that the only variables being tested when comparing projection results with the 2008 VPA results were items (i)–(v). The forward projection calculations when F was specified are given by

$$(2) \quad \begin{aligned} N_{1,y+1} &= \text{eq. 1} \\ N_{a+1,y+1} &= N_{a,y} \exp(-M_{a,y} - F_{a,y}) \\ N_{A,y+1} &= N_{A-1,y} \exp(-M_{A-1,y} - F_{A-1,y}) + N_{A,y} \exp(-M_{A,y} - F_{A,y}) \end{aligned}$$

In eq. 2, y refers to year, and A designates the plus-group.

Projection cases

Twenty-six different projection cases were explored to evaluate accuracy, possible sources of bias, and the potential to correct for bias. These are enumerated and described below and in Table 1. As noted, each projection case is run for each of the 1000 bootstrap iterations generated from each VPA retro model. Bias was determined by calculating the relative error (RE) between projected values from each bootstrap ($\hat{\theta}_{b,y}^*$) and the true value from the full VPA (θ_y) as

$$(3) \quad \text{RE}_{b,y} = (\hat{\theta}_{b,y}^* - \theta_y) / \theta_y$$

where the subscript b refers to the bootstrap iteration ($b = 1, 2, \dots, 1000$), y refers to the projection year ($y = 1, 2, \dots, p$), and p depends on the number of years between the retro model and the year 2007 (the final “true” year).

1. “Null” case. This is the proof of concept case to demonstrate that our approach works. Instead of using the $N_{a,T+1}$ as estimated in the retro model VPAs for the projections, the “true” N_a (i.e., the estimated N_a from the 2008 VPA) were input. Also, instead of using a recent 5-year mean for weights at age, maturity at age, and selectivity at age, the age- and year-specific values from the 2008 VPA were used. Finally, the year-specific recruitment estimates from the 2008 VPA were also specified instead of drawing randomly from the empirical recruitment cdf. This case demonstrates that the forward projection equations are consistent with the backward solutions in the VPA estimation equations and that when the true values are projected, one ends up with the same result as the VPA estimates.

2. "Status quo" case. For this scenario, no true values (except the fishery specification, F) were used, and the projection approach described above was implemented. This scenario allows examination of all sources of uncertainty without accounting for retrospective patterns, similar to past practice for New England groundfish stocks.
3. "Bootstrap bias correction" case. The difference between the mean of the bootstrap estimates for a given $N_{a,T+1}(\hat{\theta}_y^*)$ and the original VPA point estimate ($\hat{\theta}_y$) was used to calculate an age-specific bootstrap bias correction factor (BBC):

$$(4) \quad \begin{aligned} \text{Bias} &= \bar{\theta}_y^* - \hat{\theta}_y \\ \text{BBC} &= 1 - [(\bar{\theta}_y^* - \hat{\theta}_y) / \hat{\theta}_y] \end{aligned}$$

In the Projections, the distribution of initial $N_{a,T+1}$ was adjusted for the bootstrap bias as

$$(5) \quad N_{a,T+1,b}^{\text{bc}} = N_{a,T+1,b} \text{BBC}$$

where the subscript b refers to the b th bootstrap estimate, and the superscript bc refers to the bias correction. When bias is negligible, the BBC is near 1.0.

For cases where an age is not estimated in year $T + 1$ (typically plus-groups and the youngest age), no bias correction was made. Also, in four stock-age-retro model combinations, the proportion bias was a value greater than 1 (this happens when the bootstrap mean is $>2 \hat{\theta}$), which would have produced a negative N_a when bias-corrected. In these four cases, we scaled the N_a to an arbitrarily low value ($0.1 \hat{\theta}$).

- 4–8. "One true" cases. For these five scenarios, one true value was specified for the duration of the projection horizon: true $N_{a,T+1}$, true recruitment, true maturity, true selectivity, or true weights at age. These five cases will serve to pinpoint which of the projection specifications is responsible for the most bias. This was evaluated by finding the median relative error (MRE) for all bootstraps across a given factor (eq. 3). The factor with the smallest MRE (when projected at its true value) was deemed to be responsible for the most bias.
- 9–12. "Two true" cases. These four scenarios build upon the results of cases 4–8. In all scenarios, projections were initiated from the true $N_{a,T+1}$, rather than the bootstrap-estimated value, and then one of the remaining four specifications (recruitment, weights at age, selectivity at age, or maturity at age) was sequentially fixed at the true value. These cases allowed determination of the most important specification in the stock projections given perfect starting conditions. We note that for 8 out of 12 stocks, the maturity at age did not vary over time; thus, there was no misspecification in the projections, and for those stocks, maturity was not considered when determining the most important specification.

The next 14 cases explored adjustments to $N_{a,T+1}$ based on a measure of bias observed in previous years. For example, in Fig. 2, a typical retrospective pattern is seen by comparing the terminal year estimates from the full model with the terminal year estimates from each retro model, which has successively 1 year less of data. One can look at this pattern for any model estimate (e.g., SSB_y or F_y or $N_{a,\text{year}}$). A common measure of this pattern is to summarize the mean relative difference between some number (P) of retro models:

$$(6) \quad \rho_P = \frac{1}{P} \sum_{i=1}^P \left(\frac{\hat{\theta}_{T-i}^R - \hat{\theta}_{T-i}}{\hat{\theta}_{T-i}} \right)$$

where T is the terminal year of data in the model, P is the number of retro models, $\hat{\theta}_{T-i}^R$ is the terminal year estimate in the retro model with i years of data removed, and $\hat{\theta}_{T-i}$ is the estimate in year $T - i$ from the full model. This bias adjustment statistic, ρ , is often referred to as Mohn's ρ after Mohn (1999).

Projections are made from each retro model, and to bias-adjust these models by past patterns of bias, it was necessary to perform a retrospective analysis on each retro model. We therefore removed 1, 2, ..., 7 years of data from each retro model and ran the seven new VPA models to estimate ρ values for each retro model. For example, the retro model with data to 2000 was used to create seven secondary retro models (hereinafter "retro2 models") with data through 1993, 1994, ..., 1999, while the retro model with data to 2004 was used to create seven retro2 models with data through 1997, 1998, ..., 2003. There is no prescient reason for only removing ≤ 7 years of data. However, it provides a range of values for the number of retro2 models to use in the ρ calculation and should provide sufficient opportunity to observe the development of a pattern. Seven years also exceeds the average age of entry into the fishery, as well as the age of maturity for all stocks considered. And for technical reasons, removing more than 7 years would take the retro2 models back to 1991, 1989, 1987, etc., for each additional year of data removed, which does not leave enough data for stocks with time series that begin in the early 1980s.

- 13–19. N_a bias adjustment cases. Every age in the initial vector of $N_{a,T+1}$ for the projections was adjusted based on an age-specific ρ pattern ($\rho_{\text{age},P}$) for ages 1, 2, ..., A ; each of these seven cases corresponded to calculating $\rho_{\text{age},P}$ from relative differences for $P = 1, 2, \dots, 7$ retro2 models. For example, seven separate projection cases were made from the year 2000 retro model, and in each case, every age in the starting $N_{a,2001}$ was adjusted by a $\rho_{\text{age},P}$ factor. The same seven projection cases, for each $\rho_{\text{age},P}$, were also made for the retro models with data through 2001, 2002, ..., 2006. The bias adjustment is applied as follows:

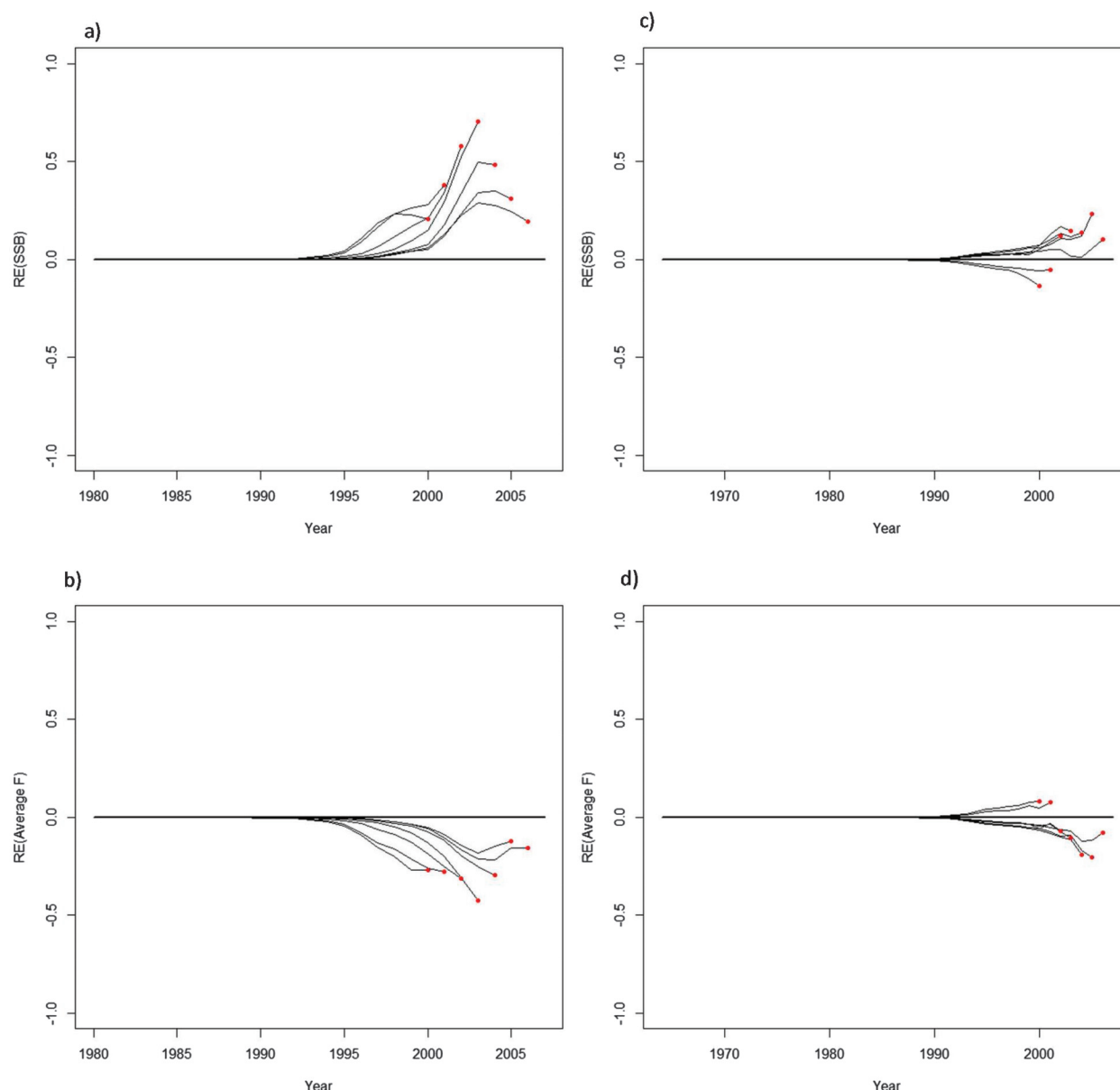
$$(7) \quad N_{a,T+1}^{\rho_{\text{age},P}} = N_{a,T+1} \frac{1}{1 + \rho_{\text{age},P}}$$

We note that the bias adjustment was applied regardless of magnitude or sign of ρ . When the estimated ρ values were positive (indicating a pattern of overestimation), the resulting ρ adjustment reduced the estimated N_a . If ρ was negative (pattern of underestimation), then the resulting ρ adjustment factor was larger than 1 and therefore increased the N_a .

- 20–26. SSB bias adjustment cases. For these last seven cases, we calculated seven mean $\rho_{\text{SSB},P}$ values for each retro model by averaging relative differences from terminal year estimates of SSB in each of the retro2 models. In these cases, the initial $N_{a,T+1}$ are all adjusted by the same value ($\rho_{\text{SSB},P}$) rather than an age-specific value (as in the previous seven cases):

$$(8) \quad N_{a,T+1}^{\rho_{\text{SSB},P}} = N_{a,T+1} \frac{1}{1 + \rho_{\text{SSB},P}}$$

Fig. 2. Retrospective patterns for two stocks with 7 years removed from the full model (Georges Bank cod (a–b) and Georges Bank haddock (c–d)). Each thin line is the relative difference between a retro model and the full virtual population analysis (VPA) (thick horizontal line at 0); the solid circle denotes the estimate in the final year of a retro model. Mohn's ρ was calculated by averaging the relative error (RE) of all seven solid circles. In panel (a), spawning stock biomass (SSB) is consistently overestimated ($\rho_{SSB} = 0.41$), while in panel (b), F is consistently underestimated ($\rho_F = -0.27$). Panels (c) and (d) show very little retrospective pattern, as points are above and below the line where RE = 0 ($\rho_{SSB} = 0.08$, $\rho_F = -0.07$).

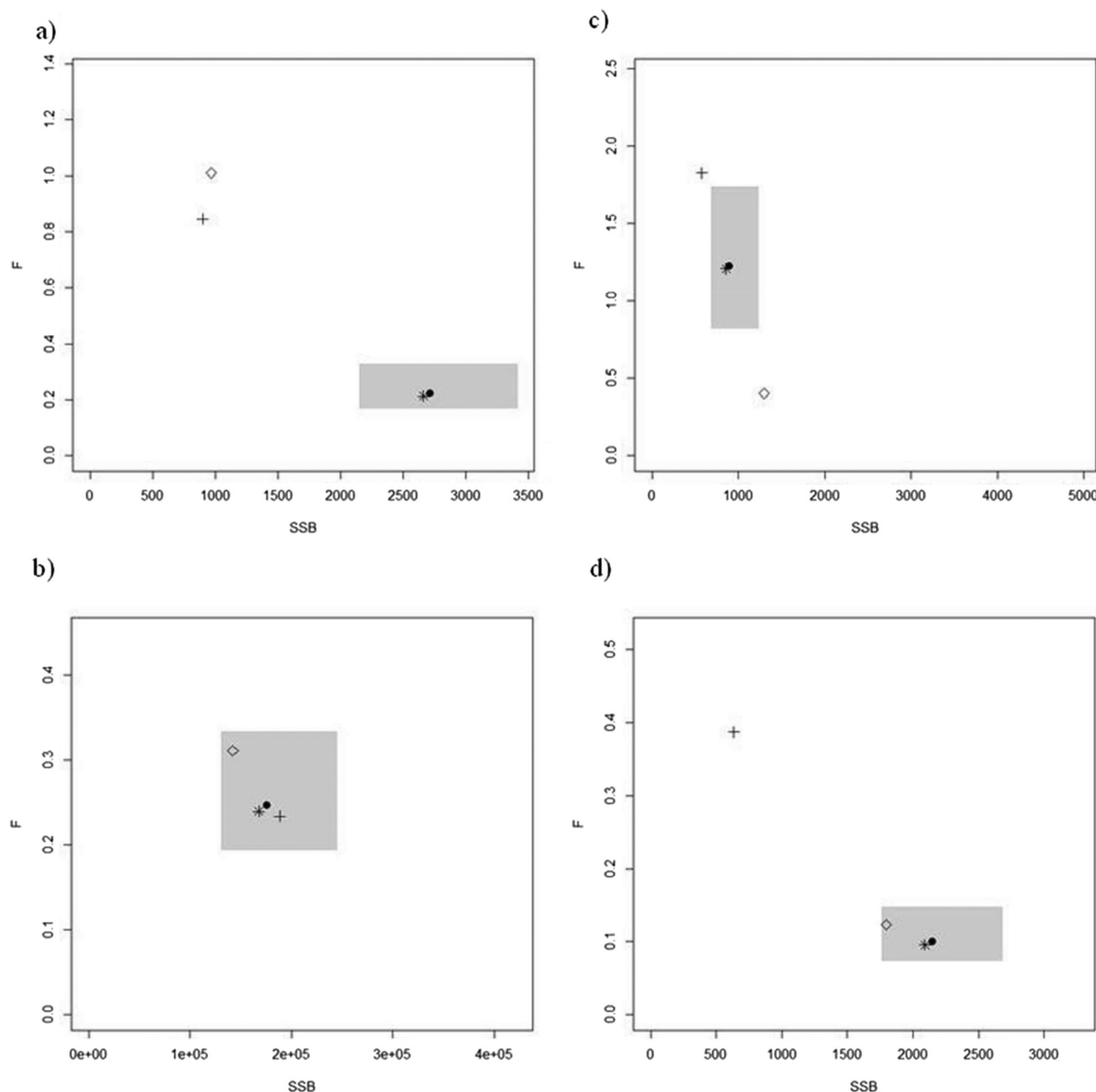


Diagnostics

The ρ statistic described above is one method to characterize model performance. A large ρ value indicates either a consistent direction of bias in subsequent terminal year estimates or several years with very large differences when compared with the full assessment. A small ρ value could either be the result of small relative differences, with or without consistent bias, or it could result from a balancing of large positive and large negative relative differences. The determination of whether a given ρ is large can be a subjective decision. A rule of thumb that has been used to reduce that subjectivity is to plot the terminal year estimate of SSB_T versus F_T along with bootstrap percentiles and compare that with the point estimate when SSB_T and F_T are adjusted by $\rho_{SSB,7}$ and $\rho_{F,7}$, respectively (Fig. 3). According to this

diagnostic, if the ρ -adjusted point estimate falls outside of the bootstrap percentiles on either axis, then the initial conditions for projections, $N_{a,T+1}$, should be adjusted using either $\rho_{SSB,P}$ or $\rho_{age,P}$. The basis for this rule of thumb is that if ρ -adjusted values are outside of the range spanned by the bootstrapped confidence interval, then it indicates the presence of strong additional bias. On the other hand, if the ρ -adjusted values fall within the bootstrapped confidence interval, then it is interpreted to mean that the bootstrap procedure adequately reflects uncertainty associated with the model, and no additional adjustment is needed to the point estimate. It is precisely because the Mohn's ρ and bootstrap uncertainty measure different sources of variability that this rule of thumb is applied.

Fig. 3. Illustrating the “rule of thumb” for when to apply a ρ adjustment. The solid circle is the point estimate from a retro model, the asterisk (*) is the bootstrap bias-corrected point estimate, and the + symbol is the ρ -adjusted point estimate. The “true” value from the full VPA is a diamond. The shaded region spans the bootstrapped 5th and 95th percentiles around the SSB, F point estimates from the retro model. The rule of thumb is to use a ρ adjustment if the + is outside of the shaded region. Panels (a) and (b) demonstrate cases where the rule of thumb works, while panels (c) and (d) demonstrate cases where the rule of thumb does not work. Panel (a) is Gulf of Maine winter flounder retro model 2004; panel (b) is Georges Bank haddock retro model 2005; panel (c) is southern New England yellowtail retro model 2003; and panel (d) is Gulf of Maine winter flounder retro model 2006.



We evaluated this rule of thumb by comparing the VPA point estimate and 90% confidence region (rectangle defined by 5th–95th bootstrap percentiles for SSB_T and F_T) for each retro model with the ρ -adjusted point estimate and determined whether applying this diagnostic produced better results. Based on existing protocol (NEFSC 2008) and results from cases 13–26, the ρ adjustment applied was a 7-year mean of Mohn’s ρ for SSB and F ($\rho_{SSB,7}$ and $\rho_{F,7}$, respectively). The statistic calculated to evaluate whether the original or ρ -adjusted point was closer to the “true” coordinates for SSB_T and F_T was the normalized Euclidean distance (NED):

$$(9) \quad NED = \sqrt{\left(\frac{\widehat{SSB} - SSB}{SSB}\right)^2 + \left(\frac{\hat{F} - F}{F}\right)^2}$$

where $\widehat{SSB}$ and $\hat{F}$ are the estimates and SSB and F are the true values. In this evaluation, we compared NED for the point estimate and the ρ -adjusted point estimate for two cases for each stock-specific retro model: (i) when the ρ -adjusted point estimate was **within** the 90% confidence region, indicating **no** adjustment should be made; (ii) when the ρ -adjusted point estimate was **outside** of the 90% confidence region, indicating an adjustment should be made.

Results

Null case

Projections for the null case confirmed for all stocks that projecting from the retro models with true values produces results that match the 2008 VPA (Table 1). Although the terminal year estimate from each retro model differs from the 2008 VPA, this point is replaced with the true $N_{a,T+1}$ before projecting, and all calculations from that point incorporate the true values, which causes the projections to align with the 2008 results (Figs. 4a–4b). In a few cases, there was a negligible difference between the result from the 2008 VPA and the projection owing to slight inconsistencies between the forward and backward calculations involving the plus-group, but those differences were centered on zero and the vast majority of this bias was $\pm 5\%$. In particular, we note that despite concerns on this very issue for Gulf of Maine cod, as raised in Butterworth and Rademeyer (2008, Suppl. A), the bias was $\pm 0.6\%$ for that stock. The fact that estimated F values were generally large could explain why this bias was negligible.

Status quo case

These projections revealed substantial bias for some, but not all stocks (Figs. 4c–4f). In general, the bias resulted in overestimates of SSB. Because the true F was specified in the projections, we calculated the projected catch that resulted from those F values and compared those with the true catches in the 2008 VPAs. Catch was also overestimated and generally to a greater extent than SSB (Tables 2, 3). Both SSB and catch would be biased because of misspecification in $N_{a,T+1}$ as well as weights at age. SSB would also be biased by misspecification in maturity at age; however, this was time-invariant for some stocks (resulting in no misspecification for projections) or exhibited only slight variation in the age at 50% maturity in the fitted logistic function. Trends in weights at age typically reflected changes in size at age, which carries over into misspecification in selectivity at age. In addition, changes in management also impact realized selectivities. These selectivity vectors were more variable than changes in maturity because there was no functional shape assumed, and this is likely the reason for generally greater bias in catch than in SSB.

On a stock-specific basis, MRE (eq. 3) in SSB between the bootstrapped retro model projections and the full VPA over all projection years ranged from a low of 0.06 (Georges Bank haddock) to a high of 2.6 (witch flounder), while MRE in catch ranged from a low of 0.20 (Georges Bank winter flounder) to a high of 2.0 (witch flounder) (Tables 2, 3).

The initial bias in VPA estimates of $N_{a,T}$ persists in the bootstrap distribution of $N_{a,T+1}$, and that bias tended to increase with the length of the projection horizon, even for stocks that had low initial bias. For example, Georges Bank haddock had very little bias in terminal year estimates (Fig. 2c), but the longer-term projections became very negatively biased (Fig. 4c). The negative bias is the result of an extraordinary year class in 2003 (at the time, the highest ever observed and 45 times larger than median recruitment). Projections made from retro models with data through 2000, 2001, 2002, and 2003 ended up with negative bias because there were no survey observations prior to 2003 that would have predicted this recruitment event, and only one out of 40 previous recruitment estimates was close to the 2003 year class (putting that observation in the 97.5th percentile when drawing from the empirical cdf). Projections made from 2004 and later developed a large positive bias because there was a strong density-dependent response in weights at age. The projection used a recent 5-year mean for weights at age, so that the mean was always larger than the true weights at age, resulting in overestimates of SSB. The slower growth also meant that fish were smaller at age, and hence they recruited to the fishery almost a year later than in the years before 2003; thus, the projected catch had even more positive bias

than projected SSB because of combined misspecification in weights at age and selectivity at age.

Bootstrap bias correction (BBC)

The BBC (eq. 4) was always close to one (i.e., bias was near 0) for abundant cohorts, resulting in not much change from the point estimate (Fig. 3). The directional change was to reduce both SSB and F slightly. These changes caused the projections to perform similarly to the base case.

One true case

MRE was used to rank factors in the projection with respect to bias reduction. When only one true value at a time was projected, we found that the bias in $N_{a,T+1}$, when summarized over all retro model projections, was the most important factor responsible for biased SSB projections in 8 out of 12 cases and was close to the most important factor in two other cases (Table 4a). Specifying true recruitment reduced bias slightly more than specifying true numbers at age for Georges Bank cod, Cape Cod – Gulf of Maine yellowtail flounder, and southern New England yellowtail flounder. For Georges Bank haddock, the status quo approach (no true values) performed best overall, which is a consequence of the very strong decrease in weights at age — projections are better using biased N_a to trade off for the bias in weights at age. For the most part, the same ranking of factors held for the catch bias (Table 4b).

Summarizing “one true” results by the length of the projection leads to a few generalizations, namely that $N_{a,T+1}$ was typically more important for projections up to about 3 or 4 years, at which point recruitment tended to be more important, although there were some exceptions (Table 5a). One explanation for this general result is that short-term projection estimates of SSB are more influenced by $N_{a,T+1}$, whereas longer-term projections will be more susceptible to bias in predicted recruitment owing to the schedule of maturation. Most of the stocks analyzed achieved 50% maturity between the ages of 2 and 3 years, although American plaice and Gulf of Maine winter flounder were close to 4 years of age, and witch flounder was between 5 and 6 years of age at 50% maturity. Conclusions for catch were similar to those for SSB, although there are a few more cases where selectivity at age was an important factor in reducing bias (Table 5b).

Two true case

The second most important factor in projections, after fixing starting values at the true $N_{a,T+1}$, depended on the stock and the length of projection horizon. In many cases, specifying true weights at age was more important for reducing bias in SSB in short-term projections (1–3 years), whereas specifying true recruitment was more important in projections longer than 3 years for the same reason mentioned above. Specifying true maturity often made no difference because a number of the stocks assumed time-invariant maturity at age, but for stocks with time-varying maturity (e.g., witch flounder), it did influence bias in projected SSB.

N_a and SSB bias adjustment

Applying a ρ adjustment to the status quo case tended to reduce bias, and a ρ adjustment calculated from 5–7 years generally performed best. We suspect this is related to the development of the retrospective pattern itself. For stocks with large estimates of retrospective bias, we noted that ρ values took 5 or more years to stabilize, whereas for stocks with smaller retrospective bias, the ρ value stabilized more quickly. Thus, when applying a ρ based on retro models with more years of data removed, it helped the stocks with large retrospective patterns and performed about the same as a ρ adjustment calculated from retro models that removed fewer years of data for stocks with small or negligible retrospective patterns.

For stocks with large retrospective patterns, adjustments to $N_{a,T+1}$ using ρ_{SSB} tended to perform better (i.e., reduce bias in

Fig. 4. Projected 90% confidence polygons for SSB (top) and catch (bottom) for (a–b) case 1 (all true values) for Georges Bank cod; (c–d) a partially “good” case 2 (status quo) for Georges Bank haddock; (e–f) a “bad” case 2 (status quo) for southern New England winter flounder. The solid circle is the point estimate in year T from each VPA retro model, and the solid line joins the point estimate to the median of the projected distribution in year $T + 1$ through year 2007. The lightest shaded polygon is for the retro model with data through 2000, while the darkest shaded polygon is the retro model with data through 2006. [Colour online.]

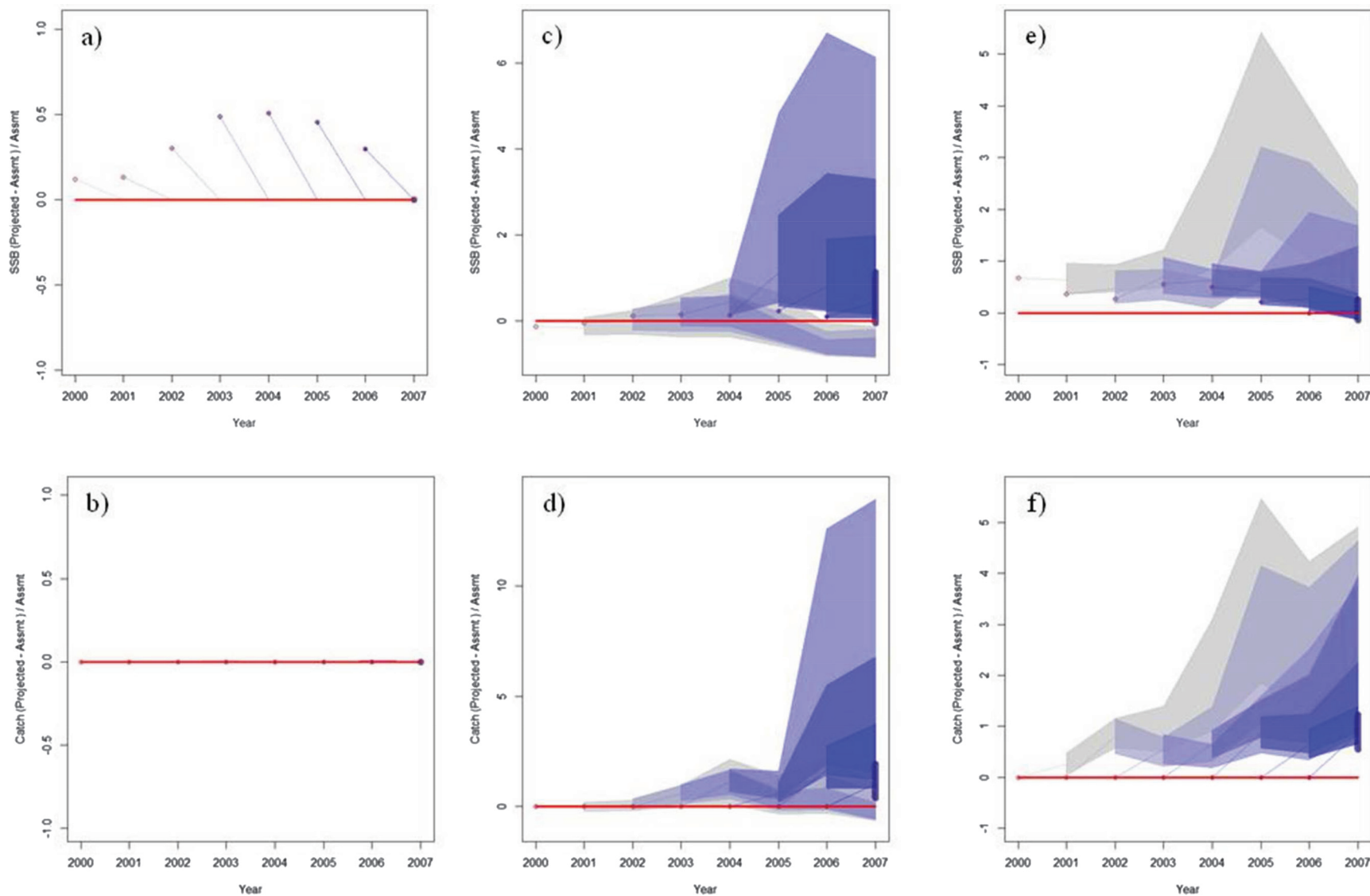


Table 2. Median relative error (eq. 3) for projected spawning stock biomass (SSB) by stock and by length of projection (over all projections in all years or for a specific length of projection (1 year through 7 years)).

Stock	Case	Length of projection							
		All years	1 year	2 years	3 years	4 years	5 years	6 years	7 years
APL	Status quo	0.41	0.53	0.50	0.44	0.41	0.22	0.15	0.01
	$\rho_{SSB,7}$	0.23	0.32	0.30	0.26	0.24	0.08	0.06	-0.04
CCY	Status quo	0.31	0.10	-0.07	0.34	0.68	0.71	0.60	0.45
	$\rho_{SSB,7}$	0.02	-0.19	-0.36	-0.01	0.50	0.67	0.59	0.45
GBC	Status quo	0.63	0.49	0.34	0.51	0.98	1.15	1.37	1.61
	$\rho_{SSB,7}$	0.38	0.18	0.06	0.28	0.82	1.08	1.34	1.59
GBH	Status quo	0.06	0.32	0.37	0.17	-0.17	-0.64	-0.72	-0.75
	$\rho_{SSB,7}$	0.04	0.37	0.43	0.14	-0.20	-0.65	-0.73	-0.76
GBW	Status quo	0.11	0.00	-0.06	0.04	0.20	0.38	0.49	0.59
	$\rho_{SSB,7}$	-0.07	-0.23	-0.30	-0.17	0.06	0.28	0.43	0.56
GBY	Status quo	0.78	2.11	1.23	0.86	0.50	0.20	0.03	-0.20
	$\rho_{SSB,7}$	0.07	0.03	-0.02	0.13	0.19	0.10	0.00	-0.21
GMC	Status quo	0.58	0.41	0.62	0.77	1.26	0.63	0.33	-0.03
	$\rho_{SSB,7}$	0.40	0.21	0.39	0.54	1.05	0.53	0.25	-0.07
GMH	Status quo	0.67	0.39	0.54	0.82	0.99	1.14	0.99	0.54
	$\rho_{SSB,7}$	0.90	0.49	0.75	1.11	1.30	1.43	1.24	0.70
GMW	Status quo	1.08	2.43	1.80	1.38	0.91	0.60	0.44	0.30
	$\rho_{SSB,7}$	-0.11	0.10	-0.11	-0.23	-0.24	0.00	0.16	0.19
SNW	Status quo	0.57	0.47	0.48	0.46	0.91	1.43	1.43	1.21
	$\rho_{SSB,7}$	0.03	-0.08	-0.07	-0.10	0.55	1.29	1.36	1.18
SNY	Status quo	1.10	0.27	0.56	2.10	4.52	5.37	3.82	2.49
	$\rho_{SSB,7}$	0.60	-0.12	-0.01	1.29	4.13	5.23	3.78	2.48
WTF	Status quo	2.60	1.93	2.26	2.78	3.37	3.51	3.56	2.70
	$\rho_{SSB,7}$	1.30	0.84	1.06	1.36	1.77	1.96	2.14	1.71

Note: "Status quo" is described as case 2 in the text, and for " $\rho_{SSB,7}$ " all projection $N_{a,T+1}$ are scaled by $\rho_{SSB,7}$ as in eq. 8. Stock abbreviations are as follows: APL, American plaice; CCY, Cape Cod – Gulf of Maine yellowtail flounder; GBC, Georges Bank cod; GBH, Georges Bank haddock; GBW, Georges Bank winter flounder; GBY, Georges Bank yellowtail flounder; GMC, Gulf of Maine cod; GMH, Gulf of Maine haddock; GMW, Gulf of Maine winter flounder; SNW, southern New England winter flounder; SNY, southern New England yellowtail flounder; WTF, witch flounder.

Table 3. Median relative error for projected catch by stock and by length of projection.

Stock	Case	Length of projection							
		All years	1 year	2 years	3 years	4 years	5 years	6 years	7 years
APL	Status quo	0.66	0.74	0.85	0.81	0.65	0.37	0.33	0.02
	$\rho_{SSB,7}$	0.44	0.49	0.59	0.58	0.44	0.21	0.20	-0.04
CCY	Status quo	0.43	0.21	-0.02	0.40	0.80	0.90	0.95	0.62
	$\rho_{SSB,7}$	0.11	-0.13	-0.33	0.00	0.58	0.85	0.94	0.61
GBC	Status quo	0.68	0.52	0.45	0.57	1.10	1.02	1.08	1.14
	$\rho_{SSB,7}$	0.42	0.19	0.17	0.29	0.86	0.95	1.04	1.12
GBH	Status quo	0.55	0.66	0.92	0.88	0.51	-0.01	-0.13	-0.42
	$\rho_{SSB,7}$	0.56	0.74	1.02	0.82	0.49	-0.05	-0.16	-0.44
GBW	Status quo	0.20	0.08	-0.05	0.06	0.32	0.51	0.78	1.28
	$\rho_{SSB,7}$	0.00	-0.16	-0.29	-0.19	0.13	0.39	0.70	1.24
GBY	Status quo	1.24	2.40	1.63	1.02	1.02	0.47	0.35	-0.12
	$\rho_{SSB,7}$	0.25	0.25	0.26	0.14	0.51	0.32	0.30	-0.13
GMC	Status quo	0.93	0.77	0.71	0.99	0.96	1.19	1.20	1.01
	$\rho_{SSB,7}$	0.73	0.53	0.53	0.75	0.73	1.01	1.06	0.91
GMH	Status quo	1.63	1.11	1.33	1.60	2.18	2.65	3.75	2.62
	$\rho_{SSB,7}$	1.96	1.23	1.53	1.98	2.62	3.22	4.37	3.02
GMW	Status quo	0.86	2.97	1.80	1.07	0.74	0.54	0.46	0.65
	$\rho_{SSB,7}$	-0.07	0.29	-0.11	-0.32	-0.33	-0.14	0.17	0.52
SNW	Status quo	0.99	0.68	0.79	0.98	1.63	2.33	2.46	2.88
	$\rho_{SSB,7}$	0.35	0.05	0.16	0.42	1.23	2.15	2.39	2.83
SNY	Status quo	1.86	0.36	0.47	1.88	4.94	8.95	9.29	8.38
	$\rho_{SSB,7}$	1.15	-0.07	-0.07	0.99	4.28	8.65	9.16	8.35
WTF	Status quo	2.00	1.48	1.59	1.99	2.39	2.77	3.14	2.66
	$\rho_{SSB,7}$	0.90	0.56	0.64	0.88	1.13	1.42	1.77	1.59

Note: Stock abbreviations as in Table 2. "Status quo" is described as case 2 in text, and for " $\rho_{SSB,7}$ " all projection $N_{a,T+1}$ are scaled by $\rho_{SSB,7}$ as in eq. 8.

projected SSB) than adjustments using ρ_{age} . For stocks with small or negligible retrospective patterns, MRE was similar among the status quo, bias-corrected, or ρ -adjusted cases — because the retrospective pattern was small in those cases, the adjustment did

not make much difference. When considering the length of projection horizon, the MRE for ρ_{SSB} -adjusted cases was usually lower than that for the status quo cases for the first few years, whereas quite a few projections of 4 years or more performed very poorly.

Table 4. Median relative error (MRE) of all projections from all retro models by stock for (a) SSB or (b) catch when no true values are used (None) or when one true value is used: true numbers at age (N_a), true recruitment (Recr), true weights at age (W_a), true selectivity at age (Se_{1a}), or true maturity at age (Mat_a).

Stock	None	N_a	Recr	W_a	Se_{1a}	Mat_a
(a) SSB						
APL	0.41	0.06	0.36	0.27	0.52	0.41
CCY	0.31	0.11	-0.01	0.23	0.36	0.31*
GBC	0.63	0.16	0.07	0.46	0.72	0.54
GBH	0.06	0.11	0.45	-0.11	0.11	0.06*
GBW	0.11	-0.04	-0.24	0.22	0.19	0.11*
GBY	0.78	0.00	0.50	0.68	0.96	0.78*
GMC	0.58	0.06	0.49	0.61	0.44	0.46
GMH	0.67	0.20	0.59	0.23	0.92	0.67*
GMW	1.08	-0.06	0.93	1.25	1.38	1.08*
SNW	0.57	-0.01	0.23	0.64	0.61	0.57*
SNY	1.10	0.27	0.21	0.88	1.07	1.10*
WTF	2.60	0.28	2.28	2.39	2.45	2.14
(b) Catch						
APL	0.66	0.23	0.60	0.47	0.44	0.66
CCY	0.43	0.27	0.07	0.32	0.34	0.43*
GBC	0.68	0.36	0.25	0.53	0.61	0.68
GBH	0.55	0.49	1.04	0.23	0.27	0.55*
GBW	0.20	0.02	-0.10	0.35	0.08	0.20*
GBY	1.24	0.07	0.89	1.16	1.21	1.24*
GMC	0.93	0.36	0.81	0.83	0.58	0.94
GMH	1.63	0.86	1.60	0.95	0.79	1.63*
GMW	0.86	-0.03	0.66	1.07	1.33	0.86*
SNW	0.99	0.38	0.55	1.00	0.64	0.99*
SNY	1.86	0.96	0.60	1.60	1.25	1.86*
WTF	2.00	0.07	1.79	1.84	1.91	2.00

Note: The median relative error (MRE) closest to zero is in bold font. For many stocks, maturity was time-invariant, so its assumed value in the projection had no misspecification and the MRE for the true maturity case matches the "None" case. An asterisk (*) in the Mat_a column indicates time-invariant maturity at age. Stock abbreviations are as in Table 2.

Although the various adjustments explored tended to result in some reduction in bias, for the most part, there was still substantial residual bias. As seen in Table 2 (rows labeled $\rho_{SSB,7}$), the remaining bias can be as high as 50%–75% in the early years of a projection; for some stocks, the correction is too large and results in an underestimate of stock size. Comparing the estimated retrospective bias from each retro model with the "true" retrospective bias calculated from the full VPA, it was clear that the pattern of relative bias in the past does not appear to be a good predictor of future relative bias (Fig. 5).

Diagnostic rule of thumb for when to use a ρ adjustment

We evaluated the rule of thumb for when to use a ρ adjustment for all seven retro models for all 12 stocks (84 cases) and summarized performance based on two possible outcomes: (i) cases where the ρ -adjusted point estimate was within the 90% confidence region from the bootstrap, indicating no adjustment should be made; and (ii) cases where the ρ -adjusted point estimate was outside the 90% confidence region, indicating an adjustment should be made. For the first outcome, there were 25 cases where the ρ -adjusted retro VPA point estimate was inside of the 90% confidence region, and in 12 out of 25 cases the unadjusted point estimate was closer to the true value (Fig. 6). Many of these 25 cases were for stocks with small retrospective patterns, so the unadjusted retro model point estimates were not expected to have too much bias. Also, the performance of the ρ adjustment in these 12 cases was often similar because the ρ adjustment value was close to 1.0 (Fig. 6). For the second outcome, the ρ -adjusted retro VPA point estimate was outside of the 90% confidence region 59 out of 84 times, and about 2/3 of those adjusted SSB_T and F_T

Table 5. Factor reducing the most bias in (a) projected SSB or (b) projected catch by stock and by length of projection (1 year to 7 years).

Stock	1 year	2 years	3 years	4 years	5 years	6 years	7 years
(a) Projected SSB							
APL	N	N	N	N	W, N	W	M
CCY	N	N	R	R	R	R	R
GBC	N	N	R	R	R	R	R
GBH	W	W	W	S	R	R	R
GBW	R,S	W, S	N, S	N	R, N	R	R
GBY	N	N	N	N	N, R	N	R
GMC	N	N	N	N	S	N, S	W, M
GMH	W	N	N	N	N	W, N	N
GMW	N	N	N	N	N	R	R
SNW	N	N	N	R	R	R	R
SNY	N	N	R	R	R	R	R
WTF	N	N	N	N	N	N	N
(b) Projected catch							
APL	N	N	N	N	W	W	M
CCY	N	S, W	R	R	R	R	R
GBC	N	N	R	R	R	R	R
GBH	S	S	S	W	S	S, N	S
GBW	N, S	W, R	N, S	S	R	R	R
GBY	N	N	N	N	R	R, W	R, S
GMC	N	N	N	N	N	S	S
GMH	S	S	N	W	N	N	W
GMW	N	N	N	N	R	R	R
SNW	N	N	N	R	R	R	R
SNY	N	N	R	R	R	R	R
WTF	N	N	N	N	N	N	N

Note: When more than one factor is listed, the second factor was close (± 0.02) to the first factor in reducing bias. Abbreviations are as follows: N, numbers at age in ($T + 1$); R, recruitment; W, weights at age; S, selectivity at age; M, maturity at age. Stock abbreviations are as in Table 2.

were closer to the true values (Fig. 6). When the adjusted retro model estimates were outside of the confidence region, the retrospective bias (calculated from seven retro2 models) was typically large (Table 6). Applying the ρ adjustment to the 59 cases outside of the confidence region reduced the NED, on average, by 0.35, compared with a reduction of only 0.02 for the 25 cases where the ρ -adjusted points remained inside the confidence region.

Discussion

Our investigation identified that the largest source of bias in stock projections was due to the VPA estimate of $N_{a,T+1}$. This is consistent with earlier findings by Rivard and Foy (1987). Examining projection bias by the length of projection horizon identified a general pattern where the earliest projected estimates of SSB and catch were most impacted by bias in $N_{a,T+1}$, whereas longer projections (3 or more years) showed increasing sensitivity to misspecification or bias in recruitment and weights at age (for SSB) and recruitment, selectivity, and weights at age (for catch). This makes sense because predicted recruitment in the first few forecast years is typically not mature nor is it fully selected by the fishery. However, 3 or more years out, projected recruitment begins to comprise the bulk of those two quantities. Because SSB and catch in the longer projections are dominated by statistically generated recruitment predictions, rather than from fish that were observed in the assessment time series, they are colloquially referred to as "paper fish", and the uncertainty of those numbers is expected to be greater. Thus, projections beyond the first 3 years that are heavily reliant on these "paper fish" will be highly uncertain and should be treated differently than short-term projections for use in management.

When trying to correct for the bias in $N_{a,T+1}$, we found that an adjustment to initial conditions based on ρ_{SSB} usually performed as well as, or better than, an adjustment calculated from retrospective patterns in numbers at age (eqs. 8 and 7, respectively).

Fig. 5. Two cases comparing historical versus current retrospective bias (American plaice (a) and Gulf of Maine cod (b)). The “true” relative bias, between the full model and each retrospective model, is the solid black line. The value of ρ_{SSB} for retro models with data through years 2000–2006 is calculated by averaging relative differences between each retro model and the corresponding retro2 models that removed 1–7 years of data. For example, the symbols “1” through “7” are stacked vertically at each year; the year on the x axis corresponds to the retro model and the numeric symbols refer to the number of retro2 models used to calculate ρ_{SSB} . In some cases, there is little correspondence between the black line and the ρ_{SSB} values, demonstrating that historical estimates of retrospective bias are not a good predictor of future pattern or magnitude of retrospective bias. [Colour online.]

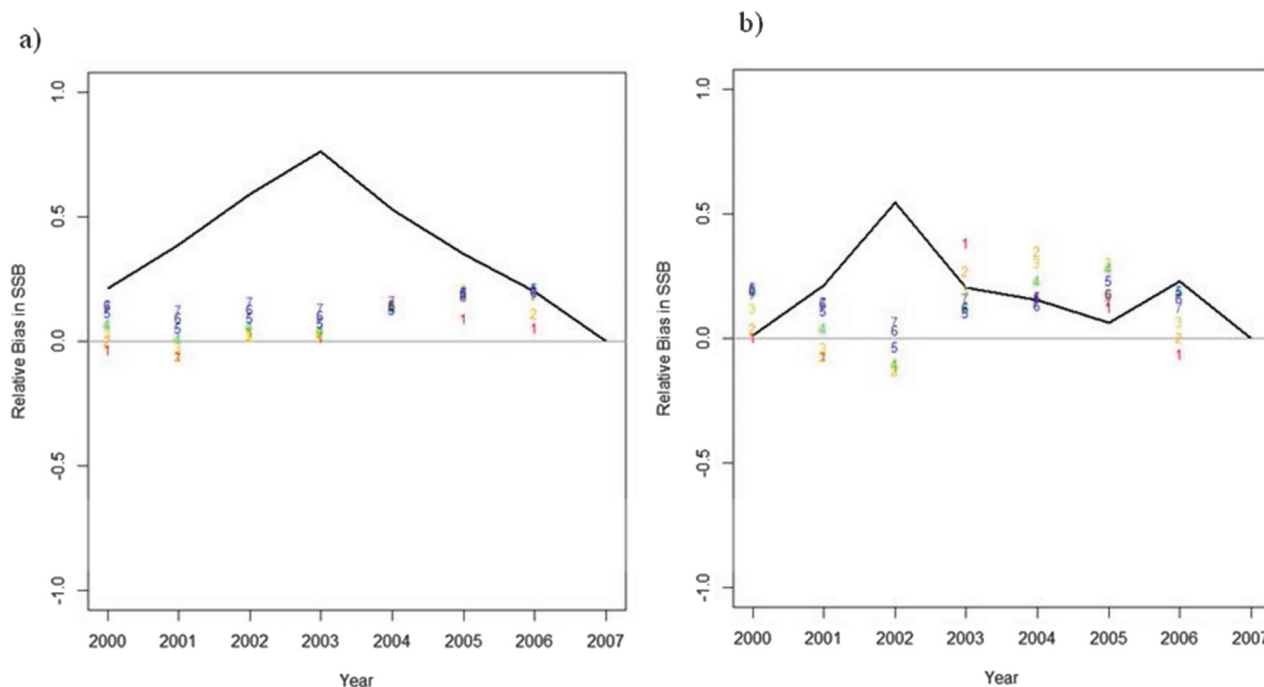


Fig. 6. Bagplots of the point estimate of relative error in SSB and relative error in F in the terminal model year (grey shading with grey dashed line), bootstrap bias-corrected (green shading with green dashed line), and adjusted by $\rho_{SSB,7}$ (blue shading with blue dashed line). The symbols (solid circle, asterisk (*), + symbol) refer to the bivariate median for the point estimate, bootstrap bias-corrected, or ρ -adjusted, respectively. The shaded region is the bivariate equivalent of the interquartile range, while the lines around the shaded regions are the equivalent of whiskers in standard boxplots (Rousseeuw et al. 1999). Panels refer to (a) results for all 84 cases (12 stocks $\times$ 7 retro models); (b) the 59 cases where the ρ -adjusted point was outside of the bootstrapped CI region; (c) the 25 cases where the ρ -adjusted point was inside the bootstrapped CI region. The distributions in panels a–b demonstrate the negative correlation between SSB and F point estimates (SSB is overestimated when F is underestimated). The ρ -adjusted distribution and median shows less bias overall compared with the point estimate or bootstrap bias correcting. Note the difference in both x-axis and y-axis scale in panel c. [Colour online.]

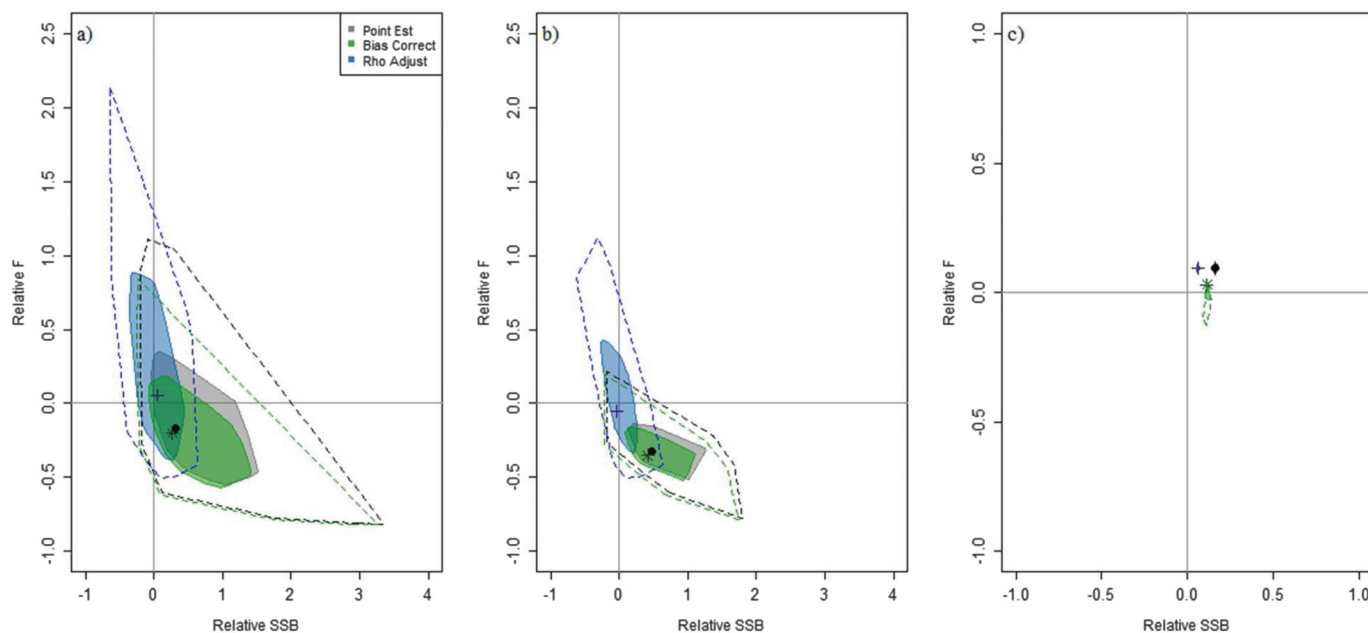


Table 6. Retrospective bias in (a) SSB or (b) mean F , calculated as the 7-year mean of relative differences between each retro model and the seven corresponding retro2 models.

Stock	2006	2005	2004	2003	2002	2001	2000
(a) SSB							
APL	0.19	0.20	0.18	0.15	0.15	0.13	0.15
CCY	0.21	0.26	0.28	0.44	0.60	0.61	0.52
GBC	0.22	0.17	0.20	0.25	0.33	0.33	0.32
GBH	0.02	-0.07	-0.09	-0.08	-0.04	0.04	0.06
GBW	0.13	0.09	0.10	0.22	0.39	0.48	0.54
GBY	2.24	2.21	2.16	2.02	1.23	0.64	0.49
GMC	0.13	0.15	0.15	0.17	0.07	0.15	0.18
GMH	0.05	0.07	0.00	-0.09	-0.16	-0.15	-0.14
GMW	2.38	2.30	2.02	2.06	2.09	2.16	2.23
SNW	0.53	0.56	0.61	0.65	0.66	0.55	0.43
SNY	0.17	0.23	0.35	0.55	0.77	0.71	0.48
WTF	0.64	0.59	0.57	0.58	0.65	0.59	0.58
(b) Mean F							
APL	-0.14	-0.19	-0.15	-0.06	-0.05	0.00	0.20
CCY	-0.08	-0.19	-0.28	-0.40	-0.41	-0.42	-0.40
GBC	-0.46	-0.41	-0.47	-0.49	-0.53	-0.46	-0.40
GBH	0.01	0.06	0.07	0.02	0.03	0.03	0.14
GBW	-0.27	-0.28	-0.32	-0.31	-0.21	-0.04	0.07
GBY	-0.63	-0.65	-0.70	-0.74	-0.65	-0.49	-0.43
GMC	0.18	0.06	-0.01	-0.18	-0.19	-0.33	-0.40
GMH	0.76	0.75	0.72	0.41	0.44	0.27	0.17
GMW	-0.74	-0.75	-0.74	-0.69	-0.65	-0.63	-0.66
SNW	-0.42	-0.44	-0.46	-0.48	-0.47	-0.44	-0.40
SNY	0.34	0.27	0.01	-0.33	-0.38	-0.37	-0.21
WTF	-0.20	-0.19	-0.27	-0.25	-0.23	-0.03	0.16

Note: Column headers indicate the last year of data in each retro model. Table values in bold indicate that ρ -adjusting the terminal year model estimates resulted in the ρ -adjusted point being outside the 90% confidence region, along one or both axes (SSB and (or) F). Stock abbreviations are as in Table 2.

Furthermore, the historical patterns of ρ required a different number of years to stabilize depending on the stock and retro model — in models with little retrospective pattern, ρ stabilized sooner than in models with stronger retrospective patterns. Given the range of years we considered for calculating ρ (1–7 years), we recommend $\rho_{SSB,7}$ as the best correction factor overall. To some extent, the maximum number of years considered was constrained by the data and assessment models we used. However, it was clear that in these analyses the adjustment asymptotes before or around year 7. If used with other assessment models or stocks, it may be useful to verify the number of years used for the calculation of the ρ statistic.

The diagnostic proposed for evaluating when to ρ -adjust was determined to be a useful rule of thumb. While our results suggest that ρ adjustments usually reduce bias and it is better to correct than not to correct, the adjustment does not remove all bias, and for 12% of ρ -adjusted cases, the adjustment was in the wrong direction. This was attributed to the result that past retrospective patterns did not necessarily characterize the magnitude of future retrospective bias, in part because the calculated relative bias in a given year is in fact an estimate that changes when additional years of data are added. Effectively, trying to calculate past performance is like trying to hit a moving target because the basis used in the calculation is continually being updated. Furthermore, whatever is causing the retrospective pattern to occur is almost certainly not a constant factor, leading to the changes in Mohn's ρ over time.

This rule of thumb diagnostic intrinsically has a sharp edge effect (i.e., if a ρ -adjusted point falls on the border of the confidence region, or just inside, the rule would indicate that no adjustment should be made). Similarly, if a smaller or larger confidence region was used (e.g., 80% or 95% instead of 90%, which we examined), then the result of the rule of thumb could poten-

tially vary for any given case. We believe that such “binary” decision points are not adequate for conveying risk. Instead, we recommend that the rule of thumb always be evaluated and presented along with the standard suite of model diagnostics. In cases where the model point estimate of SSB and F are very close to the ρ -adjusted values (i.e., when the retrospective pattern is small), then the question of bias-correcting is not so challenging because the impact on management advice will be minimal (Fig. 6c). However, if the ρ -adjusted values are close to, or on, the border of the 90% confidence region, then it would be important to communicate that uncertainty with respect to potential bias. Analysts could compare the rule of thumb test for several confidence regions to evaluate sensitivity to the percentile used. A further step would be to perform two sets of projections, with and without the ρ adjustment, to characterize the impact on catch advice and SSB trajectories (e.g., NEFMC 2012).

The ρ statistic described above is one method to characterize model performance. Although the ρ statistic cannot identify the cause, a large value does indicate that there is something misspecified in the data, the model, or both. More importantly, it identifies that there is uncertainty in the results beyond what can be characterized by the model. Frequent candidate causes include misspecified catch, misspecified survey catchability, or misspecified natural mortality (M). With respect to catch, misspecification could be due to unreported or underestimated discards, for example due to observer effects, where fishing location, gear configuration, and discarding behavior differ when a fisheries observer is on board. Unreported landings, or catch that is attributed to the wrong stock area (whether by fishermen, dealers, or catch assignment algorithms), could also contribute to misspecified catch. While underreported catch is a global problem (Patterson 1998; Pitcher et al. 2002; Hammond and Trenkel 2005) and is acknowledged to have been problematic locally (NEFMC 1980; Gavaris and Van Eckhaute 1998; King and Sutinen 2010), it would be difficult to characterize the magnitude on an annual basis. Models have been developed that attempt to estimate missing catch or to identify periods where misreporting is indicated, but they typically require that natural mortality is known perfectly, that surveys have reasonable precision, and that the age structure of unreported catch is similar to the reported catch or that the catch at age can be partitioned into age subsets where the age structure of missing catch is similar to that in the reported subset (Patterson 1998; Darby 2005; Bousquet et al. 2010). The methodological hurdles make it difficult to characterize misspecified catch. As a result, fishermen, managers, and scientists rarely discuss the issue without antagonism.

Regarding the remaining factors that are usually implicated when a retrospective pattern is present, unmodeled variability in catchability in a fishery-dependent index of stock abundance (Parma 1993) or in a research survey, or potentially year effects (Mohn 1999), could impact the scaling of population abundance. This possibility has been investigated in modeling frameworks by estimating a random walk in catchability or by splitting indices of abundance at several candidate breakpoints and adopting the break in the year that provided the best reduction in retrospective pattern (Legault 2009). Distributional shifts of the resource could explain the introduction of breakpoints, and changes in fishing practice or technological improvement of commercial fishing gear could lead to trends in catchability of fisheries-dependent indices (Wilberg et al. 2009). For the groundfish stocks explored herein, no mechanisms have been identified to justify introducing breakpoints into fisheries-independent surveys; hence, this approach to fixing retrospective patterns has been interpreted as “aliasing an unknown mechanism” (NEFSC 2008).

Given the difficulty of estimating M , it is commonly assumed to be age- and year-invariant, and in many if not most assessment applications, that constant value is assumed to be known. This is clearly a simplification of reality and therefore an understandable

point for investigating potential misspecification leading to retrospective bias. As with catchability, allowing flexibility in M has been attempted with a random walk, a step-function, or with a "ramp" that estimates a linear increase between two otherwise constant values (NEFSC 2013). There is rarely empirical evidence that can justify the increases in natural mortality; thus, it can leave the impression that the change was simply a technique to reduce the retrospective pattern rather than to address biological realism. A challenging consequence of allowing M to vary within the assessment is that it complicates selection of reference points (and resulting stock status) and projections for catch advice. Debate as to whether fishing mortality reference points should increase or decrease given an increase in M can become a management morass, and trying to reach agreement over what M value (or values) to use in projections for catch advice can prove difficult (Legault and Palmer in 2016).

The above factors (missing catch, variable catchability, or M) are not an exclusive set of potential misspecifications and their fixes. For example, Sinclair et al. (1991) considered alternative specification of F on the plus-group (effectively implementing dome-shaped selectivity), and Stewart and Martell (2014) found differences in retrospective statistics depending on how selectivity was modeled. Whatever the misspecification(s), simply increasing one of these factors (for example, doubling M for all years and ages) is not sufficient to fix a retrospective pattern; rather, a temporal change from a previous value is necessary. This suggests an underlying structural mismatch of assuming something is constant when in reality, it has an underlying trend.

Ultimately, the approaches described to reduce retrospective bias reflect different structural mechanisms for making the model more flexible. As we have demonstrated, the direction and magnitude of retrospective patterns are not constant, and future updates of an assessment with one of these structural fixes may show a re-emergence of a new retrospective pattern. In cases where changes in basic data or parameters cannot be justified, we re-emphasize that an alternative approach to reduce bias in terminal year model estimates is to adjust the $N_{a,T+1}$ based on a measure of recent retrospective bias. An advantage of ρ -adjusting $N_{a,T+1}$ is that the adjustment evolves with the assessment (i.e., the calculated ρ values are updated each time the assessment is updated; therefore, the adjustment reflects the current retrospective pattern). A disadvantage is that the bias for the most recent retro models is almost certainly underestimated and may be part of the reason that the 7-year mean ρ does not fully correct the bias. The other approaches to correct for bias (estimating missing catch or allowing variability in catchability or natural mortality) can also be updated at each assessment iteration, but the implications for perceived reliability of surveys, appropriate M for reference points and catch advice, and integrity of catch statistics raise more contentious issues. Furthermore, we compared projection performance between ρ -adjusted $N_{a,T+1}$ using $\rho_{SSB,7}$ or splitting all indices for five stocks that exhibited strong retrospective patterns (Georges Bank cod, Georges Bank yellowtail flounder, Gulf of Maine winter flounder, southern New England winter flounder, and witch flounder). In four out of five cases, the ρ -adjusted projections reduced bias more than splitting the series, and in the fifth case, both approaches performed similarly.

Given the increased focus on topics related to climate change, it is natural to wonder whether unmodeled factors are contributing to the observed bias. Changes in the environment will likely manifest as changes in biological parameters, and all assessments included in this analysis examined annual catch and survey data to derive annual weights at age and maturity at age; thus, these biological aspects were reflective of the contemporaneous environmental state. Furthermore, by fixing biological parameters at a recent 5-year mean, projections are implicitly assuming that the most recent environmental conditions will persist in the near future. However, a 5-year mean will overestimate (underestimate)

predicted weights at age when that trend is consistently declining (or consistently increasing) more so than a mean based on fewer years. We examined the status quo projections using a 3-year mean instead of a 5-year mean for weights at age, maturity at age, and selectivity at age. For several stocks with very strong declining trends (Georges Bank haddock, Gulf of Maine haddock, American plaice, and somewhat for Georges Bank cod), bias in predicted catch for the first 3 years was less with a 3-year mean than a 5-year mean. Thus, we recommend that analysts make specifications for projections based on the magnitude and persistence of trends observed in their own stocks.

With respect to recruitment, if the environment is influencing reproductive output or survival of recruits, then sampling from the entire recruitment cdf does not necessarily address the potential trend for lower or higher productivity for sustained periods of time in the future. A variant of the status quo case was performed where only the most recent 10 estimated recruitments were used to define the empirical cdf. Results for the first 3 projection years were very similar, reinforcing that both SSB and catch in the first 3 years of projections are driven by the current estimated cohorts. Longer-term projections (>3 years) were improved for stocks that had recently seen a prolonged period of low recruitment, because those projections are driven by predicted rather than observed cohorts. This result suggests that refining the recruitment estimates used for projections should be evaluated on a stock-specific basis, similar to the conclusion regarding recent trends in biological parameters and the appropriate number of years to average. However, the overall poor performance of projections beyond 3 years for the stocks explored here suggests that at the point where projected recruitment begins to contribute to future SSB and catch (paper fish), the assessment should be updated anyway.

Although this research focused on VPA assessments, retrospective bias is an issue that afflicts other types of assessment models as well. A variety of models, from surplus production to statistical catch at age, have been applied to stocks that exhibit severe retrospective patterns without resolving the problem. In a large, international simulation exercise, models that permitted additional flexibility through a parameter with temporal variability (M , catchability, intrinsic rate of growth, or selectivity), as well as "standard" VPA models, achieved inconsistent trends and divergent estimates of absolute magnitude and performed poorly in tests of self-consistency (i.e., assessment models applied to simulated data with the same model structure), particularly in the most recent years (Deroba et al. 2015). We therefore recommend that evaluation of retrospective bias be a standard diagnostic tool when conducting any stock assessments, regardless of the model.

Given experiences over the past decade with persistent retrospective bias in stock assessments for some of the stocks explored here and repeated occurrences of projected catches leading to overfishing as a result of overestimated stock size, we recommend the following: (1) provide catch advice only for short-term projections (≤ 3 years); (2) develop a harvest control rule (HCR) that explicitly addresses buffers in terms of unmodeled, unmeasured, and underestimated uncertainty. The recommendation for making projections no more than 3 years out would require that assessments be updated every 2–3 years, depending on lags between when data are available and when assessments are performed. In developing a candidate HCR, we recommend that diagnostics and ρ adjustments such as we have presented be incorporated in the framework. However, as we found that substantial bias remained even after ρ -adjusting, we also recommend that the HCR should include adjustments to the distribution to increase the variance (because the projected distributions did not span the true value in quite a few cases) and that a percentile other than the median of the projected distributions should be recommended for providing catch advice and tracking trends in SSB.

Preliminary analyses (unpublished) suggest that adjusting projected distributions to have a CV of 100% and then basing advice

on the 20th to 40th percentiles performed well as an alternative to using the median for catch advice and for tracking the trajectory of SSB. Future work could expand on this. For example, the recommended percentile could be based on consideration of the variability and direction of relative differences among retro models, the magnitude of ρ_{SSB} , sustained declines in weights at age, condition factor, and (or) a series of weak year classes, as well as the amount of rebuilding required (i.e., how far below B_{MSY} the current stock estimate is). We note that this strawman HCR blends some existing elements of the current control rule for the New England Fishery Management Council and the Mid-Atlantic Fishery Management Council and that management advice from recent groundfish assessments has been based on ρ -adjusted distributions if indicated by the rule of thumb (NEFSC 2008, 2012).

While it is a requirement for federal US fisheries managers to rebuild stocks that are found to be "overfished", that often requires long projections of 10 years or more. The reliability of long-term projections is dubious, particularly for stocks exhibiting a retrospective pattern. An alternative approach to rebuilding would be to aim for short-term milestones that increase SSB above the assessment estimate of current biomass. For example, a milestone might be to estimate short-term catch that permits SSB to increase by 10% or 20% within 3 years (NEFMC 2012). This factor was considered during the recalculation of the Georges Bank yellowtail flounder rebuilding plan (NEFMC 2012). For such an approach to replace consideration of a rebuilding period, a change in statutory language may be necessary.

The evaluations made in this paper have been based on the assumption that the 2008 assessments for each stock represent the truth. However, for stocks that have exhibited retrospective patterns, the 2008 assessment estimates would be expected to change with additional years of data if the retrospective pattern continued. Indeed, this has been observed in subsequent updates for some of these assessments. We maintain that this assumption is still a reasonable approach because it is the issue faced in reality whereby a given stock assessment estimates values for the terminal model year and must provide projected catch and SSB (and convey the projection uncertainty) based on only past observation of bias patterns. We have demonstrated that the magnitude of the retrospective pattern can change over time and so are interested in what sort of adjustment could be considered in a realistic situation of not being able to predict future changes in the retrospective pattern. Our results indicate that when a strong retrospective pattern is observed, adjusting the starting population for projections is likely to improve management advice. We expect that if the retrospective pattern does in fact continue, then this adjustment may be insufficient, but will probably at least be in the correct direction. An additional adjustment in catch advice, along the lines outlined in the HCR discussion above, may be necessary to fully capture the uncertainty in a stock assessment exhibiting a strong retrospective pattern.

Finally, decisions about future catches often consider projected economic benefits and losses. We suggest that it would be useful to compare projected versus realized economic consequences to evaluate the accuracy of those analyses to refine those models (if needed) or to identify data that are needed to produce more accurate accounting. Obviously, economic analyses would be biased if they use biased catch projections. But an approach similar to what was done here, using projections to match catches for a period where we have observations, would be a way to evaluate the economic methods without the distracting fact that they are using catches that are biased from the assessment. Similarly, if ecosystem and climate predictions begin to be provided for management advice, performance evaluations should be conducted for those models as well. In general, any considerations that are expected to inform management decisions, be they biological, economic, or ecosystem, should undergo a formal evaluation of predictive ability, and those results should be conveyed to man-

agers for characterizing uncertainty and defining acceptable risk (Brander et al. 2013; Gregr and Chan 2015).

Acknowledgements

The authors are grateful for very helpful comments provided by Tom Nies and Mark Terceiro and for comments from the Associate Editor and three anonymous reviewers that improved the focus and clarity. We acknowledge our colleagues in the Population Dynamics Branch at the Northeast Fisheries Science Center whose VPA assessments were used in this analysis (Larry Alade, Lisa Hendrickson, Ralph Mayo, Paul Nitschke, Loretta O'Brien, Michael Palmer, Gary Shepherd, Mark Terceiro, and Susan Wigley). We also acknowledge our colleagues in the Ecosystems Survey Branch for their dedication to the fisheries-independent surveys and our colleagues in the Population Biology Branch for their dedication to ageing catch and survey samples. This work originated when the authors were temporarily added to the New England Fishery Management Council groundfish Plan Development Team to evaluate historical performance of projections.

References

- Beverton, R.J.H., and Holt, S.J. 1957. On the dynamics of exploited fish populations. Fisheries Investment Series 2. Vol. 19. UK Ministry of Agriculture and Fisheries, London.
- Bousquet, N., Cadigan, N., Duchesne, T., and Rivest, L.-P. 2010. Detecting and correcting underreported catches in fish stock assessment: trial of a new method. *Can. J. Fish. Aquat. Sci.* 67(8): 1247–1261. doi:10.1139/F10-051.
- Brander, K., Neuheimer, A., Andersen, K.H., and Hartvig, M. 2013. Overconfidence in model projections. *ICES J. Mar. Sci.* 70: 1065–1068. doi:10.1093/icesjms/fst055.
- Butterworth, D.S., and Rademeyer, R.A. 2008. Statistical catch-at-age analysis vs. ADAPT-VPA: the case of Gulf of Maine cod. *ICES J. Mar. Sci.* 65(9): 1717–1732. doi:10.1093/icesjms/fsn178.
- Caddy, J.F., and McGarvey, R. 1996. Targets or limits for management of fisheries? *N. Am. J. Fish. Manage.* 16: 479–487. doi:10.1577/1548-8675(1996)016<0479:TOLFM0>2.3.CO;2.
- Clark, W.G. 1999. Effects of an erroneous natural mortality rate on a simple age-structured stock assessment. *Can. J. Fish. Aquat. Sci.* 56(10): 1721–1731. doi:10.1139/f99-085.
- Conser, R.J., and Powers, J.E. 1989. Extensions of the ADAPT VPA tuning method designed to facilitate assessment work on tuna and swordfish stocks. ICCAT Working Doc. SCRS/89/43.
- Darby, C.D. 2005. Estimating systematic bias in the North Sea cod landings data. ICES ACFM WGNSSK Appendix 4.
- Deroba, J.J. 2014. Evaluating the consequences of adjusting fish stock assessment estimates of biomass for retrospective patterns using Mohn's Rho. *N. Am. J. Fish. Manage.* 34: 380–390. doi:10.1080/02755947.2014.882452.
- Deroba, J.J., and Schueller, A.M. 2013. Performance of stock assessments with misspecified age- and time-varying natural mortality. *Fish. Res.* 146: 27–40. doi:10.1016/j.fishres.2013.03.015.
- Deroba, J.J., Butterworth, D.S., Methot, R.D., Jr., De Oliveira, J.A.A., Fernandez, C., Nielsen, A., Cadrin, S.X., Dickey-Collas, M., Legault, C.M., Ianelli, J., Valero, J.L., Needle, C.L., O'Malley, J.M., Chang, Y.-J., Thompson, G.G., Canales, C., Swain, D.P., Miller, D.C.M., Hintzen, N.T., Bertignac, M., Ibaibarriaga, L., Silva, A., Murta, A., Kell, L.T., de Moor, C.L., Parma, A.M., Dichmont, C.M., Restrepo, V.R., Ye, Y., Jardim, E., Spencer, P.D., Hanselman, D.H., Blaylock, J., Mood, M., and Hulson, P.-J.F. 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(1): 19–30. doi:10.1093/icesjms/fst237.
- Evans, G.T. 1996. Using the elementary operations of sequential population analysis to display problems in catch or survey data. *Can. J. Fish. Aquat. Sci.* 53(2): 239–243. doi:10.1139/f95-191.
- Gavaris, S. 1988. An adaptive framework for the estimation of population size. Canadian Atlantic Fisheries Science Advisory Committee Res. Doc. 88/29.
- Gavaris, S., and Van Eeckhaute, L. 1998. Diagnosing systematic errors in reported fishery catch. In *Fishery stock assessment models*. Alaska Sea Grant College Program, AK-SG-98-01. pp. 399–412.
- Goodyear, C.P. 1996. Variability of fishing mortality by age: consequences for maximum sustainable yield. *N. Am. J. Fish. Manage.* 16: 8–13. doi:10.1577/1548-8675(1996)016<0008:VOFMBA>2.3.CO;2.
- Gregr, E.J., and Chan, K.M.A. 2015. Leaps of faith: how implicit assumptions compromise the utility of ecosystem models for decision-making. *BioScience*, 65(1): 43–54. doi:10.1093/biosci/biu185.
- Hammond, T.R., and Trenkel, V.M. 2005. Censored-catch data in fisheries stock assessment. *ICES J. Mar. Sci.* 62: 1118–1130. doi:10.1016/j.icesjms.2005.04.015.
- Hulson, P.-J.F., and Hanselman, D.H. 2014. Tradeoffs between bias, robustness,

- and common sense when choosing selectivity forms. *Fish. Res.* **158**: 63–73. doi:10.1016/j.fishres.2013.12.016.
- Hurtado-Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J., Johnson, K.F., Licandeo, R., McGilliard, C.R., Monnahan, C.C., Muradian, M.L., Ono, K., Vert-Pre, K.A., Whitten, A.R., and Punt, A.E. 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. *ICES J. Mar. Sci.* **72**(1): 99–110. doi:10.1093/icesjms/fsu198.
- ICES. 2002. Report of the Working Group on Methods of Fish Stock Assessments. ICES Headquarters, Copenhagen, Denmark, 3–7 December 2001. ICES CM 2002/D:01.
- ICES. 2007. Report of the Working Group on Methods of Fish Stock Assessments (WGMG). Woods Hole, Mass., 13–22 March 2007. ICES CM 2007/RMC:04.
- ICES. 2008. Report of the Working Group on Methods of Fish Stock Assessments (WGMG). Woods Hole, Mass., 7–16 October 2008. ICES CM 2008/RMC:03.
- King, D.M., and Sutinen, J.G. 2010. Rational noncompliance and the liquidation of Northeast groundfish resources. *Mar. Pol.* **34**: 7–21. doi:10.1016/j.marpol.2009.04.023.
- Legault, C.M. 2009. Report of the Retrospective Working Group, January 14–16, 2008, Woods Hole, Mass. NEFSC Reference Doc. 09-01.
- Legault, C.M. 2014. The ability of two age composition error distributions to estimate selectivity and spawning stock biomass in simulated stock assessments. *Fish. Res.* **158**: 172–180. doi:10.1016/j.fishres.2013.12.007.
- Legault, C.M., and Palmer, M.C. 2016. In what direction should the fishing mortality target change when natural mortality increases within an assessment? *Can. J. Fish. Aquat. Sci.* **73**(3): 349–357. doi:10.1139/cjfas-2015-0232.
- Legault, C.M. and Restrepo, V.R. 1999. A flexible forward age-structured assessment program. ICCAT SCRS/98/58.
- Maunder, M.N. 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: estimating the effective sample size. *Fish. Res.* **109**: 311–319. doi:10.1016/j.fishres.2011.02.018.
- Maunder, M.N., and Piner, K.R. 2015. Contemporary fisheries stock assessment: many issues still remain. *ICES J. Mar. Sci.* **72**(1): 7–18. doi:10.1093/icesjms/fsu015.
- Mertz, G., and Myers, R.A. 1997. Influence of errors in natural mortality estimates in cohort analysis. *Can. J. Fish. Aquat. Sci.* **54**(7): 1608–1612. doi:10.1139/f97-067.
- Miller, T.J., Das, C., Politis, P.J., Miller, A.S., Lucey, S.M., Legault, C.M., Brown, R.W., and Rago, P.J. 2010. Estimation of Albatross IV to Henry B. Bigelow calibration factors. Northeast Fish. Sci. Cent. Ref. Doc. 10-05.
- Mohn, R. 1993. Bootstrap estimates of ADAPT parameters, their projection in risk analysis and their retrospective patterns. *Can. Spec. Publ. Fish. Aquat. Sci.* **120**: 173–184.
- Mohn, R. 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. *ICES J. Mar. Sci.* **56**: 473–488. doi:10.1006/jmsc.1999.0481.
- NEFMC. 1980. Compendium of preliminary descriptive materials in support of the interim groundfish plan for cod, haddock, and yellowtail flounder. New England Fishery Management Council. [Available upon request.]
- NEFMC. 2012. Framework Adjustment 47 to the Northeast Multispecies Fishery Management Plan. New England Fishery Management Council.
- NEFSC. 2008. Assessment of 19 Northeast groundfish stocks through 2007: report of the 3rd groundfish assessment review meeting (GARM III). Northeast Fisheries Science Center. Ref. Doc. 08-15.
- NEFSC. 2012. Assessment or Data Updates of 13 Northeast Groundfish Stocks through 2010. US Dept. Commer., Northeast Fisheries Science Center. Ref. Doc. 12-06.
- NEFSC. 2013. 55th Northeast Regional Stock Assessment Workshop (55th SAW) Assessment Report. Northeast Fisheries Science Center. Ref. Doc. 13-11.
- Parma, A.M. 1993. Retrospective catch-at-age analysis of Pacific halibut: implications on assessment of harvesting policies. In *Management strategies of exploited fish populations*. Alaska Sea Grant College Program, AK-SG-93-02. pp. 247–265.
- Parrack, M.L. 1986. A method of analyzing catches and abundance indices from a fishery. ICCAT Coll. Vol. Sci. Pap. **24**: 209–221.
- Patterson, K.R. 1998. Assessing fish stocks when catches are misreported: model, simulation tests, and application to cod, haddock, and whiting in the ICES area. *ICES J. Mar. Sci.* **55**: 878–891. doi:10.1006/jmsc.1998.0351.
- Patterson, K., Cook, R., Darby, C., Gavaris, S., Kell, L., Lewy, P., Mesnil, B., Punt, A.E., Restrepo, V., Skagen, D.W., and Stefansson, G. 2001. Estimating uncertainty in fish stock assessment and forecasting. *Fish. Fish.* **2**: 125–157. doi:10.1046/j.1467-2960.2001.00042.x.
- Pitcher, T.J., Watson, R., Forrest, R., Valtysson, H., and Guenette, S. 2002. Estimating illegal and unreported catches from marine ecosystems: a basis for change. *Fish. Fish.* **3**: 317–339. doi:10.1046/j.1467-2979.2002.00093.x.
- Prager, M.H., Porch, C.E., Shertzer, K.W., and Caddy, J.F. 2003. Targets and limits for management of fisheries: a simple probability-based approach. *N. Am. J. Fish. Manage.* **23**: 349–361. doi:10.1577/1548-8675(2003)023<0349:TALFMO>2.0.CO;2.
- Punt, A.E., Hurtado-Ferro, F., and Whitten, A.R. 2014. Model selection for selectivity in fisheries stock assessments. *Fish. Res.* **158**: 124–134. doi:10.1016/j.fishres.2013.06.003.
- Rivard, D., and Foy, M.G. 1987. An analysis of errors in catch projections for Canadian Atlantic fish stocks. *Can. J. Fish. Aquat. Sci.* **44**(5): 967–981. doi:10.1139/f87-115.
- Rousseeuw, P.J., Ruts, I., and Tukey, J.W. 1999. The Bagplot: A Bivariate Boxplot. *Am. Stat.* **53**(4): 382–387. doi:10.2307/2686061.
- Sampson, D.B. 2011. The accuracy of self-reported fisheries data: Oregon trawl logbook fishing locations and retained catches. *Fish. Res.* **112**: 59–76. doi:10.1016/j.fishres.2011.08.012.
- Shertzer, K.W., Prager, M.H., and Williams, E.H. 2008. A probability-based approach to setting annual catch levels. *Fish. Bull.* **106**: 225–232.
- Sinclair, A., Gascon, D., Rivard, D., and Gavaris, S. 1991. Consistency of some northwest Atlantic groundfish stock assessments. *NAFO Sci. Coun. Stud.* **16**: 59–77.
- Stewart, I.J., and Martell, S.J.D. 2014. A historical review of selectivity approaches and retrospective patterns in the Pacific halibut stock assessment. *Fish. Res.* **158**: 40–49. doi:10.1016/j.fishres.2013.09.012.
- Thompson, G.G. 1994. Confounding of gear selectivity and the natural mortality rate in cases where the former is a nonmonotone function of age. *Can. J. Fish. Aquat. Sci.* **51**(12): 2654–2664. doi:10.1139/f94-265.
- Wang, Y., O'Brien, L., Clark, K., Gavaris, S., and Hatt, B. 2009. Assessment of Eastern Georges Bank Atlantic Cod for 2009. TRAC Ref. Doc. 2009/01.
- Wilberg, M.J., Thorson, J.T., Linton, B.C., and Berkson, J. 2009. Incorporating time-varying catchability into population dynamic stock assessment models. *Rev. Fish. Sci.* **18**: 7–24. doi:10.1080/10641260903294647.
- Yin, Y., and Sampson, D.B. 2004. Bias and precision of estimates from an age-structured stock assessment program in relation to stock and data characteristics. *N. Am. J. Fish. Manage.* **24**: 865–879. doi:10.1577/M03-107.1.