Data Rich but Model Resistant: An Evaluation of Data-Limited Methods to Manage Fisheries with Failed Age-based Stock Assessments

Christopher M. Legault[1], John Wiedenmann[2] Jonathan J. Deroba[1], Gavin Fay[3], Timothy J. Miller[1], Elizabeth N. Brooks[1], Richard J. Bell[4], Joseph A. Langan[5], Jamie M. Cournane[6], Andrew W. Jones[1], Brandon Muffley[7]

[1]National Marine Fisheries Service, Northeast Fisheries Science Center, Woods Hole, MA, USA

[2]Department of Ecology, Evolution and Natural Resources, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

[3]University of Massachusetts Dartmouth, School for Marine Science and Technology, MA, USA

[4]The Nature Conservancy, RI, USA

[5]University of Rhode Island, Graduate School of Oceanography, Narragansett, RI, USA

[6]New England Fishery Management Council, Newburyport, MA, USA

[7]Mid-Atlantic Fishery Management Council, Dover, DE, USA

Corresponding author: C.M. Legault (chris.legault@noaa.gov)

Competing interests: The authors declare there are no competing interests.

Abstract

Age-based stock assessments are sometimes rejected by review panels due to large retrospective patterns. When this occurs, data-limited approaches are often used to set catch advice, under the assumption that these simpler methods will not be impacted by the problems causing retrospective patterns in the age-based assessment. This assumption has never been formally evaluated. Closed-loop simulations were conducted where a known source of error caused a retrospective pattern in an age-based assessment. Twelve data-limited methods, an ensemble of a subset of these methods, and a statistical catch-at-age model with retrospective adjustment were all evaluated to examine their ability to prevent overfishing and rebuild overfished stocks. Overall, none of the methods evaluated performed best across the scenarios. A number of methods performed consistently poorly, resulting in frequent and intense overfishing and low stock sizes. The retrospective adjusted statistical catch-at-age assessment performed better than a number of the alternatives explored. Thus, using a data-limited approach to set catch advice will not necessarily result in better performance than relying on the age-based assessment with a retrospective adjustment.

Keywords

closed-loop simulation, data-limited methods, retrospective analysis, management advice

Introduction

In the U.S., age-based, integrated, fisheries stock assessment models are frequently used to estimate annual stock abundance (biomass), fishing mortality rates, and management reference points [@maunder2013review]. These models must undergo peer review, where an independent panel of experts determines whether or not results from the model are suitable as the basis for determining stock status and for setting catch advice. There are a number of model diagnostics that are used to evaluate uncertainty and stability of assessment model results, but one that is commonly used and carries substantial weight during review is the retrospective pattern. A retrospective pattern is a systematic inconsistency among a series of sequential assessment estimates of population size (or other related assessment variables), based on increasing time periods of data used in the model fitting [@mohn1999rho]. These inconsistencies in assessment estimates are indicative of one or more mismatches between model assumptions and patterns in the data used to fit the model. Large or persistent retrospective patterns indicate an instability in model results, and may therefore be the basis for a peer review panel to determine that model results are not suitable for management purposes [@punt2020reject].

Many stock assessments in the Northeast U.S. have a history of strong retrospective patterns, whereby estimates of biomass are typically revised downward and estimates of fishing mortality rate are revised upward as new data are added to the model (i.e., implying systematic overestimation of biomass and underestimation of fishing mortality; [@ices2020wkforbias]). NOAA Fisheries, the New England Fishery Management Council, the Mid-Atlantic Fishery Management Council, and the Atlantic States Marine Fisheries Commission manage these stocks, and retrospective issues remain a challenge for managers when setting catch advice and tracking stock status. This problem has been particularly acute for, but not limited to, stocks in the New England groundfish complex [@nefsc02a; @nefsc05; @nefsc08; @nefsc15a; @nefsc15b; @nefsc17; @nefsc19; @deroba2010mackerel], managed under NOAA Fisheries and the New England Council's Northeast Multispecies (Groundfish) fishery management plan. Stock assessments exhibiting retrospective patterns can be found around the world and can be associated with a wide range of assessment approaches [@ices2020wkforbias].

The magnitude of the retrospective pattern is typically measured with a statistic called Mohn's rho [@mohn1999rho]. Mohn's rho can be used to adjust terminal year estimates of biomass in anticipation that the retrospective pattern will persist, and some accounting for the pattern will provide a more accurate estimate. Stock assessments where the so-called rho-adjusted value is outside the 90% confidence interval of the terminal year estimate of spawning stock biomass (SSB) or fishing mortality rate are classified as strong retrospective patterns. In these cases, the rho-adjusted values are used for status determination and to modify the starting population for projections used to provide catch advice [@brooks2016retroforecast].

There are many possible causes for retrospective patterns, but typically there is a temporal change in either the data or a model parameter that is not accounted for in the stock assessment model [@legault2020rose; @hurtado2014rearview; @deroba2014retro]. The strong retrospective patterns seen in the region under study have required very large magnitudes of change in order to remove the retrospective pattern. For example, the scale of missing catch needed to be three to five times the reported catch, or natural mortality needed to increase from 0.2 to near 1.0 to reproduce observed retrospective patterns; the scales of these changes have not been deemed believable by review panels. Some approaches have been used to estimate missing catch [@vanbeveren2017catch; @perretti2020sim] and increased natural mortality [@cadigan2016ss; @rossi2019inferring]. However, identifying the correct source of the retrospective pattern is difficult and using the wrong fix can lead to poor management advice [@szuwalski2017retro]. This is clearly an area where more research is needed, but currently addressing strong retrospective patterns is challenging.

There is no formal criteria in the region for rejecting an assessment based on Mohn's rho, but large, positive values of rho for SSB, especially those persisting across several assessments, have played an important role in the rejection of recent age-based assessments, including Atlantic mackerel (*Scomber scombrus*), Georges Bank Atlantic cod (*Gadus morhua*), Georges Bank yellowtail flounder (*Limanda ferruginea*), and witch flounder (*Glyptocephalus cynoglossus*) [@deroba2010mackerel; @legault2014tracgbytf; @nefsc15a; @nefsc15b]. In each of these cases, and another where the assessment rejection was not based on the retrospective pattern [black sea bass, *Centropristis striatus*, @nefsc12], the Councils have relied on a variety of data-limited approaches for setting catch advice for these stocks [@mcnamee2015mafmc; @nefsc15a; @nefsc15b; @wiedenmann2015mackerel]. These approaches have all been ad-hoc, and a recent analysis suggested that some of the data-limited approaches may not be suitable for stocks in the Northeast U.S. with a history of high exploitation rates [@wiedenman2019dlm]. In addition, large, positive retrospective patterns in SSB persist for a number of other stocks in the region [@nefsc19], raising concerns that additional stocks may rely on data-limited approaches in the future.

Current practice in the region requires identification of a back-up assessment approach for all age-based assessments in case the age-based assessment is rejected during peer review. These back-up approaches are required to be simple enough that only minor review is needed so that management advice can continue to be provided for the stock. While these DLMs cannot provide stock status determinations in our study because they rely on ad hoc setting of reference points, they all can provide catch advice. Therefore, there is an immediate need to identify suitable data-limited approaches for setting catch advice for stocks with age-based assessments that did not pass review.

We developed a closed-loop simulation [e.g., @punt2016mse; @huynh2022retro] to evaluate the suitability of alternative data-limited methods (DLMs) for setting target catches when age-based stock assessments fail. In particular, focus was placed on methods that use survey indices of abundance. The closed-loop simulation was designed to test the two most common hypothesized sources of retrospective pattern (missing catch or increases in natural mortality), and to evaluate performance of various methods relative to exploitation history and changes in fishery selectivity. Results of this factorial simulation study are summarized for quantities of interest that impact fisheries management advice. The goal of this work is to examine the hypothesis that catch advice from DLMs is more robust to under-reported catch or changes in natural mortality than from a rho-adjusted statistical catch at age model.

Methods

Overview

A closed-loop simulation was designed to approximate a process where an age-based assessment was rejected due to a retrospective pattern, requiring catch advice to be determined using a DLM. As such, the operating model (OM) used to define the "true" underlying biological and fishery dynamics was also age-based. The OM was run for an initial 50 year period of time (called the base period) that controls the historical population dynamics and fishing pressure, and allows for sufficient data to be simulated in the observation model to be used in the different DLMs. After the base period, a given management approach (i.e., DLM) was applied to set the target catch for the stock, which is then removed from the population. This process is repeated at a fixed interval for 40 years in what is called the feedback period. Multiple OMs were developed so that the performance of the DLMs could be compared among several sources of uncertainty that are especially common in the northeast U.S., but relevant more broadly. The set of OMs featured one of two possible patterns of time varying dynamics in the last 20 years of the base period, that if left misspecified as time invariant, would be sufficient to generate retrospective patterns resulting in the rejection of an age-based stock assessment, requiring transition to a DLM. The details of these dynamics, and the suite of factors explored in the closed-loop simulation, are described in sections below.

Operating and Observation Models

The Woods Hole Assessment Model [WHAM, @miller2020wham; @stock2021wham] was used as the basis for the OM in the closed-loop simulations. WHAM is an R package and the general model is built using the Template Model Builder package [@kristensenetal2016TMB]. While WHAM can serve as a stock assessment model used to estimate parameters, it can also simulate the data needed for age-based stock assessments and DLMs given a range of input parameters. WHAM was used to simulate data with known properties during the base and feedback periods. Catch and index observations upon which the DLMs largely relied were simulated according to user supplied biological and fishery parameters for each scenario (see below). Catches during the feedback period were iteratively updated based on a DLM and harvest control rule that used the simulated observations to produce catch advice. Catch advice from a given combination of DLM and control rule was specified in two year blocks, a typical catch specification timeframe for New England and Mid-Atlantic Council managed fisheries. WHAM used these catches, along with the user supplied biological and fishery inputs, to have the simulated population respond to the DLM, thereby completing the closed-loop simulation aspect. A limit was placed on the maximum fishing mortality rate when the fishery attempted to remove the catch advice from the population during the feedback period. There was no implementation error in the removal of the catch advice otherwise, except when missing catch was the source of the retrospective pattern as described below.

The age-structured OM had ten ages, with the oldest age being a plus group. Maturity- and weight-at-age were time and simulation invariant and reflected values observed for groundfish in the region (Table 1). The OM simulated catch and age composition data for a single fishery with logistic selectivity (Table 1; see below). Annual, total catch observations (metric tons) were simulated as lognormal deviations from the underlying “true” catches with a coefficient of variation (CV) equal to 0.1. Fishery age composition data were assumed to follow a multinomial distribution with an effective sample size (ESS) equal to 200. Two fishery independent surveys were simulated and were intended to represent the spring and fall, coastwide bottom trawl surveys conducted in the region. Both surveys were assumed to have time invariant logistic selectivity and constant catchability. Annual survey observations were simulated as lognormal deviations from the underlying "true" survey catches with a CV of 0.3 in the spring survey and 0.4 in the fall. Survey age composition data were assumed to follow a multinomial distribution with an ESS equal to 100 in both seasons.

Annual recruitment was simulated as autoregressive, lag-1 (AR-1) deviations from an underlying Beverton-Holt stock-recruitment relationship with steepness equal to 0.74. The degree of correlation in the AR-1 process equaled 0.4 with a conditional standard deviation about this relationship equal to 0.5. Unfished recruitment was time- and simulation invariant and equaled 10-million age-1 fish. These stock-recruitment values were based on an average of groundfish parameters estimated for the region.

Data-Limited Methods Explored

The range of DLMs evaluated was generally constrained to those that have been used or were considered plausible (e.g., based on data requirements) for the Northeast Shelf. Ultimately, thirteen DLMs were selected for evaluation. Although catch-curve analyses are not currently applied in the region, they were included here since age information is available for most of the stocks, and because @wiedenman2019dlm showed they performed well in application to groundfish stocks. Two additional DLMs (Islope and Itarget) not currently used in the region were also evaluated, as these have been tested in other applications and shown promise [@geromont2015complex; @geromont2015datapoor; @carruthers2016simpleMPs; @wiedenman2019dlm]. An ensemble of models was also considered based on recent findings that improved performance can result from combining the results from multiple models [@anderson2017superensemble; @rosenberg2018ensemble; @spence2018combineecomodels; @stewart2018ensemble]. The catch advice from the ensemble approach equaled the median of the catch advice resulting from the range of methods included in the ensemble (Table 2). This assumes an equal weighting of ensemble members. The DynLin approach was excluded from the ensemble due to the relatively long computing time required. Other methods were excluded (CC-FM, ES-FM, ES-Fstable) because they were slight variations of a more generic DLM (i.e., CC- and ES-) and including them all may have unduly overweighted the performance of the ensemble towards these methods. For the methods with multiple variations, the variant retained in the ensemble had superior performance than the alternatives based on preliminary results, or had already been considered for application in the region. The full range of methods included in this analysis were detailed below with equations (Table 2). Each method was applied to data that would lead to retrospective patterns in an age-based stock assessment and performance was evaluated using a range of metrics (see below).

Each of the methods evaluated produces a single target catch value that was fixed over a two year interval. If the methods were being applied in year y, then target catches are set for years $y + 1$ and $y + 2$ (denoted $C_{targ,y+1:y+2}$). In practice, the timing of setting target catches in the region generally occurs in late summer or early fall in between the spring and fall surveys, and before complete catch data are available. Therefore, in year y complete catch data are available through year $y - 1$, and survey data are available for the spring survey through year y and for the fall survey through year $y - 1$. Applications of DLMs in this region have used an average of the spring index in year y ($I_{spr,y}$) and the fall index in year $y - 1$ ($I_{fall,y-1}$) to reflect average abundance at the start of year y ($\bar{I}_y$). For this study, the same 1 year lag was implemented for methods that use the average of both simulated indices to generate catch advice:

$\bar{I}_y = \frac{I_{fall,y-1}+I_{spr,y}}{2}$.

Control Rules

Most DLMs do not have the ability to estimate a biomass reference point (e.g., B_{MSY}), which made consideration of so called biomass-based harvest control rules that reduce F or catch in response to estimated changes in relative stock status impossible. Although reference points can be created for DLMs, they typically rely on local expert judgment

[@harford2021harvest] and are geared towards either keeping the stock about where it is or else increasing it towards a relative amount that was thought to be good. Neither of these provide a proxy for maximum sustainable yield reference points, but might instead provide pretty good yield [@hilborn2010pgy].

Lack of clarity exists, however, on whether the catch advice from DLMs should be used directly or reduced to account for uncertainty. In the U.S. management system, an overfishing limit is the catch that would result from applying F_{MSY}, whereas an acceptable biological catch is a catch reduced from the overfishing limit to account for scientific uncertainty. Each DLM was evaluated using two harvest control rules: 1) the catch advice from a given DLM was applied directly and assumed to serve as a proxy for the catch associated with F_{MSY} (catch multiplier = 1), and 2) the catch advice from a given DLM was reduced by 25% to account for unspecified scientific uncertainty (catch multiplier = 0.75). The case where catches were reduced by 25% was intended to reflect a common default control rule in the region that uses $0.75F_{MSY}$.

Application of a Statistical Catch-at-Age Assessment (SCAA)

A SCAA model was also applied to all scenarios to generate catch advice for comparison with the DLMs. Although virtual population analysis (VPA) is also used for some age-based assessments in the region, SCAA models are more widely used. Applications of the SCAA model assumed that the assessment had the correct underlying structure for selectivity, and CVs and ESS were specified at their true underlying values. The SCAA model estimated annual recruitment deviations assuming no underlying stock-recruit relationship, annual fully-selected fishing mortality rates, fishery and survey selectivity parameters (logistic), abundance-at-age in year one of the period being assessed, and survey catchabilies. Mohn's rho was calculated (7 year peels) for abundance at age for all model fits during the feedback period and used to retro-adjust abundance at age for projections (divided by one plus Mohn's rho; [@brooks2016retroforecast]). Catch advice was determined by specifying fully-selected $F = 0.75F_{40\%}$, always assuming M=0.2. All life history parameters were fixed at their correct value, except for the natural mortality rate when it was the source of the retrospective pattern.

Study Design

In addition to the two control rules applied for each DLM described above, three aspects of the OM were varied in a full factorial study design: fishing history, fishery selectivity, and cause of the retrospective pattern (Table 3). Two variants of fishing history were considered, with fully selected fishing mortality during the base period either constant at a level equal to $2.5F_{MSY}$ (always overfishing) or equaling $2.5F_{MSY}$ in the first half of the base period then a knife-edged decline to F_{MSY} for the second half of the base period. These patterns in fishing mortality rate were based on observed patterns for Northeast groundfish [@wiedenman2019dlm]. These two different fishing intensities during the latter half of the base period led to different starting conditions for the feedback period.

Two variations of the OM were considered with either time invariant, asymptotic, fishery selectivity in the base and feedback periods, or a change in selectivity after the first half of the base period so that the age at 50% selectivity increased from approximately 3.7 to 5 (Table 1). The asymptotic selectivity pattern was based on Northeast groundfish fishery selectivity patterns. The change in the selectivity pattern when selectivity varied through time approximated an increase in mesh size in the fishery to avoid younger fish.

Two different sources of stock assessment misspecification leading to retrospective patterns were considered, temporal changes in natural mortality and misreported catch. The degree to which natural mortality and unreported catch changed through time was determined by attempting to achieve an average Mohn's rho of approximately 0.5 for SSB when an SCAA model (i.e., configured using WHAM) was used to fit the simulated data. We also fit the same SCAA configuration to data without misspecified M or catch to verify that retrospective patterns were not present on average (see Supplemental Materials Figure S1).

A third source of misspecification was also attempted, time varying survey catchability, but this source of misspecification was unable to produce severe enough retrospective patterns and was abandoned.

For the natural mortality misspecification, the true natural mortality changed from 0.2 to 0.32 in scenarios where the fishing history was always overfishing or from 0.2 to 0.36 when the fishing history included a reduction from overfished to F_{MSY}, with the differences between fishing histories necessary to produce the desired retrospective pattern severity (see Supplemental Materials Figures S2 and S3). In each case, natural mortality trended linearly from 0.2 to the higher value between years 31 and 40 of the base period and held constant at the higher level for years 41-50. Natural mortality remained constant at the higher level throughout the feedback period. Those DLMs that required natural mortality as an input parameter used the value from before any change in natural mortality (0.2) because the change in natural mortality is meant to be unknown.

For catch misspecification, a scalar multiple of the true catch observation is provided as the observed catch to the DLMs. The scalar is 0.2 when fishing intensity was always overfishing and for both selectivity patterns, 0.44 when the fishing history included a reduction to F_{MSY} and with time variant selectivity, or 0.40 when the fishing history included a reduction to F_{MSY} and selectivity was time invariant. The shift in scalar trended linearly from 1 to the lower value between years 31 and 40 of the base period and remained at the lower value for years 41-50. These scalars were applied only to the aggregate catch so that they affect all catches at age equally. When catch misspecification was applied in conjunction with a DLM during the feedback period, the true catch in the OM equaled the catch advice provided by the DLM multiplied by the inverse of the scalar multipliers (i.e., the true catches were higher than the DLM catch advice). Thus, when the scalar multipliers were applied to the true catch from the OM in order to provide observed catches at the next application of the DLM, the observed catch equaled the catch advice from the previous application of the DLM, on average. In other words, managers and analysts would be given the perception that the DLM catch advice was being caught by the fishery, when in fact the true catches were always higher. This meant that the source of the retrospective pattern continued in the feedback period. The magnitude of the retrospective pattern in the feedback period varied due to the observation error applied in each realization (See Supplemental Materials Figure S4).

Fourteen methods for setting catches were explored (13 DLMs and the SCAA) and were applied to all 16 scenarios, which created 224 factorial combinations in the study design. For each element of the full factorial combinations, 1,000 simulations were conducted. The simulations used the same random number seeds across all combinations in the study design resulting in the same patterns of recruitment deviations and observation errors. Two DLMs (AIM and ES-Fstable) had two failed simulations each, which were caused by relatively high catch advice (i.e., requiring relatively high F) that triggered errors in the Newton-Raphson iterations used to determine the F that would produce the desired catch. This small number of failures was unlikely to effect results and conclusions, and so were not considered further.

Performance Metrics

Six metrics thought to be of broad interest were reported here, each calculated and reported separately for a short-term (i.e., first six years of the feedback period) and long-term (i.e., last 20 years of the feedback period) period. These metrics were selected to represent the tradeoffs in terms of benefits to the fishery and risks to the stock. The specific metrics reported were: $\frac{SSB}{SSB_{MSY}}$, $\frac{F}{F_{MSY}}$, catch relative to MSY, interannual variation in catch [@amar2010mse], number of years of overfishing ($F > F_{MSY}$), and number of years of the stock being overfished ($SSB < 0.5SSB_{MSY}$).

Results

Overall performance varied widely across methods, and the individual performance of a method was sensitive to the different scenarios explored. Performance for each method was sensitive to the source of the retrospective pattern (missing catch or M), the exploitation history, when in the feedback period the metric was calculated (short- or long-term), and whether or not a 25% buffer was applied when setting the catch advice from a given method. Overall, similar results occurred for the scenarios with one or two selectivity blocks, so the impact of the selectivity scenarios was not discussed further.

Aggregate performance

In Figure **??**, the inner quartiles and medians for all performance measures are shown, calculated across all scenarios combined. In general, methods that resulted in high mean F/F_{MSY} (Figure **??**B) resulted in lower stock biomass (Figure **??**A), more years of overfishing (Figure **??**E) and of being overfished (Figure **??**F), and vice-versa. Higher F values were also associated with higher catches (Figure **??**C), on average, and a greater variability in catch, but there were some methods that produced lower F values that also resulted in high catch variability (CC-FM, CC-FSPR; Figure **??**D).

A number of methods performed poorly overall, resulting in high exploitation rates and low stock size, on average (Figure **??**). These methods include AIM, three of the four expanded survey biomass methods (ES-FM, ES-FSPR, and ES-Fstable), and the Skate method. The Itarget and ensemble methods also resulted in $SSB < SSM_{MSY}$ and $F > F_{MSY}$, on average, though departures from the MSY levels were not as severe as the other methods (Figure **??**). The remaining methods (CC-FM, CC-FSPR, DynLin, ES-Frecent, Islope, Ismooth, and SCAA) were able to limit overfishing and keep biomass above SSB_{MSY}, on average, although for four of these methods (CC-FM, CC-FSPR, DynLin, and Ismooth) biomass was more than 50% higher than SSB_{MSY} (Figure **??**). Principal components analysis of the median values for all methods and metrics resulted in groupings similar to those noted above (see Supplemental Materials Figure S5).

Scenario-dependent performance

The source of the retrospective pattern had a large impact on results for a given method.

The relationship between SSB/SSB_{MSY} and C/MSY is shown across scenarios for the different sources of retrospective error. Stock size and catch (relative to MSY levels) are clustered for many of the methods with no overlap between M and unreported catch sources (AIM, ES-FM, ES-FSPR, ES-Fstable, Itarget, Skate, Ensemble, and SCAA). For all of these methods, SSB/SSB_{MSY} was lower when unreported catch was the source of the retrospective pattern, and C/MSY was also lower except for the Itarget and the SCAA methods compared to the scenarios when increased natural mortality was the source of the retrospective pattern (Figure **??**). The source of the retrospective pattern also had a large impact on the other performance measures (Figure **??**). In general, when unreported catch was the source of the retrospective pattern, interannual variability in catch was higher, overfishing was more frequent and with a larger F/F_{MSY}, and the stock had a higher risk of being overfished compared to the scenarios when increased natural mortality was the source of the retrospective pattern (Figure **??**). Six methods (AIM, ES-FM, ES-FSPR, ES-Fstable, Itarget, Skate, Ensemble) resulted in overfishing in nearly every year of the feedback period (often with very high F/F_{MSY}) when missing catch was the source of the retrospective pattern (Figure **??**B, **??**E). In contrast, all methods except Skate, AIM, and ES-Fstable had low F/F_{MSY}, high SSB/SSB_{MSY}, and few years of being overfished when increased natural mortality was the source of the retrospective pattern (Figure **??**B, **??**A, **??**F). The C/MSY when increased natural mortality was the source of the retrospective pattern varied widely with some DLMs well below 1.0 and others well above (Figure **??**C). The SCAA method also resulted in frequent overfishing in the missing catch scenario, but less so when the stock was more depleted at the start of the feedback period (Figure **??**F).

Exploitation history also impacted the performance of many of the other methods. For four methods (Islope, Ismooth, DynLin and ES-Frecent), exploitation rates were higher when the stock experienced overfishing for the entire base period, but the impact was more dramatic in the short-term. Over time as these methods were used, F declined and remained below F_{MSY} in the long-term (Figure **??**A), allowing stock recovery. The majority of the other methods also resulted in greater exploitation rates in the short-term, though some methods kept $F/F_{MSY} < 1$ regardless of the time-period (CC-FM, CC-FSPR, and SCAA), while others (AIM, ES-Fstable, Skate, Ensemble) kept $F/F_{MSY} > 1$ over the short- and long-term (Figure **??**A). For the ES-FM and ES-FSPR methods, there was not a consistent pattern in exploitation rates when comparing the short- and long-term periods (Figure **??**A).

As expected, application of a buffer to the catch advice resulted in lower exploitation rates compared to no buffer across all methods, but the magnitude of the impact differed by method (Figure **??**B). For poor-performing methods where $F/F_{MSY} >> 1$, the use of a buffer tended to result in greater reductions in F than other methods. Methods like AIM, ES-FM, ES-FSPR, ES-Fstable and Skate all had large reductions in F when the buffer was applied, but the reduction was insufficient to reduce $F/F_{MSY} < 1$ (Figure **??**B). For some methods (CC-FM, CC-FSPR, SCAA), the median F/F_{MSY} was always below 1 with or without the buffer, whereas for other methods (DynLin, ES-Frecent, Islope, Ismooth, Itarget, and Ensemble) there were instances where using a buffer pushed F/F_{MSY} below 1 (though it depended on the exploitation history; Figure **??**B).

The median and interquartile range performance measures reported thus far do not express the full range of results across individual runs, however. When all the simulations are plotted, there is clearly a wide range of possible outcomes for the population, indicating that performance for a particular series of environmental conditions, expressed through recruitment deviations, can vary widely. For example, Figure **??** shows the long-term average SSB/SSB_{MSY} and C/MSY relationship across runs for a single scenario. Different patterns in the relationship between the SSB and catch ratios resulted, with methods falling into two groups. In the first group, there is a near linear relationship between SSB/SSB_{MSY} and C/MSY (AIM, ES-Fstable, ES-FSPR, ES-FM, Itarget, Skate, Ensemble, and SCAA; Figure **??**). In the second group (CC-FSPR, CC-FM, DynLin, ES-Frecent, Ismooth, and Islope) the relationship is more diffuse, with a wide range of C/MSY for a given SSB/SSB_{MSY}. The linear or diffuse relationships persisted across scenarios, although the upper limit of C/MSY

was greatly reduced for the diffuse methods when the buffer was applied to the catch advice. (See Supplemental Figures S6-S21 for these plots across all 16 scenarios and Figures S22-S37 for similar plots showing F/F_{MSY} versus SSB/SSB_{MSY}).

Discussion

A range of data-limited methods for setting catch advice were evaluated for stocks where assessment models may be rejected due to strong, positive retrospective patterns. A method was considered to perform well if it limited overfishing without resulting in light exploitation rates ($F << F_{MSY}$), thereby allowing depleted stocks to recover to SSB_{MSY} (or for healthy stocks to remain there), and for high and stable catches (close to MSY).

Overall, none of the methods evaluated performed best across the scenarios exploring the different sources of the retrospective pattern (unreported catch or increasing M) and different levels of historical fishing intensity. A number of methods did perform well in many cases, however, while others performed consistently poorly, resulting in frequent and intense overfishing ($F >> F_{MSY}$). We performed simulations for a couple of scenarios with no source of retrospective patterns and found the expected result that all DLMs and the SCAA performed better (SSB, F, and catch were all closer to the MSY reference points) than when either source of retrospective patterns was present. Due to the focus of this study, we did not examine the no retrospective source in detail and do not comment on it further.

Currently, in the Northeast U.S., if an assessment model is rejected due to a large rho value in SSB, the catch advice from that model is ignored and some data-limited approach is used. However, the rho-adjusted SCAA model performed better than a number of the alternatives explored here. Therefore, there should not necessarily be an expectation that a data-limited method will perform better than the rejected assessment model. The SCAA only resulted in high exploitation rates ($F >> F_{MSY}$) when unreported catch was the source of the retrospective pattern and for the scenario where $F = F_{MSY}$ at the end of the base period that left the stock in relatively good condition ($SSB \sim SSB_{MSY}$). In contrast, this method was particularly effective when the stock was depleted and there was unreported catch. When M was the source of the retrospective pattern, the rho-adjusted SCAA method typically resulted in light exploitation rates, on average. The light exploitation rates in these cases were likely driven by the combination of using a rho-adjustment, but also using the lower M from the beginning of the base period rather than the higher M that occurred during the feedback period. Using an M value that is too low in a stock assessment will typically bias estimates of biomass and reference points too low, resulting in catch advice that is below target levels [@Johnsonetal2014; @Puntetal2021M]. The consequences of using a value for M that is too low versus too high is also asymmetrical [@Johnsonetal2014], with negative consequences being more severe when M is assumed too high than low, and the results here are consistent with these previous conclusions.

The methods that adjusted recent average catches based on trends in the survey (Ismooth and Islope) performed well overall in terms of catch, stock status, and variation in catch. The method using the expanded survey biomass with the recent exploitation rate (ES-Frecent) also performed well and similarly to Ismooth. The performance of these methods was also generally robust among scenarios, with the exception of when there were unreported catches and the stock was depleted (see below). The generally positive performance of these methods was consistent with @Hilbornetal2002 and @CoxKronlund2008, both of which evaluated a variant of a "hold-steady" DLM. In the case of @Hilbornetal2002, the "hold-steady" DLM policy was designed to adjust catches in order to keep rockfish (*Sebastes spp.*) populations at recently observed index levels, and did so by functioning as a constant escapement harvest control rule where target catches were set to zero below some pre-specified index level. In the variant used by @CoxKronlund2008, catches were adjusted to maintain a sablefish (*Anoplopoma fimbria*) population at a pre-specified index level thought to be sustainable and desirable in terms of meeting fishery objectives (e.g., high catch), but never permitted target catches of zero and so functioned as a constant exploitation rate control rule. The "hold-steady" DLM of @CoxKronlund2008 performed similarly in terms of catch, stock depletion, and variation in catch, as a constant exploitation rate policy where target catch was specified as the product of desired exploitation rate and an estimate of biomass from a SCAA model. This result was robust to uncertainty in initial stock status and steepness [@CoxKronlund2008]. The SCAA model was always correctly specified (i.e., expected to produce unbiased estimates on average), however, and no comparison to the results of this research in the presence of retrospective patterns is possible [@CoxKronlund2008]. The "hold-steady" policy of @Hilbornetal2002 performed similarly to or better in terms of catch and stock status than other harvest control rules that relied on assessment estimates of biomass (i.e., 40:10 and constant F). The performance of the "hold-steady" DLM was also more robust to uncertainty in steepness and to the presence of unreported catch [@Hilbornetal2002]. The performance of the two harvest policies that relied on assessment estimates of biomass (i.e., constant exploitation rate and a "40:10" biomass-based policy) also degraded when the estimates of biomass were biased, which is an issue that does not effect the "hold-steady" DLM [@Hilbornetal2002]. The bias in the assessment estimates considered in @Hilbornetal2002 were not necessarily induced by a retrospective pattern, however, and no consideration of making a rho-adjustment was possible in that study.

The Ismooth method is currently used to set catches for Georges Bank cod [@nefsc19] and red hake (*Urophycis chuss*; @nefsc20). Variations of the ES-Frecent have been used for witch flounder and Georges Bank yellowtail flounder. While the findings here generally support the continued use of the Ismooth and ES-Frecent methods, they may not be well suited for depleted stocks where unreported catches are believed to be an issue. The Ismooth, Islope, and ES-Frecent DLMs produced high Fs and limited stock recovery with unreported catches and when the stock was depleted. While @Hilbornetal2002 and @CoxKronlund2008 did not reach the same conclusion about the "hold-steady" DLM, those studies did not consider initial levels of depletion as low as in this study. These results highlight the importance of accurate catch reporting, as unreported catch can create a negative feedback loop with perpetually high Fs being produced by a management system that seemingly should result in sustainable catch advice.

Three methods were consistently risk-averse across scenarios, limiting the frequency and magnitude of overfishing and resulting in high stock biomass. These methods were the two catch curve options (CC-FM and CC-FSPR) and DynLin. The catch curve methods produced a wider range of average catches across scenarios, and also had greater interannual variability in catches compared to DynLin. While the lower exploitation rates from these approaches may be undesirable due to forgone yield, there may be circumstances where they are preferred. For example, for stocks that are believed to be heavily depleted, low exploitation rates would allow for a more rapid recovery.

A number of methods performed poorly, particularly when catches were unreported. These methods include three of the expanded survey biomass approaches (ES-Fstable, ES-FM, ES-FSPR), AIM, and Skate. The AIM model has been widely used across stocks in the region [@nefsc02a; @nefsc05; @nefsc08], although there is a decreasing trend in its use across model resistant stocks [@nefsc19]. The findings here suggest that alternative approaches should be considered in cases where AIM is still used and there is concern over unreported catches. The Skate method is used to manage the skate complex in the Northeast U.S. (a group of seven co-managed species). Interestingly, six of the seven species are considered in good condition with high survey biomass indices in recent years [@nefmc20]. That the Skate method performed poorly in our analysis but performs well for the skate complex illustrates how the performance of methods in this analysis may be sensitive to the scenarios and species life history considered. As may be the case for the Skate method, the performance of some methods may depend on the condition of the stock when the method is first applied, and less so on life-history. Therefore, care is needed when trying to generalize these results across stocks that may have different life histories, exploitation histories, and without unreported catches or increases in M.

In addition to the analytical differences among the thirteen DLMs, most of the DLMs and control rules had multiple options that could be adjusted to make them more or less risk averse. DynLin had a large number of user defined decision points. Given the large range of options already explored in the study, one suite of options was selected for each DLM-control rule and kept constant for all simulations. Further studies could explore the different options within an individual DLM to understand how they might affect performance.

Many other data-limited methods exist for setting catch advice that were not included in this evaluation, and they vary widely in complexity, data inputs, and assumptions required [e.g., @carruthers2018dlm]. Length based methods were not evaluated to keep the overall number of methods tractable, and due to the availability of age based information in the region. Methods that require only catch data or snap shots of survey data were not considered due to the availability of the relatively long and contiguous Northeast Fisheries Science Center's spring and fall, coastwide bottom trawl surveys, and the fact that "catch only" methods have been shown to perform poorly [e.g., @carruthers2014eval]. Complete catch histories are not available for stocks in the region (i.e., from the inception of fishing). Consequently, methods that required complete catch histories or required assumptions about relative depletion [e.g., DCAC in @maccall2009dca; DB-SRA in @dick2022dsra] were also omitted from consideration. The need for short run-times and the desire for methods that could be reviewed quickly prevented the use of modern state-space production models such as SPiCT [@pedersen2017spict] and JABBA [@winker2018jabba].

The SCAA was confronted with inconsistent data in this study, while the DLMs typically used only a single source of data and thus did not encounter inconsistencies. A recent examination of the data used in assessments in this region similarly found inconsistencies in data streams even before modeling. @wiedenmann2022strange found a negative relationship between relative F (catch/survey) and survey Z for stocks with strong retrospective patterns but the expected positive relationship for stocks without a retrospective pattern. It is exactly this sort of tension that creates retrospective patterns in integrated models, but is not found in DLMs that only use one type of data.

Despite conducting hundreds of thousands of simulations, there are still limitations to our study. We only examined one life history representative of groundfish in the region. We acknowledge that best practice is to select a DLM for a specific life history and fishery condition [e.g., @fischer2020dlm]. As is typically the case with large simulation studies, we were not able to tune any of the DLMs or the SCAA in any given realization, which would occur in practice for an actual stock assessment. We also examined only scenarios that started with Mohn's rho values near 0.5 for spawning stock biomass. This is a strong retrospective pattern, but some stocks in the region have even stronger retrospectives. Performance of the DLMs and SCAA would be expected to degrade with stronger retrospectives, but by how much is still an open area for research. Similarly, sources of retrospective patterns that create different relationships between the true values and estimated values should also be explored [see @deroba2014retro]. To make the results interpretable, we only examined a single source for the retrospective pattern at a time. In reality, there may be more than one factor leading to an observed retrospective pattern. How the multiple sources would interact to influence performance is another topic for future research. Development of harvest control rules specifically for situations where retrospective patterns are found in age-based assessments would also be beneficial. The large number of scenarios examined and the large number of realizations gives us confidence that our results are meaningful in general, but that the performance of any of the DLMs may differ in actual practice.

An interesting finding of this study is the linear versus diffuse patterns between SSB and catch across methods. These patterns have implications for the trade-offs among methods, with linear relationships resulting in more consistent exploitation rates across stock sizes. Therefore, these methods have higher certainty of a given catch at a given stock size. However, they also tended to result in lower stock sizes, on average, across methods. The more diffuse relationships resulted in more variable exploitation rates across stock sizes, with some situations where the population biomass was quite high but the catch was low (relative to

MSY), resulting in a very low F. The reasons behind these different patterns remain unclear, and future work to explore these patterns is warranted.

One of the reasons for the difference in performance between the catch and natural mortality retrospective sources was how the reference points were calculated. In all cases, the initial conditions, including the natural mortality rate, were used to compute the reference points. This decision was made based on the fact that the increase in natural mortality was assumed to be unknown in the simulations. If the increase in natural mortality was known, the age-structured assessments would have accounted for it, different reference points might have been computed [@legault2016increaseM] and there may not have been a retrospective pattern at all [@legault2020rose], and no need to consider alternative DLMs. The reference points for the increased M scenarios would have been different if they were computed using the values from the final year of the base period, but the overall conclusions regarding the different DLMs would not change as this just results in a rescaling of the axis. These results are not shown to reduce confusion regarding the simulations.

Closed-loop simulation is a common tool for examining performance of catch advice from various stock assessment approaches in a feedback setting. It is often used as part of a full management strategy evaluation when working with stakeholders to develop management regulations that make trade offs between near term and long term catches, risk to the fish population, and mixed-fleet allocations [@carruthers2016simpleMPs; @goethel2019mse; @harlyan2019hcr]. We did not conduct a full management strategy evaluation with stakeholder input [@goethel2019stakeholder], but see that as a fruitful next step that could build on the conclusions from our closed-loop work. Using a generic groundfish life-history and monitoring standard performance metrics related to stock status and catch stability, we were able to cull the herd of potential DLMs and we would not carry the consistent poor performers forward for further study. The wide range of expertise reflected in the authorship was by design so that the simulation specifications and performance metrics were broadly useful. Before undertaking a full management strategy evaluation and engaging regional stakeholders, we would want to select a specific stock and jointly identify specific management regulations to be tested [@deroba2019dream]. Results of this work have been presented to both local fishery management councils, with generally positive feedback about the utility of the conclusions for identifying appropriate model approaches when an SCAA is rejected. Our work was similar to all other closed-loop simulations in that it was designed to address a specific situation, including much recent work comparing the performance of data-limited and data rich assessment approaches [e.g., @fulton2016datarich; @sagarese2019dlm; @bouch2020datapoor; @li2022dlm].

This study is a first attempt to identify suitable methods for setting catch advice when stock assessment models are rejected due to large, positive retrospective patterns. Although no single method performed best across scenarios, a number of generally suitable and unsuitable methods were identified under specific conditions. The results of this work can help scientists and managers select a subset of possible options for consideration to set catch advice when assessment models are rejected. The approach developed here can, and should be expanded to consider other cases not explored here, as performance of individual methods are very likely case-dependent.

Acknowledgements

We thank the Index-Based Methods and Control Rules Research Track review panel of Paul Rago (chair), Yong Chen, Robin Cook, and Paul Medley for feedback on preliminary results, three anonymous reviewers and the associate editor for reviewing an earlier version of this work. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect those of NOAA or the Department of Commerce.

Data and Code Availability

All data and code used in this work are available at https://github.com/cmlegault/IBMWG.

References

Tables

Table 1. Maturity-, weight-, and selectivity-at-age of the simulated fish population.

Age	Maturity	Weight (kg)	Fishery Selectivity (before change if applicable)	Fishery Selectivity (after change if applicable)
1	0.04	0.15	0.07	0.02
2	0.25	0.5	0.17	0.05
3	0.60	0.9	0.36	0.12
4	0.77	1.4	0.61	0.27
5	0.85	2.0	0.81	0.50
6	0.92	2.6	0.92	0.74
7	1.00	3.2	0.97	0.89
8	1.00	4.1	0.99	0.96
9	1.00	5.9	1.00	0.99
10+	1.00	9.0	1.00	1.00

Table 2. Naming convention and details of the data-limited methods evaluated.

Method	Details
Ismooth	$C_{targ,y+1:y+2} = \overline{C}_{3,y}(e^{\lambda})$ where $\overline{C}_{3,y}$ is the most recent three year average; $\overline{C}_{3,y} = \frac{1}{3}\sum_{t=1}^{t=3} C_{y-t}$ and λ is the slope of a log linear regression of a LOESS-smoothed average index of abundance (spring and fall) with span = 0.3: $\hat{I}_y = loess(\hat{I}_y)$ and $LN(\widehat{I_y}) = b + \lambda y$
Islope	$C_{targ,y+1:y+2} = 0.8\overline{C}_{5,y}(1 + 0.4e^{\lambda})$ where $\overline{C}_{5,y}$ is the most recent five-year average catch through year $y-1$: $\overline{C}_{5,y} = \frac{1}{5}\sum_{t=1}^{t=5} C_{y-t}$ and λ is the slope of a log-linear regression of the most recent five years of the averaged index.
Itarget	$C_{targ,y+1:y+2} = \left[0.5C_{ref}\left(\frac{\overline{I}_{5,y} - I_{thresh}}{I_{target} - I_{thresh}}\right)\right] \overline{I}_{5,y} \geq I_{thresh}$ $C_{targ,y+1:y+2} = \left[0.5C_{ref}\left(\frac{\overline{I}_{5,y}}{I_{thresh}}\right)^2\right] \overline{I}_{5,y} < I_{thresh}$; C_{ref} is the average catch over the reference period (years 26 through 50): $C_{ref} = \frac{1}{25}\sum_{y=26}^{y=50} C_y$; I_{target} is 1.5 times the average index over the reference period: $I_{target} = \frac{1}{25}\sum_{y=26}^{y=50} \overline{I}_y$; $I_{thresh} = 0.8\ I_{target}$, and is the most recent five year average of the combined spring and fall index: $\overline{I}_{5,y} = \frac{1}{5}\sum_{t=1}^{t=5} \overline{I}_{y-t+1}$
Skate	$C_{targ,y+1:y+2} = F_{rel}\overline{I}_{3,y}$ where $F_{rel} = median\left(\frac{\overline{C}_{3,\mathbf{Y}}}{\overline{I}_{3,\mathbf{Y}}}\right)$ is the median relative fishing mortality rate calculated using a 3 year moving average of the catch and average survey index across all available years ($\mathbf{Y}$): $\overline{C}_{3,y} = \frac{1}{3}\sum_{t=1}^{t=3} C_{y-t}$ and $\overline{I}_{3,y} = \frac{1}{3}\sum_{t=1}^{t=3} I_{y-t+1}$

Method	Details
An Index Method (AIM)	AIM first calculates the annual relative F: $F_{rel,y} = \frac{C_y}{\frac{1}{3}\sum_{t=1}^{t=3}\bar{I}_{y-t+1}}$ and the annual replacement ratio: $\Psi_y = \frac{\bar{I}_y}{\frac{1}{5}\sum_{t=1}^{t=5}\bar{I}_{y-t}}$. These values are used in a regression: $LN(\Psi_y) = b + \lambda LN(F_{rel,y})$ to determine $F_{rel,*}$, which is the value of $F_{rel,y}$ where the predicted $\Psi = 1$ or $LN(\Psi) = 0$. $F_{rel,*}$ is called either the “stable” or “replacement” F, and is used to calculate the target catch: $C_{targ,y+1:y+2} = \bar{I}_y F_{rel,*}$.
Dynamic Linear Model (DynLin)	@Langan2021DLM.
Expanded survey biomass method 1 $F_{40\%}$ (ES-FSPR)	$C_{targ,y+1:y+2} = B_{\bar{I},y}\mu_{targ}$ where $B_{\bar{I}}$ is the average of estimated fully-selected biomass from each survey: $B_{\bar{I},y} = \frac{1}{2}\left(\frac{I_{spr,y}}{q_{spr}} + \frac{I_{fall,y-1}}{q_{fall}}\right)$ and target exploitation fraction, μ_{targ} is calculated as: $\mu_{targ} = \frac{F_{targ}}{Z_{targ}}\left(1 - e^{-Z_{targ}}\right)$; $F_{targ} = F_{40\%}$ and $Z_{targ} = F_{targ} + M$
Expanded survey biomass method 2 $F =$ AIM replacement (ES-Fstable)	Same as the above expanded survey method, but with μ_{targ} equal to the stable exploitation fraction $F_{rel,*}$ calculated using the AIM approach (see above).
Expanded survey biomass method 3 $F = M$ (ES-FM)	Same as the above expanded survey methods, but with the target exploitation rate set to the assumed M: $F_{targ} = M$.
Expanded survey biomass method 4 $F =$ recent average (ES-Frecent)	Same as the above expanded survey methods, but with the target exploitation fraction set to the most recent three year average exploitation fraction: $\mu_{targ} = \frac{\sum_{y-2}^{y}\mu_y}{3}$ $\mu_y = \frac{C_{y-1}}{B_{\bar{I},y}}$

Method	Details
Catch curve Method 1 $F_{40\%}$ (CC-FSPR)	$C_{targ,y+1:y+2} = \frac{F_{targ}}{Z_{avg,y}} B_{cc,y} \left(1 - e^{-Z_{avg,y}}\right)$ where B_{cc} is the estimated biomass: $B_{cc,y} = \frac{C_{y-1}}{\frac{F_{avg,y}}{Z_{avg,y}}\left(1 - e^{-Z_{avg,y}}\right)}$ with $Z_{avg,y} = \frac{Z_{spring,y} + Z_{fall,y-1}}{2}$; $F_{avg,y-1} = Z_{avg,y-1} - M$ and, $F_{targ} = F_{40\%}$. Survey catch at age used in catch curve to estimate Z.
Catch curve Method 2 M (CC-FM)	Same as catch curve method 1 above, but with $F_{targ} = M$.
Ensemble	Median of catch advice provided by AIM, CC-FSPR, ES-Frecent, ES-FSPR, Islope, Itarget, Ismooth, and Skate methods.

Table 3. Summary of the scenarios evaluated within the study design.

Factors	Variants
retrospective source	catch or natural mortality
fishing history	F_{MSY} in second half of base period or overfishing throughout base period (2.5xF_{MSY})
fishery selectivity blocks	constant selectivity or selectivity changes in second half of base period
catch advice multiplier	applied as is from DLM (1) or reduced from DLM (0.75)

List of Figures