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

Christopher M. Legault Richard Bell Elizabeth Brooks Jamie Cournane
Jonathan J. Deroba Gavin Fay Andy Jones Joe Langan
Timothy J. Miller Brandon Muffley John Wiedenmann

Introduction

In the U.S., integrated fisheries stock assessment models that are most frequently age-structured are used to estimate annual stock abundance (biomass), fishing mortality rates, and management reference points (Maunder and Punt 2013). 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 (Mohn 1999). 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 (Punt et al. 2020).

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. 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 (Northeast Fisheries Science Center (NEFSC) 2002, 2005, 2008, 2015a, 2015b, 2017, 2019; Deroba et al. 2010), managed under NOAA Fisheries and the New England Council's Northeast Multispecies (Groundfish) fishery management plan.

The magnitude of the retrospective pattern is typically measured with a statistic called Mohn's rho (Mohn 1999). Mohn's rho can be used to adjust terminal year estimates of biomass in anticipation that the retrospective pattern will persist, and so 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 (Brooks and Legault 2016).

There is no formal criteria in the region for rejecting an assessment based on Mohn's rho, but large, positive values of rho (especially those persisting) 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*) (Deroba et al. 2010; Legault et al. 2014; Northeast Fisheries Science Center (NEFSC) 2015a, 2015b). In each of these cases, and another where the assessment rejection was not based on the retrospective pattern (black sea bass, *Centropristis striatus*; Northeast Fisheries Science Center (NEFSC) 2012), the Councils have relied on a variety data-limited approaches for setting catch advice for these stocks (McNamee et al. 2015; Northeast Fisheries Science Center (NEFSC) 2015a, 2015b; Wiedenmann 2015). 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 (Wiedenmann et al. 2019). In addition, large, positive retrospective patterns persist for a number of other stocks in the region (Northeast Fisheries Science Center (NEFSC) 2019), raising concerns that additional stocks may rely on data-limited approaches in the future. Therefore, there is an immediate need to identify suitable data-limited approaches for setting catch advice and for stocks with age-based assessments that did not pass review.

We developed a management strategy evaluation (MSE; e.g., Punt et al. 2016) to evaluate the suitability of alternative data-limited methods for setting target catches when age-based stock assessments fail. In particular, focus was placed on methods that use survey indices of abundance, or more generally, index based methods (IBMs).

Methods

Overview

The MSE used here attempted to approximate a process where an age-based assessment was rejected due to a retrospective pattern, requiring catch advice to be determined using an IBM. 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 IBMs. After the base period, a given management approach (i.e., IBM) was applied to set the target catch for the stock, which is then removed from the population with some degree of implementation error. 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 IBMs could be compared among several sources of uncertainty that are especially common in the northeast US, but relevant more broadly. The set of OMs included two versions with time varying dynamics in the last 10 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 an IBM. The details of each of these components are described in sections below.

Operating and Observation Models

The Woods Hole Assessment Model (WHAM; Miller and Stock 2020; Stock and Miller 2021) was used as the basis for the OM in the MSE. WHAM is an R package and the general model is built using the Template Model Builder package (Kristensen et al. 2016). 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 IBMs 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 IBMs 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 an IBM and harvest control rule that used the simulated observations to make catch advice. Catch advice from a given combination of IBM 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 data, to have the simulated population respond to the IBM, thereby completing the closed-loop simulation aspect of an MSE.

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 equaled values intended to be groundfish-like for 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 was 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. All these stock-recruitment values were based on an average of groundfish parameters estimated for the region.

Index Based Methods Explored

The range of IBMs 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 IBMs 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 Wiedenmann et al. (2019) showed they performed well in application to groundfish stocks. Two additional IBMs (Islope and Itarget) not currently used in the region were also evaluated, as these have been tested in other applications and shown promise (Geromont and Butterworth 2015a, 2015b, Carruthers et al. 2015, Wiedenmann et al. 2019). An ensemble of models was also considered based on recent findings that improved performance can result from combining the results from multiple models (Anderson et al. 2017, Rosenberg et al. 2017, Spence et al. 2018, Stewart and Hicks 2018). The catch advice from the ensemble approach equaled the median of the catch advice from a range of other methods (Table 2). The DLM 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 IBM (i.e., CC- and ES-) and including them all may have unduly overweighted the performance of the ensemble towards these methods. In these cases, the methods 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). The performance of each method was compared using a range of metrics with data that would lead to retrospective patterns in an age-based stock assessment (see below).

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., Carruthers and Hordyk, 2018). 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. Complete catch histories are not available for stocks in the region (i.e., from the inception of fishing). Furthermore, assumptions of surplus production models are likely violated due to time varying productivity (e.g., in recruitment or natural mortality), and surplus production model fits resulted in different estimates of biomass over time compared to age-based assessments for many stocks (Wiedenmann et al. 2019). Consequently, methods that required complete catch histories, assumed underlying surplus production population dynamics, or required assumptions about relative depletion (e.g., DCAC in MacCall 2009; DB-SRA in Dick and MacCall 2011) were also omitted from consideration.

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 $y-1$. In practice, the data-limited methods that have been applied have used an average of the spring and fall index, and that approach was followed here.

If a method for setting catches uses an average of spring and fall, the average index in year y included the spring data in year y and the fall data in year $y-1$:

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

Control Rules

Most IBMs 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. Lack of clarity exists, however, on whether the catch advice from IBMs should be treated as an overfishing limit (OFL) or an acceptable biological catch (ABC). OFLs are equated to the catch that would result from applying F_{MSY}, whereas an ABC is a catch reduced from the OFL to account for scientific uncertainty. Each IBM was evaluated using two "harvest control rules": 1) the catch advice from a given IBM was applied directly and assumed to serve as a proxy for the catch associated with F_{MSY}, thereby being equated to an OFL (catch multiplier = 1), and 2) the catch advice from a given IBM was reduced by 25% to account for unspecified scientific uncertainty, thereby being equated to an ABC (catch multiplier = 0.75). Catches were reduced by 25% to approximate an ABC because using the catch associated with 0.75 F_{MSY} is a common default ABC control rule in the region.

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 IBMs. Although virtual population analysis (VPA) are 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). Catch advice was determined by specifying fully-selected $F = 0.75F_{40\%}$, always assuming M=0.2.

Study Design

In addition to the two control rules applied for each IBM 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. 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; referred to as "*OF*" below) 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 (referred to as "*KF*" below). These patterns in fishing mortality rate were based on observed patterns for Northeast groundfish (Wiedenmann et al. 2019). 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, asympototic, fishery selectivity in the base and feedback periods (referred to as "*S1*" below), 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 (referred to as "*S2*" below; 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 (Figure 1.2). 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 for the *OF* fishing history or to 0.36 for the *KF* fishing history, with the differences between fishing histories necessary to produce the desired retrospective pattern severity. In each case, natural mortality trended linearly from 0.2 to the higher value between years 30 and 40 of the base period. Natural mortality remained constant at the higher level throughout the feedback period. Those IBMs that required a natural mortality rate 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 IBMs. The scalar is 0.2 for fishing intensity *OF* and both selectivity patterns, 0.44 for fishing intensity *KF* and selectivity scenario *S2*, or 0.4 for fishing history *KF* and selectivity *S1*. The shift in scalar trended linearly from 1 to the lower value between years 30 and 40 of the base period. These scalars were applied only to the aggregate catch so that they affect all catches at age equally.

Fourtneen methods for setting catches were explored (13 IBMs 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. Two IBMs (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 that 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. A naming convention was developed to more easily label and track results among scenarios (Table 3).

Some sensitivity runs were also conducted with all sources of retrospective pattern removed for two of the scenarios. All the IBMs, except DLM and SCAA were applied to these sensitivity runs.

Performance Metrics

A total of 50 performance metrics were recorded during the simulations, but many were redundant and displayed similar tradeoffs among the IBMs and SCAA model. So 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: mean catch relative to MSY, mean interannual variation in catch (A'mar et al., 2010), mean $\frac{SSB}{SSB_{MSY}}$, mean number of years among simulation with SSB less than half SSB_{MSY}, mean number of years among realizations that fully-selected fishing mortaity was greater than the F_{MSY}, and mean $\frac{F}{F_{MSY}}$.

Results

Which Figures do we want to include? Take away from peer review? Which performance measures do we want to focus on? (need to pick probably ≤ 6)

Are we OK with the coarse overview of results, or are more details needed?

Metrics to focus on:

Catch status (realtive to something)

variability over time in catch advice

Biomass status (relative to something

Risk of low population size

Frequency of overfishing

Magnitude of F

Ratios of MSY reference points

The ratio of the mean SSB, F, or catch to its respective MSY reference point showed differences among the IBMs by scenarios, with some factors having larger impact than others. Figures 4.3-4.5 show the mean SSB, F, and catch ratios in the long term (i.e., final 20 years), respectively, while Figures 4.6-4.8 show the ratios in the short term (ie., first six years). All six figures have the IBMs sorted so that the best (largest SSB and catch ratios, smallest F ratios) are at the top based on the mean across all 16 scenarios. The plots show similar patterns for 1 or 2 selectivity blocks in both the short and long term. The plots also generally show similar patterns for fishing histories in the long term, but there are differences in the short term, as expected due to the different starting conditions. The catch multiplier often had the expected effect of reducing catch in the short term, but could sometimes result in higher average catch in the long term due to the larger SSB and lower F. The retrospective source had a large impact on the ordering of the IBMs, with groups of IBMs having either high or low performance for either catch or M, but rarely both. One group of IBMs contains the CC-FSPR, CC-FM, DLM, PlanB, ES-Frecent, and Islope which performed well in terms of both SSB and F in both the short and long term, while the other group contains Skate, AIM, ES-Fstable, ES-FSPR, ES-M, Ensemble, and Itarget which performed well in terms of catch in both the short and long term. In the long term, the SSB ratio was above 1 for the M retrospective source for all IBMs, while the catch retrospective source depended on IBM group as to whether it was above 1 or not. The IBM group that performed well for SSB ratios was able to rebuild the stock above SSBmsy on average in the long term, while the other IBM group was not. Thus, if a stock is thought to be in poor condition, the IBMs in the group that performed well in terms of rebuilding would be preferred to the IBMs in the other group.

The distribution of 16 scenarios by IBM or 13 IBMs by scenario can be used to summarize the metrics. For example, the short and long term SSB/SSBmsy distributions are shown in Figures 4.9-1.10. Similar figures for all the metrics are available in Appendix 6. These plots show the groupings of IBMs and influence of different scenarios on those groupings, but in a more concise way than the 8 panel plots. This allows all the metrics to be presented in Appendix 6.

The distributions of mean values do not express the full range of results, however. When all the simulations are plotted, there is clearly a wide range for each ratio, indicating that performance for a particular series of environmental conditions, expressed through recruitment deviations, can vary widely. For example, Figure 4.11 shows the SSB/SSBmsy and catch/MSY relationship for scenario CF1A (ie., catch retrospective source, Fmsy in second half of base period, constant selectivity block, and catch multiplier equal to 1.0) in the long term for the 1,000 simulations. Note the plots for the remaining 15 scenarios as well as the are available in Appendix 6. The same groups of IBMs as noted above display different patterns in the relationship between the SSB and catch ratios. While all IBMs have large ranges for both ratios, Skate, AIM, ES-Fstable, ES-FSPR, ES-M, Ensemble, and Itarget have nearly linear relationship while CC-FSPR, CC-FM, DLM, PlanB, ES-Frecent, and Islope have a much more diffuse relationship. This pattern by IBM group is consistent across the different scenarios. These linear or diffuse relationships have implications for the trade-offs among IBMs, with linear relationships having higher certainty of performance but lower population sizes on average. The more diffuse relationships can also result in situations where the population is quite high but the catch is low relative to MSY, meaning the F is quite low.

Examination of the simulation plots in Appendix 6 also demonstrates some of the changes in results by the factors. For example, toggling between pages 27 and 28 (A vs R catch advice multipliers) shows that reducing the catch advice has a big impact on the vertical distribution of the diffuse relationship IBMs (much lower for R than A), while the linear relationship IBMs don't change as much but do appear to move a little to the right and maybe even up. This might occur because the diffuse relationship IBMs with reduced catch multipliers are seeing a population bouncing around an average value, meaning catch advice should be about the same, but the catch advice multiplier of 0.75 keeps reducing it.

Another way to explore the impact of the factors is to make so-called "confetti plots" where the mean value of a metric is shown for each IBM and scenario combination but the points are colored by the factor. For example, Figure 4.12 shows the mean value from the 1,000 simulations for six SSB metrics for the 208 combinations of IBM and scenario with the color of the point determined by the retrospective source. Here the differences are clearly seen between catch and M as the retrospective source for most of the metrics. In contrast, the same plot except the points are colored by the fishing history during the base period shows much more interspersed results (Figure 4.13). The full set of "confetti plots" by metric and factor are provided in Appendix 6.

Risk issues

The average SSB and F relative to their MSY reference points are indicative of the expected status of population under different combinations of IBM and scenario, but other metrics can also be used to examine risk. Specifically, the "*is*" metrics can be used to examine the probability that an event will occur at least once during the period. For example, the average value of the SSB metric l_is_less_05_bmsy from the 1,000 simulations provides the probability that the SSB falls below half SSBmsy, meaning the stock would be declared overfished, at least once during during the last 20 years of the simulation. Similarly, the average value of the F metric l_is_gr_fmsy from the 1,000 simulations provides the probability that the F falls above Fmsy, meaning the stock would be declared undergoing overfishing, at least once during the last 20 years of the simulation. The number of times that overfished or overfishing status happened can be found using the associated "*n*" metrics. This allows consideration of how often such an event happened on average. The use of the "avg" metrics of SSB and F relative to their MSY reference points then includes the magnitude of the difference as well, but not the number of years. Consideration of the metrics together allows for a more complete understanding of the performance of the IBMs across scenarios than using only a single metric. These results can be seen in the R Shiny app as well as through a number of different plots in Appendix 6. The IBMs that have the diffuse relationship between SSB/SSBmsy and catch/MSY performed better than the IBMs that have the linear relationship for these metrics.

Catch stability

While overfished and overfishing status are regulatory issues, there are other aspects of performance that may be of interest to managers. One commonly mentioned is the stability of catch advice. This was explored in these simulations through the use of the "*iav*" metrics for catch. The interannual variability tries to distinguish between an IBM and scenario combination that has little change from one assessment to the next compared to an IBM and scenario that fluctuates wildly from one assessment to the next, even if they have the same mean value. These results can be seen in the R Shiny app as well as through a number of different plots in Appendix 6. Generally, the IBMs that have the diffuse relationship between SSB/SSBmsy and catch/MSY performed better with lower catch variability than the IBMs that have the linear relationship for this metric. The exceptions to this general rule are the two CC methods, which performed poorly according to this metric.

Ensemble

By design, the Ensemble model generally had performance that fell in the middle of the orderings for metrics. It had an equal number of IBMs from the two groups (diffuse or linear relationships) of IBMs. This resulted in having an overall performance more similar to the IBMs with the linear relationship because the variability in the diffuse relationship IBMs could offset each other. The Ensemble did perform better than the other linear relationship IBMs in terms of catch stability, as would be expected. So there could be benefits to using an Ensemble approach if managers are interested in trying to trade off the benefits from both types of IBMs, although it generally followed the results of the linear relationship IBMs so the amount of trade off is limited in these simulation results. The performance of the Ensemble can be seen in the R Shiny app as well as through a number of different plots in Appendix 6.

No Retrospective

The no retrospective sensitivity analysis consists of the scenarios CF1A, CO1A, MF1A, MO1A, NF1A, and NO1A for all the IBMs except DLM. The performance of IBMs did not always improve when there was no source of retrospective error. Some of this was due to the fact that the starting conditions were different from the M retrospective source due to the changing reference points for the latter scenarios. In the long term, the average SSB/SSBmsy and catch/MSY were generally closer to 1.0 than either the catch or M retrospective sources (Figure 4.14). This demonstrates a weakness with the scoring algorithm used in this study, values well above SSB or MSY reference points are scored higher than values close to the reference points. This could be taken into account by developing alternative algorithms for deriving the score, such as mean distance from the reference point with a penalty for being on the bad side of the reference point. This would require additional input from managers about their preferences, so was not pursued in this study, but could be done in future analyses.

Despite the shortcomings of the scoring algorithms, there was some change in the ordering of the IBMs when only the no retrospective scenarios were considered, but generally the same groupings held as were seen in the base analyses. See Appendix 6 for some sample scores using the noretro set and the scorer app to create additional results using other sets of metrics.

The performance of the IBMs when no retrospective source is present can perhaps be most clearly seen in the equivalent of Figures 4.3-4.8, where the points represent the mean values from the 1,000 simulations for each IBM and scenario (Figures 4.15-4.20). Note that due to the limited number of scenarios, there are fewer panels in these plots. The long term SSB/SSBmsy for the no retrospective source showed generally good performance among IBMs, although the Skate, AIM, and ES-Fstable methods resulted in a mean value below 0.5 for the fishing history of overfishing throughout the base period. Surprisingly, the long term F/Fmsy mean values were above 1.0 for all the IBMs in the no retrospective source scenarios. This may be due to the averaging across years and the fact that F could go well above Fmsy, but was limited at 0 in how far below Fmsy it could go. Despite the high mean values of F/Fmsy, the no retrospective source performed better than the catch retrospective source for nearly all IBMs. The M source performed better than the no retrospective source for F/Fmsy, but this is most likely due to the high Fmsy values associated with the increased M rate. The long term catch/MSY for the no retrospective source did not have any of the very low values seen for some of the IBMs in the catch retrospective source, and did generally similar to the M retrospective source despite having much higher MSY values. The three short term plots demonstrate the importance of the starting conditions as the fishing history scenarios were often quite different.

The 1,000 point plots for the no retrospective source scenarios (pages 27-62 of tables_figures_noretro.pdf in tables_figs folder) (or bag plots if we can generate them) were not that different from the associated catch and natural mortality retrospective source. The diffuse patterns tended to be less so, and the linear patterns were moved so that they more closely intersected the (1,1) point. These plots are provided in Appendix 6, along with a large number of plots similar to those from the base scenarios.

SCAA

The SCAA sensitivity analysis used scenarios CF1A, CO1A, MF1A, and MO1A. Note, the no retrospective source scenarios were not included due to time limitations. The SCAA model performed better than all the IBMs when the long term SSB, F, and catch relative to their MSY reference points was used as the scoring metric (Figure 4.X1). While the superior performance of the SCAA model held for many metrics, it did not hold for them all. For example, the set of metrics containing the interannual variability during the entire feedback period and the short term catch/MSY had SCAA in the lower half of the IBMs order (Figure 4.x.2).

The performance of the SCAA model can perhaps be most clearly seen in the equivalent of Figures 4.3-4.8, where the points represent the mean values from the 1,000 simulations for each IBM and scenario (Figures 4.X3-4.X8). In the long term, the SCAA model performed near the top of the ordered list, with no IBM consistently performing better than it. In the short term, the SCAA model's performance varied by the fishing history, with some metrics doing well for one fishing history but not the other, leading to a middling performance across these three metrics.

The SCAA model had a near linear relationship between the SSB/SSBmsy and catch/MSY points, with better performance for the M than catch retrospective source (Figure 4.X9). The full suite of 1,000 point plots for the SCAA scenarios are available in Appendix 6 (pages 27-62 in tables_figures_scaa.pdf in the tables_figs folder). The full suite of figures for the SCAA sensitivity analysis is available in Appendix 6.

Candidates for consideration

Overall, none of the IBMs considered in these simulations performed better than the rho-adjusted SCAA model. So in situations where an SCAA model is rejected due to a strong retrospective pattern, there should not be an expectation that an index based method will perform better than the rejected model. These simulations were by necessity limited in scope, so it is not clear that this will always be the case, especially if the retrospective pattern is much larger than examined in this study.

There were two groups of IBMs that performed similarly. In situations where the stock is felt to be in poor condition, CC-FSPR, CC-FM, DLM, PlanB, ES-Frecent, and Islope should be candidates for consideration because they had better performance rebuilding an overfished stock. In situations where the stock is felt to be in good condition, Skate, AIM, ES-Fstable, ES-FSPR, ES-M, Ensemble, and Itarget should be candidates for consideration because they had higher short term catch.

Discussion

Acknowledgements

References

Brooks, E.N., and Legault, C.M. 2016. Retrospective forecasting – evaluating performance of stock projections for New England groundfish stocks. Canadian Journal of Fisheries and Aquatic Sciences **73**(6): 935–950. doi:10.1139/cjfas-2015-0163.

Deroba, J., Shepherd, G., Gregoire, F., and P. Rago, J.N. amd. 2010. Stock assessment of Atlantic mackerel in the Northwest Atlantic for 2010. Transboundary Resources Assessment Committee, Reference Document 2010/01. 59 p.

Legault, C.M., Alade, L., Gross, W.E., and Stone, H.H. 2014. Stock Assessment of Georges Bank Yellowtail Flounder for 2014. TRAC Ref. Doc. 2014/01. 214 p. Available from http://www.nefsc.noaa.gov/saw/trac/.

Maunder, M.N., and Punt, A.E. 2013. A review of integrated analysis in fisheries stock assessment. Fisheries Research **142**: 61–74. doi:10.1016/j.fishres.2012.07.025.

McNamee, J., Fay, G., and Cadrin, S. 2015. Data Limited Techniques for Tier 4 Stocks: An alternative approach to setting harvest control rules using closed loop simulations for management strategy evaluation.

Miller, T.J., and Stock, B.C. 2020. The Woods Hole Assessment Model (WHAM). Available from https://timjmiller.github.io/wham/.

Mohn, R. 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. ICES Journal of Marine Science **56**(4): 473–488. doi:10.1006/jmsc.1999.0481.

Northeast Fisheries Science Center (NEFSC). 2002. Assessment of 20 Northeast groundfish stocks through 2001: a report of the Groundfish Assessment Review Meeting (GARM), Northeast Fisheries Science Center, Woods Hole, Massachusetts, October 8-11, 2002. Northeast Fisheries Science Center, Woods Hole, Massachusetts, October 8-11, 2002. Northeast Fish. Sci. Cent. Ref. Doc. 02-16. Available from National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Northeast Fisheries Science Center (NEFSC). 2005. Assessment of 19 Northeast groundfish stocks through 2004. 2005 Groundfish Assessment Review Meeting (2005 GARM), Northeast Fisheries Science Center, Woods Hole, Massachusetts, 15-19 August 2005. Northeast Fish. Sci. Cent. Ref. Doc. 05-13; 499 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026 or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Northeast Fisheries Science Center (NEFSC). 2008. Assessment of 19 Northeast Groundfish Stocks through 2007: Report of the 3rd Groundfish Assessment Review Meeting (GARM III), Northeast Fisheries Science Center, Woods Hole, Massachusetts, August 4-8, 2008. Northeast Fish. Sci. Cent. Ref. Doc.08-15; 884 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026 or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Northeast Fisheries Science Center (NEFSC). 2012. 53rd Northeast Regional Stock Assessment Workshop (53rd SAW) Assessment Report. Northeast Fish. Sci. Cent. Ref. Doc.12-05; 559 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Northeast Fisheries Science Center (NEFSC). 2015a. Stock Assessment Update of 20 Northeast Groundfish Stocks Through 2014. Northeast Fish. Sci. Cent. Ref. Doc.15-24; 251 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Northeast Fisheries Science Center (NEFSC). 2015b. 60th Northeast Regional Stock Assessment Workshop (60th SAW) Assessment Report. Northeast Fish. Sci. Cent. Ref. Doc.15-08; 870 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Northeast Fisheries Science Center (NEFSC). 2017. Operational Assessment of 19 Northeast Groundfish Stocks, Updated Through 2016. Northeast Fish. Sci. Cent. Ref. Doc.17-17; 259 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Northeast Fisheries Science Center (NEFSC). 2019. Operational Assessment of 14 Northeast Groundfish Stocks, Updated Through 2018. Northeast Fish. Sci. Cent. Ref. Doc.XX-XX; XXX p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Punt, A.E., Butterworth, D.S., Moor, C.L. de, De Oliveira, J.A.A., and Haddon, M. 2016. Management strategy evaluation: Best practices. Fish and Fisheries **17**(2): 303–334. doi:10.1111/faf.12104.

Punt, A.E., Tuck, G.N., Day, J., Canales, C.M., Cope, J.M., de Moor, C.L., De Oliveira, J.A., Dickey-Collas, M., Elvarsson, B.Þ., Haltuch, M.A., Hamel, O.S., Hicks, A.C., Legault, C.M., Lynch, P.D., and Wilberg, M.J. 2020. When are model-based stock assessments rejected for use in management and what happens then? Fisheries Research **224**: 105465. doi:10.1016/j.fishres.2019.105465.

Stock, B.C., and Miller, T.J. 2021. The woods hole assessment model (wham): A general state-space assessment framework that incorporates time- and age-varying processes via random effects and links to environmental covariates. Fisheries Research **240**: 105967. doi:10.1016/j.fishres.2021.105967.

Wiedenmann, J. 2015. Application of data-poor harvest control rules to Atlantic mackerel. Final report to the Mid-Atlantic Fishery Management Council. Final report to the Mid-Atlantic Fishery Management Council.

Wiedenmann, J., Free, C.M., and Jensen, O.P. 2019. Evaluating the performance of data-limited methods for setting catch targets through application to data-rich stocks: A case study using northeast u.s. Fish stocks. Fisheries Research **209**: 129–142. doi:10.1016/j.fishres.2018.09.018.

Tables

Table 1.

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	0.6	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	1.0	3.2	0.97	0.89
8	1.0	4.1	0.99	0.96
9	1.0	5.9	1.0	0.99
10+	1.0	9.0	1.0	1.0

Table 2.

Method	Details
loess smooth	$C_{targ,y+1:y+2} = \overline{C}_{3,y}(e^{\lambda})$ 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 $\overline{I}_{3,y} = \frac{1}{3}\sum_{t=1}^{t=3} I_{y-t+1}$
An Index Method (AIM)	AIM first calculates the annual relative F: $F_{rel,y} = \frac{C_y}{\frac{1}{3}\sum_{t=1}^{t=3} \overline{I}_{y-t+1}}$ and the annual replacement ratio: $\Psi_y = \frac{\overline{I}_y}{\frac{1}{5}\sum_{t=1}^{t=5} \overline{I}_{y-t}}$. 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: $C_{targ,y+1:y+2} = \overline{I}_y F_{rel,*}$.
Dynamic Linear Model (DLM)	Supplement XX or citation.
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, $F_{targ} = F_{40\%}$.
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, CCFSPR, ES-Frecent, ES-FSPR, Islope, Itarget, PlanB, and Skate methods.

Table 3.

Position	Factors	Values
1	retrospective source	C = catch M = natural mortality N = none
2	fishing history	F = Fmsy in second half of base period O = overfishing throughout base period
3	fishery selectivity blocks	1 = constant selectivity 2 = selectivity changes in second half of base period
4	catch advice multiplier	A = applied as is from IBM R = reduced (multiplied by 0.75) from IBM

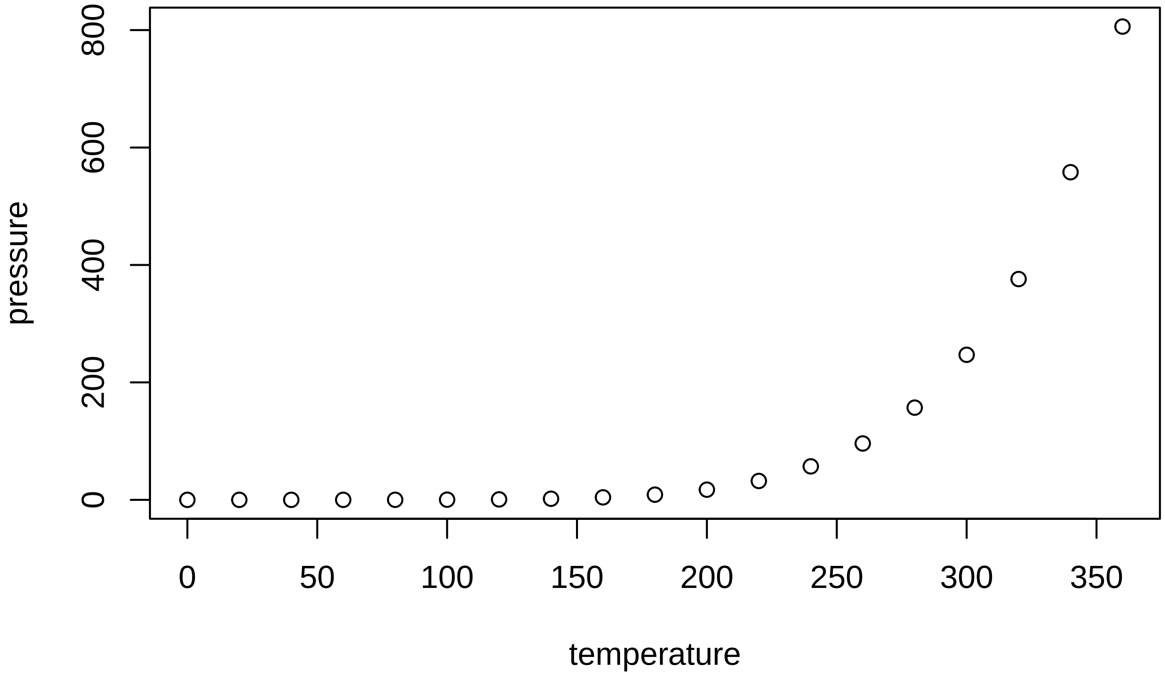
Example embedding R table

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Example Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.