Draft report for peer review only

Draft Report of the Index Based Methods Working Group

Submitted for peer review November 20, 2020

Executive Summary

TOR 1. Develop methods to create data that if assessed with standard age-based approaches (e.g., VPA or ASAP) could exhibit a strong retrospective pattern.

The stock assessment program WHAM (Woods Hole Assessment Model) was used to generate data that exhibited strong retrospective patterns when assessed with an age-based approach. There were two alternative sources for the retrospective pattern; unaccounted catch and an unknown increase in natural mortality. The magnitudes associated with each source were modified according to the specific scenario to produce a Mohn's rho of approximately 0.5 for spawning stock biomass. This is a strong retrospective pattern and comparable to retrospective patterns seen in some stock assessments in the region. There were eight combinations of retrospective source, fishing history, and number of fishery selectivity blocks (each factor had two levels) that formed the 50 year base period. An index-based assessment was applied every other year during the next 40 years of the feedback period. The data available to the index-based methods contained noise in the observations, as would occur in actual stock assessments. For each scenario, index-based method, and control rule combination, 1,000 closed-loop simulations were performed.

TOR 2. Identify a number of index-based methods and a range of harvest control rules for use in closed-loop simulation, using index-based data resulting from ToR 1.

Thirteen index-based methods (12 individual methods and one ensemble approach) were selected from the large number of available methods for use in the closed-loop simulations. The selected methods reflect local use, common use elsewhere, or a newly developed approach being proposed for the region. When applied, the index-based methods used the natural mortality rate from the initial base period to reflect that the change in natural mortality was not known in the age-based assessment. There were two harvest control rules used. The first method simply applied the catch advice resulting from the index-based method directly. The second method

reduced the catch advice (multiplied it by 0.75) to approximate a decision that the original catch advice was an overfishing limit and that the acceptable biological catch should be reduced from it. In neither case did the size of the population modify the harvest control rule, as is sometimes done with age-based assessments, because many of the index-based methods do not have a way of determining current population abundance relative to a reference point.

Table of Index based methods (IBM) with short description. A * in the description indicates the IBM is used in the Ensemble, while a # indicates that the IBM is used in a local assessment.

IBM	Description
AIM	An Index Method *#
CC-FSPR	Catch Curve F40%SPR *
CC-FM	Catch Curve F=M
DLM	Dynamic Linear Model
Ensemble	Combination of models
ES-FM	Expanded survey biomass F=M
ES-Frecent	Expanded survey biomass recent F*#
ES-FSPR	Expanded survey biomass F40%SPR *#
ES-Fstable	Expanded survey biomass stable F
Islope	common trend based IBM *
Itarget	common level based IBM *
PlanB	survey smoother *#
Skate	catch/B driven *#

TOR 3. Identify metrics from the index-based assessment results that could be used in evaluations of trade-offs in performance among harvest control rules and index-based methods.

A total of 50 performance metrics were collected from the simulation results. All were based on the true population values that are known because this is a simulation exercise (would not be known in application of the index-based methods to real data). There were sets of performance metrics for spawning stock biomass, fishing mortality rate, and catch. The metrics were collected for either the short term, the first 6 years of the feedback period, or the long term, the last 20 years of a 40 year feedback period. The metrics included average values compared to the associated reference point, classification of whether an event occurred during the time period (e.g., F goes above Fmsy), or the number of years that an event occurred during the time period. Examples of how these performance metrics could be used to score the performance of the index-based methods are provided, but these are management decisions and are shown for demonstration purposes only.

TOR 4. Evaluate the combinations of index-based methods and control rules using the metrics in ToR 3 to determine candidates for consideration by the Councils or other management authorities.

The index based methods (IBMs) generally formed two groups in terms of performance. The first group consists of CC-FSPR, CC-FM, DLM, PlanB, ES-Frecent, and Islope, while the second group consists of Skate, AIM, ES-Fstable, ES-FSPR, ES-FM, Ensemble, and Itarget. The IBMs within these groups performed similarly in terms of both the mean values and the type of relationship between SSB/SSBmsy and catch/MSY. The first group generally had lower short term catch, lower long term F/Fmsy and higher SSB/SSBmsy than the second group. The second group had a more linear relationship between long term SSB/SSBmsy and catch/MSY than the first group. IBMs in the first group may be more suited for stocks that are thought to be in good condition, while IBMs in the second group may be more suited for stocks that are thought to need rebuilding.

Multiple lines of evidence pointed to the retrospective source and the catch advice multiplier as the most important factor in determining performance of the IBMs. For some performance metrics, fishing history and interaction terms among factors were also important.

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 IBM will perform better than the rejected model.

The reasons for why the IBMs grouped as they did according to the performance metrics is not understood at this time and in need of future research. The results of this study provide a basis for these explorations. Additional questions can also be addressed by the framework developed in this study.

TOR 5. Provide guidance on specific situations that are and are not well-suited for a particular control rule or index-based method identified in ToR 4.

For several IBMs, the variability among metrics was primarily due to the factor related to the cause of the retrospective pattern. Thus, if you cannot identify the likely (or dominant) source of a retrospective pattern (catch or M), then using an IBM sensitive to the retrospective source would be risky to the stock and fishery, and this risk could be avoided by using a method robust to this uncertainty. In this regard, the following methods were more robust to retro type: DLM, PlanB, ES-Frecent, Islope, and to some extent the two catch curve methods. We anticipate these methods would have more robust performance in situations where a retrospective pattern exists similar to that simulated in this project.

A rebuilt stock (long term SSB/SSBmsy > 1) was greatest for the IBMs that were the least sensitive to the source of the retrospective pattern (CC-FSPR and CC-FM, DLM, PlanB, ES-Frecent and Islope). DLM, PlanB, and CC-FM and CC-FSPR achieved the greatest median SSB/SSBMSY, but catch was lowest for CC-FSPR and CC-FM. PlanB, Islope and ES-Frecent had the highest median catch among the methods that achieved rebuilding more than 50% of the time.

Trade-offs in risk (overfishing and rebuilding) and rewards (catch) are inherent in management decisions. Balancing median catch close to MSY in the short-term while still maintaining a probability of at least 50% of achieving rebuilding was possible for the

ES-Frecent, PlanB, DLM, and Islope, although long-term median catch with these methods was far below MSY. This could indicate that these four methods are appropriate short-term models for management advice, but while they are employed other efforts should be invested to return to an age-based model. As noted in TOR 4, none of the IBMs consistently performed better than the SCAA with rho-adjustment, supporting the recommendation to use age-based approaches despite retrospective patterns.

Caveats for the conclusions relate to the large, but by necessity limited, number of IBMs, scenarios, and simulations that could be conducted. The biological and fishery characteristics were based on local groundfish, other species or fisheries may have different performance for the IBMs. A single source, catch or M, and magnitude, 0.5, of retrospective pattern was considered, multiple, changing, and stronger retrospective patterns may lead to different performance of IBMs. The IBMs were applied every other year in the feedback period to mimic local application, but missing surveys could require 3 years before the next assessment could be conducted. By necessity, the IBMs were applied formulaically within the simulations, so review of data meeting IBM assumptions or poor diagnostics were not evaluated during this study, these results should be considered minimum performance of the IBMs. As with all research, many questions were raised during this study. The framework developed for these simulations is well suited to address many of them immediately or with minor modifications. The IBMWG recommends this work be continued to explore the results generated during this study as well as building on these results to address new questions.

TOR 6. Create guidelines for setting biological reference points for index-based stocks.

This term of reference requires additional study. Generally, it is challenging to generate appropriate biological reference points for index-based stocks because they lack production functions that allow examination of trade offs between catch and future population size. For this reason and due to time constraints, only general guidelines and ideas for future research are provided at this time. The framework developed for this project is well-suited for this additional research.

Working Group Process

The working group (Appendix 1) met weekly from March to November via webinar due to travel restrictions associated with the Covid-19 global pandemic. All meetings of the working group were open the public (Appendix 2), with announcements made regarding connection information prior to each meeting through the dedicated website for the working group (https://www.fisheries.noaa.gov/new-england-mid-atlantic/population-assessments/stock-assessment-working-group-index-based-methods-and-control-rules). For each of the 41 meetings (Appendix 3), the chair prepared an agenda and often a set of prompts to allow working group members to contribute thoughts before the meeting. This asynchronous sharing of ideas through Google Docs allowed rapid progress to be made. The working group used a GitHub repository to share code used to run and analyze the simulations (https://github.com/cmlegault/IBMWG). A number of computer resources were used to run the actual simulations. Due to the large size of the resulting files (~300 GB in total), these files could not be stored on the GitHub repository, so were instead stored on Google Drive.

Introduction

In the U.S., age-structured stock assessment models are used when possible to estimate annual stock abundance and fishing mortality rates, as well as management reference points. 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 in the 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 determining 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, positive retrospective patterns in biomass estimates, whereby estimates of biomass are 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 (NEFSC 2002a, 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 known as Mohn's rho (Mohn 1999). Stock assessments where the rho-adjusted (divide the terminal year estimate by one plus Mohn's rho) value is outside the 90% confidence interval of the terminal year estimate of spawning stock biomass or fishing mortality rate are classified as strong retrospectives and the rho-adjusted values 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 for stocks 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; 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;* NEFSC 2012), the Councils have relied on a variety data-limited approaches for setting catch advice for these stocks (McNamee et al. 2015; 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 (NEFSC 2019), raising concerns that additional stocks may rely on data-limited approaches in the future. Additionally, many of these stocks are in rebuilding plans without a mechanism to track rebuilding progress, in the absence

of reference points. A recent challenge also emerged in the 2020 stock assessments when index-based methods in place for some stocks were rejected and alternative index-based methods were pursued (i.e., red hake and Northern windowpane flounder). Therefore, there is an immediate need to identify suitable data-limited approaches for setting catch advice and stock status determination for stocks with age-based assessments that did not pass review.

Stocks assessments in the region are classified as either management track or research track assessments (see Box 1 for more details on the distinction). Research track assessments can focus on individual stocks or they can be topic based. Topic based assessments are meant to provide utility and application in future management track assessments. The Index-Based Methods Working Group (IBMWG) was formed to conduct a topic based research track assessment to evaluate the suitability of a range of data-limited methods for setting target catches for stocks with assessments with strong retrospective patterns. The Terms of Reference for the Index-Based Methods Research Track Stock Assessment are:

1. Develop methods to create data that if assessed with standard age-based approaches (e.g., VPA or ASAP) could exhibit a strong retrospective pattern.
2. Identify a number of index-based methods and a range of harvest control rules for use in closed-loop simulation, using index-based data resulting from ToR 1.
3. Identify metrics from the index-based assessment results that could be used in evaluations of trade-offs in performance among harvest control rules and index-based methods.
4. Evaluate the combinations of index-based methods and control rules using the metrics in ToR 3 to determine candidates for consideration by the Councils or other management authorities.

5. Provide guidance on specific situations that are and are not well-suited for a particular control rule or index-based method identified in ToR 4.
6. Create guidelines for setting biological reference points for index-based stocks.

To address the Terms of Reference the IBMWG developed a management strategy evaluation (MSE) simulation model (e.g., Punt et al. 2016). MSE models are used to evaluate the tradeoffs associated with different management options in the face of uncertainty, and consist of a series of linked submodels (Figure 0.1). The foundation of MSE is the operating model that controls the population and fishery dynamics in the system. The operating model is run for an initial period of time (called the *Base* period here) that controls the historical population dynamics and fishing pressure, and allows for sufficient data to be generated in the observation model to be used in the different assessment / management options being explored in the MSE. After the *Base* period, a given management approach is applied to set the target catch for the stock, which is then removed from the population. This process is repeated at a fixed interval over a number of years in what is called the *Feedback* period (Figure 0.2). In many MSEs, catch advice is based on a stock assessment model that estimates current abundance and management reference points. In data-limited cases, or when an assessment model is rejected, the model output is not used, and catch advice is based on the available data required by a given data-limited approach. Details of the operating model are provided in response to ToR 1, and the data-limited methods explored are provided in response to ToR 2 below. Because the Northeast U.S. has a long time series of data from a fishery-independent survey, the IBMWG focused on methods that can utilize this index of abundance, and we broadly refer to these approaches as index-based methods, or IBMs.

Box 1. New Stock Assessment Process in the New England and Mid-Atlantic Region

The Northeast Region Coordinating Council (NRCC) is comprised of leadership from the Atlantic States Marine Fisheries Commision (ASMFC), Greater Atlantic Regional Fisheries Office (GARFO), Mid-Atlantic Fishery Management Council (MAFMC), New England Fishery Management Council (NEFMC), and the Northeast Fisheries Science Center (NEFSC). The NRCC coordinates science and management activities and resources in the region. A significant responsibility of the NRCC is setting stock assessment priorities and schedules for the region.

Historically, the process to establish stock assessment priorities and schedules was conducted annually on an ad hoc basis and, at times, influenced by outside pressure and demands. In 2018, the NRCC agreed to a new process in determining the region's assessment needs in order to provide for a more strategic and long-term approach with the goal of improving the overall quality of the regions' stock assessments (NRCC 2018). The new process identified two types of stock assessments: management track and research track. Management track assessments provide routine and updated advice on a specified schedule to directly inform management actions. Generally, management track assessments utilize the previously peer reviewed stock assessment for a particular species; however, the new process allows for greater flexibility to incorporate new or updated information and approaches to provide for continued improvements. Research track assessments are comprehensive and complex efforts that are typically carried out over several years in order to develop and consider new research. These assessments consider new data sources and new or different modeling approaches. An approved research track assessment can then be applied in future management track assessments and eventually inform management.

Research track assessments may focus on a single species and individual stocks or may examine a topic or issue that could apply to a broad range of species or assessment types. Topic based research track assessments, should demonstrate utility and application to future management track assessments or management actions. The terms of reference for a research track assessment should specify what and how outputs and outcomes could apply to future management track assessments. The index-based methods research track assessment is the first topic based assessment to be developed, conducted, and peer reviewed through the new NRCC process.

TOR 1. Develop methods to create data that if assessed with standard age-based approaches (e.g., VPA or ASAP) could exhibit a strong retrospective pattern.

Operating Model

We used the Woods Hole Assessment Model (WHAM; Stock and Miller, in review, Miller and Stock 2020; Appendix 4) as the operating model for the simulation study. WHAM is an R package and the general model is built using the Template Model Builder package (Kristensen et al. 2016). WHAM is a stock assessment model used to estimate parameters from real data, but it can also simulate data, including during a projection period, given a set of parameters and projection specifications. We used WHAM operating models to simulate data with known properties during the base period. Catches during the feedback period were iteratively updated based on an index-based assessment method (IBM) and harvest control rule that used the simulated observations to make catch advice during the projection period. We specified the initial base period to be 50 years, labeled 1970-2019, where the age-structured population and catch and index observations were simulated according to user supplied biological and fishery parameters for each scenario. The following period of 40 years, labeled 2020-2059, normally viewed as a projection period in a WHAM assessment model, is here specified as a feedback period. From the beginning of the feedback period, an IBM was applied to generate catch advice for 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.

There were a few different assumptions that defined different operating model scenarios. First there were two different assumptions on history of fully-selected fishing mortality rates:

1) 2.5 x Fmsy(based on M=0.2 and terminal selectivity) for the entire base period (ie., always overfishing) or

2) 2.5 x Fmsy for first half of base period and Fmsy for second half of base period (ie., overfishing the first half and then reduced to Fmsy for second half).

Second, there were two different assumptions about variation in fishery selectivity:

1) constant during the base and feedback period or
2) a change in selectivity after the first half of the base period so that the age at 50% selectivity increased to 5 from approximately 3.7.

Third, we considered two types of misspecification:

1) an unknown change in natural mortality (M) or
2) catch is underreported.

The biological and fishery characteristics used in the simulations were derived from local groundfish stocks. The high level of fully-selected fishing mortality that we assumed for overfishing (F>Fmsy) was based on work by Wiedenmann et al. (2019) (Figure 1.1) looking at harvest rates relative to fishing mortality reference points for Northeast groundfish stocks. These two fishing intensities during the latter half of the base period lead to different starting conditions for each of the simulations.

The two selectivity patterns are based on local groundfish fishery selectivity patterns. The changed selectivity pattern when two selectivity blocks occur reflects an increase in mesh size of the fishery to avoid younger fish (Table 1.1).

The values we used to adjust natural mortality and unreported catch were derived by attempting to achieve a Mohn's rho of approximately 0.5 for spawning stock biomass (SSB) when a statistical catch at age model configuration of WHAM was used to fit the simulated data (Table 1.2). We also fit the same SCAA configuration of WHAM to data without mis-specified M or catch to verify that retrospective patterns were not present on average (Figure 1.2).

For the natural mortality misspecification, the true natural mortality used to simulate the population and observations changes from 0.2 to 0.32 (for fishing intensity history type 1) or 0.36 (for fishing intensity history type 2) linearly between years 2000 and 2009 during the base period and remains at the higher level throughout the feedback period. The IBM uses the

observations and, those IBMs that require a natural mortality rate use the value (0.2) from before any change in natural mortality because the change in natural mortality is meant to be unknown.

For the catch misspecification, a scalar multiple of the true catch observation is provided to the IBM. The scalar is 5 for fishing intensity history type 1 and both selectivity patterns, 2.25 (for fishing intensity type 2 and a change in selectivity) or 2.5 (for fishing intensity type 2 and constant selectivity) linearly from year 2001 to 2010. Note that this scalar is applied only to the aggregate catch so that it affects all catches at age equally.

We used the same set of 1,000 random seeds to initialize simulations for all scenarios. For every random seed, we applied each of the IBMs so that the base period is identical for each of them, but the population and observations during the feedback period vary by IBM due to the differences in catch advice provided by each of the IBMs (Figure 1.3).

In all scenarios, when catch advice provided by the IBM, and increased due to catch misreporting when appropriate, was impossible to catch given the size of the stock at the time, we instead imposed a fully selected fishing mortality of 2. This prevented the population from going negative, as could have happened if the catch advice was simply subtracted from the population.

All operating model scenarios assumed autoregressive deviations between the realized recruitment and that expected from a Beverton-Holt stock recruitment curve (Table 1.3). This stochasticity in recruitment is the only source of variation in the population for a given random seed. The initial numbers at age are specified at equilibrium values with total mortality defined by natural mortality in the first year, fishery selectivity in the terminal year of the base period, and corresponding Fmsy. Therefore, they are the same for all 1,000 seeds of a given scenario, but they differ between scenarios when there are differences in terminal year fishery selectivity. Consistent seeds were used across all the scenarios so that differences in performance among IBMs can be attributed to characteristics of the IBMs themselves and not different sequences of observation or process errors, and such standardization is considered best practice (Punt et al., 2016).

There were a number of scenarios considered in a full factorial design as well as a number of additional scenarios denoted one-offs. The factors included in the full factorial design were: cause of the retrospective pattern, fishing history, fishery selectivity, IBM, and catch multiplier (Table 1.4). This resulted in a total of 2 x 2 x 2 = 8 operating models where 13 IBMs and 2 catch multipliers were applied during the feedback period resulting in 208 scenarios for each of the 1,000 random seeds. Several one-offs were considered (Table 1.5), however due to time limitations, only the no retrospective pattern and statistical catch at age model analyses could be conducted. The statistical catch at age model was configured and estimated using WHAM, but specified to closely match ASAP (Figure 1.4). The purpose of the extra analyses was to examine specific aspects of the simulations without requiring the full number of runs if they had been part of the full factorial design.

Biological reference points

We used the true values of the population and fishery to define biological reference points and associated metrics to evaluate IBM performance (Table 1.6). The MSY reference points were derived for each year during the base period using the fishery and biological characteristics of that year. This resulted in changes between the first and second halves of the base period due to changes in M and fishery selectivity in appropriate scenarios. Similar to the definition of initial numbers at age, the fully selected fishing mortality during the base period was based on Fmsy defined using M = 0.2 and the selectivity during the terminal years of the base period (Figure 1.5). For example, in scenarios with F history pattern = 2, fully selected fishing mortality begins at 2.5 times this Fmsy and reduces to this Fmsy value during the last portion of the base period. Since the changes in the second half of the base period were held at these values during the feedback period, the MSY reference points from the terminal year of the base period, including changes in M, were used for performance metrics in the analyses.

This difference between the reference points used to set the fishing history and those to determine performance metrics led to some counterintuitive starting conditions. Specifically, the

large change in natural mortality required to achieve the desired retrospective pattern led to the fishing mortality ratio compared to the terminal year reference points being below (in some cases well below) one instead of equal to or well above one (Figure 1.5). This led to the SSB ratios being above one for F history scenarios of overfishing then Fmsy in the second half and SSB ratios equal or above one for F history scenarios of overfishing throughout (Figure 1.6). In contrast, the reference points associated with catch as the retrospective source had only minor changes due to fishery selectivity changes (Figure 1.7). Had we used the terminal year MSY reference points when setting the fishing mortality rate histories, the overfishing levels would have been so high as to collapse the population. So instead, we used the F histories from the catch retrospective scenarios directly in the M retrospective scenarios and recognize that the terms associated with the F histories do not necessarily apply for the M retrospective scenarios. The changes in MSY reference points associated with M as the retrospective source have consequences for the performance metrics computed for the feedback period and could impact the IBMs that rely on stationarity assumptions when deriving catch advice (such as knowing the M rate or being able to approximate the biomass reference point).

Feedback period procedure

Within the MSE process, the IBMs were conducted in the middle of the calendar year resulting in different years of information being available for different data inputs. Two surveys were conducted each year to produce an index of abundance, the first occurred 0.25 into the year (spring like survey) and the second at 0.75 of the year (fall like survey). The IBM method conducted in year y had the spring like survey in year y available to it as well as the fall like survey from year y-1 and the catch data from year y-1 (e.g., an assessment that occurred in summer 2014 would have the spring survey data from 2014, the fall survey data from 2013 and the catch from 2013). IBM methods that utilized only a single survey were provided with the average of the two surveys calculated as the mean of the spring survey in year y and the fall survey in year y-1. As no fall survey took place prior to year one, the first year of mean survey index was the mean of the spring survey in year two and the fall survey in year one approximating the index on January 1st of year two. IBM methods that evaluated the catch over

the mean survey index lined up the catch in year y with the survey in year y with survey in year y defined as the mean of the spring survey in year y+1 and the fall survey in year y (e.g. the total catch for the year 2018 would be lined up with the survey index approximated as Jan 1, 2019 (mean of fall 2018 survey and spring 2019 survey)).

Annual aggregate catch and index observations were assumed to be normal after log-transformation. The associated age composition observations were assumed to be multinomial distributed. Coefficients of variation and effective sample sizes are provided in Table 1.7.

TOR 2. Identify a number of index-based methods and a range of harvest control rules for use in closed-loop simulation, using index-based data resulting from ToR 1.

A large number of data-limited methods exist for setting catch advice, and they vary widely in complexity, data inputs, and assumptions required (e.g., Carruthers and Hordyk, 2018). The northeast U.S. is a data rich region with a range of information available including the Northeast Fisheries Science Center (NEFSC) spring and fall coastwide bottom trawl surveys as well as historical catch data. As a result, the Index Based Methods Working Group (IBMWG) focused on a subset of data-limited methods known as index-based methods (IBMs) that utilize time series of fishery catch and CPUE from a survey or fishery, and omitted methods that require only catch data, snap shots of survey data or length data (e.g., constant catch methods or length-based methods). The IBMWG also omitted methods that required complete catch histories (from the inception of fishing), assumed an underlying surplus production population dynamics, or required assumptions about relative depletion. Complete catch histories are not available for stocks in the region, and surplus production Example methods that meet these criteria for omission include Depletion-Corrected Average Catch (DCAC; MacCall 2009), Depletion Based Stock Reduction Analysis (DB-SRA; Dick and MacCall 2011), and the Simple Method for Estimating MSY (SMSY; Martell and Froese 2012). The IBMWG also omitted methods that assume an underlying surplus production model where changes in productivity are driven solely by changes in biomass (e.g., Martell and Froese 2012). Changes in recruitment (Miller et al. 2017; Xu et al. 2018; Tableau et al. 2019) or natural mortality (Pershing et al. 2015) for many stocks in the region violates this assumption, and surplus production fits to survey and catch data result in very different estimates of biomass over time compared to age-based assessments for many stocks in the region (Wiedenmann et al. 2019).

Given the long time-series of trawl survey data, the group focused largely on survey IBMs. The relatively short timeline for the IBMWG restricted the total number of methods that could be evaluated to those that have been used or would be considered plausible for the Northeast Shelf. A list of methods currently used in the region is provided in Table 2.1, with thirteen IBMs selected by the IBMWG for evaluation. The IBMWG deliberately did not include

every approach to limit the overall number of methods explored, but also to allow for exploration of additional options not currently used but that have been explored or applied elsewhere. 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 a retrospective application to groundfish stocks. We did not include length based methods to reduce the overall number of methods explored, and due to the availability of age based information. We also included two additional IBMs (Islope and Itarget) not currently used in the region, 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 full range of methods included in this analysis are detailed below with equations provided in Table 2.2. Each method was examined for their ability to produce sustainable catch advice with data that would lead to large retrospective patterns.

Methods that only use the survey index and catch

Plan B smooth - The Plan B smooth approach has been used to set catch advice for Georges Bank cod since the rejection of the 2015 age-based assessment (NEFSC 2015, 2017, 2019). The Plan B approach combines the spring and fall surveys into an average index, then a LOESS smoother is applied to the average index (with a span = 0.3). The predicted LOESS smoothed values in the final three years are used in a log-linear regression to estimate the slope, and this slope (transformed back to the linear scale) is used to adjust the most recent three year average catch to generate catch advice (Table 2.2).

Islope - The Islope method is similar to the Plan B smooth approach in that it uses recent trends in the combined average index (spring and fall) to adjust the recent average catch up or down. This general method was proposed by Geromont and Butterworth (2015a) and has been evaluated in a number of studies (e.g., Geromont and Butterworth 2015a, 2015b, Carruthers et al. 2015, Wiedenmann et al. 2019). A log-linear regression is applied to the final five years of the

unsmoothed average index, and the slope is used to adjust a multiple (< 1) of the most recent five year averaged catch to generate catch advice (Table 2.1). Four formulations of Islope were proposed by Geromont and Butterworth (2015a), and we used the least conservative version here (version 1), although flexibility was incorporated to allow for exploration of the other versions. Key differences between Islope and Plan B smooth are the use of an unsmoothed average index, a longer recent time period (5 vs. 3 years), and a buffered catch (80% of the recent average in version 1; Table 2.2).

Itarget - The Itarget method was proposed by Geromont and Butterworth (2015a), and has been evaluated in a number of studies (Geromont and Butterworth 2014, 2015, Carruthers et al. , Wiedenmann et al. 2019). Instead of using trends in the recent index, catch advice is determined by comparing the most recent five year average index to some target based on a reference period. The reference period does not change, and we used the last 25 years of the base period (years 26-50) as our reference period. The target index of abundance is set to some multiple (≥ 1) of the average index value over the reference period, and the catch advice is based on the average catch over the reference period, adjusted up or down based on the recent five year average index relative to target index (Figure 2.1; Table 2.2). We combined the spring and fall surveys into an average index to use in the Itarget method. As with Islope, Geromont and Butterworth (2015a) proposed four formulations, and we used the least conservative option (version 1; Table 2.2).

Skate method - The skate control rule method was developed by the New England Council's Skate Plan Development Team and endorsed by its Scientific and Statistical Committee to produce catch advice for the skate complex using only a time-series of catch and a survey index for each skate species (Skate FMP Amendment 3). A single time-series of the survey index is calculated as the mean of the fall and spring surveys. The median value of the annual catch divided by the annual index over the entire time series (except the most recent years) provides the relative fishing mortality to produce catch advice. A moving average smoother is applied to the catch and survey index prior to dividing the two. The relative fishing mortality times the terminal year of the smoothed survey index is the proposed catch advice. The entire time-series is used and there is no comparison to a pre-specified reference time period.

An Index Method (AIM) - AIM was developed by Dr. Paul Rago at the Northeast Fisheries Science Center (NEFSC 2002b) and the executable with GUI can be found in the NOAA Fisheries Toolbox (https://nmfs-fish-tools.github.io/AIM/). For the current project, the methods were implemented in R (comparisons were performed to confirm that the NFT program and the R implementation produce the same results; notation herin is consistent with that in the NFT program help files). The AIM model requires two inputs: an index of relative stock biomass and total catch in biomass, and seeks to identify a relationship between two calculated series: Replacement Ratio (Ψ) and Relative F (F_{REL}). Ψ is calculated as a ratio of the survey index in a given year divided by a weighted average of user-defined survey values; for this application, the numerator of Ψ is the current index year and the denominator was the average of the previous 5 years of the index. Assuming no density-dependence in the population dynamics, this time series of replacement ratio reflects changes in annual population biomass due to recruitment, growth, natural and fishing mortality. F_{REL} is the ratio of catch in a given year (y) to an average of recent index values (for this implementation, a 3 year average of the biomass index, centered on y, was used), and in principle reflects the relative magnitude of catches relative to average biomass; the catchability of the index is unknown so the scale of F_{REL} is not simply the proportion of biomass loss. The AIM model performs a regression of $\ln(\Psi)$ on $\ln(F_{REL})$ to identify the value of F_{REL} where Ψ=1 (i.e. where $\ln(\Psi)$=0, a level of fishing that allows replacement, which we'll refer to as F_{REL*}). Assuming the regression is satisfactory, and modeling assumptions are met, then the value of F_{REL*} is a relative fishing rate that allows the population biomass to replace itself. To arrive at an approximation of stable catch advice, F_{REL*} is multiplied by the most recent index value.

Dynamic Linear Model (DLM) - The DLM is a state-space approach requiring an index of relative stock biomass and total catch in biomass to produce an estimated target catch level. For a single or multiple survey indices and a vector of catch data, the method fits a dynamic linear model on the log scale with at least two components: 1) a smoothly evolving mean abundance, and 2) a dynamic regression on catch. Because survey indices and catch time series are often strongly correlated, multicollinearity can cause difficulties in model fitting. However, this linear correlation represents redundant statistical information, a reflection of spawning stock biomass, contained in both time series. Therefore, the average relative exploitation rate is differenced out

of the catch time series using a linear regression between the log survey indices and the log catch to create a time series of catch anomalies. Essentially, the catch anomalies values reflect whether more or less catch occurred than one would expect given the survey indices and catch history. Catch is chosen to be modified because it is assumed that the survey indices represent a more reliable measure of stock abundance. If the survey indices and catch time series for a given stock were not correlated, this procedure would only de-mean the catch time series and thus not impact model fitting. To track the mean survey abundances in this simulation experiment, the method incorporates a dynamic linear trend. A dynamic linear trend is estimated by two state variables: 1) an initial, static intercept at time 0, and 2) a random walk whose value is added to the intercept variable at each time t. The second state variable is therefore an estimate of the abundance trend in the survey, after the effect of catch is accounted for, at any point in time. This is advantageous in cases where population productivity is nonstationary (ex. *M* or recruits/spawner change over time) in that forecasts account for the recent trajectory of the population. Once the DLM is fit, the survey indices are projected forward in time with the estimated parameters and an optimization routine is used to determine the annual catch levels that would result in the reference biomass level on average. For this project, the biomass reference level was set as the 75th percentile of the survey indices and a 10-year rebuild to the reference level was desired. The first two years of catch advice returned by the optimization function is averaged with the catch levels in the two years prior to the forecast period in order to smooth out changes in harvest and produce a single value for catch advice. Notably, the quantification of forecast uncertainty and the ability to test different model structures was not employed due to the time constraints of the project. Additional options that could be explored in future work include testing different model structures and using the estimated forecast uncertainty to assess risk in setting different catch advice tailored to the stock of interest. Finally, the DLM structure could also be augmented with additional model components (e.g. other covariates, an autocorrelated component) and/or length or age information.

Swept area biomass methods

Expanded Survey Methods - Estimates of swept area biomass have been used to determine catch advice for a number of stocks in the region. The estimation of swept area biomass was

approximated by scaling the simulated spring and fall survey observations to units of biomass using their respective catchabilities. These "scaled-up" biomass indices were then averaged using the spring observation from year *y* and the fall observation in year *y*-1, as is common practice. Catch advice was determined as the product of the averaged biomass index from the most recent year and one of four target exploitation rates based on: 1) $F_{40\%}$, from a Spawner Per Recruit analysis 2) the exploitation rate associated with stock replacement (F_{REL*} calculated as detailed above for the AIM method), 3) the target Fishing mortality set equal to the assumed natural mortality (F=M), and 4) the average of the relative F values, defined as catch divided by the "scaled-up" biomass indices, from the most recent three years.

Catch curve methods - The catch curve methods explored by the working group utilized numbers at age of fully-selected fish in each survey to estimate total mortality (Z) using catch curve analysis. An aggregate abundance at age was calculated by summing across the most recent three years for a given survey to create a single numbers-at-age vector. The age class with the largest abundance at age in each survey is set to the age at full selection. Ages below the age at full selection are dropped, as is the plus group, and the remaining values are used in a log-linear regression to estimate Z in each survey (Z = the inverse of the slope). An average Z across surveys was calculated, and used with the assumed M to calculate an average F ($Z = M + F$), which is used with the most recent catch estimate to estimate total biomass (Table 2.1). Target catch values are then set using a specified target F, and we explored two options (F_{targ} = the assumed M, and $F_{targ} = F_{40\%}$ from a SPR analysis). It is possible for the estimated Z from the catch curve analysis to be quite low, and if $Z \leq M$, it would result in nonsensical estimates of recent biomass because the recent F would be < 0 . As a result, we set a constraint such that $Z = max(Z, M + 0.05)$ so that the estimated recent F would never be < 0.05.

Combination of methods

Ensemble method - An ensemble approach was evaluated that determined catch advice as the median of the advice produced by other IBMs. Only eight of the twelve IBMs were included in the ensemble. Variants of the catch curve and expanded survey methods were excluded because

each variant is not independent of the others and including all of them would skew the performance of the ensemble towards these two general methods, effectively weighting these approaches more heavily than other IBMs. Ultimately, two variants of the expanded survey method were included because preliminary results suggested that their performance was distinct enough as to negate any concerns about unduly weighting the expanded survey method in the ensemble. The DLM method was excluded because it had relatively long computational times that would have prohibited any evaluation of an ensemble approach in the time available. The IBM methods included in the ensemble were: AIM, catch curve using $F_{40\%}$, expanded survey using $F_{40\%}$, expanded survey using the average relative F values from the three most recent years, Islope, Itarget, Plan B smooth, and skate method.

Traditional age-based assessment methods

Statistical Catch-At-Age (SCAA) - An SCAA model and harvest control rule were used in simulations for a subset of operating model scenarios. The SCAA model was configured in the Woods Hole Assessment Model (WHAM). The correct CVs and effective sample sizes were assumed for all catch and index data in the SCAA model. No stock-recruit model was assumed and there were no random effects. Natural mortality was mis-specified at 0.2 for the scenarios where M changed. 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 F equal to 0.75 of the estimated $F_{40\%}$ (M=0.2).

Application of the methods

Each of the methods produces a single target catch value that was fixed over a two year interval. If the methods are 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 we followed that approach here. If a method for setting catches uses an average of spring and fall, the average index in year *y,* $\bar{I}_y$ includes 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 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. So, 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 $0.75F_{MSY}$ is a common default ABC control rule in the region.

TOR 3. Identify metrics from the index-based assessment results that could be used in evaluations of trade-offs in performance among harvest control rules and index-based methods.

The IBMWG identified a broad range of metrics to use in evaluating trade-offs that are typical of similar simulations, such as in management strategy evaluations (Punt 2017). The metrics were chosen because they directly represent management or biological interests, or were of legislative relevance (e.g., overfishing and overfishing metrics related to Magnuson Stevens Act; MSA). The metrics fell into five categories: 1) catch, 2) variability in catch, 3) biomass of the target species, 4) fishing mortality, and 5) miscellaneous (e.g., ecosystem, legislative). Generally, each category corresponds to some fundamental objective typical of fisheries management, and tradeoffs among categories are likely to be of greatest interest (Tables 3.1-3.2). Metrics related to catch and variability in catch can also be considered reasonable proxies for economic metrics, such as profit or revenue, and societal metrics, such as stability in production, although the relationship between catch metrics and these societal and economic metrics may not be linear. Ecosystem and multi-species metrics were considered, but fell outside the Terms of Reference and time frame available. Thus, all the metrics had a single species focus. Ultimately, not all of the metrics discussed by the IBMWG were used in analyses See Table 3.2), and may be evaluated in future research.

Some of the metrics within each of the categories are redundant and displayed similar tradeoffs. The redundancies were expected, but the IBMWG preferred to maintain a broader list to hopefully capture more metrics of specific interest to varied stakeholders, managers, and legislative matters. Thus, the IBMWG expects that the majority of tradeoffs and subsequent decisions will be made using a subset of available metrics.

Performance metrics also varied as to whether they took a short-term (first 6 years of the projection period) vs a long-term (last 20 years of a 40 year projection period) perspective. The IBMWG felt this critical as tradeoffs among short and long-term objectives can be expected, but both are important to consider when comparing strategic performance of control rules.

Performance metrics could also be split into risk-based metrics that used probability statements (e.g., frequency of simulations over which something happens) vs quantity-based metrics (e.g., the value of spawning stock biomass relative to a reference point). Risk-based metrics are useful for summarizing and comparing scenarios because a single numeric value bounded by zero and one captures the space of the outcomes. However, these metrics are often more difficult to interpret as they require some acceptable tolerance level for the probability, which can be more challenging for managers to define explicitly. Quantity-based metrics tend to be more easily interpretable and tolerance levels more easy to discuss (but the IBMWG recognizes that explicit transparency about what is 'enough' from decision-makers is also commonly a challenge for these metrics too). However, quantity-based metrics must be reported as summary from simulations, and so choices need to be made when communicating them about how the values and range of values of outcomes across simulations are summarized (e.g. mean/median over simulations, vs median plus/minus some quantile interval, vs. some lower/upper quantile that reflects the preferred direction of the metric).

Shorthand definitions for each metric were created for easier graphical displays used in addressing other Terms of Reference. These definitions are below (Table 3.1), but follow the convention that: "l" is long-term, "s" is short-term, "avg" is an average, "is" represents a metric recorded as a probability statement, "n" represents the number of years that a situation occurred during a time period, "less" represents a value being less than some reference point, "ge" represents a value being greater than or equal to some reference point, "gr" represents a value being greater than some reference point, "f" indicates a fishing mortality quantity, "ssb" and "b" are used interchangeably and indicate spawning stock biomass, "catch" is self evident, "dot" indicates a metric that compares realized fishing mortality to an equilibrium fishing mortality rate that would drive the stock to some given level of biomass (i.e., either 10% Bmsy or 50%Bmsy), 01 or 1 indicate 10% of Bmsy, 05 or 5 indicate 50% Bmsy, msy indicates a maximum sustainable yield reference point (i.e., MSY, Fmsy, or Bmsy).

TOR 4. Evaluate the combinations of index-based methods and control rules using the metrics in ToR 3 to determine candidates for consideration by the Councils or other management authorities.

Base and Sensitivity Runs

There were 208 factorial combinations of IBMs and scenarios examined in the base runs, where the catch advice multiplier (1 or 0.75) was considered part of the scenario definition (Table 4.1). Most of these factorial combinations had the full 1,000 simulations produce results (Table 4.2). There were two IBMs with two scenarios that had individual simulations fail and the DLM had a much lower number of runs due to time limitations (it took significantly longer to run than all the other IBMs).

There were two sets of sensitivity runs. The first applied an SCAA to four of the scenarios. The SCAA applied a rho-adjustment to account for the estimated retrospective pattern in each assessment of each simulation. This required considerable computing time, so only four scenarios were examined. The second sensitivity run removed all sources of retrospective pattern for two of the scenarios. All the IBMs, except DLM and SCAA due to time constraints, were applied to these scenarios.

There were a total of 230,147 successful simulations produced.

Linear Models

The simulations generated a massive amount of results. Therefore, evaluating the performance of the IBMs using only graphical displays would be both time consuming and subjective. So linear models were conducted to help identify the most important elements of the study design that affected the performance of the IBMs, an approach that has been used in management strategy evaluations (Punt et al., 2008; Fay et al., 2011). The objective for the linear

models was to supplement the graphical displays and help focus additional analyses. So while some consideration was given to the validity of the linear models, a thorough evaluation of their assumptions and statistical rigor was not conducted.

A linear model was conducted independently for each of the metrics (TOR 3). Metrics recorded as frequencies or proportions were arcsine square root transformed, while all other metrics were log transformed (Punt et al., 2008; Fay et al., 2011), $tran(metric)$. Explanatory variables included source of the retrospective error, $retro_{type}$, fishing history, $F_{history}$, number of selectivity blocks, $num_{selectivity\,blocks}$, index based method, IBM, multiplier on the catch advice, $catch_{mult}$, and all two-way interactions:

$$tran(metric) = \mu + retro_{type} + F_{history} + num_{selectivity\,blocks} + IBM + catch_{mult} + two\text{-}way\ interactions$$

where μ is the overall intercept. Results were summarized by creating a table noting which explanatory variables were significant for each metric at the 0.05 level, Sig, or not, NS. The proportion of times an explanatory variable was significant among metrics grouped by spawning stock biomass, fishing mortality, or catch, was reported.

The explanatory variables $retro_{type}$, IBM, $catch_{mult}$, and the two-way interactions between $retro_{type}$ and IBM , and $F_{history}$ and IBM were significant for all metrics (Table 4.3). The interaction of $retro_{type}$ and $catch_{mult}$ was also significant for the vast majority of metrics. The interaction of $catch_{mult}$ and IBM was consistently significant for long-term metrics and metrics related to catch. $F_{history}$ was significant for a majority of metrics, most consistently for short-term spawning stock biomass and catch metrics. The interaction between $F_{history}$ and $retro_{type}$ was significant for a majority of metrics with no discernible difference between long- and short-term metrics. The $num_{selectivity\,blocks}$ variable was significant for the majority of the spawning stock biomass metrics and all but one of the catch metrics. The remaining explanatory variables were generally significant for less than half of metrics. Given these results, greater emphasis was placed on understanding the effects of $retro_{type}$, IBM, and

$catch_{mult}$ because these variables were the most consistently significant relative to the other variables.

Scoring

The mean value of each performance metric for all IBMs was computed across the 1,000 simulations of the 16 scenarios. Since some metrics are better when values are larger and others are better when values are smaller, these mean values cannot easily be combined. So metrics where smaller values were better were multiplied by negative 1 to create a set of values where bigger is better (Table 4.4). This mean and adjustment approach to make bigger values better was also applied to the SCAA scenarios (Table 4.5, more detail about SCAA scenarios provided below).

Scores were generated from these values in two ways. The first was simply ranking them giving the largest value the number of IBMs considered, the smallest value the number 1, and integer values in between (except in the case of ties). The second was to subtract the mean and divide by the standard deviation of each metric across all the IBMs. Both approaches result in scores where bigger values are better for each metric and can be easily summed across selected metrics to determine which IBMs perform best relative to those metrics.

Two examples sets of metrics are provided here to demonstrate that the ordering of the IBMs depends strongly on the metrics selected. The first set of metrics contains the mean ratios of SSB, F, and catch to their respective MSY reference points in the long term (Figure 4.1). The second set of metrics contains the interannual variability in catch across the entire feedback period and the short term mean catch/MSY (Figure 4.2). These sets of metrics produce different orderings of the IBMs. For example, both catch curve (CC) IBMs score highly using the first set of metrics, while these IBMs score poorly using the second set of metrics. A number of additional sets of metrics are provided in Appendix 6 to demonstrate the changes in IBM ordering when different metrics are selected.

To allow users to easily evaluate a large number of sets of metrics, an R Shiny app was developed (see the scorer_app folder in the GitHub repository). This R Shiny app can be copied to a local machine and run to examine the Rank and Resid scores for whatever combination of metrics is desired. The sensitivity runs, denoted noretro and scaa in the app, are also available in the R Shiny app. The app was created because specific metrics were not provided as the basis for determining performance of the IBMs. The app allows users to pick among the 50 metrics to see how the IBMs performed relative to each other.

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-FM, 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 equivalent short term plots are available in Appendix 6. The long and short term relationships can also be visualized through bagplots (Rousseeuw et al. 1999). For example, Figure 4.12 shows both the long term and short term SSB/SSBmsy and catch/MSY for scenario CF1A. The full set of bagplots 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 in both the plots showing all the simulations and the bagplots. While all IBMs have large ranges for both ratios, Skate, AIM, ES-Fstable, ES-FSPR, ES-FM, 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 Figure A6.39 and A6.40 (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 (see Appendix 6). 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 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 4.16). 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 4.17).

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 (see Appendix 6). 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 4.18). 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 SCAA performed with respect to long term probability of the stock being overfished or undergoing overfishing compared to the IBMs (Figure 4.19). 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-FM, Ensemble, and Itarget should be candidates for consideration because they had higher short term catch.

TOR 5. Provide guidance on specific situations that are and are not well-suited for a particular control rule or index-based method identified in ToR 4.

ANOVA for evaluating sensitivity of IBMs to simulation factors

For each iteration of each IBM in the simulation, the performance metrics of SSB, F, and catch relative to their true MSY values from the OM were calculated. Short- and long-term metrics were summarized over a number of years for SSB/SSBMSY, F/FMSY, and Catch/MSY: the average of each metric over the first 6 years of the projection period (short-term) and the average over the last 20 years of the projection (long-term). The variability among these metrics was analyzed in an ANOVA to determine which factors in the simulation explained the greatest proportion of total variance by IBM. Simulation factors evaluated were: retro_type (catch or M), Fhist (always overfishing or overfishing followed by $F=F_{MSY}$), n_selblocks (1 or 2 fishery selectivity blocks), catch.mult (multiplier of 1.0 or 0.75 on 2-year catch advice), time.avg (whether the metric was short- or long-term). The following interactions were also included in the analysis: Fhist:time.avg, retro_type:Fhist, retro_type:catch.mult, retro_type:time.avg, catch.mult:time.avg, and Fhist:catch.mult. Factors explaining a large proportion of total variance are interpreted as something that the metric for a particular IBM is sensitive to, while factors explaining relatively small fractions of the total variance are interpreted as factors that a metric for a particular IBM is relatively insensitive to. This approach is similar to explorations of reference point sensitivity to biological parameters in Brooks et al. (2008).

Before performing the ANOVA, the distributions of the metrics were evaluated for normality. The data were found to be quite skewed, and a square root transformation generally performed better than the natural logarithm transformation in reducing skewness, so all values were square-root transformed (see distributions of data with and without transforms and qq plots of residuals to the fitted linear models in Appendix 6). For the transformed data, all of the above listed factors and 2-way interactions were evaluated in separate ANOVAs for each IBM for each

of the 3 metrics. The ANOVA was done separately for each IBM so that differences in performance across methods could be summarized. The value in each table reflects the proportion of total variance explained by that factor, and the final row is the total fraction explained by those factors. “NA” indicates a non-significant factor.

The proportion of total variance explained by each factor is reported in Tables 5.1 – 5.3 for SSB/SSB_{MSY}, F/F_{MSY}, and catch/MSY. Most factors were significant at the 0.01 or 0.001 level for all 3 metrics, though some of those explained less than 1% of the variability. The two factors explaining the most for SSB/SSB_{MSY} and F/F_{MSY} were retro_source and time.avg, while for Catch/MSY time.avg explained the most followed by catch.mult and Fhist.

Sensitivity of IBMs to the different factors tended to fall out in two distinct groups for the SSB/SSB_{MSY} and F/F_{MSY} metrics. Specifically, DLM, PlanB, ES-Frecent and Islope were relatively insensitive to the cause of the retrospective (retro_type) and variability in their metrics were primarily explained by whether the value reflected a short- versus long-term average (time.avg). IBMs most sensitive to the cause of the retrospective were Itarget, Skate, AIM, ES-FSPR, ES-FM, ES-FStable, and the Ensemble. The two catch curve methods (CC-FSPR and CC-FM) were intermediate to these two IBM groupings.

Catch/MSY variability was largely explained by whether the value was a short- or long-term average (time.avg) for all IBMs but Itarget and the two catch curves. After the factor time.avg, the same two IBM groupings identified for SSB/SSB_{MSY} and F/F_{MSY} held: Islope, ES-Frecent, PlanB and DLM had the catch multiplier as the second most important factor, while Itarget, Skate, AIM, ES-FSPR, ES-FM, ES-FStable, and the Ensemble had the interaction retro_type:time.avg and Fhist:time.avg as the second or third most important factor.

Conclusions on suitability of IBMs for different scenarios:

1. For several IBMs, the variability among metrics was primarily due to the factor related to the cause of the retrospective pattern (retro_type). Thus, if you cannot identify the likely (or dominant) source of a retrospective pattern (catch or M), then using an IBM sensitive to the retrospective source would be risky to the stock and fishery, and this risk could be avoided by using a method robust to this uncertainty. In this regard, the following

methods were more robust to retro type: DLM, PlanB, ES-Frecent, Islope, and to some extent the two catch curve methods. We anticipate these methods would have more robust performance in situations where a retrospective pattern exists similar to that simulated in this project.

2. DLM variability in SSB/SSBMSY and Catch/MSY is mostly explained by short vs long term means; this is not surprising since the method aimed to rebuild in 10 years, which is longer than the cut-off for short-term metrics; catch multiplier followed by retro_type were the other main explanatory terms.
3. PlanB, ES-Frecent, and Islope had similar results to the DLM IBM. PlanB and Islope are similar (use slopes to adjust catch advice), but simpler, methodologically, so this result is not surprising.
4. The ES-Frecent IBM generally performed well, and differently from the other ES methods. This is due to the survey catchability coefficient canceling out in the application of catch advice. The catch advice is derived as Frecent times the expanded survey. Since Frecent is calculated as catch/expanded survey, the effect of the survey catchability is essentially removed. So this method will generally keep the stock at the current fishing mortality rate, although uncertainty in the survey and catch observations can cause it to change. The other ES methods would be expected to depend strongly on the survey catchability estimate. Future research could explore how the ES methods perform with uncertainty and/or bias in the catchability estimate.
5. Different sensitivity of IBMs to the different simulation factors could be useful for exploring future ensemble composition. Based on the scenarios explored herein, future analyses could consider combining the two catch curve methods with the four IBMs that were insensitive to the source of retrospective pattern for comparison with the results from the current Ensemble to see if an overall improvement (consistent rebuilding, higher and less variable catch) is achieved.

Heatmaps for identifying similar performance of median performance metrics

Results from the ANOVAs highlighted which factors explained variability in IBM performance, and served as the basis for data subsetting and generation of heatmaps that show groupings of IBMs that exhibit similar performance across the factor groupings. The heatmaps were produced by the function heatmap.2 in the package *gplots* (Warnes et al. 2020).

As with the ANOVA analyses, X/X_{MSY} metrics for each iteration of each IBM in the simulation, were summarized over short- and long-term time periods. At the highest level of aggregation, medians by IBM were calculated across all simulations (13 categories (IBMs) with 16,000 values per IBM if all iterations were conducted and/or converged). More granular analyses calculated medians by IBM and single factor (26 categories with 8,000 values for complete cases) and then medians by IBM and 2-way factors (52 categories with 4,000 values for complete cases).

In the heatmap.2 function, the median values for all 3 metrics are first normalized so that values are on the same scale. Dissimilarities are calculated across the categories and dendrograms reorder the categories. In the heatmaps for this analysis, categories are the factors into which all of the data were subsetted (13 IBMs, or 26 IBMxFactor, or 52 IBMxFactor1xFactor2). In the top left of each heatmap, a key indicates the colors that distinguish where the normalized score falls within the distribution, and a superimposed cyan histogram within that key indicates how the data were distributed over the normalized distribution. The title of each heatmap indicates the subsetting of the data, and the right-side y-axis gives the IBM x Factor category for the median value of the metric in that row. The rows are arranged by the dendrogram on the left-side y-axis, and the color in each cell of the heatmap corresponds to the normalized metric and matches the colors in the key at the top left of the heatmap. Comparing the contrasting colors within a given row shows trade-offs between achieving a higher median SSB/SSBMSY (value >1 indicates rebuilding was achieved) versus the fraction of MSY that could be caught; similarly, the F/FMSY median indicates whether overfishing occurred more or less than 50% of the time (values >1 indicate overfishing). A table of the median performance metric values was produced for each heatmap, and the row order in the tables matches the right-hand side y-axis labels in the Figures to facilitate easier association between the figure and corresponding table.

The median X/Xmsy metric by IBM indicates that >50% of iterations achieved rebuilding ($SSB/SSB_{MSY} > 1$) for CC-FSPR and CC-FM, DLM, PlanB, ES-Frecent and Islope, in order of greatest median SSB/SSB_{MSY} (Figure 5.1 and Table 5.4). This is seen on the heatmap as the block in the SSB column with the darkest red shading. These same methods avoided overfishing in >50% of iterations (F/F_{MSY} was 0.31-0.87). The trade-off for rebuilding and avoiding

overfishing is seen by the corresponding lighter color blocks in the Catch column, however among the IBMs achieving rebuilding >50% of the time, PlanB, ES-Frecent, and Islope achieved median catch that was 65-70%, respectively, of MSY. By contrast, the Skate, AIM, ES-FStable had median SSB/SSBMSY of 0.53-0.7, while median catch was > MSY.

The heatmap for median performance metrics by IBM and retrospective source indicates that rebuilding was more successful when the retrospective source was M rather than catch (Figure 5.2 and Table 5.5). Scenarios where rebuilding was not achieved at least 50% of the time were all in cases where catch was the retrospective source for the IBMs Skate, ES-FM, ES-FSPR, AIM, ES-Fstable, Ensemble, and Itarget. Several IBMs that achieved at least 50% rebuilding also achieved catch that was greater than MSY: ES-FSPR, ES-FM, ES-FStable, AIM, and Skate. Partitioning the results by retrospective source helped identify that methods that performed poorly at the overall IBM level (AIM and Skate) actually can have good performance for some of the scenarios. This finding is consistent with the ANOVA, which identified that the largest source of variability in metrics for AIM and Skate was the retrospective source.

The factor for retrospective source explained a lot of variability in some IBMs, while in others, the time horizon of metric summary explained the most. The heatmap for median performance by IBM and time horizon not unexpectedly shows that the greatest rebuilding is achieved in the long-term (Figure 5.3 and Table 5.6). However for the methods PlanB, DLM, Islope and ES-Frecent, as well as the two catch curves, median rebuilding was greater than 1 even in the short-term (slightly greater for some of these), while long-term median metrics were greater than 2. This factor explained the greatest amount of variability in the metrics for these IBMs. For these same IBMs, catch in the short-term was greater than long-term catch.

When median metrics for IBMs and catch multipliers are analyzed, rebuilding probability was greatestest when the catch advice was scaled by 0.75 rather than no scaling–and as a result, median catch was far below MSY (Figure 5.4 and Table 5.7).

Median metrics for IBM by Fhistory (1=always overfishing then $F=F_{MSY}$, 2=always overfishing) reveal that higher median SSB/SSBMSY is achieved by Fhistory=1 rather than 2, although for the IBMs PlanB, DLM, Islope, and Frecent rebuilding is achieved at least 50% of

the time under either Fhistory (Figure 5.5 and Table 5.8). The two catch curves were insensitive to this for the SSB metric but median catch was nearly double when Fhistory=1 rather than 2.

Heatmaps for median metrics by IBM and 2 factors were also examined, and the figures and corresponding tables can be found in Appendix 6. This finer granularity of data subsetting did not alter the bigger picture result that a subset of IBMs consistently outperformed others (PlanB, DLM, Islope, ES-Frecent and the two-catch curve methods).

Conclusions:

1. Looking at IBM performance across all iterations, rebuilding was greatest for the IBMs that were the least sensitive to source of the retrospective pattern (CC-FSPR and CC-FM, DLM, PlanB, ES-Frecent and Islope). DLM, PlanB, and CC-FM and CC-FSPR achieved the greatest median SSB/SSB_{MSY}, but catch was lowest for CC-FSPR and CC-FM. PlanB, Islope and ES-Frecent had the highest median catch among the methods that achieved rebuilding more than 50% of the time.
2. The IBMs ES-FSPR, ES-FM, Skate, Ensemble, ES-Fstable, AIM, and Itarget on the whole had less success rebuilding, though it depended on the scenario. If the retrospective source was M, some of these achieved median rebuilding.
3. The Ensemble method slightly underperformed at rebuilding, with median SSB/SSBMSY of 0.92, but 90% of MSY was attained for median catch. As the ensemble included several of the IBMs that did not achieve rebuilding, the median in the ensemble was weakened by their inclusion. Future research could look at optimizing ensemble composition based on the performance in scenarios most similar to that expected in a given assessment setting.
4. Trade-offs in risk (overfishing and rebuilding) and rewards (catch) are inherent in management decisions. Balancing median catch close to MSY in the short-term while still maintaining a probability of at least 50% of achieving rebuilding was possible for the ES-Frecent, PlanB, DLM, and Islope, although long-term median catch with these methods was far below MSY. This could indicate that these four methods are appropriate short-term models for management advice, but while they are employed other efforts should be invested to return to an age-based model.

Caveats

Although a large number of IBMs, scenarios, and simulations were conducted, there are always limits to the scope of simulation studies. These caveats should be kept in mind as the results of this study are applied in other situations. The biological and fishery characteristics used in this study were based on local groundfish parameters and other life histories or fishing histories may show different performance for the IBMs. This framework is well-suited to application of other biological and fishery characteristics and should be used in the future to do so.

There was only a single source causing the retrospective patterns in the age-based data that remained at the same magnitude throughout the feedback period in these simulations. Recent experience in the region has demonstrated this is not how retrospective patterns typically behave, they either increase or decrease in magnitude over successive assessments. While the framework is currently not formulated to handle such a situation of changes in the forcing function for the retrospective pattern, it could be modified to do so with relative ease. This is an area for future research. The framework is well suited to handle either multiple sources of retrospective patterns or different magnitudes of retrospective patterns (with constant forcing functions). These issues should be explored in the future using the framework.

The simulations applied the IBMs every other year in the feedback period to mimic the common use of index-based methods in the region. The IBMs are generally easy to apply with little tuning needed, so could be applied every year. However, this would create a burden on the management process that would need to be considered before actual application. Large improvements in performance would need to be demonstrated to justify this additional management cost. The reliance of the IBMs on the surveys means that missing years in the time series can have larger impacts than in age-based models. The impact of missing surveys could be examined through this framework by occasionally using the catch advice for 3 years and delaying an assessment. The framework is well-suited to explore this issue, but it was not considered in this study due to time limitations.

Due to the nature of closed-loop simulations, the IBMs had to be applied formulaically with no ability to modify the IBM formulation to a specific set of observations. This is not how

actual index-based assessments are conducted. Whoever is conducting a real index-based assessment will examine diagnostics and look for issues with the particular IBM being applied to the actual data available. This may raise questions about either the IBM suitability or the data themselves. Peer review processes are also used to ensure that assumptions for an IBM are met in actual applications, something that could not be easily considered in this study. The results presented here should be considered minimum performance of the IBMs given the formulaic application of the IBMs without consideration of the data available in any given assessment.

The Dynamic Linear Model (DLM) approach makes no explicit assumptions about the population dynamics, life history, or condition of a stock. It requires at least one index of abundance and catch information for use, though it could be employed in a manner similar to PlanBSmooth if no reliable catch information was available. If missing values are present in the abundance index(ices), the DLM fitting procedure generates a probabilistic imputation in those years based upon all of the available data and thus provides an estimate of what would have been observed with quantified uncertainty. If catch information is included as a covariate and missing values are present, those values will need to be imputed prior to fitting the DLM. The time series length necessary to fit the DLM depends upon the number of available abundance indices and their signal-to-noise ratio(s), but at least 25 years of data should generally be used. Because the DLM "learns" more about the behavior of a stock over time as more data is included in fitting, it can be expected that the model fit achieved will improve over subsequent updates. Due to the constraints of this simulation experiment, a single model structure and weakly informative prior distributions were used. In a management setting, different model components and prior assumptions could be compared to determine the optimal structure for a given stock. Similarly, catch advice in this work was set such that the mean forecasted stock trajectory was on track for a 10-year rebuild (or decline) to the 75th percentile of the observed abundance index. This decision rule was only used for this experiment and is not required by the DLM. The DLM produces a probabilistic forecast of future abundance observations based upon assumed harvest levels (when catch information is included). This forecast and uncertainty can be used by managers to aim toward any desired stock level while assessing the risk of different strategies. While still in development, the DLM structure can also be augmented to include additional covariates (e.g. climate information), length- or age-structure, or multispecies information.

The Skate method was developed using the median catch over biomass ratio assuming that this would provide a reasonable relative fishing intensity. This is appropriate for stocks that have not been undergoing overfishing for half or more of the years considered. It would not be expected to perform well in situations where the stock has been undergoing fishing for the majority of time. In such a situation, a different quantile of the distribution of catch over biomass would be preferred. This aspect was not examined in this study due to time constraints, but could be an interesting avenue of future research.

The IBMs that change the catch advice based on recent trends in the surveys (e.g., PlanB, Islope, DLM) do not appear well suited to applying a reduction to the catch advice. This is because when the stock rebuilds, the surveys do not change much yet the catch advice continues to decline due to the reduction (catch multiplier of 0.75 in this study). This can lead to the situation where the stock is well above SSBmsy but the catch advice is well below MSY. This would clearly be a frustrating situation for fishers and managers.

TOR 6. Create guidelines for setting biological reference points for index-based stocks.

The Index Based Methods (IBM) evaluated here used a variety of techniques to derive catch advice. The methods ranged from simple deterministic calculations to fitted state-space models, but generally fell into three basic categories: 1) IBMs examined the trend in the time series to determine if catch should be increased or decreased (e.g. Plan B Smooth); 2) IBMs compared the current survey index with with some reference time period of the survey index to set catch advise (e.g. Itarget). These methods often calculate the relative fishing mortality (catch/survey index) to determine what level of fishing would result in the reference index level; and 3) IBMs estimated total biomass (e.g. catch curve) then used a proxy fishing mortality reference level derived from other sources to calculate catch advice. True MSY reference points are based on the population growth rate of a stock. None of the IBMs estimate the population growth rate and thus can not provide true MSY reference points. Instead, the IBMs attempt to link the raw survey indices and catch data with thresholds that might serve as reference levels, though they should not be considered proxies for MSY reference points. In addition to the structural challenges associated with IBMs, a full exploration of the results to provide guidance on setting reference points with IBMs was not possible due to time constraints. A review of the results however, highlighted a number of interesting observations.

The trend IBMs (Plan B Smooth, Islope) examine the slope of the survey and catch in the most recent years and then modify catch based on the slope of the survey index. The performance of the two trend methods with regard to maintaining the population near MSY based references points differed considerably. Plan B Smooth consistently ranked as one of the top methods for maintaining SSB above SSBmsy and keeping F below Fmsy in the short and long term. The method tended to produce relatively conservative catch advice however. Islope tended to perform moderately on these same metrics and was consistently ranked in the middle of the different IBMs. While the two methods utilize a similar concept to derive catch advice based on recent trends in catch and survey, the specifics of the methods result in different performance. Islope contains multiple options specifying how conservative the catch advise

would be. The IBMWG ran the simulations with the least conservative options suggesting that alternative formulations could have resulted in different outputs.

The survey reference level methods (Itarget, AIM, Skate method, and DLM) utilized a range of techniques to set catch advice in order to realize a target survey level. Some methods such as AIM and the Skate control rule have a prespecified method for establishing the reference level and produce catch advice to realize that goal while others were designed to incorporate a user defined reference level. For example, the IBMWG set the reference period for Itarget as the last twenty-five years of the base period (years 25 - 50) and the reference level for DLM to the 75th percentile of the survey index. Both of these methods were then evaluated on their performance based on these decisions, when any time period or any percentile could have been selected as the reference level. Because of the range of functional forms within the methods and the range of potential reference levels, the performance of this group of methods for maintaining SSB and F near MSY reference points varied widely. DLM consistently ranked near the top for SSB and F metrics, however it was quite conservative and typically resulted in lower catches. Had a different reference level been selected such as the mean of the survey index for a particular set of years or the 55th percentile of the survey index, both the catch advice and the performance of the DLM method could have been different. This group also contained methods that consistently ranked in the middle of the IBMs and those that tended to rank near the bottom (e.g., AIM, Skate control rule). For the Skate control rule, there is an underlying assumption that the time series of data captures a period when the stock is in a reasonable condition. If the survey index only captures the stock when it is overfished, the Skate control rule will provide catch advice that maintains it at that overfished level. When setting the reference period for many of these methods, at least qualitative information on the status of the stock is very important. Without some information on an appropriate biomass target level, it will be difficult for many of the methods to produce catch advice to achieve the target.

Two methods estimated total biomass with two different techniques (catch curve analysis and the expanded biomass method) and utilized proxy target fishing mortality values to produce catch advice. Despite using the same fishing mortality estimates that could be considered proxy Fmsy values (F40% from a spawner per recruit analysis and F=M), the two methods performed very differently based on the SSB, F, and catch metrics. The catch curve analysis consistently

ranked among the top methods when producing catch advice with F40% or F=M, while the expanded biomass method ranked in the middle and low end of the IBMs. Within a single method, such as within the catch curve analysis, the performance was similar when using F40% or F=M, though not identical. The same was true for the expanded biomass method. The expanded biomass method also evaluated the performance of simply using the mean of the most recent catch to provide catch advice (Frecent). This method tended to perform well, typically in the top half of all methods, however, the starting conditions were likely very important for this method. In many of the scenarios, the IBMs started providing catch advice to a population that was near its MSY reference points. Harvesting at MSY when the biomass is at Bmsy should maintain the population. The use of this method when the stock is above or below Bmsy, however, would maintain the stock at a low level or forgo potential yield.

Despite a huge volume of simulations and results, the output did not produce consistent guidelines for developing IBM reference levels. A few factors to consider did emerge, however. At least a qualitative knowledge of the status of the stock (e.g. good, bad, ok) is important for setting a reference level. Some examples of qualitative stock indicators include increases in exploitable biomass from surveys, expansion in size or age structure in fishery-dependent and independent data sources, and tracking and monitoring the progress of year classes over time. While there are numerous caveats around why a stock might have been in an ok state at some point in the past (e.g. high productive period, limited harvest), without at least some historical knowledge, a reference level could be inadvertently set that maintains a stock in an overfished condition. The actual functional form of the index based method can be as important as the reference level itself. Plan B smooth and Islope both examined the trend of the time series in the most recent years, however their performance was quite different. Similarly the catch curve analysis and the expanded biomass method both estimated total biomass and used the same fishing mortality rate, but produced different results. DLM and Itarget have very different functional forms and were run with different reference levels. It would be interesting to see how they performed if given the same reference level. Future work based on this study could focus on two aspects: 1) conducting a similar simulation study with a subset of the IBMs and examine a range of different reference levels and different initial biomass conditions for each method to determine when different reference levels work. 2) Further exploration of the results of this and other simulation and real word examples for potential relationships between the survey and catch

time series and MSY reference points. As always, all results need to be considered in light of the assumptions, starting conditions, and decisions made by the working group. Different assumptions and starting conditions could produce different results.

Conclusions and Recommendations

General conclusions and recommendation

- For stocks that have had an age-based assessment rejected due to a strong retrospective pattern, there is no expectation that an index-based assessment will perform better than a rho-adjusted statistical catch at age analysis.
- The performance of an index-based assessment in a specific situation can be analyzed through the framework developed for this project, but requires specific hypotheses about possible sources of the retrospective pattern.
    - The IBMWG recommends this framework be used for all assessments that have changed from age-based to index-based due to retrospective patterns, using biological and fishery settings appropriate for that stock to ensure the selected index-based method has a high probability of providing reasonable catch advice.
- The IBMWG recommends future research be conducted to both analyze the results of this study in more detail as well as build on this study to address other questions
    - Examples of future research on the results of this study include
        - Filtering the results to have a common value for one metric, say probability of overfishing in the long term to see how the other metrics compare across IBMs. This is conceptually similar to tuning the IBMs to produce a common risk outcome, but is easier to apply across multiple scenarios.
        - Digging into the detailed results to look for reasons why IBMs of similar type did not always perform similarly while IBMs of different types sometimes did.
        - Exploring the detailed results to look for observed data that could be used to set biological reference points. This is a big task that will probably require additional simulations to fully address.
    - Examples of future research building on this study include
        - Examining life history and fishery characteristics beyond groundfish

        - Applying multiple or additional sources to create the retrospective patterns
        - Applying different magnitudes of retrospective patterns
        - Using state-space models to see if performance is better than SCAA
        - Tuning the framework to specific data for a stock to examine performance of IBMs in a specific situation

Process related conclusions and recommendations

- More time needs to be allocated to topic-based research tracks in order to fully address the TORs.
- Covid-19 travel limitations required weekly meetings with remote collaborations. This may be a useful approach for future topic-based research tracks, but may be less useful for stock-specific research tracks. The time certain weekly meetings helped track the programming and decision making progress necessary for the large simulation study. This may not apply well to stock-specific research tracks.
- GitHub was helpful for coordinating coding among multiple programmers.
    - Training session early on would ensure everyone able to work together efficiently
- This was a big project that required lots of computing power. The cooperation of network users not involved in the project to free up computing time was greatly appreciated.
- Fast internet speed an issue for moving large files and large numbers of files
    - Cloud computing would have been helpful
- Google docs with prompts before meeting allowed asynchronous contributions and then could build on it during meeting
- Google docs handy for meeting notes but not great for report writing
    - Need workflow all the way through to final report (508 compliance)
    - Would be helpful to be able to use Rmarkdown so don't have to update tables and figures in report by hand
- Project management software (e.g., Jira) could be useful but would require training

Literature Cited

A'mar, Z.T., Punt, A.E., and Dorn, M.W. 2010. Incorporating ecosystem forcing through predation into a management strategy evaluation for the Gulf of Alaska walleye pollock fishery. Fisheries Research 102: 98-114.

Anderson, S. C., Cooper, A. B., Jensen, O. P., Minto, C., Thorson, J. T., Walsh, J. C., Afflerbach, J., Dickey-Collas, M., Kleisner, K. M., Longo, C., Osio, G. C., Ovando, D., Mosqueira, I., Rosenberg, A. A., and Selig, E. R. 2017. Improving estimates of population status and trend with superensemble models. Fish and Fisheries, 18: 732-741.

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: 935–950.

Brooks, E.N., Shertzer, K.W., Gedamke, T., and Vaughan, D.S. 2008. Stock assessment of protogynous fish: evaluating measures of spawning biomass used to estimate biological reference points. Fish. Bull 106:12–23.

Carruthers, T.R., and Hordyk, A. 2018. The Data-Limited Methods Toolkit (DLMTool). An R package for informing management of data-limited population. Meth. Ecol. Evol. 9: 2388-2395.

Carruthers, T.R., Punt, A.E., Walters, C.J., MacCall, A., McAllister, M.K., Dick, E.J., and Cope, J. 2014. Evaluating methods for setting catch limits in data-limited fisheries. Fish. Res. 153: 48-68.

Carruthers, T., Kell, L., Butterworth, D., Maunder, M., Geromont, H., Walters, C., McAllister, M., Hillary, R., Levontin, P., Kitakado, T., Davies, C. 2015. Performance review of simple management procedures. ICES J. Mar. Sci. 73(2): 464−482. doi:
602 10.1093/icesjms/fsv212.

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

Dick, E.J., and MacCall, A.D. 2011. Depletion-Based Stock Reduction Analysis: a catch- based method for determining sustainable yields for data-poor fish stocks. Fish. Res.110: 331-341.

Fay, G., Punt, A.E., and Smith, A.D.M. 2011. Impacts of spatial uncertainty on performance of age-structure based harvest strategies for blue eye trevalla. Fisheries Research 110: 391-407.

Geromont, H.F., and Butterworth, D.S. 2015a. Generic management procedure for data-poor fisheries: forecasting with few data. ICES J. Mar. Sci. 72: 251-261. doi:10.1093/icesjms/fst232.

Geromont, H.F., and Butterworth, D.S. 2015b. Complex assessments or simple management procedures for efficient fisheries management: a comparative study. ICES J. Mar. Sci. 72: 262–274. doi: 10.1093/icesjms/fsu017.

Kristensen, K., Nielsen, A., Berg, C., Skaug, H., and Bell, B.M. 2016. TMB: Automatic differentiation and Laplace approximation. Journal of Statistical Software 70: 1–21. doi:10.18637/jss.v070.i05.

Legault, C.M., Alade, L., Gross, W.E., 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/

MacCall, A.D. 2009. Depletion-corrected average catch: a simple formula for estimating sustainable yields in data-poor situations. ICES J. Mar. Sci. 66: 2267-2271.

Martell, S., and Froese, R. 2012. A simple method for estimating MSY from catch and resilience. Fish Fish. doi: 10.1111/j.1467-2979.2012.00485.x.

McNamee, J., Fay, G., 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. Final report to the Mid Atlantic Fishery Management Council. Available: https://static1.squarespace.com/static/511cdc7fe4b00307a2628 ac6/t/55a661a5e4b060ebc9d03cf0/1436967333432/DLanalysis_bsb_final.pdf

Miller, T.J., Hare, J.A., Alade, L.A. 2016. A state-space approach to incorporating environmental effects on recruitment in an age-structured assessment model with an application to southern New England yellowtail flounder. Can. J. Fish. Aquat. Sci. 73: 1261–1270. dx.doi.org/10.1139/cjfas-2015-0339.

Miller, T.J., and Stock, B.C. 2020. The Woods Hole Assessment Model (WHAM), version 1.0. 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 J. Mar. Sci. 56: 473-488. doi:10.1006/jmsc.1999.0481.

Northeast Fisheries Science Center (NEFSC). 2002a. 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 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). 2002b. Re-evaluation of biological reference points for New England groundfish. Northeast Fish. Sci. Cent. Ref. Doc. 02-04; 395 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). 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. U.S. Dep. Commer., Northeast Fish. Sci. Cent. Ref. Doc. 05-13; 499 pp. 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. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 08-15; 884 pp. 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. US Dept Commer, 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. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 15-XXXX; 238 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. US Dept Commer, 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/publications/

Northeast Fisheries Science Center (NEFSC). 2017. Operational Assessment of 19 Northeast Groundfish Stocks, Updated Through 2016. US Dept Commer, 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/publications/

Northeast Fisheries Science Center (NEFSC). 2019. Operational Assessment of 14 Northeast Groundfish Stocks, Updated Through 2018. US Dept Commer, Northeast Fish Sci Cent Ref. 205

p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/publications/

Northeast Region Coordinating Council (NRCC). 2018. Description of New England and Mid-Atlantic Region Stock Assessment Process. 16 p. Available at: https://s3.amazonaws.com/nefmc.org/Stock-assessment-process-June2020.pdf

Pershing, A.J., Alexander MA, Hernandez CM, Kerr LA, Le Bris A, Mills KE, Nye JA, Record NR, Scannell HA, Scott JD, Sherwood GD, Thomas AC. Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery. Science. 2015; 350: 809-812. doi:10.1126/science.aac9819.

Punt, A.E. 2017. Strategic management decision-making in a complex world: quantifying, understanding, and using trade-offs. ICES Journal of Marine Science 74: 499-510.

Punt, A.E., Dorn, M.W., and Haltuch, M.A. 2008. Evaluation of threshold management strategies for groundfish off the US west coast. Fisheries Research 94: 251-266.

Punt, A.E., Butterworth, D.S., de Moor, C.L., De Olivier, J.A.A., and Haddon, M. 2016. Management strategy evaluation: best practices. Fish and Fisheries 17: 303-334.

Punt, A. E., Tuck, G. N., Day, J., Canales, C. M., Cope, J. M., de Moor, C. L., De Oliveira, J. A. A., Dickey-Collas, M., Elvarsson, B., 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.

Rosenberg, A. A., Kleisner, K. M., Afflerbach, J., Anderson, S. C., Dickey‐Collas, M., Cooper, A. B., Fogarty, M. J., Fulton, E. A., Gutiérrez, N. L., Hyde, K. J. W., Jardim, E., Jensen, O. P., Kristiansen, T., Longo, C., Minte‐Vera, C. V., Minto, C., Mosqueira, I., Osio, G. C., Ovando, D., Selig, E. R., Thorson, J. T., Walsh, J. C., and Ye, Y. 2017. Applying a new ensemble approach to estimating stock status of marine fisheries around the world. Conservation Letters, doi: 10.1111/conl.12363.

Rousseeuw, P.J., Ruts, I., and Tukey, J.W. 1999. The Bagplot: A Bivariate Boxplot. The American Statistician 53(4):382-387.

Spence, M. A., Blanchard, J. L., Rossberg, A. G., Heath, M. R., Heymans, J. J., Mackinson, S., Serpetti, N., Speirs, D. C., Thorpe, R. B., and Blackwell, P. G. 2018. A general framework for combining ecosystem models. Fish and Fisheries, 19: 1031-1042.

Stewart, I. J., and Hicks, A. C. 2018. Interannual stability from ensemble modelling. Canadian Journal of Fisheries and Aquatic Sciences, 75: 2109-2113.

Stock, B. C., Miller, T. J. *In review*. 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.

Tableau, A., Collie, J.S., Bell, R.J., Minto, C. 2019. Decadal changes in the productivity of New England fish populations. Can. J. Fish. Aquat. Sci. 76: 1528-1540.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R. Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. 2020. gplots: Various R Programming Tools for Plotting Data. R package version 3.1.0. https://CRAN.R-project.org/package=gplots (heatmap.2 function)

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

Wiedenmann, J., and Jensen, O.P. 2018. Uncertainty in stock assessment estimates for New England groundfish and its impact on achieving target harvest rates. Can. J. Fish. Aquat. Sci. 75(3): 342-356, https://doi.org/10.1139/cjfas-2016-0484

Wiedenmann, J., C.M. Free, and O.P. Jensen. 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.

Xu, H., Miller, T.J., Hameed, S., Alade, L., Nye, J.A. 2018. Evaluating the utility of the Gulf Stream Index for predicting recruitment of Southern New England-Mid Atlantic yellowtail flounder. Fish. Ocean. 27(1): 85-95.