Center for Independent Experts (CIE) Independent Peer Review of the Index Based Methods and Harvest Control Rules Research Track

7-10th December 2020

Woods Hole, Massachusetts

Prepared by

Robin Cook

Center for Independent Experts

Contents

Executive summary 2

Background 4

Reviewer's role 4

Summary of findings 4

General comments 4

TOR1: Develop methods to create data that if assessed with standard age-based approaches (e.g., VPA or ASAP) could exhibit a strong retrospective pattern. 5

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. 7

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....................8

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. 8

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. 11

TOR#6 Create guidelines for setting biological reference points for index-based stocks. 11

Conclusions 12

Recommendations 12

Reference 13

Appendix 1. Bibilography 14

Appendix 2: Statement of Work 15

Appendix 3. Panel Memebership 28

Executive summary

i. This was an ambitious and complex research track project. Given time limitations the WG produced an impressive report that goes a long way to establishing procedures for testing Index Based Methods (IBMs) and understanding their performance.

ii. Perhaps the most important result from the study is that IBMs do not necessarily perform any better than rho-adjusted SCAA models in the presence of retrospective patterns. Furthermore, the data free of a retrospective pattern did not materially improve the performance of IBMs.

iii. The operating model and scenarios were well-conceived and tailored to address the retrospective problem. They ensured that the IBMs would be challenged with a high level of retrospective pattern over the entire feedback period.

iv. The time available to the WG was limited and this restricted the range of potential retrospective sources that could be investigated. In particular, survey catchability is often considered a possible source but was absent from this study. It needs to be investigated in the future.

v. The group chose 12 individual IBMs for investigation, plus an ensemble method. The IBMs exhibit a variety of approaches that make use of survey indices in different ways ensuring a good range of model properties was considered. More research on the ensemble and DLM models, which appeared promising, may prove worthwhile.

vi. The large range of metrics selected means that model performance can be evaluated in a number of ways depending on the interests of relevant stakeholders. At present, the state of the project has not reached a point where a definitive set of performance criteria had been established. The results saved from the simulations will enable significant flexibility to consider alternative ways of evaluating IBM performance in the future.

vii. IBM performance was evaluated using a scoring procedure that ranked the IBMs based on the assumption that "bigger is better" in terms of the magnitude of the metric. This has the advantage of simplicity and is easy to understand. However, it does mean that an IBM could achieve high rank even if a stock is under-exploited in the MSY sense. Clearly this is a topic that needs further research to identify a suitable way of evaluating performance.

viii. It remains difficult to identify clear winners and losers among IBMs. Unfortunately, the WG had insufficient time to develop a wider range of scenarios and analyse outputs fully. Results may be difficult to generalise beyond the conditioning assumptions of the simulations, especially in relation to retrospective source, population biology and fishery characteristics.

ix. It proved difficult to provide guidance on the suitability of models to given situations (TOR#5). In view of the issues raised in relation to identifying "best" performing models, it was difficult for the WG to fully address this question. While some guidelines do emerge, reaching more specific conclusions requires working with managers to specify the main metrics of interest, and more analyses of outputs to develop an agreed framework for scoring IBM performance.

x. With regard to creating guidelines for setting biological reference points for index-based stocks, the WG drew the following conclusion: “Despite a huge volume of simulations and results, the output did not produce consistent guidelines for developing IBM reference levels”. I fully concur with that statement.

xi. As a priority, the conclusion that the rho-adjusted SCAA performs well compared to IBMs should be further investigated. This should encompass all the scenarios used to test the IBMs. There is a case for investigating the best ways of implementing the rho-adjustment given its apparently good performance.

xii. In this simulation study it would have been informative to demonstrate that a SCAA with a retrospective problem does actually perform worse than an IBM, and if so, by how much. This might inform the criteria for rejecting a SCAA assessment in the presence of a retrospective pattern.

Background

Strong retrospective patterns, indicative of model mis-specification, can lead to the rejection of age structured stock assessments, compromising the provision of robust catch advice for fishery management. Index-based methods (IBMs) that make use of catch and survey abundance indices are sometimes used in place of a full assessment to provide catch advice. The implication of this is that such methods mitigate the problems of the retrospective pattern and may offer more robust advice. A variety of IBMs have been used in the Northeast when retrospective behavior has been identified, but there remains a question as to how useful advice to management is in instances where a more complex model has revealed fundamental problems of model mis-specification.

In this review the Panel was asked to evaluate progress on a comprehensive research project that addresses the problem of retrospective patterns in age-based assessments. The review examines the results of the Index-Based Methods Working Group (IBMWG) that addresses this fundamental problem. The research task was designed to be able to derive guidelines on the use of IBMs in situations where a retrospective problem has been identified in a statistical catch-at-age (SCAA) stock assessment model. The IBMWG had worked over a period of 8 months during 2020 and prepared a report for review on 22nd November 2020. The review panel convened on the 7th December, working remotely via Google Meets.

Reviewer's role

Prior to the meeting the IBMWG report was reviewed. This was a substantial document comprising text, tables, figures and 6 appendices running to approximately 750 pages. A preliminary meeting of the Review panel was held prior to the main review to discuss and identify any issues that may arise during the plenary meeting. During the review meeting the IBMWG chair made a number of presentations to the Panel that summarized the WG report. The reviewer actively engaged with the WG chair to seek further clarification and discuss the issues arising. Following the formal presentations and public comment, the review panel discussed its initial findings. The reviewer assisted the Panel chair in the preparation of a presentation of the review panel's initial conclusions and recommendations. Following the end of the meeting, the reviewer worked by correspondence with the chair and other panel members to prepare a final summary report which was agreed by the 21st December.

Summary of findings

General comments

This was an ambitious and complex research track project to try to identify IBMs that would perform well when retrospective problems undermine the quality of a fully age based stock assessment. Initially, the project was expected to take 18 months but was subsequently compressed into a shorter 8-month period. The project was also carried out during the COVID pandemic when meetings could only take place virtually. Given these limitations, the WG produced an impressive report that goes a long way to establishing procedures for testing IBMs and understanding their performance.

The time pressures do appear to have restricted the amount of analyses that the Group could undertake and with the present state of the art only tentative conclusions can be made and the ability to generalize is limited. The review report also would benefit from further editing to clarify some of the methods. In particular, I would suggest:

a) A clearer description of the way catch misreporting was included and why the scaling constants of 2.5-5 are realistic.
b) Provide clearer information on the DLM model with some of the essential equations added to Table 2.2.
c) Identify the relevant Appendix 6 figures that are referred to in the main report text with the appropriate numbers.
d) Discuss the issues relating to survey catchability and why this source of retrospective problem was not included in the present study.

The nature of the research and its magnitude means that the direction of the work needs to be able to respond to emerging findings as the program proceeds and may require amendments to the terms of reference. I felt that TOR#6 was problematic and not really achievable once the IBMs had been selected. There may be an argument for a process at critical points of the work to review progress and update TORs accordingly.

Perhaps the most important result from the study is that IBMs do not necessarily perform any better than rho-adjusted SCAA models in the presence of retrospective patterns. This finding deserves further investigation but it highlights the issue about when an IBM should replace a SCAA assessment that shows retrospective problems. There appears to be an assumption, that if an assessment is affected by a large retrospective pattern, an IBM should then be used for catch advice. While apparently reasonable, this may not necessarily be the case especially if there is bias in the data (such as catch under-reporting) being used by the IBM. In this simulation study it would have been informative to demonstrate that a SCAA with a retrospective problem does actually perform worse than an IBM, and if so, by how much. Establishing a baseline performance using an unadjusted SCAA would provide a benchmark against which any improvement due to the use of an IBM or rho-adjusted SCAA model could be evaluated.

TOR1: Develop methods to create data that if assessed with standard age-based approaches (e.g., VPA or ASAP) could exhibit a strong retrospective pattern.

This ToR was fully met. The WG created simulated data based on a generalized groundfish stock that had been overfished historically but where fishing mortality rates had reduced in recent years. There were two fishing regimes and up to two selectivity blocks. Biological characteristics (growth, maturity etc.) were fixed but recruitment was a stochastic process based on a Beverton-Holt curve and time correlated deviations. Retrospective patterns were created either from historical under-reporting of catches or a change in natural mortality (M). Other possible sources of retrospective patterns were considered but not investigated due to lack of time.

Strengths

There was careful design of the simulation environment for testing. The operating model and scenarios were well-conceived and tailored to address the retrospective problem. The retrospective pattern was confirmed to occur throughout the feedback period, thereby validating that the IBMs were addressing a persistent, not a transient retrospective pattern.   The scenarios chosen ensured that the IBMs would be challenged with a high level of retrospective pattern over the entire feedback period. The same Mohn's rho characterized the retrospective pattern for both the mis-reported catch and M scenarios and this facilitates comparisons.

Comments

The time available to the WG was limited and this restricted the range of potential retrospective forcing that could be investigated. In particular, survey catchability is often considered a possible source. The WG did consider catchability but was not able to produce a retrospective pattern comparable in magnitude to catch and M, and hence was not subjected to analysis. While this is understandable, since IBMs explicitly make use of survey indices, it would have been useful to explore scenarios where survey year effects or trends in catchability were considered, even if catchability, *per se*, was not the source of the retrospective pattern.

Observation errors in the simulated catch and survey age composition data were assumed to have a multinomial distribution but with a lognormal distribution for the total number observed. For the multinomial a fixed sample size was set at 100 for the surveys and 200 for the catch. During the meeting additional analysis on real data suggested these effective sample sizes were too high and might therefore generate unrealistically low observation errors in the simulations. This is unlikely to be a major issue but may give the impression that IBMs perform better than might be achieved with real data.

The retrospective pattern caused by M was related to a change in its value during the burn-in period of the simulations. This is clearly one possible cause but M here was specified as age invariant whereas in reality it is almost certainty size (and hence age) dependent. This misspecification interacts with model estimates of selectivity and can contribute to a retrospective pattern especially if the shape of the selectivity function is itself mis-specified. It would be useful to have a brief discussion of these issues in the final IBMWG report to provide context that allows the reader to understand the generality of M as retrospective source.

IBMs make use of survey indices to provide information on the direction of change for catch advice. This reflects stock biomass trends which will be heavily influenced by recruitment variability. Hence the performance of an IBM will to some degree depend on signal to noise ratio in the survey index. Stocks with high recruitment variability may therefore be better suited to some IBMs while stocks with less variable recruitment are better suited to others. The restriction of the simulated data to a single recruitment model may therefore limit the generality of the results.

During the review meeting it appeared that some stakeholders were interested in IBM performance for those stocks where no SCAA assessment is possible. In this situation retrospective issues are unknown and may not exist. In such a case the design of simulation scenarios is likely to be different and focus more on the precision and bias in the survey data as sources of uncertainty.

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.

This TOR was fully met. The WG considered a range of IBMs that are currently in use in the region as well as others tested and reported in the peer reviewed literature. In addition, a more experimental IBM based on dynamic linear modelling (DLM) was included.

Strengths

The group chose 12 individual IBMs for investigation, plus an ensemble method (a subset of the 12), which is a manageable number given the very large range of potential candidates. The IBMs exhibit a variety of approaches that make use of survey indices in different ways. Some use the index purely on its raw relative scale, others expand the index to estimate absolute biomass, while some use the age structure of the index to estimate mortality rates. Hence a good range of properties was considered.

Comments

Clearly in selecting IBMs difficult choices have to be made in order to restrict the simulation testing to a workable project in the time available. The ensuing comments are made therefore simply as observations which may be useful. Given TOR#6 which asks for guidance on reference points it may have been useful to consider at least one surplus production model as these make use of the same data but explicitly estimate MSY reference points. These models are not without their weaknesses but I did not feel the reasons given by the WG for their exclusion were entirely convincing. One argument was that these models require an estimate of depletion for the stock at the time of the first observations. However, an informative prior will often suffice. Such a prior could readily be estimated from a full SCAA even in the presence of a retrospective problem.

With the exception of the catch curve IBMs, few of the models exploit the information in the age structure of the survey index and hence miss information about mortality rates and recruitment variability. It is possible to fit a fully age structured model to survey index data alone to estimate fishing mortality and relative biomass (e.g., Cook, 2013). Indeed, fitting an age structured model to the surveys alone and investigating this for retrospective patterns may give clues about whether catch data are the source of the retrospective problem in an SCAA assessment.

Many of the IBMs selected include some form of smoothing that averages data over a period of years. The way smoothing is configured is usually *ad hoc* based on expert judgement. Inevitably by running automated simulations with no user intervention the configurations of the IBMs used in this study may not be optimized for any particular scenario and this needs to be considered when evaluating their performance.

The inclusion of an ensemble model reflects the current interest in using multiple models to improve assessments. This is a welcome addition to the project but comes with a number of questions that require further research. Firstly, which and how many IBMs should be included in the ensemble and secondly how should the models be weighted? The approach applied in this study represented an initial exploration of the topic and treated all the selected models equally, but this would benefit from further investigation.

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.

This TOR was met. Approximately 50 potential metrics for model performance were identified and cover MSY considerations, variability in catch advice and measures of risk.

Strengths

The metrics chosen reflect well-established measures of stock status in relation to MSY reference points as well as the interests of managers and harvesters. MSY reference points are the accepted standard for good management practice, while metrics of inter-annual variability, especially in relation to catch, are often of concern to harvesters. Likewise, metrics of risk (e.g., probability of falling below a biomass threshold) are an important measure of sustainability.

The large range of metrics selected means that model performance can be evaluated in a number of ways depending on the interests of relevant stakeholders. At present, the state of the project had not reached a point where a clear set of performance criteria had been established but the results saved from the simulations will enable significant flexibility to consider a range of ways of evaluating IBM performance.

Comments

The large number of metrics makes objective interpretation of model performance challenging. This means there is still work to be done to make best use of the metrics recorded to judge IBM performance. Unfortunately, it is also too easy to think of additional metrics that may be relevant. For example, during the meeting it was suggested that the frequency with which realizations produced F>2 (and where a cap was placed on the advised TAC) should be recorded as a metric to indicate the rate of model failure. This would undoubtedly prove a useful indicator but illustrates the difficulty in forming a closed list of important metrics.

It may be possible to avoid redundancy in the metrics by using multivariate methods such as PCA to reduce dimensionality but this simply avoids the issue of what is important to managers. In any case, the choice of metric is likely to be both stock and fishery specific which means that there is unlikely to be a universal panacea to the problem in a generalized study of the kind discussed here.

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.

This TOR was met. Substantial analyses and outputs are provided from simulations to characterise model performance under different retrospective and fishery assumptions.

Strengths

Simulation experiments were conducted using a factorial design with 2 fishing histories, 2 selectivity scenarios, 2 sources of retrospective forcing and 2 catch advice multipliers, hence incorporating a range of likely conditions to which IBMs would be subjected. A target of 1000 simulations was attempted for each IBM/scenario leading to a large volume of output. This was saved and made available for post processing. Since the results are all saved a large repository of information is available for further analysis. In addition to the main experiment simulations were done on a subset of scenarios treating the SCAA as an IBM but with an automated bias correction on the retrospective pattern based on Mohn's rho. This is a crucial control experiment that has important implications for the conclusions of this study and is discussed further below.

A further control experiment was performed where IBMs were tested with data that had no retrospective forcing. This too is an important control that helps interpret the value of IBMs since the results tended to show that model performance did not improve markedly in the absence of a retrospective pattern.

Comments

A number of methods were used to evaluate the IBMs. This included linear models that related the metrics identified in TOR#3 to the experimental factors. The analysis showed that most metrics were significantly influenced by the main factors, including interactions, considered in the experimental design.

IBM performance was evaluated using a scoring procedure that ranked the IBMs. Two main methods were used, both based on the assumption that "bigger is better" in terms of the magnitude of the metric. Some metrics, such as the ratio of F/Fmsy would be considered better for smaller rather than larger values and to overcome this issue the WG multiplied such metrics by -1. An alternative would be simply to invert the ratio.

In applying the bigger is better approach, IBMs were ranked for each metric and then ranks were summed across a set of metrics to obtain a combined score. Two obvious questions arise from this procedure. Firstly, does bigger=better really capture model performance appropriately? Secondly, which metrics should be used in the summation to obtain a combined score?

There was considerable discussion of the question of bigger=better both within the WG and during the review. It has the advantage of simplicity and is easy to understand. However, it does mean that an IBM could achieve high rank even if a stock is driven to high biomass with very low catches. Generally, this would not be regarded as satisfactory since it rewards conservation of biomass at the expense of economic performance. The difficulty is that it is matter of judgement how far above Bmsy (for example) a stock should be and is likely to be stock and fishery specific. Clearly this is a topic that needs further research to identify a suitable way of evaluating performance.

On the issue of which metrics to use in ranking the models, the WG adopted the conventional ratios of SSB/SSBmsy, F/Fmsy and catch/MSY as the principal candidates which is a sensible and pragmatic approach. Inevitably there will be questions about how ranks would change if different metrics were chosen. Helpfully the WG developed a scoring approach that allows the user to select their preferred metrics to obtain a measure of performance. This still relies, however, on the bigger is better assumption.

Model performance was evaluated both over the short term and long term. It is important to note that short term performance will be heavily influenced by the initial conditions at the start of the feedback period and hence determined by the fishery scenario (F history and selection pattern). The long term performance should be less determined by initial conditions and may give insights into equilibrium behavior. Long term projections should not be seen as indicative of future stock status since states of nature are likely to change and modelling errors accumulate.

The WG devoted considerable effort to try to draw general conclusions from a highly complex simulation experiment and presented a variety of innovative approaches to visualize IBM performance by scenario and metric. While these provide very useful initial insights into performance, it remains difficult to assimilate the information and arrive at clear conclusions, not least because of the problem of identifying a closed set of preferred metrics, how to rank them and the extent to which the results are conditioned on the limited range of scenarios that could be investigated.

One of the analyses presented shows the relationship of the catch ratio (C/Cmsy) to the biomass ratio (SSB/SSBmsy) for each IBM accumulated over the scenarios. This tends to group the IBMs into two, with one group showing a near linear relationship and the other more diffuse. This grouping tended to reflect results from other analyses and may be indicative of the classes of IBM. A particular feature of these plots was that when a catch multiplier of 0.75 was applied to the catch advice the diffuse group of IBMs tended to provide almost constant catch advice regardless of the biomass ratio suggesting that even when a stock is rebuilt, advised catches remain low. Clearly it would not be desirable to apply such a model in perpetuity, though it might be useful for stock rebuilding scenarios.

The bivariate plots discussed above illustrate some of the trade-off between catch and biomass. Another consideration is the inter-annual variability of catch and similar plots could be produced with the variability of catch versus the variability of F. For a stock with natural variation due to recruitment there will be a trade-off between constant catch and constant F, neither of which is simultaneously achievable with the other. Constant catch offers more predictability for returns at the expense of variable fishing activity and may not therefore be a sole criterion for performance.

Two important results emerged from the “control” simulations when the SCAA was applied to the scenarios as an IBM, and when IBMs were tested in the absence of retrospective forcing in the data. In the former, the SCAA appeared to perform at least as well as the IBMs while in the latter IBM performance was not noticeably improved by the absence of a retrospective pattern. These conclusions are based on a smaller subset of scenarios and need further investigation but they strongly suggest that where a retrospective pattern is identified, reverting to an IBM may not be an automatic choice.

It remains difficult to identify clear winners and losers among IBMs due to the sheer volume and diversity of the various analyses. Unfortunately, the WG had insufficient time to develop a wider range of scenarios and analyse outputs fully. Results may be difficult to generalise beyond the conditioning assumptions of the simulations, especially in relation to retrospective source, population biology and fishery characteristics.

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.

This TOR was partially met. In view of the issues raised in relation to TOR#4, where identifying "best" performing models proved elusive, it was difficult for the WG to fully address this question. While some guidelines do emerge, reaching more specific conclusions requires working with managers to specify the main metrics of interest and more analyses of outputs to develop a framework for scoring IBM performance.

Strengths

Similar to the use of linear models in TOR#4 the WG applied ANOVA models to identify factors that account of the largest amount of variance in the metrics. This provides an initial focus for understanding how different IBMs may perform. Heat maps were also presented that help visualize how IBMs that perform similarly may be grouped according to chosen metrics and scenarios. The information in these diagrams is dense and requires careful consideration given the wide range of factors and performance metrics chosen for analysis.

Comments

The ANOVA and heat map analyses undoubtedly offer useful insights into how the models perform but the complexity of the study makes it extremely difficult to draw straight-forward conclusions. For example, some IBMs will perform better than others when catch misreporting is the source of the retrospective pattern. However, when retrospective problems are found in practice, the source is generally unknown making it a moot point about how a suitable IBM may be chosen. Perhaps significantly, when the rho adjusted SCAA model was run as an IBM it showed good performance regardless of the retrospective source. This may be because the approach directly addresses the symptom by explicitly making a bias correction. The IBMs simply use less data without cognizance of the retrospective problem. They rely on locally weighted data in the expectation that it better reflects current conditions.

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

This TOR was partially met. The WG drew the following conclusion: "Despite a huge volume of simulations and results, the output did not produce consistent guidelines for developing IBM reference levels". I fully concur with that statement.

The IBMs considered in the study were not designed for the estimation of biological reference points as typically understood in MSY based management. Some, such as the catch curve IBMs, could use the age structure from the index to derive yield-per-recruit proxies, however. Surplus Production Models that use the same data as the IBMs analysed in the study were not considered appropriate though they do allow the estimation of MSY reference points and are widely used. A reason not to use them was that there is a need to specify depletion at the start of the time series.

While some IBMs permit the estimation of reference points, many simply moderate the projected catch using any trend in the survey index. Some use a reference index level to ensure biomass does not fall below a threshold, but such levels amount to little more than expert judgement and may have minimal biological meaning. It should be possible to calculate reference points even from a SCAA with a retrospective pattern and develop an IBM that is responsive to that reference point. Nevertheless, given the good performance of the bias corrected SCAA when used as an IBM, this may well be the best approach with current knowledge.

Conclusions

This is an important study that examines a pernicious problem affecting many stock assessments. It illustrates how difficult it is to comprehensively test and draw clear and unambiguous conclusions about the performance of IBMs in the presence of retrospective forcing. The working group did an impressive job in the limited time available to construct realistic scenarios that reveal something of the performance of a useful range of IBMs. In its present state the research would benefit from more time to digest the voluminous output from the simulations and to review the initial objectives of the study given the results obtained thus far.

Apart from the complexity of the problem, the way IBM performance was judged using the "bigger is better" assumption, makes understanding model performance problematic since the assumption might reward poor performing IBMs that leads to a stock being under-exploited in the MSY sense.

The observation that the rho-adjusted SCAA performed at least as well as IBMs in most instances is significant. Even if this result is conditioned on the particular characteristics of the simulated scenarios, the fact the performance was no worse than IBMs should give pause for thought before abandoning SCAA assessments in the light of a retrospective pattern.

Recommendations

1. The final report of the IBMWG should
    a. Include a clearer description of the way catch misreporting was included and why the scaling constants of 2.5-5 are realistic.
    b. Provide clearer information on the DLM model with some of the essential equations added to Table 2.2.
    c. Identify the relevant Appendix 6 figures that are referred to in the main report text with the appropriate numbers.
    d. Discuss the issues relating to survey catchability and why this source of retrospective problem was not included in the present study.
2. The IBMWG should be given more time to further analyse the output from the simulation studies and consider alternative ways of ranking model performance that does not depend on the "bigger is better" approach.
3. The problem of trends or strong year effects in survey catchability needs to be investigated both as a source of retrospective patterns and uncertainty that affects the performance of IBMs.
4. As a priority, the conclusion that the rho-adjusted SCAA performs well compared to IBMs should be further investigated. This should encompass all the scenarios used to test the IBMs.
5. The criteria for rejecting a SCAA assessment in the presence of a retrospective pattern should be reconsidered in the light of the apparently good performance of the rho-adjustment. The best ways of implementing the rho-adjustment should be investigated.
6. For large research track projects of this type, thought should be given to a mechanism of periodic review during the project that would allow objectives to be revised on the basis of emerging findings.

Reference

Cook, R. M. (2013). A fish stock assessment model using survey data when estimates of catch are unreliable. Fisheries Research, 143:1–11.

Appendix 1. Bibliography

Documentation for the meeting that included the working group report, its associated tables, figures and appendices, and PowerPoint presentations summarizing the report were provided via Google Share Drive https://drive.google.com/drive/u/0/folders/1VqyaTfGzod5rCwuqiHhJXW4C8hvJfhBt and the NEFSC Data Portal https://apps-nefsc.fisheries.noaa.gov/saw/sasi/sasi_report_options.php. They are listed in the table below. All relevant source code and model outputs were available via a GitHub site https://github.com/cmlegault/IBMWG developed by Chris Legault. An Excel spreadsheet with all of the tables in the Working Group report was also available.

Table A1. List of documents made available for the review

Document type	File
Assessment Report	1_Report Text.pdf
Figures	0_Readme.txt
Figures	3_Report Figures.pdf
Figures	4_Appendices_1-5.pdf
Figures	5_Appendix_6_part1.pdf
Figures	5_Appendix_6_part2.pdf
Tables	2_Report Tables.pdf
Tables	6_Report_Tables_Excel.zip
Background	Background_190628_Groundfish_Assessment_Regs_Summary_through_December_2018.pdf
Background	Background_A4_200410_Groundfish_FW59_FINAL_Affected_Environment_Excerpt.pdf
Background	Background_SSC-Sub-Panel-Review-of-Report_-Eval_of_Inaccurate-Catch.pdf
Background	Background_Stock-assessment-process-June2020.pdf
Background	Background_WKFORBIAS_2019.pdf
Presentations	70_Introduction.pptx
Presentations	71_TOR_1_Make_Data.pptx
Presentations	72_TOR_2_Pick_IBMs.pptx
Presentations	73_TOR_3_Select_Metrics.pptx
Presentations	74_TOR_4_Crank_Sims.pptx
Presentations	75_TOR_5_Advise.pptx
Presentations	76_TOR_6_Ref_Points.pptx
Presentations	81_Homework_Day_1.pptx
Presentations	82_Langan_IBMWG_Review.pdf
Presentations	83_Homework_Day_2.pptx
Presentations	Summary Report of the Stock Assessment Review Committee-final.pptx

Appendix 2: Statement of Work

Performance Work Statement (PWS)

National Oceanic and Atmospheric Administration (NOAA)
National Marine Fisheries Service (NMFS)
Center for Independent Experts (CIE) Program
External Independent Peer Review

Index Based Methods and Harvest Control Rules
Research Track Peer Review

Dec. 7 -11, 2020

Background

The National Marine Fisheries Service (NMFS) is mandated by the Magnuson-Stevens Fishery Conservation and Management Act, Endangered Species Act, and Marine Mammal Protection Act to conserve, protect, and manage our nation's marine living resources based upon the best scientific information available (BSIA). NMFS science products, including scientific advice, are often controversial and may require timely scientific peer reviews that are strictly independent of all outside influences. A formal external process for independent expert reviews of the agency's scientific products and programs ensures their credibility. Therefore, external scientific peer reviews have been and continue to be essential to strengthening scientific quality assurance for fishery conservation and management actions.

Scientific peer review is defined as the organized review process where one or more qualified experts review scientific information to ensure quality and credibility. These expert(s) must conduct their peer review impartially, objectively, and without conflicts of interest. Each reviewer must also be independent from the development of the science, without influence from any position that the agency or constituent groups may have. Furthermore, the Office of Management and Budget (OMB), authorized by the Information Quality Act, requires all federal agencies to conduct peer reviews of highly influential and controversial science before dissemination, and that peer reviewers must be deemed qualified based on the OMB Peer Review Bulletin standards[1]. Further information on the Center for Independent Experts (CIE) program may be obtained from www.ciereviews.org.

Scope

The Northeast Regional Stock Assessment Review Committee (SARC) meeting is a formal, multiple-day meeting of stock assessment experts who serve as a panel to peer-review tabled stock assessments and models. The SARC peer review is the cornerstone of the Northeast Stock Assessment Workshop (SAW) process, which includes assessment development, and report preparation (which is done by SAW Working Groups or Atlantic States Marine Fisheries Commission (ASMFC) technical committees),

[1] http://www.cio.noaa.gov/services_programs/pdfs/OMB_Peer_Review_Bulletin_m05-03.pdf

assessment peer review (by the SARC), public presentations, and document publication. This review determines whether or not the scientific assessments are adequate to serve as a basis for developing fishery management advice. Results provide the scientific basis for fisheries within the jurisdiction of NOAA's Greater Atlantic Regional Fisheries Office (GARFO).

The purpose of this meeting will be to provide an external peer review of index based stock assessment methods and harvest control rules. The requirements for the peer review follow. This Performance Work Statement (PWS) also includes: **Appendix 1**: TORs for the research track, which are the responsibility of the analysts; **Appendix 2:** a draft meeting agenda; **Appendix 3:** Individual Independent Review Report Requirements; and **Appendix 4:** Peer Reviewer Summary Report Requirements.

Requirements

NMFS requires three reviewers under this contract (i.e. subject to CIE standards for reviewers) to participate in the panel review. The chair, who is in addition to the three reviewers, will be provided by either the New England or Mid-Atlantic Fishery Management Council's Science and Statistical Committee; although the chair will be participating in this review, the chair's participation (i.e. labor and travel) is not covered by this contract.

Each reviewer will write an individual review report in accordance with the PWS, OMB Guidelines, and the TORs below. All TORs must be addressed in each reviewer's report. No more than one of the reviewers selected for this review is permitted to have served on a SARC panel that reviewed this same species in the past. The reviewers shall have working knowledge and recent experience in the use and application of both index-based and age-based stock assessment models, including familiarity with retrospective patterns and how catch advice is provided from stock assessment models. In addition, knowledge and experience with simulation analyses is required

Tasks for Reviewers

- Review the background materials and reports prior to the review meeting
  - Two weeks before the peer review, the Assessment Process Lead will electronically disseminate all necessary background information and reports to the CIE reviewers for the peer review.
- Attend and participate in the panel review meeting
  - The meeting will consist of presentations by NOAA and other scientists, stock assessment authors and others to facilitate the review, to provide any additional information required by the reviewers, and to answer any questions from reviewers
- Reviewers shall conduct an independent peer review in accordance with the requirements specified in this PWS and TORs, in adherence with the required formatting and content guidelines; reviewers are not required to reach a consensus.
- Each reviewer shall assist the SARC Chair with contributions to the Peer Reviewer Summary Report
- Deliver individual Independent Reviewer Reports to the Government according to the specified milestone dates
- This report should explain whether each research track Term of Reference was or was not completed successfully during the SARC meeting, using the criteria specified below in the "Tasks for SARC panel."

- If any existing Biological Reference Points (BRP) or their proxies are considered inappropriate, the Independent Report should include recommendations and justification for suitable alternatives. If such alternatives cannot be identified, then the report should indicate that the existing BRPs are the best available at this time.
- During the meeting, additional questions that were not in the Terms of Reference but that are directly related to the assessments and research topics may be raised. Comments on these questions should be included in a separate section at the end of the Independent Report produced by each reviewer.
- The Independent Report can also be used to provide greater detail than the Peer Reviewer Summary Report on specific stock assessment Terms of Reference or on additional questions raised during the meeting.

Tasks for Review panel

- During the SARC meeting, the panel is to determine whether each research track Term of Reference (TOR) was or was not completed successfully. To make this determination, panelists should consider whether the work provides a scientifically credible basis for developing fishery management advice. Criteria to consider include: whether the data were adequate and used properly, the analyses and models were carried out correctly, and the conclusions are correct/reasonable. If alternative assessment models and model assumptions are presented, evaluate their strengths and weaknesses and then recommend which, if any, scientific approach should be adopted. Where possible, the SARC chair shall identify or facilitate agreement among the reviewers for each research track TOR.
- If the panel rejects any of the current BRP or BRP proxies (for B_{MSY} and F_{MSY} and MSY), the panel should explain why those particular BRPs or proxies are not suitable, and the panel should recommend suitable alternatives. If such alternatives cannot be identified, then the panel should indicate that the existing BRPs or BRP proxies are the best available at this time.
- Each reviewer shall complete the tasks in accordance with the PWS and Schedule of Milestones and Deliverables below.

Tasks for SARC chair and reviewers combined:

Review the Report of the Index Based Methods and Harvest Control Rules Working Group.

The SARC Chair, with the assistance from the reviewers, will write the Peer Reviewer Summary Report. Each reviewer and the chair will discuss whether they hold similar views on each research track Term of Reference and whether their opinions can be summarized into a single conclusion for all or only for some of the Terms of Reference of the SAW. For terms where a similar view can be reached, the Peer Reviewer Summary Report will contain a summary of such opinions. In cases where multiple and/or differing views exist on a given Term of Reference, the Peer Reviewer Summary Report will note that there is no agreement and will specify - in a summary manner – what the different opinions are and the reason(s) for the difference in opinions.

The chair’s objective during this Peer Reviewer Summary Report development process will be to identify or facilitate the finding of an agreement rather than forcing the panel to reach an agreement. The chair will take the lead in editing and completing this report. The chair may express the chair’s opinion on each research track Term of Reference, either as part of the group opinion, or as a separate minority opinion. The Peer Reviewer Summary Report will not be submitted, reviewed, or approved by the Contractor.

Place of Performance

The place of performance shall be held remotely, via Google Meets video conferencing.

Period of Performance

The period of performance shall be from **01 November 2020 through 31 January 2021**. Each reviewer's duties shall not exceed **14** days to complete all required tasks.

Schedule of Milestones and Deliverables: The contractor shall complete the tasks and deliverables in accordance with the following schedule.

Schedule	**Milestones and Deliverables**
Within 2 weeks of award	Contractor selects and confirms reviewers
Approximately 2 weeks later	Contractor provides the pre-review documents to the reviewers
December 7-11, 2020	Panel review meeting
Approximately 2 weeks later	Contractor receives draft reports
Within 2 weeks of receiving draft reports	Contractor submits final reports to the Government

* The Peer Reviewer Summary Report will not be submitted to, reviewed, or approved by the Contractor.

Applicable Performance Standards

The acceptance of the contract deliverables shall be based on three performance standards:

(1) The reports shall be completed in accordance with the required formatting and content (2) The reports shall address each TOR as specified (3) The reports shall be delivered as specified in the schedule of milestones and deliverables.

Travel

No travel is necessary, as this meeting is being held remotely.

Restricted or Limited Use of Data

The contractors may be required to sign and adhere to a non-disclosure agreement.

NMFS Project Contact

Michele Traver, NEFSC Assessment Process Lead
Northeast Fisheries Science Center
166 Water Street, Woods Hole, MA 02543
Michele.Traver@noaa.gov
Phone: 508-495-2195

Appendix 1. Index Based Methods and Harvest Control Rules Research Track Terms of Reference and Background

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.

Background

There are two reasons stock are assessed with index-based approaches. Either the data are not available to support an age-based assessment, e.g., ocean pout, or the age-based assessment was rejected and replaced by an index-based approach, e.g., Georges Bank yellowtail flounder. In recent years, the number of index-based assessments due to the latter reason has increased. This research track is focused on how to deal with this situation because the presence of a strong retrospective pattern is an indication of an inconsistency in the data and model that prevents standard simulation testing approaches to be used.

The Councils are charged with setting harvest control rules for each stock. The work conducted during this research track is meant to inform this decision by testing a range of harvest control rules against simulated data that would generate strong retrospective patterns in an age-based assessment.

Many of the index-based approaches currently used do not have the ability to generate biological reference points because they do not have an underlying population dynamics model. The creation of reference points for such situations requires expert knowledge about the fish and fishery. The guidelines created to address ToR 6 cannot be formulaic because of this dependency. Instead, the guidelines can be considered more of a checklist of items to consider when setting the biological reference points for a particular stock. The National Standard 1 technical guidance working group (subgroup 1) will provide some of the information to support this effort.

Simulation will be the approach used to address the ToR. If time permits, historical data may be used to see how the catch advice resulting from any recommended harvest control rules compares to what was used, particularly for situations where retrospective adjustments were made to analytical models in the past. The most recent data for any stock will not be used to prevent the creation of a “new” assessment that could require action by a Council.

Index-based approaches can be more impacted by missing survey data than age-based assessments, in some situations. This research track is not intended to examine the challenges associated with missing or partial survey data, or any other logistical issues associated with the generation of an index to be used.

SAW Research Track TORs:

General Clarification of Terms that may be

used in the Research Track Terms of Reference

Guidance to SAW Research Track Working Group about "Number of Models to include in the Peer Reviewer Report":

In general, for any TOR in which one or more models are explored by the Working Group, give a detailed presentation of the "best" model, including inputs, outputs, diagnostics of model adequacy, and sensitivity analyses that evaluate robustness of model results to the assumptions. In less detail, describe other models that were evaluated by the Working Group and explain their strengths, weaknesses and results in relation to the "best" model. If selection of a "best" model is not possible, present alternative models in detail, and summarize the relative utility each model, including a comparison of results. It should be highlighted whether any models represent a minority opinion.

On "Acceptable Biological Catch" (DOC Nat. Stand. Guidelines. Fed. Reg., v. 74, no. 11, 1-16-2009):

Acceptable biological catch (ABC) is a level of a stock or stock complex's annual catch that accounts for the scientific uncertainty in the estimate of Overfishing Limit (OFL) and any other scientific uncertainty…" *(p. 3208) [In other words, OFL ≥ ABC.]*

ABC for overfished stocks. For overfished stocks and stock complexes, a rebuilding ABC must be set to reflect the annual catch that is consistent with the schedule of fishing mortality rates in the rebuilding plan. *(p. 3209)*

NMFS expects that in most cases ABC will be reduced from OFL to reduce the probability that overfishing might occur in a year. (p. 3180)

ABC refers to a level of ''catch'' that is ''acceptable'' given the ''biological'' characteristics of the stock or stock complex. As such, Optimal Yield (OY) does not equate with ABC. The specification of OY is required to consider a variety of factors, including social and economic factors, and the protection of marine ecosystems, which are not part of the ABC concept. (p. 3189)

On "Vulnerability" (DOC Natl. Stand. Guidelines. Fed. Reg., v. 74, no. 11, 1-16-2009):

"Vulnerability. A stock's vulnerability is a combination of its productivity, which depends upon its life history characteristics, and its susceptibility to the fishery. Productivity refers to the capacity of the stock to produce Maximum Sustainable Yield (MSY) and to recover if the population is depleted, and susceptibility is the potential for the stock to be impacted by the fishery, which

includes direct captures, as well as indirect impacts to the fishery (e.g., loss of habitat quality).” (p. 3205)

Participation among members of a Research Track Working Group:

Anyone participating in SAW meetings that will be running or presenting results from an assessment model is expected to supply the source code, a compiled executable, an input file with the proposed configuration, and a detailed model description in advance of the model meeting. Source code for NOAA Toolbox programs is available on request. These measures allow transparency and a fair evaluation of differences that emerge between models.

Appendix 2. Draft Review Meeting Agenda

Index Based Methods and Harvest Control Rules

Research Track Assessment Peer Review Meeting

December 7 – 11, 2020

Google Meet link: TBD

Phone: TBD

DRAFT AGENDA

**All times are approximate, and may be changed at the discretion of the SARC chair. The meeting is open to the public; however, during the Report Writing sessions we ask that the public refrain from engaging in discussion with the SARC.*

Monday, December 7th, 2020

Time	Topic	Presenter(s)	Rapporteur
1:00 – 1:30pm	Welcome/Description of Review Process Introductions/Agenda/Conduct of Meeting	Michele Traver, Assessment Process Lead TBD, Chair	
1:30 – 3:00pm	TOR #1	Chris Legault, WG Chair	TBD
3:00 – 3:15pm	Break		
3:15 – 4:15pm	TOR #1 cont.	Chris Legault, WG Chair	TBD
4:15 – 4:45pm	Discussion/Review/Summary	Review Panel	TBD
4:45 – 5:00pm	Public Comment	Public	TBD
5:00pm	Adjourn		

Tuesday, December 8th, 2020

Time	Topic	Presenter(s)	Rapporteur
8:30 – 8:45am	Welcome/Logistics	Michele Traver, Assessment Process Lead TBD, Chair	
8:45 – 10:15am	TOR #2	Chris Legault, WG Chair	TBD
10:15 – 10:30am	Break		
10:30 – 11:30am	TOR #2 cont.	Chris Legault, WG Chair	TBD
11:30 – 12:00pm	Discussion/Review/Summary	Review Panel	
12:00 – 12:15pm	Public Comment	Public	

12:15 – 1:15pm	Lunch		
1:15 – 3:00pm	TOR #3	Chris Legault, WG Chair	TBD
3:00 – 3:15pm	Break		
3:15 - 4:15pm	TOR #3 cont.	Chris Legault, WG Chair	TBD
4:15 – 4:45pm	Discussion/Review/Summary	Review Panel	TBD
4:45 – 5:00pm	Public Comment	Public	TBD
5:45pm	Adjourn		

Wednesday, December 9th, 2020

Time	Topic	Presenter(s)	Rapporteur
8:30 – 8:45am	Welcome/Logistics	Michele Traver, Assessment Process Lead TBD, Chair	
8:45 – 10:15am	TOR #4	Chris Legault, WG Chair	TBD
10:45 – 10:30am	Break		
10:30 – 11:30am	TOR #4 cont.	Chris Legault, WG Chair	TBD
11:30 – 12:00pm	Discussion/Review/Summary	Panel	TBD
12:00 – 12:15pm	Public Comment	Public	TBD
12:15 – 1:15pm	Lunch		
1:15 – 3:00pm	TOR #5	Chris Legault, WG Chair	TBD
3:00 – 3:15pm	Break		
3:15 – 4:15pm	TOR #5 cont.	Chris Legault, WG Chair	TBD
4:15 – 4:45pm	Discussion/Review/Summary	Review Panel	TBD
4:45 - 5:00pm	Public Comment	Public	TBD
5:00pm	Adjourn		

Thursday, December 10th, 2020

Time	Topic	Presenter(s)	Rapporteur
8:30 – 8:45am	Welcome/Logistics	Michele Traver, Assessment Process Lead TBD, Chair	
8:45 – 10:15am	TOR #6	Chris Legault, WG Chair	TBD
10:45 – 10:30am	Break		
10:30 – 11:30am	TOR #6 cont.	Chris Legault, WG Chair	TBD
11:30 – 12:00pm	Discussion/Review/Summary	Panel	TBD
12:00 – 12:15pm	Public Comment	Public	TBD
12:15 – 1:15pm	Lunch		
1:15 – 2:15pm	Discussion of Key Points	Review Panel	TBD
2:15 – 5:00pm	Report Writing	Review Panel	
5:00pm	Adjourn		

Friday, December 11th, 2020

Time	Topic	Presenter(s)	Rapporteur
8:30 – 5:00pm	Report Writing	Review Panel	

Appendix 3. Individual Independent Peer Reviewer Report Requirements

1. The independent Peer Reviewer report shall be prefaced with an Executive Summary providing a concise summary of whether they accept or reject the work that they reviewed, with an explanation of their decision (strengths, weaknesses of the analyses, etc.).

2. The report must contain a background section, description of the individual reviewers' roles in the review activities, summary of findings for each TOR in which the weaknesses and strengths are described, and conclusions and recommendations in accordance with the TORs. The independent report shall be an independent peer review, and shall not simply repeat the contents of the Peer Reviewer Summary Report.

   a. Reviewers should describe in their own words the review activities completed during the panel review meeting, including a concise summary of whether they accept or reject the work that they reviewed, and explain their decisions (strengths, weaknesses of the analyses, etc.), conclusions, and recommendations.

   b. Reviewers should discuss their independent views on each TOR even if these were consistent with those of other panelists, but especially where there were divergent views.

   c. Reviewers should elaborate on any points raised in the Peer Reviewer Summary Report that they believe might require further clarification.

   d. The report may include recommendations on how to improve future assessments.

3. The report shall include the following appendices:

   Appendix 1: Bibliography of materials provided for review

   Appendix 2: A copy of this Performance Work Statement

   Appendix 3: Panel membership or other pertinent information from the panel review meeting.

Appendix 4. Peer Reviewer Summary Report Requirements

1. The main body of the report shall consist of an introduction prepared by the SARC chair that will include the background and a review of activities and comments on the appropriateness of the process in reaching the goals of the SARC. Following the introduction, for each assessment /research topic reviewed, the report should address whether or not each Term of Reference of the Research Track Working Group was completed successfully. For each Term of Reference, the Peer Reviewer Summary Report should state why that Term of Reference was or was not completed successfully.

   To make this determination, the SARC chair and reviewers should consider whether or not the work provides a scientifically credible basis for developing fishery management advice. If the reviewers and SARC chair do not reach an agreement on a Term of Reference, the report should explain why. It is permissible to express majority as well as minority opinions.

   The report may include recommendations on how to improve future assessments.

2. If any existing Biological Reference Points (BRPs) or BRP proxies are considered inappropriate, include recommendations and justification for alternatives. If such alternatives cannot be identified, then indicate that the existing BRPs or BRP proxies are the best available at this time.

3. The report shall also include the bibliography of all materials provided during the SAW, and relevant papers cited in the Peer Reviewer Summary Report, along with a copy of the CIE Performance Work Statement.

   The report shall also include as a separate appendix the assessment Terms of Reference used for the SAW, including any changes to the Terms of Reference or specific topics/issues directly related to the assessments and requiring Panel advice.

Appendix 3. Panel Membership

Review Panel

Paul Rago, chair
Yong Chen, CIE
Robin Cook, CIE
Paul Medley, CIE

Participants
Andrew Jones - NEFSC
Brandon Muffllley - MAFMC
Brian Linton - NEFSC
Brian Stock - NEFSC
Burton Shank - NEFSC
Charles Adams - NEFSC
Charles Perretti - NEFSC
Chris Kellogg - NEFMC
Chris Legault - NEFSC
Chris Tholke - NEFSC
Corinne Truesdale - RIDEM
David Richardson - NEFSC
Deb Lambert - NOAA Fisheries HQ
Gavin Fay - SMAST
Jamie Cournane - NEFMC
Jennifer Couture - NEFMC
John Wiedenmann - Rutgers University
Jon Deroba - NEFSC
Karen E Greene - NOAA Fisheries HQ
Kathy Sosebee - NEFSC
Kelly Whitmore - MADMF
Kiersten Curti - NEFSC
Larry Alade - NEFSC
Liz Brooks - NEFSC
Liz Sullivan - GARFO
Mackenzie Mazur - Gulf of Maine Research Institute
Mark Grant - GARFO
Mark Terceiro - NEFSC
Mike Simpkins - NEFSC
Michele Traver - NEFSC
Paul Nitschke - NEFSC
Quang Huynh - University of British Columbia
Robin Frede - NEFMC
Russ Brown - NEFSC

Steve Cadrin - SMAST
Susan Wigley - NEFSC
Tim Miller - NEFSC
Tom Nies - NEFMC
Toni Chute - NEFSC
Tony Wood - NEFSC
Tyler Pavlowich - NEFSC