Center for Independent Experts (CIE) independent peer report of the American Plaice Research Track

Dr. Massimiliano Cardinale

Executive Summary

This document is the individual Center for Independent Experts CIE) Reviewer report of the American Plaice Research Track conducted during July-August 2022 and provided at the request of the Center for Independent Experts (CIE) (see Appendix 2).

This report solely represents the view of the independent reviewer (Dr. Massimiliano Cardinale). The text in this report is mainly based on the original assessment report (i.e., Report of the American Plaice Research Track Working Group, thereafter referred also as the Report) and background documents provided to the reviewer in advance of the meeting, particularly WP17 Hennen & Hansell (Stock Synthesis) and WP18 Hart et al. (WHAM) as the WHAM model is used to provide management advice and Stock Synthesis to verify that the results from two different model platforms were consistent. Additional comments based on discussions during the American Plaice Research Track Panel Peer Review web meeting and presentation of alternative model configurations are also included in the meeting report. The draft of the web meeting agenda is included in Appendix 2.

The Assessment team tackled all the assigned terms of reference (TORs, see under Appendix 2).

The reviewer considers that the Working Group has done a satisfactory job in carrying out the assessment, analyzing most of the available data and using most of them, modelling part of the uncertainty, and providing extensive sensitivity analyses of the different data sources and alternative model structure. As such, the reviewer accepts the work that he has reviewed and considers it suitable for providing management advice for American plaice under the current stock status.

However, there are several weaknesses in the model structure and model selection process that require future work. First, important data series, albeit presented, are not fully integrated in the WHAM assessment model used to provide advice. Those are among others age and time varying M, separated landings and discards length compositions by fleet and their associated age length keys (ALKs), separated length compositions and ALKs of the surveys with their associated uncertainty, precision in age reading, the use of a SR relationship, the possibility of estimating M within the model, defining spatial units with specific biology, use of historical data and others. Second, the reviewer considers that several aspects of the process that leads to the "best case" model configuration used to provide advice as described in the Report can be greatly improved. The revieweconsiders that the process of model selection could be improved.

Diagnostics should be used in combination (and not in isolation) to compare and select models, including navigating between different model configuration and their pruning. to The development of the alternative model configurations could benefit to a factorial structure instead of a linear one, where alternative hypotheses are represented as ramification or evolution of the original or base configuration.

Criteria for model selection and pruning should not be based on derived quantities as SSB or reference points but should be centered on diagnostics that allow comparison between models with different weight of the model components and different data sources, which is the norm in modern stock assessment. Thus, AIC is not recommended, while MASE, Mohn's rho, and quantitative analysis of the residuals should be preferred.

Objective criteria as above should be augmented by first principles. First principles are particularly useful to build base case scenarios from which model exploration should be derived. Establishing a base scenario would also facilitate navigation between the different model configurations by external readers. Alternative model configurations should be based on hypothesis testing.

The reviewer considers that a "best model" cannot be singled out to be used for advice given the diagnostics presented here. Several WHAM model configurations and even different model platforms achieve comparable performances in the terms of model diagnostics but sometimes different stock status in terms of depletion (e.g., SS_Model_Run_BASE14fixFleet). In addition, the current and possible alternative structure of one of the model platforms, Stock Synthesis, has several aspects that I find more appropriate by first principles and are supported by literature simulations studies. These include among others the integration of length compositions and ALKs within the model, the use of ageing accuracy and precision by age and time, modelling selectivity by length, defining spatial units with explicit biology, estimating M within the model, numerous time varying options, and many others not listed here (but see Detailed Comments on the Report of the American Plaice Research Track Working Group hereafter referred as Detailed Reviewer Report). Thus, I recommend that in the future an ensemble of different plausible configurations and model platforms selected and weighed by a comprehensive diagnostic against performance criteria agreed beforehand is developed to provide stocks status and management advice for American plaice. This is particularly important given the uncertainties in the data used as input, and in key biological parameters and processes in the context of providing probabilistic statements of stock status (see also Recommendations section).

The reviewer also considers that the diagnostics tools used by the Working Group to evaluate the robustness of the model are appropriate but still incomplete and should be augmented following recent publications (see details in the Detailed Reviewer Report).

Findings that are reported in the American Plaice Research Track Working Group Report are not necessarily fully repeated in this individual report. This report focuses on clarification of elements contained in the American Plaice Research Track Working Group Report (including also information presented in all background documents) and some additional views of the individual reviewer about how available data could have been better exploited and model selection improved to derive more robust estimates of the exploitation rate and stock status of American plaice stock.

Further recommendations aimed at improving the assessment of American plaice as presented in the American Plaice Research Track Working Group Report were made and included in the Detailed Reviewer report below.

Introduction

The American Plaice Research Track Working Group Report, associated background documents containing detailed information on the data used in the assessment and input files of the assessment models were provided to the independent reviewer (Dr. Massimiliano Cardinale) well in advance of the deadline. The reports and documentations were reviewed at the request of the Center for Independent Experts (CIE).

Description of reviewer activities

This review was undertaken by Dr. Massimiliano Cardinale during July-August 2022 at the request of the Center for Independent Experts (CIE) (see Annex 1).

Relevant documents (see Annex 2, with the list of background documents and how to access those via web) and background information were made available two weeks prior to the deadline through email and via a link to the American Plaice Research Track Working Group Report data portal (https://apps-nefsc.fisheries.noaa.gov/saw/sasi/uploads/Readme%20file_Please%20Read%20first.pdf). The Report was made available two weeks' prior the deadline via a weblink (https://apps-nefsc.fisheries.noaa.gov/saw/sasi/sasi_report_options.php). Details of the WHAM models tested were available at https://drive.google.com/drive/u/0/folders/1qOp68jfubFrHTss0vx-O8hhOyUBtD_1S. The documentation was reviewed prior to the deadline and the deadline was met. The background information and assessment Report of American plaice was presented through several documents (see bibliography in Annex 2). Background information relevant to this review is presented in a series of appendices, including a bibliography (Appendix 1) and Performance Work Statement and associated Terms of Reference (Appendix 2) Comments included here are provided following the TORs and are those of the independent reviewer only.

Summary of the main findings

Important data series, albeit presented, are not fully integrated in the WHAM assessment model used to provide advice as for example separated landings and discards with their respective uncertainty, precision of the ageing estimates by year, time varying natural mortality, historical data, and many others. Also, age length keys (ALKs) and length compositions are combined outside the assessment model to estimate number at age of the catches and in the surveys, which are then used as a direct input to the different WHAM model configurations. Instead, integrating ALKs and length compositions within the model as done in the Stock Synthesis configuration would have the advantage to allow tracking changes in growth over time (and distinguish those from changes in condition) and to integrate uncertainty in both length and age composition (i.e., ageing precision and accuracy by age and time when available), which is then in turn translated into uncertainty of the derived quantities and stock status as described by the Kobe plot (Figure 5.8 of the Report).

The process that leads to the “best case” model configuration used to provide advice can be greatly improved. The assessment Team should decide a priori the criteria for model selection and use them in combination to compare and select models. Those criteria should be clearly listed at the beginning of the model selection process and referred each time a model is selected or discarded, which would make the process much easier to follow than it is currently. In particular, the use of AIC is not recommended, as it does not allow comparison between models with different weight of the model components and different data sources, which is the norm in modern stock assessment. Finally, the alternative model configurations should be presented in the context of hypothesis testing clearly indicating which alternative aspect of the biology, ecology or fisheries is being tested.

Based on the diagnostics presented in the Report, a “best model” cannot be singled out to be used for advice. Thus, although current stock status is reasonably robust to different hypotheses on data and model structure, in the future an ensemble should be developed to provide stocks status and management advice for American plaice. The ensemble might include different plausible configurations of the WHAM model and model platforms as Stock Synthesis. The alternative configurations should be selected and weighed by a comprehensive diagnostic made against performance criteria agreed beforehand. This is particularly important given the uncertainties in the data used as input, and in key biological parameters and processes in the context of providing probabilistic statement of stock status.

In addition, one of the model platform, Stock Synthesis has several aspects that I find more appropriate by first principles and are supported by simulations studies as for example the possibility of integrating length compositions and age length keys within the model, fitting selectivity by length, defining spatial units with specific biology as growth, natural mortality and maturity, estimating M within the model, numerous time varying options, and many others not listed here (but see Detailed comments on the Report of the American Plaice Research Track Working Group).

Conclusions and recommendations

Not all data available and presented are used and some are underutilized, which might affect model results, particularly impacting uncertainty, and probabilistic statements on stock status. Thus, I recommend that in the future those are fully integrated in the assessment model used to provide advice. In this context, it is recommended that historical data as far back in time as possible are collated and used in future assessments.

It is recommended to integrate time varying biology, in particular M, and the spatial dimension in the long term to be line with findings and conclusions of TOR1 (Ecosystem and climate influences). Also, as plaice display sexually dimorphic growth, it is reasonable to assume that sex ratio changes with depletion rate of the population linked to periods of high and low F. This will in turn impact M and growth in a single sex model. Thus, developing a sex separated model would also be recommended in the long term.

Model selection process is inconsistent and challenging to follow. It is recommended that the criteria for model selection are established a priori and used in combination (and not in isolation) to compare and select models, including navigating between different model configurations and their pruning. In particular, the use of AIC is not recommended.

The model diagnostic toolbox should be expanded to include as a minimum runs test of the residuals, hindcasting MASE for all models, ASPM and MCMC.

Presentation of the numerous runs tested, and their diagnostic could be further improved, for example using a shiny app (see for example https://maxcardinale.shinyapps.io/Ensemble_2022/).

Several WHAM model configurations and even different model platforms achieve comparable performances in the terms of model diagnostics but sometimes different stock status in terms of depletion (e.g., depletion rate is less than the reference point (SSB*40%*) for the SS model BASE14fixFleetCode compared to WHAM final model used for advice). In this context, the role of the sensitivity analysis is unclear. As the stock status is described in a probabilistic manner, integrating structural uncertainty would have significant effects on the probabilities estimated in the Kobe plot. Thus, I recommend that an ensemble of different plausible model configurations selected using hypothesis testing and weighed by a comprehensive diagnostic against performance criteria agreed beforehand should be used to provide stocks status and management for American plaice in the future.

A SR relationship is not included in the existing WHAM model used to provide advice. Although this has most likely a negligible effect on the current stock status and short-term forecast, ignoring SR has its largest impact when modelling long term dynamic as for example through Management Strategy Evaluations (MSE). Assuming average recruitment at all levels of SSB runs the risk of overestimating recovery potential when the stock is low, which has important consequences for rebuilding plans. In terms of the SPR target levels and how F_{SPR} relates to F_{MSY}, for SPR fraction = 0.4, F_{SPR} exceeds F_{MSY} at steepness levels below 0.65. Thus, given the assumed best estimate of steepness being less than 0.65, there are some risks associated to an $F_{SPR40\%}$.

TORs (In *italics* is a condensed answer of the reviewer to each specific TOR; detailed elaborations of each identified issue can be found in the detailed Reviewer Report below)

1. Identify relevant ecosystem and climate influences on the stock. Characterize the uncertainty in the relevant sources of data and their link to stock dynamics. Consider findings, as appropriate, in addressing other TORs. Report how the findings were considered under impacted TORs.

The Working Group fully addressed and met this TOR.

Ecosystem indicators were not included in the stock assessment models used for providing advice for the American plaice. However, an attempt to integrate some of the key environmental variables as temperature was made but those model configurations were discarded based primarily on first principles (i.e., expected relationship between recruitment and temperature conflicted with analysis under TOR1 (Ecosystem and climate influences)) and the qualitative analysis of the residuals. Biology, except weight at age, is considered time unvarying. Nevertheless, time varying biology as for example yearly estimates of natural mortality by age based on changes in size at age (i.e., growth) over time is presented, although not used in any of the assessment models. Time varying natural mortality (M) represents the natural link between ecosystem and climate influence on key productivity parameters as it is strongly associated with growth and maturation. It exemplifies the realized effect of the environment on the biology of the species and its link with growth and maturity in fish has a robust literature underpinning. These key productivity parameters are also the focus of TOR1 and therefore the use of time and age varying M would represent the obvious and easier way *of integrating realized ecosystem effect on the American plaice in the stock assessment. Instead, M is assumed time and age unvarying in WHAM model configurations used for providing advice, which conflicts with findings and conclusions of TOR1. Thus, given the observed increase in water temperature occurring in the area coupled with the large decline in growth (i.e., size at age) of age classes 6 and older, it is conceivable that M has increased over time, which should be the primary hypothesis in the assessment models.*

2. Estimate catch from all sources including landings and discards. Describe the spatial and temporal distribution of landings, discards, and fishing effort. Characterize the uncertainty in these sources of data.

The Working Group fully addressed and met this TOR.

Catch and discards data are well described and presented in background documents and in the Report. The data presented included spatial and temporal distribution of landings and discards and associated size and age compositions. However, size compositions data of the landings are not presented by fleet and all commercial catches were pooled into a single pseudo-fleet when used in the assessment models. This is in theory fine if the pooled fleets have similar selectivity and/or if the proportion between the fleets is approximately constant between years. However, this does not seem to be the case when analyzing data presented in section TOR2 of the Report, which would imply that allowing for separate fleets in the model would be more appropriate.

Uncertainty associated to the different data sources is estimated and well presented. However, important data series, albeit presented, are not integrated in the WHAM assessment model used to provide advice. For example, separated landings and discards with their respective age compositions, uncertainty of the landings and discards, precision of the ageing estimates by age and year, time varying natural mortality, and many others. Also, age length keys (ALKs) and length compositions are combined outside the assessment model to estimate number at age of the catches and in the surveys, which are then used as a direct input to the different WHAM model configurations. Instead, integrating ALKs and length compositions within the model as done in the Stock Synthesis configuration would have the advantage to allow tracking changes in growth over time and integrating uncertainty in both length and age composition (i.e., ageing precision and accuracy by age and time when available). This will be translated into uncertainty of the derived quantities and impact the stock status in probabilistic terms.

Catch data prior to 1980s are excluded from the final model. The only (putative) rationale seems to be the difference in SSB$_{40\%}$ to which no explanation is given in WD 18 although it is not repeated in the Report. Historical catch data generally are informative of the scale of the stock at low level of F and thus have an important effect on the biomass reference points and long time series of catches are informative of scale. Thus, excluding the extended time series based on difference in biomass reference points is illogical and should not be pursued. Also, as information on historical catches of American plaice prior to 1960s might exist (https://maineanencyclopedia.com/american-plaice-landings/), it is recommended that historical catches as far back in time as possible are collated and used in the future as an alternative model configuration.

3. Present the survey data used in the assessment (e.g., indices of relative or absolute abundance, recruitment, state surveys, age-length data, application of catchability and calibration studies, etc.) and provide a rationale for which data are used. Describe the spatial and temporal distribution of the data. Characterize the uncertainty in these sources of data.

The Working Group fully addressed and met this TOR.

Design based survey indices and combined spatio-temporal integrated survey index based on model standardization were estimated for American plaice. Trawl surveys data were modelled using the spatio-temporal model VAST to produce yearly estimate of relative biomass for the assessment. Data and associated uncertainty are well presented. Inshore surveys (MADMF and MENH) were excluded as separated survey time series because of conflicting signals and alleged moving of the plaice deeper with time. Thus, only NEFSC surveys were used in the WHAM assessment. As expected, there are similarities and differences between Design based and VAST based indices. The use of fully standardized VAST indices is generally considered to achieve less retrospective bias and outperform assessments with design-based indices. Also, combined standardized VAST indices, which also includes inshore surveys, are theoretically more suited for a single area model as the American plaice WHAM model. This is because the use of a single spatio-temporal standardized index avoids problematic conflicts which can arise when several design-based survey indices of the same indicator are used but do not exactly cover the same time and space.

The analysis showed that the center of gravity of the stock is variable over time with a latitudinal trend in the last two decades. This implies that any standardization of the trawl surveys needs to account for the interaction between space and time in the distribution of

American Plaice. An important result is that the water temperature at which the haul is carried out is a significant variable in determining American plaice distribution, abundance, and biomass. This implies that corresponding standardization of LPUE and CPUE from fisheries dependent data which do not account for temperature might provide a biased trend in relative abundance and biomass. Also, the same would be valid for design-based survey indices, which reinforce the idea that spatio-temporal standardized indices should be preferred over other indices of relative abundance. Instead, separated design-based indices of the offshore surveys were used in the WHAM model. The reason why combined standardized VAST index was not used in the final model is that their performances in terms of diagnostics was considered inferior when compared to separated Albatross and Bigelow survey indices. However, I cannot see significant differences between WHAM model 28B and 29F compared to models using VAST indices. As for most of the comparisons, those are based mainly on qualitative analysis of the residuals, which by nature are hard to follow, and lower (but still under the threshold) Mohn´s rho. On the other hand, MASE was not calculated for VAST models as it was not shown for most of the alternative configurations presented in the Report.

4. Use appropriate assessment approach to estimate annual fishing mortality, recruitment and stock biomass (both total and spawning stock) for the time series and estimate their uncertainty. Compare the time series of these estimates with those from the previously accepted assessment(s). Evaluate a suite of model fit diagnostics (e.g., residual patterns, sensitivity analyses, retrospective patterns), and (a) comment on likely causes of problematic issues, and (b), if possible and appropriate, account for those issues when providing scientific advice and evaluate the consequences of any correction(s) applied.

The Working Group fully addressed and met this TOR.

I consider that the data used within the presented assessment models are generally appropriate and data uncertainty sufficiently acknowledged, albeit not fully integrated (see sections above). The models used to conduct the data preparation for the assessments are suitable for the available data as well as the data series are adequate to support the assessment models used. The choice of the various analytical tools used to derive the data is well justified in the background documents presented. The models (i.e., WHAM, Stock Synthesis, VPA and ASAP) used to assess the American plaice stock are appropriate, robust and in general properly configured, and in line with standard practices. However, the process of selection of the final model is difficult to follow. The reviewer considers that the process of model selection could be improved. Diagnostics should be used in combination (and not in isolation) to compare and *select models, including navigating between different model configuration and their pruning. The development of the alternative model configurations could benefit to a factorial structure instead of a linear one, where alternative hypotheses are represented as ramification or evolution of the original or base configuration. increasesoneAlso, and importantly, the process used to select between model configuration candidates and pruning (i.e., discarding certain model configurations along the path of model development) is centered on AIC, retrospective analysis and qualitative analysis of the residuals. Criteria for model selection and pruning should not be based on derived quantities as SSB or reference points but should be centered on diagnostics that allow comparison between models with different weight of the model components and different data sources, which is the norm in stock assessment. Thus, AIC is not recommended, while Mohn's rho, quantitative analysis of residuals and MASE should be preferred. Objective criteria as above can be augmented by first principles. First principles are particularly useful to build a base case scenario from which model exploration could be derived. A clear definition of a base case scenario is missing in the Report and its definition at the beginning of the model development would have been helpful to allow the reader to follow the process. Alternative model configurations should be based on well-defined hypothesis testing for each model.*

Another major concern I have is the role of the sensitivity analysis in the model development process and consequently how data and structural uncertainty is treated in the context of providing advice. A key part of the uncertainty (i.e., structural uncertainty and uncertainty for some of the data components) is not included when using "best case" philosophy for model development but it is only presented as sensitivity analysis and thus has no impact on the stock status and on the management advice. A large effort is made by the Working Group to develop and evaluate different model configurations, many of those having at least equal performances in terms of model diagnostics as the "best case", which is then proposed to be used for providing management advice. Thus, I recommend that in the future, an ensemble of different plausible model configurations and platforms, selected and weighed by a comprehensive diagnostic against performance criteria agreed beforehand, should be developed to provide stocks status and management advice for American plaice. As best practice, and as a minimum, the ensemble should integrate the three main sources of uncertainty, process uncertainty, parameter uncertainty and observation error in the data. The ensemble should also be used for deriving catch forecast scenarios, in which plausible assumptions on the productivity of the *stock (e.g., recruitment, growth, mortality, etc.) can be integrated to mimic variability of the ecosystem and possible effects of changing climate.*

The model diagnostic toolbox should be greatly expanded to include as a minimum quantitative analysis of the residuals as runs test, hindcasting and MASE of all models, ASMP and MCMC. Finally, presentation of the numerous runs tested, and their diagnostics should be further improved, for example using a shiny app.

5. Update or redefine status determination criteria (SDC; point estimates or proxies for BMSY, BTHRESHOLD, FMSY and MSY reference points) and provide estimates of those criteria and their uncertainty, along with a description of the sources of uncertainty. If analytic model-based estimates are unavailable, consider recommending alternative measurable proxies for reference points. Compare estimates of current stock size and fishing mortality to existing, and any redefined, SDCs.

The Working Group fully addressed and met this TOR.

SPR fractions tailored to current conditions in terms of stock biology are used as reference points. This is fully justified as the WHAM model does not include an SR function and weight at age is changing over time. Given the SR pairs estimated by the WHAM model and the use of SPR_0 fraction-based reference points, the absence of a SR is not too relevant for the determination of American plaice reference points at current conditions, with the stock being in healthy status. I also recognize that ignoring the existence of a functional form of the SR curve used in conjunction with average recruitment in the projections and SPR_0 fraction as reference points has limited impact on the short-term forecast advice, which span for 3 years. However, ignoring SR has its largest impact when modelling long term dynamics as for example when conducting an MSE. Assuming average recruitment at all levels of SSB runs the risk of overestimating recovery potential when the stock is low which has important consequences for rebuilding plans. Finally, as described above and in the detailed report of the American Plaice Research Track assessment review, a key part of the uncertainty (i.e., structural uncertainty and uncertainty of some of the data components) is not included in the model, which has direct consequences on the probabilistic statement of the stock status.

6. Define appropriate methods for producing projections; provide justification for assumptions of fishery selectivity, weights at age, maturity, and recruitment; and comment

on the reliability of resulting projections considering the effects of uncertainty and sensitivity to projection assumptions.

The Working Group fully addressed and met this TOR.

Methods and assumptions for short term projections are adequate and well described. The short-term forecast has a 3-year span. For short periods, which are typically 3 years or less, assuming that future recruitment, selectivity and observed weight at age will resemble recent estimated or observed values as done for American plaice has been shown to be a reasonable hypothesis by different simulations studies. However,,it is important to note that short term projections performance will deteriorate as the time interval increases (typically beyond the 3 years interval for the projections and 3-5 years assumptions for recruitment, selectivity, and biology). Unlike previous assessments, the current projection methodology in WHAM model included uncorrelated process variance in survival and recruitment. This process variance was then carried forward into the projections.

There is no evidence of a clear SR relationship when analyzing SR pairs estimated by WHAM. This justifies the non-use of steepness in the assessment model and the use of fraction of SPR_0 (i.e., SPR40%) as reference point for American plaice in the short-term forecast. In terms of the SPR target levels and how F_{SPR0} relates to F_{MSY}, for SPR fraction = 0.4 F_{SPR0} exceeds F_{MSY} at steepness levels below 0.65. Thus, given the assumed best estimate of steepness being less than 0.65, there are some (albeit small) risks associated to an FSPR40%. When looking at potential depletion level of SSB as a proxy for limit reference point (e.g., 20% B_0), at FSPR40% depletion will be less than the limit reference point only when steepness is less than 0.5 so that short term forecast based on the current set of reference points is well justified.

7. Review, evaluate, and report on the status of research recommendations from the last assessment peer review, including recommendations provided by the prior assessment working group, peer review panel, and SSC. Identify new recommendations for future research, data collection, and assessment methodology. If any ecosystem influences from TOR 2 could not be considered quantitatively under that or other TORs, describe next steps for development, testing, and review of quantitative relationships and how they could best inform assessments. Prioritize research recommendations.

The Working Group fully addressed and met this TOR.

The reviewer considers that the Working Group has done a comprehensive evaluation of all research recommendations made by the last assessment peer review. The WG concluded that *all the previous research commendations had been addressed except ageing samples for the MADMF inshore survey to which the reviewers also agree. The WG has also developed a series of new recommendations, which the reviewer supports. Among the additional recommendations included in the Summary Report of the American Plaice Research Track Stock Assessments Peer Review, the reviewer consider the development of an ensemble of different plausible configurations and model platforms as described in the* ***detailed report of the American Plaice Research Track assessment review*** *as a top priority to in future American plaice Research Track working group. Each different plausible configurations and model platforms included in the ensemble should be weighed by a comprehensive diagnostic against performance criteria agreed beforehand. As best practice, and as a minimum, the ensemble should integrate the three main sources of uncertainty, process uncertainty, parameter uncertainty and observation error in the data. The different model configurations should mimic different overarching assumptions as for example natural mortality, selectivity, time series of catches, etc. The results of the ensemble should be then used to provide stocks status and management advice for American plaice. The ensemble should also be used for deriving catch forecast scenario, in which plausible assumptions on the productivity of the stock (e.g., recruitment, growth, mortality, and others) can be integrated to mimic future variability of the ecosystem and possible effects of climate factors.*

8. Develop a backup assessment approach to providing scientific advice to managers if the proposed assessment approach does not pass peer review or the approved approach is rejected in a future management track assessment.

The Working Group fully addressed and met this TOR.

The Working Group has investigated numerous alternative assessment approaches to deliver scientific advice if the proposed assessment approach is rejected. Both different model platforms and models' configurations were indicated as suitable alternative in case the proposed WHAM assessment is not accepted. The assessment Team indicated that ASAP (Run 43) is their favorite candidate as an alternative model for providing advice for American Plaice. Thus, under the current conditions, the reviewer agrees with the WG recommended ASAP (Run 43) as a suitable candidate in the case the proposed assessment approach does not pass peer review, or the approved approach is rejected in a future management track assessment.

Appendix 1: Background material

American Plaice WG Report

Model Selection Procedure for American Plaice Research Track 2022

Ecological Influences (ToR1)

WP_14. Ecosystem and Climate Influences, by Jamie Behan, Lisa Kerr, Amanda Hart, Alex Hansell, Tyler Paklovitch and Steve Cadrin (November 16, 2021)

WP_16. Plaice Ecosystem Drivers by Jamie Behan and Lisa Kerr (June 21, 2022)

Fishery Data (ToR2)

WP_5. Fishing Industry Knowledge of American plaice, by Tyler Pavlowich, David Richardson, John Manderson and Greg DeCelles (November 9, 2021)

WP_6. Exploration of Fishery Data to Evaluate Catch Rates of American Plaice, by Max Grezlik, Lucy McGinnis, Keith Hankowsky, Gavin Fay, Steve Cadrin and Alex Hansell (November 10, 2021)

WP_7. Catch Rates of American Plaice Trawl Fishery, by Keith Hankowsky, Max Grezlik, Lucy McGinnis, Gavin Fay, Steve Cadrin and Alex Hansell (November 12, 2021)

WP_8. American plaice catch rate analysis using a spatial model, by Andy Jones, Tyler Pavlowich, David Richardson and Anna Mercer (November 13, 2021)

WP_9. Fishery Dependent Data Indices of Abundance (LPUE or CPUE) for American Plaice, by Mark Terceiro (November 16 2021)

WP_10. Electronic Monitoring Data: American Plaice, by Cate O'Keefe, Mel Sanderson and Liz Moore (December 4 2021)

WP_19. Fishery Data, by Larry Alade

Survey Data (ToR3)

WP_11. Seasonal Variation in Size-at-Age of American Plaice from Survey Data, by Steve Cadrin (November 22 2021)

WP_12. Spatio-temporal dynamics of American plaice (*Hippoglossoides platessoides*) in US waters of the northwest Atlantic, by Alexander Hansell, Larry Alade, Andrew Allyn, Lauran Brewster, Steve Cadrin and Lisa Kerr (December 1 2021; updated July 2022)

WP_13. Relative efficiency of a chain sweep and the rockhopper sweep used for the NEFSC bottom trawl survey and biomass estimates for American plaice, by Timothy J. Miller, David E. Richardson, Andrew Jones and Phil Politis (December 9 2021)

WP_20. Survey Data, by Larry Alade

Biology (ToR4)

WP_1. Size distribution analysis of American plaice, by Tyler Pavlowich (August 2021)

WP_2. Overview of American Plaice ageing in the Northwest Atlantic, by Josh Dayton and Eric Robillard (September 10 2021)

WP_3. Updating Parameters for Length and Weight Relationships and Length at Age of American Plaice, by Ashley Silver, Tyler Pavlowich and Larry Alade (September 10, 2021)

Appendix 2: Performance Work Statement (PWS)

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

American Plaice Research Track Virtual Peer Review

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].

Scope

The Research Track Peer Review 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 research track peer review is the cornerstone of the Northeast Region Coordinating Council stock assessment process, which includes assessment development, and report preparation (which is done by Working Groups or Atlantic States Marine Fisheries Commission (ASMFC) technical committees), assessment peer review (by the peer review panel), public presentations, and document publication. The results of this peer review will be incorporated into future management track assessments, which serve as the basis for developing fishery management recommendations.

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

[1] https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2005/m05-03.pdf

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. Modifications to the PWS and ToRs cannot be made during the peer review, and any PWS or ToRs modifications prior to the peer review shall be approved by the Contracting Officer's Representative (COR) and the CIE contractor. All TORs must be addressed in each reviewer's report. The reviewers shall have working knowledge and recent experience in the use and application of index-based, age-based, and state-space 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 virtually 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 Peer Review Panel (co)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 peer review meeting, using the criteria specified below in the "Tasks for Peer Review 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 peer review 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 Peer Review Panel 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 Peer Review Panel chair and reviewers combined:
Review the Report of American plaice Research Track Working Group.

The Peer Review Panel 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 peer review meeting. For terms where a similar view can be reached, the Peer Reviewer Summary Report will contain a summary of such 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 their 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 WebEx video conferencing.

Period of Performance
The period of performance shall be from the time of award through September, 2022. 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
July 18-21, 2022	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

PWS Appendix 1. Generic Research Track Terms of Reference

1. Identify relevant ecosystem and climate influences on the stock. Characterize the uncertainty in the relevant sources of data and their link to stock dynamics. Consider findings, as appropriate, in addressing other TORs. Report how the findings were considered under impacted TORs.

2. Estimate catch from all sources including landings and discards. Describe the spatial and temporal distribution of landings, discards, and fishing effort. Characterize the uncertainty in these sources of data.

3. Present the survey data used in the assessment (e.g., indices of relative or absolute abundance, recruitment, state surveys, age-length data, application of catchability and calibration studies, etc.) and provide a rationale for which data are used. Describe the spatial and temporal distribution of the data. Characterize the uncertainty in these sources of data.

4. Use appropriate assessment approach to estimate annual fishing mortality, recruitment and stock biomass (both total and spawning stock) for the time series, and estimate their uncertainty. Compare the time series of these estimates with those from the previously accepted assessment(s). Evaluate a suite of model fit diagnostics (e.g., residual patterns, sensitivity analyses, retrospective patterns), and (a) comment on likely causes of problematic issues, and (b), if possible and appropriate, account for those issues when providing scientific advice and evaluate the consequences of any correction(s) applied.

5. Update or redefine status determination criteria (SDC; point estimates or proxies for BMSY, BTHRESHOLD, FMSY and MSY reference points) and provide estimates of those criteria and their uncertainty, along with a description of the sources of uncertainty. If analytic model-based estimates are unavailable, consider recommending alternative measurable proxies for reference points. Compare estimates of current stock size and fishing mortality to existing, and any redefined, SDCs.

6. Define appropriate methods for producing projections; provide justification for assumptions of fishery selectivity, weights at age, maturity, and recruitment; and comment on the reliability of resulting projections considering the effects of uncertainty and sensitivity to projection assumptions.

7. Review, evaluate, and report on the status of research recommendations from the last assessment peer review, including recommendations provided by the prior assessment working group, peer review panel, and SSC. Identify new recommendations for future research, data collection, and assessment methodology. If any ecosystem influences from TOR 2 could not be considered quantitatively under that or other TORs, describe next steps for development, testing, and review of quantitative relationships and how they could best inform assessments. Prioritize research recommendations.

8. Develop a backup assessment approach to providing scientific advice to managers if the proposed assessment approach does not pass peer review or the approved approach is rejected in a future management track assessment.

Research Track TORs:

General Clarification of Terms that may be Used in the Research Track Terms of Reference

Guidance to Peer Review Panels 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 peer review 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.

PWS Appendix 2. Draft Review Meeting Agenda

{Final Meeting agenda to be provided at time of award}

American plaice Research Track Assessment Peer Review Meeting

July 18-22, 2022

WebEx link: TBD

DRAFT AGENDA* (v. 5/3/2022)

**All times are approximate, and may be changed at the discretion of the Peer Review Panel 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 Peer Review Panel.*

Monday, July 18, 2022

Time	Topic	Presenter(s)	Notes
9 a.m. - 9:30 a.m.	Welcome/Logistics Introductions/Agenda/ Conduct of Meeting	Michele Traver, Assessment Process Lead Russ Brown, PopDy Branch Chief Yong Chen, Panel Chair	
9:30 a.m. - 10:30 a.m.	TOR #1		
10:30 a.m. - 10:45 a.m.	Break		
10:45 a.m. - 11:45 a.m.	TOR #2		
11:45 a.m. - 12:15 p.m.	Discussion/Summary	Review Panel	
12:15 p.m. - 12:30 p.m.	Public Comment	Public	
12:30 p.m. - 1:30 p.m.	Lunch		
1:30 p.m. - 3 p.m.	TOR #3		
3 p.m. - 3:15 p.m.	Break		
3:15 p.m. - 4:15 p.m.	TOR #4		
4:15 p.m. - 4:45 p.m.	Discussion/Summary	Review Panel	
4:45 p.m. - 5 p.m.	Public Comment	Public	
5 p.m.	Adjourn		

Tuesday, July 19, 2022

Time	Topic	Presenter(s)	Notes
9 a.m. - 9:15 a.m.	Welcome/Logistics	Michele Traver, Assessment Process Lead Yong Chen, Panel Chair	
9:15 a.m. - 10:30 a.m.	TOR #5		
10:30 a.m. - 10:45 a.m.	Break		
10:45 a.m. - 11:45 a.m.	TOR #6		
11:45 a.m. - 12:15 p.m.	Discussion/Summary	Review Panel	
12:15 p.m. - 12:30 p.m.	Public Comment	Public	
12:30 p.m. - 1:30 p.m.	Lunch		
1:30 p.m. - 3 p.m.	TOR #7		
3 p.m. - 3:15 p.m.	Break		
3:15 p.m. - 4:15 p.m.	TOR #8		
4:15 p.m. - 4:45 p.m.	Discussion/Summary	Review Panel	
4:45 p.m. - 5 p.m.	Public Comment	Public	
5 p.m.	Adjourn		

Wednesday, July 20, 2022

Time	Topic	Presenter(s)	Notes
9 a.m. - 9:15 a.m.	Welcome/Logistics	Michele Traver, Assessment Process Lead Yong Chen, Panel Chair	
9:15 a.m. - 10:30 a.m.	TOR #5		
10:30 a.m. - 10:45 a.m.	Break		
10:45 a.m. - 11:45 a.m.	TOR #		
11:45 a.m. - 12:15 p.m.	Discussion/Summary	Review Panel	
12:15 p.m. - 12:30 p.m.	Public Comment	Public	
12:30 p.m. - 1:30 p.m.	Lunch		

1:30 p.m. - 3 p.m.	TOR #		
3 p.m. - 3:15 p.m.	Break		
3:15 p.m. - 4:15 p.m.	TOR #		BRPs, Projections and EGB Reference Points
4:15 p.m. - 4:45 p.m.	Discussion/Summary	Review Panel	
4:45 p.m. - 5 p.m.	Public Comment	Public	
5 p.m.	Adjourn		

Thursday July 21, 2022

Time	Topic	Presenter(s)	Notes
9 a.m. - 5 p.m.	Report Writing	Review Panel	

PWS 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.

PWS Appendix 4. Peer Reviewer Summary Report Requirements

1. The main body of the report shall consist of an introduction prepared by the Research Track Peer Review Panel 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 peer review meeting. 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 peer review panel chair and reviewers should consider whether or not the work provides a scientifically credible basis for developing fishery management advice. If the reviewers and peer review panel 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 peer review meeting, 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 peer review meeting, including any changes to the Terms of Reference or specific topics/issues directly related to the assessments and requiring Panel advice.

Appendix 3: List of participants

NEFSC - Northeast Fisheries Science Center

GARFO - Greater Atlantic Regional Fisheries Office

NEFMC - New England Fisheries Management Council

SMAST - University of Massachusetts School of Marine Science and Technology

GMRI - Gulf of Maine Research Institute

MADMF - Massachusetts Division of Marine Fisheries

Yong Chen - Chair
Steven Holmes - CIE Panel
Peter Stephenson - CIE Panel
Massimiliano Cardinale - CIE Panel

Russ Brown - NEFSC, Population Dynamics Branch Chief
Michele Traver - NEFSC, Assessment Process Lead

Alex Dunn - NEFSC
Alex Hansell - NEFSC
Alicia Miller - NEFSC
Amanda Hart - SMAST
Angela Forristall - NEFMC Staff
Charles Adams - NEFSC
Charles Perretti - NEFSC
Chris Kellogg - NEMFC Staff
Cole Carrano - SMAST
Dan Hennen - NEFSC
David McCarron - MADMF (retired)
Jackie ODell - Executive Director of Northeast Seafood Coalition
Jamie Behan - GMRI
Jamie Cournane - NEFMC Staff
Jason Boucher - NEFSC
Kathy Sosebee - NEFSC
Libby Etrie - NEFMC Member
Lisa Kerr - GMRI
Mark Alexander - Asst. Director (retired) of the Fisheries Division, Connecticut Dept. of Energy & Environmental Protection
Mark Terceiro - NEFSC
Max Grezlik - SMAST
Paul Nitschke - NEFSC
Robin Frede - NEFMC Staff
Steve Cadrin - SMAST
Tim Miller - NEFSC
Tony Wood - NEFSC

Appendix 4: Detailed report of the American Plaice Research Track assessment review

List of the TORs of the American Plaice Research Track Working Group

TOR1: Ecosystem and Climate Influences

TOR2: Fishery Data

TOR3: Survey Data

TOR4: Estimate Stock Size and Fishing Mortality

TOR5: Status Determination Criteria

TOR6: Projection Methods

TOR7: Research Recommendations

TOR8: Backup Assessment Approach

General comments (*in italics*) on each presented WP. Note that the WPs are order by number and not by to which TOR they belong, which is indicated in brackets after the WP title.

WP1 Pavlovich size distribution analysis **(TOR3)**

Size distribution analysis of American plaice

By Tyler Pavlowich

The size and abundance of the largest fish caught during bottom trawl surveys increases considerably from the inshore strata moving deeper in all areas, with deeper areas holding a higher proportion of large plaice. There is substantial year-to-year variability between size distributions for all areas and strata. Georges Bank generally harbors many fewer plaice than the rest of the Gulf of Maine although large fish were present even in Georges Bank during 1980s so that spatial depletion following intense exploitation during 1990s and 2000s could be *not discarded. Current assessment model configurations are as one area model. However, although the biology appears to be similar different between areas (WP 3 and 4), there are notable differences in growth (and thus most likely in natural mortality (M) between areas. Therefore, reference points might differ between a spatially aggregated and disaggregated model and a spatially aggregated assessment is most likely to cause depletion of the components that is more sensitive to exploitation (Okamoto et al., 2020).*

A commonly used approach to account for spatial structure is using the 'areas-as-fleets' approach in which fishery or survey selectivity and catchability are assumed to differ spatially. However, several simulation studies suggest that adopting spatial approaches to stock assessment will improve estimation performance compared to the areas-as-fleets approach or ignoring spatial structure when conducting stock assessments, although at the cost of a larger number of estimable parameters (Punt 2019).

Regional variation in growth (and consequently in natural mortality and maturation) between the Gulf of Maine and Georges Bank has been recognized in the past (NEFSC 1999a, 2001a). WHAM models used for assessing American plaice are all one fleet, one area models. However, one of the alternative models used for American plaice (e.g., Stock Synthesis) allows for multiple areas and multiple fleets, which will be able to account for the different biology (i.e., growth and maturity) and size structure observed between Georges Bank and Gulf of Maine and possible difference in exploitation rates between areas. One future recommendation would be to develop a spatial model for comparison with the current single area model.

WP2 Dayton _ Robillard Age Determination **(TOR2)**

Overview of American Plaice ageing in the Northwest Atlantic

By Dayton and Robillard

Information on age reading precision was available. It is not clear if those were available only for samples collected in 2008-2009 or also for other time periods. Information on age reading precision by age and time should be incorporated in the assessment model as they can have notable effects on the estimation of cohort strength and mortality especially when large year classes are adjacent to small ones.

WP3 Silver et al. Growth _ Length-Weight **(TOR2)**

Updating Parameters for Length and Weight Relationships of American Plaice

By Silver et al.,

There are no visible differences in the length weight relationship (LW) between areas and time. However, it would be easier to evaluate those if estimated parameters with uncertainty of the LW were reported by area and time in a tabulated form.

There are large temporal differences in growth for all areas. Even if part of the difference can be explained by the absence of large fish in the population due to higher fishing mortality (F) from 1990s to 2000s, the recent decline in F does not appear to compensate for it and therefore different processes are likely to be at work simultaneously, which determine the temporal trend in growth of American plaice and differences between areas. Even if there is less variability between areas than there is over time, there are notable differences in growth between areas. However, those differences are not always consistent over time, which might indicate either that there are no true geographical differences in growth or that several processes (e.g., spatial differences in exploitation, density dependent and density independent processes) are at work at the same time.

WP4 Goffe et al. Maturity (**TOR4)**

Maturity Analyses of American Plaice in the Georges Bank and Gulf of Maine region

By Goffe et al.,

Maturity used in the assessment was calculated for females only. Maturity used in the assessment was time invariant and averaged between areas (i.e., Georges Bank and Gulf of Maine). There is a weak temporal trend in length at maturity (about 4.5 cm difference between 1980s and 2010s) but not for age at maturity. Also, differences between areas are negligible for age at maturity but not for length at maturity, with plaice in the Gulf of Maine region maturing later than in Georges Bank. When combined, length at maturity of the stock (i.e., combined areas) is very similar to that estimated for Georges Bank. However, it would be interesting to see temporal trends in length at maturity within the same area as it might be an indicator of possible different levels of depletion between areas over time.

WP5 Pavlovich et al. Fishermen’s ecosystem knowledge **(TOR2)**

Fishing Industry Knowledge of American plaice

By Pavlovich et al.

No comments

WP6 Grezlik et al. Fishery_data_exploration **(TOR2)**

Exploration of Fishery Data to Evaluate Catch Rates of American Plaice

By Max Grezlik, Lucy McGinnis, Keith Hankowsky, Gavin Fay, Steve Cadrin

At the beginning of the time series, more pounds of plaice are caught in the spring and summer than during the rest of the year. This seasonal pattern is less apparent in the mid-2000s, and by the end of the time series the pattern has nearly reversed, with higher catch in the winter months and low catch during the summer. This are also linked to seasonal differences in size at age as shown in WP1).

Patterns in gear usage showed that almost all the plaice landed over the analyzed period was caught via otter trawl. Going forward, selecting for just otter trawl gear would retain almost all the catch of plaice and it is deemed to be appropriate.

Species composition and price of American plaice seems to be the most reliable variable to distinguish between targeting and non-targeting plaice trips (i.e., targeting strategy).

WP7 Hankowsky et al. Catch Rate Standardization **(TOR2)**

Catch Rate Standardization of American Plaice Trawl Fishery

By Keith Hankowsky, Max Grezlik, Lucy McGinnis, Gavin Fay, Steve Cadrin

Models with Δ AIC >10 have essentially no support (Burnham and Anderson, 2002) and should be discarded, otherwise the most parsimonious model should be selected. Thus, excluding mesh size as a main effect explained is warranted here as this model has an AIC only 3 points higher than the model without mesh size. GAMs were explored but discarded as the assumptions of homogeneity of variance were not met. The smooths of depth and price from the GAMs showed that the relationship was roughly linear, for the range of most of the data. However, no figures *or table is showed in support of those statements. The standardized LPUE are fairly correlated with survey CPUE. However, lack of true 0 hauls, difficulties to identify targeting and low spatial resolution of the data would increase the risk of hyperstability.*

WP8 Jones et al. Spatiotemporal CPUE **(TOR2)**

Fitting a geostatistical model sdmTMB to standardize the catch rates of American Plaice (Hippoglossoides platessoides) from the Gulf of Maine and Georges Bank.

By Andrew Jones1, Tyler Pavlowich, David Richardson, Anna Mercer

CPUE of plaice catch per tow was converted to a swept-area biomass, which is better approach than nominal CPUE if you have the specific gear dimension and speed of each observation. However, both wingspread of the net and towing speed is not reported for each observation, but those values are assumed to be constant for the entire dataset. It would be important to have indication on how much wingspread and towing speed has varied over time to verify that assuming constant values for those two key variables is not affecting estimates of the swept-area biomass. The authors recognized this as a field of improvement for the future, but I wonder if some limited data do exist to corroborate those assumptions. Also, although differences between the indices are small, comparison with the SSB estimated from the assessment might be misleading if survey indices used in the model were not estimated in the same way and did not account for spatio-temporal changes in the stock distribution (see WP12).

WP9 Terceiro Fishery Indices **(TOR2)**

Fishery Dependent Data Indices of Abundance (LPUE or CPUE) for American Plaice

By Mark Terceiro

A GLM standardization of commercial LPUE and CPUE data using the discrete variables YEAR (the 'year' effect that in a main classification factor only model serves as the index of abundance), calendar quarter (QTR), 3-digit statistical area (AREA), and vessel tonnage class (TC).

WP10 OKeefe et al. Electronic Monitoring

Audit Model Electronic Monitoring Data: American Plaice

By OKeefe et al.,

No comments

WP11 Cadrin survey size at age by season (TOR)

Seasonal Variation in Size-at-Age of American Plaice from Survey Data

By Steve Cadrin

Seasonal differences in length at age are evident, which implies that combining seasonal samples of size at age for annual age-length keys may not be appropriate for estimating age composition, particularly for young ages with large seasonal differences in size at age. This means that a time resolved model (e.g., quarterly model) would be more appropriate here.

WP12 Hansell_Plaice_VAST (**TOR3**)

Spatio-temporal dynamics of American plaice (Hippoglossoides platessoides) in US waters of the northwest Atlantic

By Alexander Hansell, Larry Alade, Andrew Allyn, Lauran Brewster, Steve Cadrin, Lisa Kerr and others

Trawl surveys data were modelled using the spatio-temporal model VAST to produce yearly estimate of relative biomass for the assessment. The analysis showed that the center of gravity of the stock is variable over time with a latitudinal trend in the last two decades. This implies that any standardization of the trawl surveys needs to account for the interaction between space and time in the distribution of American Plaice. An important result is that the water temperature at which the haul is carried out is a significant variable in determining American plaice distribution, abundance, and thus biomass. This implies that corresponding standardization of LPUE and CPUE from fisheries dependent data which do not account for temperature might provide a biased trend in relative abundance and biomass.

Regression model standardization procedures that account for an unbalanced sampling between depth and space (and many other covariates) are widely used for deriving CPUE to be used in stock assessment models (e.g., VAST; https://github.com/James-Thorson-NOAA/VAST). The use of regression model also facilitates the presentation and interpretation of the covariate effect on the CPUE. Similar spatio-temporal standardization procedures are used for example for Northeast Atlantic stocks (e.g., Berg and Kristensen 2013; Berg et al., 2014).

WP14 Behan et al. Ecosystem Profile **(TOR1)**

Ecosystem profile of American plaice

Jamie Behan, Lisa Kerr, Amanda Hart, Alex Hansell, Tyler Paklovitch, Steve Cadrin

No indication on which key environmental parameter might be integrated into the assessment model but temperature and predation as good candidates.

WP15 Cadrin Approximating M (**TOR4**)

Approximation of Natural Mortality Rate for American Plaice in US Waters Based on Life History Traits

By Steve Cadrin

Different estimates of M were derived using life history covariates. Those can be used as alternative M formulations in assessment models. Punt et al. 2021 advocated for estimating M in an integrated assessment model with priors (Punt et al. 2021). Derived values of M were by lifetime or by age and year.

WP16 Behan _ Kerr Ecosystem Drivers (**TOR1**)

Environmental Influences on American Plaice Stock Dynamics

By Jamie Behan and Lisa Kerr

Changes in ocean conditions have been documented to affect key life history processes, including recruitment, distribution, and growth of American plaice. The goal of this work was *to conduct exploratory modeling to examine the relationship between key aspects of American plaice stock dynamics (i.e., recruitment, distribution, and growth) and ocean climate variables together with SSB. Time series of relevant environmental variables included sea surface (SST) and bottom temperature anomalies, Atlantic Multidecadal Oscillation (AMO), North Atlantic Oscillation (NAO), and the Gulf Stream Index (GSI) and were related to time series of stock variables using generalized additive models (GAMs).*

While collinearity and non-significance of exploratory variables was considered, temporal autocorrelation of response variable was not accounted for, which could have included adding a s(year) effect or correlation = corAR1(form= ~ year) in the model. Also, there is no explained rationale for indices of recruitment rate (R/SSB) and SSB to be derived from survey data instead of the assessment model, which integrates all available information, including survey data.

WP17 Hennen _ Hansell Stock Synthesis (**TOR4**)

American Plaice Assessment Model Developed in Stock Synthesis

By Hennen and Hansell

An American plaice assessment model was developed in Stock Synthesis (SS; Methot and Wetzel (2013)), with the objective of providing support for the primary assessment model results. Therefore, if the results from SS using a different structural configuration align with results from the primary assessment model, the WG could feel confident that the primary assessment model is producing results that are reasonably robust to model configuration and platform.

The SS model is correctly specified. Q of the surveys is estimated but a float option could be used to reduce the number of estimable parameters (see Methot et al., 2021). Similarly to WHAM several biological parameters are fixed (e.g. M, maturity) but the current SS model has the advantage to be able to fit length distributions and age length key compositions within the model to allow tracking changes in growth over time and integrating uncertainty in both length and age composition (i.e. ageing precision and accuracy by age and time when available) and translating it to the uncertainty in the derived quantities. Also, it can keep track of changes in management regulations which are typically aimed to size and not age. Also, discards are modelled separately from landings, which is another advantage compared to WHAM.

*Diagnostics show comparable performances to WHAM selected models but MASE of the surveys, age and length compositions should be added to the diagnostic toolbox. Overall performances of the model were similar to WHAM but with a reduced number of parameters and a closer mirroring of biological processes occurring over time as changes in growth rates of the population. Given changes in the biology of the stock over time, unexploited biomass and not virgin biomass should be used (see the concept of "dynamic B_0"; Bessell-*Brown *et al., 2022 and its practical application as for Northern shrimp; ICES 2022). Also, it would be useful to add reference points as in the WHAM models (i.e., $SSB_{40\%}$ and $F_{40\%}$) for a comprehensive comparison between the results of the two different model platforms.*

WP18 Hart et al. WHAM (**TOR4**)

A state-space assessment of American plaice using the Woods Hole Assessment Model (WHAM)

By Amanda Hart, Lisa Kerr and Tim Miller

The document present settings and results from all different model configurations run with WHAM assessment software. This constitutes the core part of the Report and therefore specific comments on this WP are embedded in the detailed comments on the Report of the American Plaice Research Track Working Group which are included below.

WP19 Alade Fishery Data **(TOR 2**)

Estimate catch from all sources including landings and discards. Describe the spatial and temporal distribution of landings, discards, and fishing effort. Characterize the uncertainty in these sources of data

By Larry Alade

Total commercial landings and discards (in tonnes) are reported by main gear type (i.e., trawl, gillnet, dredges, and others) but only number at age for the aggregated pseudo fleet used in the assessment (i.e., the combination of all active fleets over the time series) are reported under TOR2. Neither number at age by fleet nor length at age are reported although length at age for the pseudo fleet is used in the Stock Synthesis model.

The overall precision associated with the allocation process of the landings is expressed in terms of the number of plaice sampled for length, number of otoliths aged, and CV is estimated to be much less than 0.1 over the entire time series (Table 3 of WP 19), which is unusually low. Sample size for the biological samples of discards is reported as number of tows sampled for discards, number of plaice sampled for length, number of otoliths aged from which the CV of the estimated numbers at age of discards is calculated. However, CV based on the number at age do not account for spatial and temporal autocorrelation of samples taken from the same haul or trip. Many aged otoliths from a limited number of trips would artificially produce a low CV as fish within the same haul and trip are highly correlated. A more appropriate estimate of the precision would be the number of trips from which individuals to be aged are taken. Ideally, in a timely (e.g., quarterly) and spatial resolved model, those would be separated by time and area, which would reduce the spatial and temporal autocorrelation effect on the CV. This might also partially explain the unusually low CV estimated by year for the landings.

Total catches and total number at age of the catches are used in the WHAM models but precision associated with that information are not used. Uncertainty associated to estimates of total landings and discards by year, area and/or time strata should be integrated in the assessment models. Similarly, associated uncertainty to number at age of landings and discards should also be integrated in the model, ideally as number of trips by year, area and/or time strata as it would allow important sources of uncertainty to be propagated into key derived quantities as SSB and F.

WP20 Alade Survey Data (**TOR3**)

Present the survey data used in the assessment (e.g., indices of relative or absolute abundance, recruitment, state surveys, age-length data, application of catchability and calibration studies, etc.) and provide a rationale for which data are used. Describe the spatial and temporal distribution of the data. Characterize the uncertainty in these sources of data

By Larry Alade

The document presents the latest estimates of relative abundance and biomass of American plaice as estimated from the three bottom trawl surveys (NEFSC, MADMF and MENH) in spring and fall. Yearly estimates of relative abundance and biomass together with associated CVs were presented for each survey and time of the year. Indices were estimated as aggregated *survey indices. For NEFSC, indices accounted also for the relative efficiency of the survey trawl with rockhopper ground gear to a chain sweep trawl. The efficiency was estimated to be dependent on the size of the plaice, declining with increasing length in chain sweep trawl. Corresponding yearly number at age per surveys and time of the year were also estimated. Inshore surveys were excluded (MADMF and MENH) because of conflicting signals and moving of the plaice deeper with time. Thus, only NEFSC surveys were used in the assessment model as separated survey per season (i.e., spring and autumn).*

WP21 Alade Projections (**TOR4**)

Define appropriate methods for producing projections; provide justification for assumptions of fishery selectivity, weights at age, maturity, and recruitment; and comment on the reliability of resulting projections considering the effects of uncertainty and sensitivity to projection assumptions

By Larry Alade

Methods and assumptions for short term projections are adequate and well described. The short-term forecast has a 3-year span. For short periods, which are typically 3 years or less, assuming that future recruitment will resemble recent recruitment as done for American plaice has been showed to be a reasonable hypothesis (e.g., Ward et al., 2014) but its performance will deteriorate as the interval increases (Van Beveren et al., 2021).

Appendix 5: **1**

Detailed comments on the Report of the American Plaice Research Track Working Group

This part of the individual review report describes the key findings and considerations on the stock assessment of American plaice. It is based on the Report of the American Plaice Research Track Working Group, but it integrates information contained in the different WPs presented. It focuses on TOR4, but links aims, data and methods presented in TOR1-3, 5 and 6 with the objective of providing a holistic illustration of strength and weakness of the data, processes and presentation that lead to proposed assessment and management advice of American plaice.

Information available under TOR2 of the American plaice research track working group but not integrated in the assessment model as described under TOR4

Several key information sources are available and presented but not fully integrated in the WHAM models. These are separated landings, discards and associated length and age compositions by fleet, uncertainty of the landings and discards, length and age compositions of the surveys with their associated uncertainty, precision of the ageing estimates by year and age, time varying natural mortality, historical data and many others. Landings and discards are combined outside the model to derive total catches, catch at age and weight at age of the catches for the pseudo fleet (i.e., the combination of all active fleets over the time series). However, it is not specified in the Report how landings and discards weight were combined to derive values reported in Table 2.13. I guess by a weighted average, but it would be valuable to specify it in the Report. Also, discards age compositions are derived from survey information and not from sampled discards for ageing which might be an issue if selectivity of the survey is very different from that of the commercial fleet.

On the other hand, there are several advantages of integrating landings and discards information separately within the model. First, landings and discards have typically very different levels of precision, with precision of landings being larger (i.e., lower CV) than discards. When dealt separately within the model, the different uncertainties can be assigned, and correctly propagated to key derived quantities as SSB and F, which will impact the probabilistic status of the stock. Also, management questions related to discards can be answered in a more wide-ranging way when discards are treated separately within the assessment model.

Also, age length keys (ALKs) and length compositions are combined outside the assessment model to estimate number at age of the catches and the surveys, which are then used as a direct input to the different WHAM model configurations. Apparently, no spatial considerations were made when combining ALKs and length compositions to derive numbers at age in the catch and in the survey. Although cohort tracking performances are good in catch at age data as shown in the Report, spatial ALKs approaches are known to reduce errors in stratified estimates of abundance at age over non-spatial approaches (Babyin et al., 2021). As an alternative, ALKs and length compositions can be integrated within the model as done in the Stock Synthesis configuration. The integration of ALKs and length compositions within the model would have the advantage to allow tracking separately changes in growth and condition over time and integrating uncertainty of length and age compositions (as CV, sample size and ageing precision and accuracy by age and time when available), which is then propagated into the uncertainty of the derived quantities.

Also, as information on historical catches of American plaice prior to 1960s seems to exists (https://maineanencyclopedia.com/american-plaice-landings/), it is recommended that historical catches as far back in time as possible are collated and used in the future as an alternative model configuration.

Therefore, I recommend development of an alternative configuration that treats landings, discards, landings and discards length compositions and age length keys by fleet, and their associated uncertainty separately in the assessment model. Also, in the long run, an effort should be made to collate historical catches of American plaice prior to 1960s.

Inclusion of fishery catch rates in the WHAM assessment model (TOR2)

Different time series of standardized commercial LPUE and CPUE were provided under WP 6-9). Differences between the CPUE and LPUE indices were small and comparison with the trawls surveys indices of biomass show a rather good agreement between commercial and survey-based indices of relative abundance. Alternative model WHAM configurations were built to explore the integration of standardized commercial LPUE and CPUE into the stock assessment model. However, given the moderate to strong correlation with survey CPUEs, there might be little need to increase the complexity of the assessment model adding several fisheries dependent time series of relative abundance, which has the risk of creating data conflicts which are mostly caused by noisy indices in a single area model. The only advantage of using commercial CPUE would be that they contain a larger proportion of large and old individuals compared to surveys. However, given the very strong cohort signals (presented by Steve Cadrin on Tuesday 19/07 under my request), including older ages, from the surveys, the exclusion of commercial CPUEs is justified.

Thus, I recommend that based on first principles as standardized trawl surveys are generally more noisy but less biased than standardized commercial LPUE and CPUE, base case scenario should exclude standardized commercial LPUE and CPUE.

Use of survey data in the assessment models (TOR3)

Design based survey indices and combined spatio-temporal integrated survey index based on VAST software were estimated for American plaice. As it would be expected, there are similarities and differences between Design and VAST based survey indices, although the two were rather similar in the general trend. Also, very strong and comparable cohort signals (as presented by Steve Cadrin on Tuesday 19/07 under my request), including older ages, are evident for both design and VAST based age compositions. As for commercial LPUE and CPUE and based on first principles, the use of fully standardized VAST indices is generally considered to achieve less retrospective bias and outperform assessments with design-based indices (Cao et al. 2017). Also, combined standardized VAST indices are theoretically more suited for a single area model such as the American plaice WHAM model. This is because the use of a single spatio-temporal standardized index avoids problematic conflicts which can arise when several design-based survey indices of the same indicator are used but do not exactly cover the same time and space. The reason why a combined standardized VAST index was not used in the final model is that its performance in terms of diagnostics was considered inferior when compared to separated Albatross and Bigelow survey indices. However, I cannot see significant differences between WHAM model 28B and 29F compared to models using VAST indices. As for most of the comparisons, those are based mainly on qualitative analysis of the residuals, which by nature are hard to follow, and lower (but still under the threshold) Mohn´s rho. On the other hand, MASE was not calculated for VAST models and was not shown for most of the alternative configurations presented in the Report.

Several of the tested model configurations were using inshore surveys as independent indices of the stock. However, the spatial distribution of those surveys is limited and much smaller than offshore survey. In a single area model, the assumption is that each single index represents the trend and age structure of the entire stock. This assumption is not fulfilled for inshore surveys, which questions their use in the assessment, irrespective of their statistical performance. Introducing such surveys only increases the risk of apparent conflicts between data sources. This is another example of the importance of developing base scenarios based on first principles when doing model development.

Thus, I recommend that the combined spatio-temporal integrated survey index based on VAST software should be used in the base case scenario.

Stock assessment models

General comments

Although the WHAM model was used as the primary assessment model, several different stock assessments platforms were run, i.e., WHAM, ASAP, VPA and Stock Synthesis, to describe the dynamic of American plaice. The data used were the same for ASAP and VPA when compared to WHAM but different from Stock Synthesis, which used extended time series of catches, abundance indices and length compositions. Also, SS differs from the presented WHAM model configuration as it integrates age length keys directly within the model and models selectivity as a length-based process. The objective of running different model platforms was to confirm results from WHAM, i.e., WHAM, the primary assessment model is producing results that are robust to different model platform and configurations. The rationale is rather circular as two or more coincident results obtained with very similar data by different models does not assure that the primary model results are robust. For example, there are several aspects of the Stock Synthesis model that I find more appropriate by first principles as landings, discards, landings and discards length compositions and age length keys, and its associated uncertainty be treated as separated components in the model, the use of a SR relationship, the possibility of estimating M within the model, defining spatial units with specific biology and many others. Moreover, it is unclear what would happen if the SS model would achieve better performances than the best WHAM model. Clearly, overall performances of the model were similar to best case WHAM models but with a reduced number of parameters and a closer mirroring of biological processes occurring over time as changes in growth rates of the population. Anyhow, if the scope of using SS is only to verify WHAM results, **it would be important to add reference points to SS as in the WHAM models (i.e., SSB$_{40\%}$ and F$_{40\%}$) for a comprehensive comparison between the results of the two different model platforms.**

WHAM model selection process

The first step of the model selection process was to import previously used VPA model data into WHAM and ASAP as the first step to bridge WHAM with the former assessment model configuration. Following this stage, I was expecting that a base case scenario would be established followed by a factorial building of alternative model configurations, which should have been based on first principles, preferred model structure, analyzed processes affecting stock dynamic and available data. However, already at this stage the process used to select between WHAM model configuration candidates and pruning (i.e., discarding certain model configurations along the path of model development) following bridging with VPA (Run 9) became ambiguous and it was difficult to follow. The main reasons are that criteria used to navigate through the process of selecting and discarding different model configurations are not well defined a priori, are often based on a semi-qualitative evaluation of the diagnostics (see detailed description and recommendations in section **Criteria for model selection and pruning** below) and are used differently for different alternative models. Also, the different model configurations tested during the model development process are not clearly associated to a hypothesis testing, which makes difficult for the reader to judge why certain models are presented and others equally plausible are not. Associating each alternative model to a clear biological, ecological or fisheries related hypothesis to be tested would greatly improve the readability of the entire process.

Criteria for model selection and pruning

As it is now, it is hard to navigate through the performances of the different model configurations simply looking at table 4.1 of WP 18 and at the Report text. It is also very difficult to understand the logic that leads from one model to the other. Ideally, alternative model configurations should be presented in the context of hypothesis testing, clearly indicating which alternative aspect of the biology, ecology or fisheries is being tested. Qualitative evaluations are repeatedly used (e.g., section 4.6. "Time-varying selectivity was also explored by implementing independent and identically distributed (iid) random effects for the fleet and both indices (run 23), **resulting in a better fit to the data and smaller age composition residuals**) or section 4.7 ("Run 27 had **improved fits to age composition data (smaller OSA residuals)**, but older fish **tended to have more negative residuals** than younger fish"), with the part in bold very hard to evaluate from the report. All sections are pervaded by qualitative evaluations, which are hard to follow even when scrutinizing the numerous plots for each model reported in https://github.com/ahart1/PlaiceWG2021/tree/main/WG_Revised_Runs.

Another example of this difficulty is the evaluation of the extended catch time series model configuration (Run 28 in section WHAM runs with Extended catch time series of the Report) which is abandoned although having apparently similar performances when evaluating Mohn´s rho and residuals compared to the selected runs (29F2, F4 and F5). The only (putative) rationale seems to be the difference in $SSB_{40\%}$ to which no explanation is given in WD 18 although it is not repeated in the Report. Historical data generally are informative of the scale of the stock at low level of F and thus have an important effect on the biomass reference points (ICES 2021) and long time series of catches are informative of scale (Hordyk et al., 2019). Thus, excluding the extended time series based on difference in biomass reference points is illogical and should not be pursued.

Another instance of model discarded but rationale of discarding not explained is for models with time-varying selectivity (Run 23) to models without (Run 25 and successive). When reading the report and looking at Table 4.1 in WP 18 and https://github.com/ahart1/PlaiceWG2021/tree/main/WG_Revised_Runs I cannot find a clear explanation why time-varying selectivity is abandoned during the process of model selection. A time varying selectivity, which can be approximated by a random walk, is to be expected when catch at age data are not reflecting true fleets as here but the alleged fleet is instead a combination of different gears and fleets. The simple variation in proportion of the catches between fleets over time will cause a perceived (artificial) change in selectivity in a one single fleet model. Most of the catches have been taken by bottom trawls since 1964, but the proportion of catches taken by vessels belonging to the different ton classes has greatly changed over time. Vessels differing largely in size, even if using similar gear, generally show rather different fishing pattern and are de facto different fleets. Also, figures 2,2 and 2.6 of the Report clearly show that different trawls typologies have been operating historically and that the proportion of the catches between those has changed over time. Therefore, based again on first principles and model structure, base case scenarios should have time varying selectivity for the aggregated fleet while time unvarying selectivity might be more appropriate if catches and associated catch at age are split between the different fleets. The situation is different for the surveys, for which if gear, timing, and haul position is constant, selectivity should be time unvarying in base case scenarios unless selectivity in WHAM is the combination of gear selectivity and fish availability and the stock distribution changes substantially over time. Run 23 explored time varying selectivity, but once again why this configuration was abandoned is not reported and neither it is possible to understand the rationale behind its exclusion analyzing plots under https://github.com/ahart1/PlaiceWG2021/tree/main/WG_Revised_Runs.

Also, sections 4.9.1 and 4.9.2 of WP 18 and relative sections in the Report are hard to navigate with numerous slightly different configurations which perform rather similarly in terms of diagnostics. Another key difficulty in evaluating the different model configurations as described in the report text is that sometimes the choice is based on one diagnostic metric (e.g., MASE in section 4.9.3 of WP 18) but then mainly on residual pattern and Mohn's rho in the successive 4.9.4 section of WP 18. Also in the report, the section describing WHAM runs with alternative stock indices or alternative age compositions models are a clear example of mixed criteria for model selection and model navigation. Although model selection is primarily AIC centered, different models are selected or discarded utilizing different criteria so that sometimes AIC is taking the precedence but in other cases Mohn's rho or a qualitative analysis of the residuals are the primary criteria used.

I also closely scrutinized diagnostics plots of numerous alternative runs under https://github.com/ahart1/PlaiceWG2021/tree/main/WG_Revised_Runs/ but still I cannot reconcile the choice made to move between runs with the presented diagnostics criteria. Numerous models have very similar performance when looking at residuals and Mohn's rho but anyhow are discarded along the way when moving from one model configuration to another based mainly on qualitative and often visual evaluations of the residuals, results in terms of derived quantities and other criteria. Also, the selection of the three candidate models is unclear. There are so many models that show similar performance as the three candidate models but the rationale behind their selection is not well explained.

The only way to avoid cherry picking in model selection is to establish a priori criteria of model performance, which the authors have done when using Mohn's rho, residuals (albeit in a qualitative manner) and MASE. However, the reviewer considers that the process of model selection could be improved. Diagnostics should be used in combination (and not in isolation) to compare and select models, including navigating between different model configuration and their pruning. The development of the alternative model configurations could benefit to a factorial structure instead of a linear one, where alternative hypotheses are represented as ramification or evolution of the original or base configuration. **. For reporting purposes and presentation of the different model configuration results and diagnostics, I recommend the development of a shiny app as for example https://maxcardinale.shinyapps.io/Ensemble_2022/.**

Diagnostics of assessment model configurations

The analysis integrates several of the recent developments in model diagnostics (Carvalho et al., 2021, Kell et al., 2021). The process used to select between model configuration candidates and pruning (i.e., discarding certain model configurations along the path of model development) is centered on AIC and retrospective analysis. However, AIC-based selection is not suitable for comparing models with different datasets and weighting, and the absence of retrospective bias alone does not ensure that a model is valid. Also, residual patterns can be removed by adding more parameters than justified by the data (i.e., overfitting), and retrospective patterns by ignoring the data (Kell et al., 2021a). An alternative is to use prediction skill and MASE to compare predictions to observation not used when fitting (i.e., out of sample), and to reject models that do not fit the data and so are inconsistent with observations. Generating predictions to be used in management advice is the key objective of any assessment model and thus a model that is unable to predict (i.e., has low prediction skill) has limited use in fisheries management (Kell et al., 2021). Prediction skill also can be used to explore model misspecification and data conflicts, and help to identify alternative hypotheses, and weight ensemble models. MASE has the desirable property of scale invariance, so it can be used to compare forecasts across data sets with different scales, predictable behavior, symmetry, interpretability, and asymptotic normality. It is also possible to calculate MASE using other data, e.g., length or age compositions (see Carvalho et al., 2021) and combine several MASE is a single statistics (see ss3diags; devtools::install_github("jabbamodel/ss3diags")).

Unfortunately, MASE is only presented for a few runs while it should be calculated for all runs and used during the model development and pruning process. Residuals are calculated but their analysis is based on visual inspection of the plots. It is difficult and highly subjective to distinguish between "good" and "bad" residuals simply from the visual inspection of the plots. Several tests of the residuals (e.g., runs test, Carvalho et al., 2021) are available and could be used.

Thus, I recommend adding single and combined MASE (Merino et al., 2022) for each model configuration tested along with a quantitative test of the residuals. Runs test, Mohn's rho and MASE are comparable across all models and should be preferred to AIC. Also, as suggested by Carvalho et al., 2021 and Kell et al., 2021, single diagnostics should not be used in isolation, but several diagnostics should be combined (for example in a pass/fail scheme to facilitate the evaluation of the different model configurations (see Masnadi et al., 2021, ICES 2022) **to select between models since different diagnostics are able to identify different type of model issues. Importantly, I recommend that estimates of reference points or derived quantities as SSB and F are not used to compare and select between different model configurations as for example done in figure 4.4 of WD 18.**

Structural uncertainty and the role of alternative model configurations and platforms

Another major concerns I have is about the use of the sensitivity analysis and how structural uncertainty is treated in the modelling context and to provide advice which is based on a best-case scenario. For American plaice a large effort is made by the assessment Team to develop and evaluate different WHAM model configurations, which are then presented only as sensitivity analysis. While parameter uncertainty and observation error are partially integrated (but see section of **Additional information available but not used in the model** presented above), a key part of the uncertainty (i.e., structural uncertainty) is not included but only presented as sensitivity analysis and thus has no effects on the stock status and the management advice. Unfortunately, many of the different model configurations have similar performance in terms of model diagnostics as the "best case". Also, it is not specified what happens in the situation that two or several model configurations (or even model platforms) achieve the same performances as here but show different results in terms of derived quantities such as SSB and F. When looking at model diagnostics of several of the WHAM tested models in Table 4.1 of WP 18 it is evident that numerous models have very similar performance as those selected for providing advice. The lack of a priori established criteria makes it difficult to justify the selection of a certain model configuration over others as presented in the Report. Alternative model platforms (i.e., SS, VPA and ASAP) are developed with the objective of providing support for the primary WHAM assessment model results. However, if SS achieves rather similar results as WHAM, it does not imply that WHAM is the most appropriate model for advice. There are several aspects of the Stock Synthesis model that I find more appropriate by first principles and are supported by simulations as for example the possibility of integrating length compositions and age length keys and estimating M within the model, fitting selectivity by length, defining spatial units with specific biology and many others not listed here. Also, results in terms of depletion are not the same between the updated SS model (BASE14fixFleetCode) and the proposed final WHAM run. Depletion rate is less than the reference point ($SSB_{40\%}$) for SS model compared to WHAM. Finally, as the stock status is described in a probabilistic manner, integrating additional sources of uncertainty would have significant effects on the probabilities estimates in the Kobe plot.

Thus, I recommend that an ensemble of different plausible configurations and model platforms selected and weighed by a comprehensive diagnostic against performance criteria agreed beforehand is developed in the future to provide stocks status and management advice for American plaice. As best practice, and as a minimum, the ensemble should integrate the three main sources of uncertainty, process uncertainty, parameter uncertainty and observation error in the data (sensu Punt et al. (2016). The ensemble should also be used for deriving catch forecast scenario, in which plausible assumptions on the productivity of the stock (e.g., recruitment, growth, mortality, and others) can be integrated to mimic variability of the ecosystem and possible effects of climate factors.

Additional diagnostic to be added to the toolbox of American plaice

Apart from quantitative diagnostics such as MASE or Mohn's rho, which are used in TOR4 of the report and in WP 18 to select between models, quantitative analysis of the residuals should be added to the diagnostic toolbox of American plaice. Analysis of the residuals is based primarily on the visual inspections of the residual's diagnostic plots (e.g. https://github.com/ahart1/PlaiceWG2021/blob/main/WG_Revised_Runs/WHAM_Run10_RevisedData-Maturity/plots_png/diagnostics/OSAresid_catch_4panel_index2.png).

Considering the large number of models and associated diagnostic residuals, I consider that quantitative analysis based on probability as common currency (as for example runs test and similar, see Carvalho et al., 2021) would greatly facilitate the comparison between the performance of different model configurations.

Model selection can also be done based on first principles. For example, the use of fully standardized and combined VAST indices is theoretically more suited for a single area model such as the American plaice WHAM model as it avoids problematic conflicts which can arise when several indices of the same indicator are used. In 4.11, selection and pruning for model configurations including environmental covariates is indeed based on first principles (i.e., the expected relationship between recruitment and temperature, although it would be more appropriate to show recruitment deviations from an assumed stock and recruitment (SR) curve). **Objective criteria such as MASE or Mohn's rho can be augmented by first principles during model selection processes. However, I consider first principles particularly useful to build base case scenarios from which construction of alternative model configurations could be derived.**

Further, MCMC and ASPM could be added to the diagnostic toolbox. MCMC is a powerful diagnostic for detecting model misspecifications in the broader sense and to regularize the model, i.e., to check that all parameters are identifiable (Monnahan et al., 2019). The age-structure production model diagnostic (ASPM, Maunder and Piner, 2015) was proposed to evaluate whether the trend in abundance, as represented by the indices, may be explained exclusively by the fishery removals. The ASPM is a powerful diagnostic that can evaluate data conflicts in information related to absolute abundance and abundance trends and detect misspecification in the population dynamics (e.g., steepness or natural mortality; Carvalho et al., 2021). **Thus, I recommend that MCMC and ASPM should be added to the diagnostic toolbox of the American plaice assessment.**

Integration of environmental aspects into stock assessment of American plaice (TOR1 of the American plaice research track working group)

Assumed biological parameters

The characterization of the climate-driven marine ecosystem changes, as well as modeling of its effect on exploited commercial fish stocks, has become the central research topic within the coupled climate-fisheries discipline (Link 2010; Tanaka 2019). In this context, the assessment Team of American plaice has made a good effort to identify environmental covariates that might be linked to several aspects of stock dynamics as requested by TOR1 of the American Plaice Research Track and reported in the Working Group Report and related WPs. The analyses conducted under TOR1 of the American Plaice Research Track showed highly dynamic environmental conditions partially linked to global climate changes. The assessment Team of American plaice attempted to integrate identified potential environmental drivers as for example sea surface temperature (SST) anomalies in the stock assessment model WHAM (i.e., Run 39-50) but those model configurations were discarded primarily based on first principles (i.e., expected relationship between recruitment and temperature conflicted with analysis under TOR1). These frequent documentations of "breaking relationships" indicate that integration of environment-recruitment relationships in stock assessment requires a clear mechanistic understanding as well as modeled relationships cross-validated with new data (Tanaka 2019). Environmental influences on population dynamics are embedded in temporal and spatial variation of key productivity parameters such as growth, condition, maturity and natural mortality. Decline in weight at age is clearly visible in Figure 2.17 of the Report, which might mirror decline in growth and/or condition over time as described in WP 3 and can be partially linked to changing environment and warming (Levangie et al., 2021). Observed accelerated growth and smaller maximum size is expected to affect natural mortality and maturity of American plaice (Levangie et al., 2021, Zheng et al., 2020). Thus, a convenient way to integrate environmental influences in stock assessment is through realized time and space variability of population key productivity parameters. The most obvious candidates in the current age-based WHAM model would be using natural mortality at age and time as estimated in WP15 and maturity at age and time as estimated in WP14.

A time unvarying maturity ogive was used for all different configurations tested, which is warranted given results by Goffe et al. (WP4). However, assumptions of natural mortality (M) are not in line with conclusions of Cadrin et al., (WP15) and Behan and Kerry (WP16), and inconsistent with objectives of TOR1. M is treated as time and age invariant in the WHAM assessment models. However, it is well established that natural mortality declines with size and age in fish (Lorenzen 2000). Also, increasing water temperatures, as observed in the distribution area of American Plaice, is predicted to reduce body size, increase natural mortality and earlier age at maturity in many marine fish species (Levangie et al., 2021; Ahti et al., 2020) and it is thought to play a role in driving growth, maturation, and thus natural mortality in American plaice (Zheng et al., 2020). Thus, given the observed increase in water temperature occurring in the area coupled with the large decline in growth (i.e., size at age) of older fish, it is conceivable that M has increased over time.

Thus, I recommend that a model configuration with a time and age varying M should be tested. Also, Punt et al., (2019) recommended that M should be estimated within integrated models. Assuming that WHAM is chosen to provide advice, M could be estimated within the Stock Synthesis model and used in WHAM as an alternative model configuration as also recommended by Cadrin et al., (WP15).

Spatial differences in key productivity parameters are described in WP 1, 3 and 4. Albeit considered less important than temporal differences, future development should also aim to verify the effect of spatial differences within the stock area in the attempt to integrate the other key aspect of the ecosystem, the space. Spatially stratified integrated population models provide a more realistic representation of true population dynamics (Punt 2019) and even if incorrectly specified, they can adequately support spatial management decisions (Bosley et al., 2021). Finally, as plaice display sexually dimorphic growth, it is reasonable to assume that sex ratio changes with depletion rate of the population linked to periods of high and low F, which will in turn impact aggregated M in a single sex model.

Thus, I recommend that in the future to develop a spatial and sex separated model for comparison with the current single area, single sex model.

Stock and recruitment in WHAM and effect on current estimation of reference points

A stock and recruitment function is not used in the WHAM model and R_0 is derived as the mean of observed recruitments. This implies that recruitment has a large probability of being large at very low level of SSB (which is evident looking at model results, see for example https://github.com/ahart1/PlaiceWG2021/blob/main/WG_Revised_Runs/WHAM_Run29A_splitNEFSC-BigUnits/plots_png/results/SSB_Rec.png). This is equivalent to fixing steepness (s) at 1, where s is defined as the ratio of recruitment to the virgin recruitment R_0 when SSB equals 20% of the unfished SSB_0. However, there is no evidence of a clear SR relationship when analyzing SR pairs estimated by WHAM. This justifies the use of fraction of SPR_0 (i.e.,

$SPR_{40\%}$) as a reference point for American plaice. Also, time varying key productivity parameters is accounted for using average current conditions when estimating B_0 fraction reference points in American plaice (but not possible for time varying steepness, see Miller et al., (2020).

Given the current stock status, the SR pairs estimated by the WHAM model and the use of SPR_0 based reference points, the absence of an SR is not relevant for the determination of American plaice target reference points at the current stock status (but it would be for limit reference points which are not defined for American plaice). Also, I recognize that ignoring the existence of a functional form of the SR curve used in conjunction with average recruitment in the projections and SPR as reference points has limited impact on the short-term forecast advice. In terms of the SPR target levels and how F_{SPR} relates to F_{MSY}, for SPR fraction = 0.4 F_{SPR} exceeds F_{MSY} at steepness levels below 0.65. Thus, given the assumed best estimate of steepness being less than 0.65, there are some risks associated with an $F_{SPR40\%}$ (Preece et al., 2012). When looking at potential depletion level of SSB as a proxy for a limit reference point (e.g., 20% B_0), at $F_{SPR40\%}$ depletion will be less than the limit reference point only when steepness is less than 0.5 (Preece et al., 2012).

General consideration on the use of steepness

Although assuming no SR and average R (i.e., steepness equal to 1) has no direct implications for management of American plaice at current stock status, it is important to note that setting h = 1 is biologically unrealistic (Martell et al. 2008; Mangel et al. 2010), obviously risk-prone, and entails precautionary conservation of an adequate minimum biomass if used as an approximation in an assessment, although it is unclear how to determine the precautionary minimum (Mangel et al. 2013).

The SR curve of a fish stock has invariably a functional form and could theoretically be estimated with the model if there is strong contrast in the spawning stock time series (Kolody et al., 2019). Simulations have demonstrated that steepness estimates are often imprecise and biased, often converging to the upper bound (i.e., close to 0.999), even when the true h is considerably lower (e.g., Magnusson and Hilborn, 2007; Lee et al., 2012). Thus, when data are not informative, steepness could be fixed based on best available knowledge (Mangel et al., 2013) (see the R package SPMpriors, which uses FishLife (https://github.com/James-Thorson-

NOAA/FishLife)), which enables straight-forward integration of available prior information on the steepness *s* of the SSR from a recent meta-analysis (Thorson 2020).

Although available data are often uninformative for steepness (Lee et al., 2012), an SR function exists, especially at low level of SSB, and therefore ignoring the SR (e.g., using a statistical catch at age approach without an SR) might still have some consequences on the fit and the predictive capability of the model. Ignoring the SR has its largest impact when modelling long term dynamics as for example when conducting an MSE (i.e., Management Strategy Evaluation). Assuming average recruitment at all levels of SSB runs the risk of overestimating recovery potential when the stock is low, which has important consequences for rebuilding plans, as shown for rockfishes, where the estimated rates of rebuilding are strongly influenced by the assumed value of steepness (Thorson et al., 2019).

It is also easy to demonstrate that due to the flat yield curve when the true h=1 as in the current WHAM models (which is anyhow biologically rarely the case), under-specifying h (i.e., assuming a lower h when the true h is indeed larger, e.g., assuming h=0.75 when the true h is close to 1) results in less lost catch than over-specifying h. There is invariably a much higher risk to long-term yields and stock biomass when specifying positively biased recruitment compensation than the equivalent negative bias (Hordyk et al., 2019). Also, the consequences of setting M to an incorrect value were reduced if stock-recruitment steepness was estimated within the model (Punt et al., 2021). **Thus, I recommend introducing an SR as an alternative WHAM model configuration for comparison. If steepness should be used but it cannot be reliably estimated within the model, it should be set based on the integration of available prior information and best available knowledge.**

Additional WHAM configurations to be tested during the plaice research Track review panel or in future American plaice assessment

1. Time and age varying M as estimated by Cadrin et al., (WP15).
2. M estimated within the Stock Synthesis model and used in WHAM as an alternative model configuration as also recommended by Cadrin et al., (WP15).
3. Landings, discards, landings and discards length compositions and age length keys, and its associated uncertainty should be treated as separated components in the assessment model.

4. Introducing a SR relationship and steepness in the WHAM model for comparison. Steepness can be estimated within the model or assumed based on best available knowledge.
5. For reporting purposes and presentation of the different and numerous model configuration results and diagnostics, I recommend the development of a shiny app as for example https://maxcardinale.shinyapps.io/Ensemble_2022/.

Long term additional configuration (to be planned for future benchmark of American plaice)

1. Develop a spatial and sex separated model for comparison with the current single area, single sex model.
2. Develop an ensemble model of different plausible configurations selected and weighed by a comprehensive diagnostic against performance criteria agreed beforehand.

Summary of the general recommendations:

1. The process of model selection could be improved. Diagnostics should be used in combination (and not in isolation) to compare and select models, including navigating between different model configuration and their pruning. The development of the alternative model configurations could benefit to a factorial structure instead of a linear one, where alternative hypotheses are represented as ramification or evolution of the original or base configuration.
2. Criteria for model selection and pruning should not be based on derived quantities as SSB or reference points but should be centered on diagnostics that allow comparison between models with different weight of the model components and different data sources, which is the norm in stock assessment. Thus, AIC is not recommended, while Mohn's rho, quantitative analysis of residuals and MASE should be preferred.
3. Objective criteria as above can be augmented by first principles. First principles are particularly useful to build base case scenarios from which model exploration could be derived.

4. The diagnostic toolbox should be augmented to include also quantitative analysis of residuals as for example runs test, MCMC and ASPM.
5. Given changes in the biology of the stock over time, unexploited biomass and not virgin biomass should be used (see the concept of "dynamic B_0"; Bessell-Brown et al., 2022 and its practical application as for Northern shrimp; ICES 2022) for Stock Synthesis.
6. There are several aspects of the Stock Synthesis model that I find more appropriate by first principles and are supported by simulations as for example the possibility of integrating length compositions and age length keys within the model, fitting selectivity by length, defining spatial units with specific biology, estimating M within the model, numerous time varying options, and many others not listed here. Thus, I recommend that an ensemble of different plausible configurations and model platforms selected and weighed by a comprehensive diagnostic against performance criteria agreed beforehand is developed in the future to provide stocks status and management advice for American plaice. This is particularly important given the uncertainties in the data used as input, and in key biological parameters and processes in the context of providing probabilistic statement of stock status.

Additional comments

Add reference points to SS which are comparable to those used in the WHAM models (i.e., based on SPR, $SSB_{40\%}$ and $F_{40\%}$) for a comprehensive comparison between the results of the two different model platforms. Ideally, a probabilistic Kobe plot should be estimated also for the best SS model.

Insert the name of the index (i.e., specific survey name) in diagnostic plots (e.g. https://github.com/ahart1/PlaiceWG2021/blob/main/WG_Revised_Runs/WHAM_Run29F-2_swapInitSel-randAlbFall/plots_png/diagnostics/Catch_age_comp_index_2_b.png).

Specify the how probability values showed in Figure 5.8 of the Report are estimated and what the contour means. Also, the shape of the probability surface is unusually flat and does not have the classic banana shape of a variance-covariance matrix.

Further investigate the reason of the shape of the Albatross fall selectivity as estimated by the final model (see Figure 4.8 of the Report).

Add individual MASE plot for each survey with annual deviations.

References

Ahti, P.A., Kuparinen, A., and Heikkilä, S.U. 2020. Size does matter — the eco-evolutionary effects of changing body size in fish. Environmental reviews, 28 (3), https://doi.org/10.1139/er-2019-0076.

Babyn, J., Varkey, D., Regular, P., Ings D., Flemming, J.M., 2021. A gaussian field approach to generating spatial age length keys. Fisheries Research, 240, 105956, https://doi.org/10.1016/j.fishres.2021.105956.

Berg, C., Kristensen, K., 2013. Spatial age-length key modelling using continuation ratio logits, 2012. Fisheries Research 129, 119–126. https://doi.org/10.1016/j.fishres.2012.06.016.

Berg, C., Nielsen, A., Kristensen, K., 2014. Evaluation of alternative age-based methods for estimating relative abundance from survey data in relation to assessment models. Fisheries Research, Volume 151, March 2014, Pages 91-99, https://doi.org/10.1016/j.fishres.2013.10.005.

Bessell-Brown, P., Punta, A. E., Tucka, G. N., Day J., Klaer, N., & Penney, A., 2022. The effects of implementing a ‘dynamic B0’ harvest control rule in Australia’s Southern and Eastern Scalefish and Shark Fishery. Fisheries Research 252. https://doi.org/10.1016/j.fishres.2022.106306.

Bosley, K.M., Schueller, A.M., Goethel, D.R., Hanselman, D.H., Fenske, K.H., Berger, A.M., Deroba, J.J., Langseth, B.J., 2021. Finding the perfect mismatch: Evaluating misspecification of population structure within spatially explicit integrated population models. Fish and fisheries, doi: 10.1111/faf.12616.

Burnham, K.P. and Anderson, D.R. 2002. Model Selection and Inference: A Practical Information-Theoretic Approach. 2nd Edition, Springer-Verlag, New York. http://dx.doi.org/10.1007/b97636.

Cao J, Thorson JT, Richards RA, Chen Y. 2017. Spatiotemporal index standardization improves the stock assessment of northern shrimp in the Gulf of Maine. Canadian Journal of Fisheries and Aquatic Sciences 74: 1781-1793.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M. et al. 2021. A cookbook for using model diagnostics in integrated stock assessments. Fisheries Research, 240: 105959.

Merino, G., Urtizberea, A., Fu, D., Winker, H., Cardinale, M., Lauretta, M.V., Murua, H., Kitakado, T., Arrizabalaga, H., Scott, R., Pilling, G., Minte-Vera, C., Xu, H., Laborda, A., Erauskin-Extraminiana, M., Santiago, J., 2022. Investigating trends in process error as a diagnostic for integrated fisheries' stock assessments. Fisheries Research, conditionally accepted.

Hordyk, A.R., Huynh, Q.C., Carruthers, T.R., 2019. Misspecification in stock assessments: Common uncertainties. and asymmetric risks. Fish and Fisheries. doi: 10.1111/faf.12382.

ICES. 2021. Benchmark Workshop on the development of MSY advice for category 3 stocks using Sur-plus Production Model in Continuous Time; SPiCT (WKMSYSPiCT). ICES Scientific Reports. 3: 20. 317 pp. https://doi.org/10.17895/ices.pub.7919. Editors Manuela Azevedo and Massimiliano Cardinale.

ICES, 2022. Benchmark workshop on Pandalus stocks (WKPRAWN). *ICES Scientific Reports*. 4:20, 249 pp. http://doi.org/10.17895/ices.pub.19714204.

Levangie P.E.L., Blanchfield P.J., Hutchings J.A. 2021. The influence of ocean warming on the natural mortality of marine fishes. Environmental Biology of Fishes https://doi.org/10.1007/s10641-021-01161-0.

Lorenzen K. 2000. Allometry of natural mortality as a basis for assessing optimal release size in fish stocking programmes. Can. J. Fish. Aquat. Sci. 57: 2374-2381.

Mangel, M., Brodziak, J., and DiNardo, G. 2010. Reproductive ecology and scientific inference of steepness: a fundamental metric of population dynamics and strategic fisheries management. Fish Fish. 11: 89–104. doi:10.1111/j.1467- 2979.2009.00345.x

Mangel, M., MacCall, A.D., Brodziak, J., Dick, E.J., Forrest, R. E. Pourzand, R., and Ralston, S., 2013. A perspective on steepness, reference points, and stock assessment Can. J. Fish. Aquat. Sci. 70: 930–940 (2013) dx.doi.org/10.1139/cjfas-2012-0372

Martell, S.J.D., Pine, W.E., and Walters, C.J. 2008. Parameterizing age-structured models from a fisheries management perspective. Can. J. Fish. Aquat. Sci. 65(8): 1586–1600. doi:10.1139/F08-055.

Masnadi, F., Carbonara, P., Cardinale, M., Scarcella, G., Milone, N., Arberi, E., Dragičević, B., Scanu, M., Ceriola, L., Hernandez, P., Sharma, R., & Falsone, F., 2021. Stock Assessment Form Demersal species - Stock assessment of common sole in GSA 17. https://doi: 10.13140/RG.2.2.32101.73441.

Maunder, M.N., Piner, K.R., 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. 72, 7–18. https://doi.org/10.1093/icesjms/fsu015.

Methot R.D., Wetzel, C.R., Taylor, I.G., and Doering, K. 2022. Stock Synthesis User Manual Version 3.30.19. NOAA Fisheries, Seattle Washington. https://nmfs-stock-synthesis.github.io/doc/SS330_User_Manual.html.

Miller, T. J., and Brooks, E. N. 2021. Steepness is a slippery slope. Fish and Fisheries, 22: 634–645.

Monnahan, C.C., Branch, T. A., Thorson, J. T., Stewart, I. J., Szuwalski, C. S., 2019. Overcoming long Bayesian run times in integrated fisheries stock assessments. ICES Journal of Marine Science, fsz059, https://doi.org/10.1093/icesjms/fsz059.

NEFSC (Northeast Fisheries Science Center). 1999. Report of the 28th Northeast Regional Stock Assessment Workshop (28th SAW): Public Review Workshop. NEFSC Ref. Doc. 99-07.

NEFSC (Northeast Fisheries Science Center). 2001. Report of the 32nd Northeast Regional Stock Assessment Workshop (32nd SAW): Public Review Workshop. NEFSC Ref. Doc. 01-04.

Okamoto, D.K., Hessing-Lewis, M., Samhouri , J.F., Shelton A.O., 2020. Spatial variation in exploited metapopulations obscures risk of collapse. Ecological Applications, 30(3), 2020, e02051.

Perälä T, Kuparinen A. 2017. Detection of Allee effects in marine fishes: analytical biases generated by data availability and model selection. doi/full/10.1098/rspb.2017.1284.

Perälä T., Hutchings J.A., Kuparinen A. 2022. Allee effects and theAllee-effect zone in northwest Atlantic cod. Biol. Lett.18: 20210439.https://doi.org/10.1098/rsbl.2021.0439.

Preece, A., Hillary, R., Davies, C. 2012. Identification of candidate limit reference points for the key target species in the WCPFC. MOW1-IP/03 06 Nov 2012 (WCPFC-SC7-2011/MI-WP-03).

Punt, A. E., Butterworth, D. S., de Moor, C. L., De Olivera, J. A. A., & Haddon, M. (2016). Management strategy evaluation: Best Practices. Fish & Fisheries, 17, 303–334. https://doi.org/10.1111/faf.12104.

Punt, A.E., 2019. Spatial stock assessment methods: A viewpoint on current issues and assumptions. Fisheries Research Volume 213, May 2019, Pages 132-143. https://doi.org/10.1016/j.fishres.2019.01.014

Punt A.E., Castillo-Jordan C., Hamel O.S., Cope J.M., Maunder M.N. and Ianelli J.N. 2021. Consequences of error in natural mortality and its estimation in stock assessment models. Fisheries Research 233: 105759.

Stewart, I.J., Hicks, A.C., Taylor, I.G., Thorson, J.T., Wetzel, C., Kupschus, S., 2013. Comparison of stock assessment uncertainty estimates using maximum likelihood and Bayesian methods implemented with the same model framework. Fisheries Research 142 (2013) 37– 46.

Thorson, J.T., Dorn, M.W., Hamel, O.S. 2019. Steepness for West Coast rockfishes: Results from a twelve-year experiment in iterative regional meta-analysis. Fisheries Research, Volume 217, September 2019, Pages 11-20 https://doi.org/10.1016/j.fishres.2018.03.014.

Thorson, J. T. 2020. Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using a data-integrated life-history model. Fish and Fisheries, 21: 237–251. John Wiley & Sons, Ltd. https://doi.org/10.1111/faf.12427.

Van Beveren, E., Benoît, H.P., Duplisea, D.E., 2021. Forecasting fish recruitment in age-structured population models. Fish and Fisheries, doi: 10.1111/faf.12562.

Ward, E. J., Holmes, E. E., Thorson, J. T., & Collen, B. (2014). Complexity is costly: A meta-analysis of parametric and non-parametric methods for short-term population forecasting. Oikos, 123(6), 652–661. https://doi.org/10.1111/j.1600-0706.2014.00916.

Zheng, N., Robertson, M., Cadigan, N., Zhang, F., Morgan, J., & Wheel, L. 2020. Spatiotemporal variation in maturation: A case study with American plaice (*Hippoglossoides platessoides*) on the Grand Bank off Newfoundland. Canadian Journal of Fisheries and Aquatic Sciences, 77(10), 1688–1699.