Fisheries Research 224 (2020) 105465

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

When are model-based stock assessments rejected for use in management and what happens then?

André E. Punt[a,b,*], Geoffrey N. Tuck[a], Jemery Day[a], Cristian M. Canales[c], Jason M. Cope[d], Carryn L. de Moor[e], José A.A. De Oliveira[f], Mark Dickey-Collas[g,h], Bjarki Þ. Elvarsson[i], Melissa A. Haltuch[d], Owen S. Hamel[d], Allan C. Hicks[j], Christopher M. Legault[k], Patrick D. Lynch[l], Michael J. Wilberg[m]

[a] *CSIRO Oceans and Atmosphere, Castray Esplanade, Hobart, TAS, Australia*
[b] *School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, USA*
[c] *Escuela De Ciencias Del Mar, Pontificia Universidad Catolica De Valparaiso, Valparaiso, Chile*
[d] *National Marine Fisheries Service, Northwest Fisheries Science Center, Seattle, WA, USA*
[e] *Marine Resource Assessment and Management (MARAM) Group, Department of Mathematics and Applied Mathematics, University of Cape Town, South Africa*
[f] *The Centre for Environment, Fisheries and Aquaculture Science, Pakefield Road, Lowestoft, NR33 0HT, United Kingdom*
[g] *ICES H. C. Andersens Boulevard 44-46, DK 1553 Copenhagen V, Denmark*
[h] *National Institute for Aquatic Resources, Technical University of Denmark, Kemitorvet 1, DK-2800 Lyngby, Denmark*
[i] *Marine and Freshwater Research Institute, Skúlagata 4, 101, Reykjavík, Iceland*
[j] *International Pacific Halibut Commission, 2320 West Commodore Way, Suite 300 Seattle, WA 98199, USA*
[k] *National Marine Fisheries Service, Northeast Fisheries Science Center, Woods Hole, MA, USA*
[l] *Office of Science and Technology, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Silver Spring, MD, USA*
[m] *Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD, USA*

ARTICLE INFO

Handled by: S. Xavier Cadrin

Keywords:
Peer review
Retrospective analysis
Stock assessment
Uncertainty

ABSTRACT

Model-based stock assessments form a key component of the management advice for fish and invertebrate stocks worldwide. It is important for such assessments to be peer-reviewed and to pass scientific scrutiny before they can be used to inform management decision making. While it is desirable for management decisions to be based on quantitative assessments that use as much of the available data as possible, this is not always the case. A proposed assessment may be found to be unsatisfactory during the peer-review process (even if it utilizes all of the available data), leading to decisions being made using simpler approaches. This paper provides a synthesis across seven jurisdictions of the types of diagnostic statistics and plots that can be used to evaluate whether a proposed assessment is 'best available science', summarizes several cases where a proposed assessment was not accepted for use in management, and how jurisdictions are able to provide management advice when a stock assessment is 'rejected.' The paper concludes with recommended general practices for reducing subjectivity when deciding whether to accept an assessment and how to provide advice when a proposed assessment is rejected.

1. Introduction

There is an increasing expectation that management advice for fisheries is to be provided for as many stocks of fish and invertebrates subject to harvest as possible, and it is now the case that advice on catch limits is expected for most of the federally-managed species in Australia, the European Union (EU), and the United States. In these regions and several others, management advice in the form of catch limits depends on having (a) a harvest control rule that relates catch (or perhaps fishing mortality) to inputs, and (b) the inputs required to apply the harvest control rule. The inputs to the harvest control rule depend on its nature and can generally be categorized into either the results from a quantitative (population model-based) stock assessment[1] or from indicators of stock status and productivity sampled directly from the fishery or using fishery-independent methods (empirical harvest control rules; see Rademeyer et al. (2007) for the advantages and

[*] Corresponding author at: CSIRO Oceans and Atmosphere, Castray Esplanade, Hobart, TAS, Australia.
E-mail address: aepunt@uw.edu (A.E. Punt).

[1] Henceforth "model-based stock assessments"

https://doi.org/10.1016/j.fishres.2019.105465
Received 25 September 2019; Received in revised form 28 November 2019; Accepted 30 November 2019
Available online 12 December 2019
0165-7836/ Crown Copyright © 2019 Published by Elsevier B.V. All rights reserved.

disadvantages of empirical harvest controls rules vs inputs from a model-based stock assessment). Management advice also often involves estimating the probability that a stock is moving towards or fluctuating about its target reference point and the probability that it is below its limit reference point, and this is now a requirement under the EU Marine Strategy Framework Directive.

Best practice for using empirical harvest control rules involves identifying a range of candidate empirical harvest control rules and using closed loop simulation (i.e., Management Strategy Evaluation; MSE: Punt et al., 2016) to assess how well each harvest control rule satisfies the (agreed) management objectives. A key aspect of MSE is to select a range of alternative scenarios that reflect the possible state of the system and to explore the performance of candidate harvest control rules to determine the range of conditions to which they are robust (i.e., perform adequately or at least in a predictable way). It is now increasingly common not only to select an empirical harvest control rule based on MSE, but also to identify situations when an empirical harvest control rule should not be used to provide management advice ('reject' the empirical harvest control rule) and what should be done when the harvest control rule is rejected. Such 'exceptional circumstances provisions' have, for example, been developed for the Cape hake and rock lobster fisheries off South Africa (Johnston and Butterworth, 2005; Rademeyer et al., 2008) amongst others. Population, fishery, and ecosystem indicators are used in these provisions to determine whether the outcomes from an empirical harvest control rule should be over-ridden, in which case catch limits are set based on alternative analyses developed by a species-specific Working Group. The expectation is that rejection of the empirical harvest control rule is rare.

It is now becoming common for MSE to be used to test the harvest control rules that are based on model-based stock assessments, but it is less common to use MSE to test the combination of the assessment method and the harvest control rules (the management strategies developed by IWC (2012), Hillary et al. (2016) and ICES (2019a) being among the exceptions in this regard, although the population models underlying some of those management strategies are quite simple). The assessment used to provide the inputs for the harvest control rule is developed on a regular basis (possibly annual) taking into account the latest information to ensure that management decisions are based on the 'best available science' regarding the size, productivity and status of the stock. The results of an assessment might be used for several years to set management arrangements such as the Total Allowable Catch.

Developing a model-based assessment generally involves the following steps: (a) identification of the data available to conduct a stock assessment, (b) selection of candidate models that can make use of the data and provide the information needed by management, (c) application of the models and selection of a 'best' model and set of specifications (and often a set of alternative models and specifications to allow uncertainty to be captured), (d) application of the selected models including production of forecasts and calculation of stock status relative to reference points and (e) peer-review of the stock assessment (which generally includes a recommendation on whether and how the assessment should be used to inform management). The specifications of a model-based stock assessment include selection of the basic form of the population dynamics equations (e.g., biomass dynamics, age-structured, length-structured, whether sex is explicitly represented or not, etc.), choices related to biological and fishery processes (such as how many fisheries and surveys to include, how growth and reproduction are to be modelled, values for fixed parameters such as the natural mortality rate, etc.), specifications for which data to use, how the raw data collected from fisheries and surveys are pre-processed for use in the stock assessment, how those data are to be included in the objective function, and how the data sources are to be weighted.

Model-based stock assessments are tailored to the stock being assessed with the aim of making the best use of the available data to provide the most accurate and precise estimates of the quantities needed for management decision making. Development of a model-based stock assessment is tied to the quantities needed for management purposes as well as the data available. However, there are many (often nuanced) decisions when conducting a model-based stock assessment, which can lead to some subjectivity in the process. For example, some analysts prefer simple models that require fewer (but more substantial) assumptions and other analysts prefer complex models. Consequently, it is reasonable to expect that different analysts will create different assessments when confronted with the same information (Deroba et al., 2015).

The stock assessment community has held workshops to identify 'best practices' for various aspects of stock assessment (e.g., Maunder et al., 2014, 2016, 2017; Sharma et al., 2019) and developed Terms of Reference for assessments (e.g., Pacific Fishery Management Council (PFMC, 2019) to reduce subjectivity when conducting stock assessments. In addition, most jurisdictions have a peer-review process to review stock assessments.

The nature of the peer-review process varies substantially among (and within) jurisdictions, with some using only reviewers with a natural science background and clear conflict of interest restrictions (e.g., in many of the regions of the US; Lynch et al., 2018) and others including a broad range of stakeholders, some of whom may be directly impacted by the outcomes of the assessment (e.g., Australia, the ICES region via the benchmark process). However, a common feature of peer-review processes is that they make a recommendation whether the reviewed stock assessment is suitable for management purposes (often expressed as the assessment being the 'best available science' or words to that effect), and if not what (a) conclusions can be drawn from the proposed assessment, and (b) additional analyses / data that could resolve the concerns that led to the non-acceptance of the assessment.

This paper first summarizes the types of information that are provided in a stock assessment to allow peer-reviewers to evaluate whether a stock assessment is 'best available science', and hence the types of considerations that are used to decide whether to 'accept' or 'reject' an assessment. Then, the types of problems that lead to rejection of an assessment are identified using a survey of assessments that were proposed for use in management but were rejected (and the reasons for rejection were documented). This approach is needed because no jurisdiction has formal quantitative guidelines for rejecting a stock assessment. It is the case that some assessments that would have been rejected were recognized as being flawed before peer review by the assessment analysts, and thus not presented for peer review; their 'rejection' is consequently not documented here. The approaches for evaluating whether an assessment is 'best available science' can be applied to decide whether an operating model intended to be used in a Management Strategy Evaluation is adequate. However, it is often the case that the standards for rejecting a candidate operating model are less stringent because the aim of operating models within MSE (particularly robustness tests) is to explore whether candidate management schemes behave as expected.

Rejection of a model-based assessment does not (necessarily) lead to there being no management advice. Rather, most jurisdictions have a 'fall-back' policy that specifies how management advice is provided even if the stock assessment using the most recent data is rejected. This paper consequently also reviews the 'fall-back' polices. Additionally, our categorization of stock assessments that were rejected does not necessarily align with how each jurisdiction would communicate their assessment results. However, we attempted to be consistent across jurisdictions to facilitate comparisons.

The focus for this paper is on seven countries / regions (Australia, Chile, the ICES region, Japan, New Zealand, South Africa, and USA) that are at the forefront of the development and application of model-based stock assessment methods, and generally have sufficient data and technical expertise to conduct model-based stock assessments as well as having well-established peer-review processes. The manuscript does not attempt to summarize all stock assessments that have been rejected. Rather, the focus is on providing a representative set of stock assessments to showcase the breadth of causes of rejection.

2. Information presented to a peer-review body that could lead to rejection

Most of the assessment reports for model-based stock assessments presented to peer-review bodies have a typical format. These reports typically include a description of (a) the stock and some of its basic biological characteristics such as stock structure, longevity, growth and maturity, (b) the fisheries that operate on the stock, (c) the data on which assessments could be based and how these data were collected, (d) the population dynamics models considered for the assessment and which model (or models) formed the basis for the assessment, and (e) the results of the assessment, including diagnostics, estimates and plots. Assessments often include a 'bridging' analysis that shows the progression from a previous assessment to the assessment under review to allow the peer reviewers to assess what features of the model structure or data has led to changes in assessment outcomes.

The types of diagnostics that are expected to be reported in stock assessments differ among jurisdictions and types of stock assessments, but the following are the most common:

1 Diagnostics related to technical problems with the assessment. Convergence diagnostics to show that the estimation procedure has converged to the global minimum of the objective function (for assessments based on maximum likelihood or more generally penalized maximum likelihood) or that the algorithm used to draw samples from a posterior distribution has done so successfully (for assessments based on Bayesian methods). Common convergence diagnostics for maximum likelihood assessments include checking that the maximum gradient among parameters is sufficiently small, ensuring that all parameters move from their starting values, checking whether the Hessian matrix is invertible, checking whether any of the parameters are on or close to bounds, and assessing whether the minimization method has converged to the global minimum of the objective function, for example by examining the results of 'jitter' analyses in which the starting values for the parameters are varied and the estimation procedure is applied to each alternative set of starting values. For Bayesian methods, there is a wide range of diagnostics designed to detect lack of convergence (e.g., those included in the *coda* package in R; (Plummer et al., 2006).

2 Diagnostics related to fits to the data. Plots showing fits to the data based on observations and model predictions (maximum likelihood analyses) or posterior predictive distributions (Bayesian analyses) as well as residual plots to assess whether there is evidence for data conflicts or model mis-specification. Considerable research has been undertaken to develop 'residual plots' for complex data sets such as growth from tagging data (e.g., Punt et al., 2017), length-frequency data (e.g., Francis, 2011), and conditional age-at-length data (available in the R package r4ss; Taylor et al., 2019).

3 Retrospective patterns. Conducting a retrospective analysis, which involves re-running the assessment after removing data for the most recent year, the two most recent years, etc. and evaluating whether there are systematic changes in key model outputs such as spawning biomass, fishing mortality and recruitment. Mohn's Rho (Mohn, 1999) has been developed to quantify retrospective patterns, and Hurtado et al. (2015) established quantitative thresholds to define 'significant' values of Mohn's Rho. An alternative approach to defining 'significant' values of Mohn's Rho is used in the northeast US that involves comparing the magnitude of the adjusted biomass and fishing mortality rate to the uncertainty in the terminal year estimates (Brooks and Legault, 2016).

4 Diagnostics to identify data conflicts. Constructing likelihood profiles for 'key parameters' (e.g., the virgin recruitment, R_0; the rate of natural mortality M; and the steepness of the stock-recruitment relationship h), and plotting the value of the negative log-likelihood in total and by data component against the profiled parameter. Data conflicts are indicated by the minimum of the negative log-likelihood function occurring at different values for the profiled parameter among data sources (Carvalho et al., 2017). Wang et al. (2014) proposed an extension of R_0 profiling to diagnose mis-specified stock assessment models. This involves constructing an R_0 profile for data components simulated without error from a known stock assessment model. The R_0 profile from the known stock assessment model is assumed to represent the 'true' information content of each data component. Any differences in subsequent models from the R_0 profile from the known stock assessment model are presumed to indicate conflict in the data or model misspecification (Carvalho et al., 2017). Results of likelihood profiles can also be used to show how key derived outputs (e.g., current spawning and relative stock status) change with the key parameters and hence the uncertainty in the derived outputs.

5 Diagnostics that determine whether the assessment is behaving as expected. Conducting sensitivity tests and looking for unexpected results and whether the results are very sensitive to 'small' changes to the specifications of the assessment; most assessments include sensitivity analyses that involve changing the values for fixed parameters, including additional, or excluding baseline, data sources, and changing the weights assigned to the various data sources. Sensitivity analysis is used primarily to quantify uncertainty in the baseline (or 'best') model configuration. The aim of sensitivity analyses in the context of evaluating assessment model performance is to determine whether the results from the assessment change 'as expected' and in particular that assessment outcomes are robust to 'small' changes to the specifications of the assessment (or at least that the extent of change is consistent with that which would be expected from assessments for other stocks).

Assessment reviews also examine the likely validity of the assumptions of the population dynamics model and the biological plausibility of the estimates (and variance) of those parameters that have physical interpretations, such as those that determine the growth curve, the values for survey catchability coefficients, and whether selectivity is strongly domed when this is not expected. The latter can be evaluated by exploring why selectivity might be domed (e.g., older fish are beyond the range of the fisheries/surveys) and reporting the proportion of total biomass unavailable to the fisheries/surveys. Whether an assumption is valid is necessarily a more subjective consideration and there are consequently no formal procedures for assessing the validity of assumptions.

Consideration of data not included in the model may indicate a problem with the model fit as well, for example, a production model that estimates current conditions similar to unexploited but the current age structure is severely truncated (e.g., Deroba et al., 2015). In addition, large changes from the previous assessment to estimates of parameters or model outputs are usually evaluated during the peer-review process; large changes that cannot be explained are considered reasons for rejecting a new assessment.

Evaluation of how the various data sources are weighted is also considered during the assessment review process. Methods such as those of McAllister and Ianelli (1997), Francis (2011) and Thorson et al. (2017) are available to determine the weights assigned to age- and length-composition data sources, and the method of Methot and Taylor (2011) to determine the value for the extent of variance in recruitment about the stock-recruitment relationship as well as the recruitment bias correction factor. Setting incorrect weights for the data can exacerbate the effects of data conflicts, under (or over) estimate the variances of the model outputs, and assign incorrect weights to alternative models (Punt, 2017).

3. What may cause a stock assessment to be rejected?

Except for some South African fisheries for which exceptional circumstances provisions are provided for both empirical and model-based harvest control rules (Rademeyer et al., 2008; Johnston and Butterworth, 2005), no jurisdiction has developed explicit rules regarding whether a model-based assessment should be accepted to provide management advice (unlike the situations for the empirical harvest control rules referred to in the introduction). As such, whether a model-based assessment is rejected or not depends largely on the expert judgement within the peer-review panel, and given the diagnostic statistics and plots presented in the assessment report, as well as analyses requested during the peer-review process.

Table 1 (USA) and 2 (other jurisdictions) provide a summary of examples of assessments that have been rejected and the reasons for their rejection. As noted above, the examples of rejected assessments reflect what is publicly available. It is likely that many other assessments are not submitted for peer-review because problems are detected

Table 1
Examples of model-based assessments in the US that have been rejected.

Stock	Reason	Reference
Caribbean region		
Queen Conch	The data were inadequate to provide information on status. This was because: (a) landings data were incomplete, missing the recreational component in most years; (b) trends in commercial CPUE were not informative of stock abundance; and (c) expansion of habitat-specific survey densities to domain-wide abundance estimates were based on low survey coverage and small sample sizes.	SEDAR (2007)
Queen Snapper	The assessment method had not been tested enough in a simulation study; the Bayesian posterior was not adequately sampled.	SEDAR (2011a)
Gulf of Mexico region		
Blacknose shark	Concerns about the assumption about an unfished status when the assessment started; marked residuals patterns in the fit to some of the indices.	SEDAR (2011b)
Southeast region		
Atlantic Large Coastal Shark Complex	Assessment cannot represent the status of a complex because of the potential for conflicting information from the various sources.	SEDAR (2006)
Goliath grouper	Treatment, and high uncertainty of the landings (catch) and the indices of relative abundance, and the structure of the chosen assessment models.	SEDAR (2016a)
Gray triggerfish (Atlantic)	Data weighting, and primarily over-fitting a survey index of abundance; also errors found in some of the basic data.	SEDAR (2016b)
Hogfish, Georgia-North Carolina	Insufficient data, and conflicts among the data; model does not fit the available indices.	SEDAR (2014)
Mid-Atlantic region		
Black sea bass, US Mid-Atlantic	Assessment assumes a completely mixed stock, while tagging analyses suggest otherwise; whether the reference points are appropriate for the species (protogynous hermaphrodite).	MAFMC (2012)
Atlantic surfclam[1]	Unable to resolve the scale of fishing mortality and biomass[1]	MAFMC (2017)
New England region		
Atlantic halibut	The assessment produced an unstable and unrealistic solution. Estimates of current stock size were highly sensitive to initial conditions and slight changes in assumed parameter values.	NEFSC (2015)
Atlantic cod, Georges Bank	Retrospective patterns (worse than in an earlier assessment) as well as poor residual patterns in the fits to the survey index of abundance.	NEFSC (2015)
Winter flounder (Gulf of Maine)	Difficulty with conflicting data trends, specifically the large decrease in the catch over the time series with very little change in the indices or age structure in both the catch and surveys. The scaling of the population estimates was sensitive to the weight imposed on the catch-at-age compositions. The within-model uncertainty did not capture the uncertainty, considering how sensitive the results were to the model formulation and weighting.	NEFSC (2011)
Witch flounder	Retrospective pattern.	Sullivan et al. (2016)
Yellowtail flounder (Georges Bank)	Strong retrospective pattern, conflicts between model estimates of biomass and independently derived ones	TRAC (2014)
Pacific region		
Arrowtooth flounder[2]	The analyses all exhibited results that were unexpected given the observed data	PFMC (2015a)
Kelp Greenling (California)	The estimates of the harvest rates for one sector were unrealistically high (model misspecification or they represent local depletion)	PFMC (2005a)
Pacific sanddab	Inconsistency between the model estimates of biomass and those based on swept area surveys.	PFMC (2013)
Stripetail rockfish	Unable to resolve the scale of fishing mortality and biomass.	Cope et al. (2013)
Vermilion rockfish	The model produced divergent results and exhibited extreme sensitivity to what should be minor changes in data or assumptions (estimates of current stock depletion ranging from over twice unfished biomass to 1 % of unfished biomass).	PFMC (2005b)
Yellowtail rockfish (southern area)	General paucity of data and a very high degree of sensitivity to several factors and parameters; also a lack of consistency among key parameters (e.g., natural mortality) between the northern and southern models.	PFMC (2017a)
North Pacific region		
Arrowtooth flounder	Significantly reduced biomass, OFL, and ABC estimates resulting from the new female maturity relationship and concerns over the method used to estimate the maturity parameters	Anon (2012)
Western Pacific region		
Main Hawaiian Islands Deep 7 Bottomfish Complex[3]	Quality of input data on catch and CPUE questionable.	Brodziak et al. (2014)

1: passed SAW/SARC peer review; the SSC had concerns about the appropriateness of the reference points, and decided not to use them for determining OFLs and ABCs; issues eventually resolved following further discussion.
2: a data-moderate assessment not used for management decision making but triggered the need for a full assessment to address the issues with the assessment.
3: the 2015 assessment was not rejected outright, but the model updates were; eventually, the previous model was updated with new data.

Table 2
Examples of rejected assessments for regions other than the USA.

Stock	Reason	Reference
ICES Region		
NEA mackerel	High sensitivity to input data	ICES (2019b)
Norwegian Spring-Spawning herring	Error in the conversion of acoustic data from the Norwegian acoustic survey	ICES, 2017a ICES(2017a)
Eastern Channel sole	Unavailability of a tuning index due to a change in the method for calculating effort; subsequently an unrealistic estimate of plus group size, leading to an almost doubling of catch advice.	ICES (2019c, d)
North Sea saithe	Error in the calculation for average F.	ICES (2019e)
North Sea turbot	Incorrect implementation of how the model specified a commercial tuning index	ICES (2018a)
Herring (27.43031)	Retrospective pattern; issues with estimates of abundance based on acoustics	ICES (2017b)
Deep pelagic redfish (Irminger Sea and adjacent areas)	Disagreement on the appropriateness of using the 2015 and 2018 survey data for biomass estimation	ICES (2019f)
Roundnose grenadier (Celtic Seas and the English Channel, Faroes grounds, and western Hatton Bank)	No reliable biomass series available	ICES (2018b)
Southeast Australia		
Silver Trevally	Changes in the assumptions of the model resulted in large changes in model outputs (e.g., depletion from 0 % to 98 %).	Anon (2007)
Pink ling	Two assessments were presented; one was selected and the other was rejected.	Anon (2013)
Western gemfish	Uncertainty about stock structure, data issues and model sensitivities, in particular: large changes in depletion estimates depending on which areas are included in the assessment, sensitivity to particular years of biological data (data quality issues), and high discarding	GABRAG minutes (2008, 2010, 2013, 2016)
Chile		
Nylon shrimp	Model overestimation in the predictions of abundance indexes (swept-area surveys and CPUE)	SUBPESCA/CCT-CD/2 (2013)
Anchovy (north Chile)	Considerable uncertainty in growth and age	SUBPESCA / CCT-PP / 6 (2017a)
King clip	Uncertainty in stock spatial structure, data and model specification	SUBPESCA/CCT-RDZSA/4 (2016)
Toothfish (southern Chile)	Considerable uncertainty in stock spatial structure and data	SUBPESCA / CCT-RDAP / 4 (2017b)

during the process of their development. For example, in some regions, assessments are developed within the context of ‘ongoing’ review within a working group / scientific committee setting prior to being submitted for ‘official’ review by, for example, an external review panel.

There are, as expected, a variety of reasons why an assessment might be rejected (Tables 1 and 2). Some assessments are rejected due to errors in the basic data. The assessment of gray triggerfish (*Balistes capriscus*) in the USA southeast (Atlantic stock) was rejected for a variety of reasons, including a data provider discovering an error in the age-composition data associated with the survey used when fitting the model. A similar problem occurred for Norwegian Spring-spawning herring in 2017; there was an error in the conversion of acoustic data from the Norwegian acoustic survey on the spawning grounds for the period 1988–2008, which implied that abundance indices from this period were significantly underestimated (ICES, 2017a). Errors in the data included in assessments are not necessarily fundamental flaws with the assessments and can be corrected before the assessment is reviewed again.Alternatively, data with known errors can remain in the assessment and be accounted for through increased uncertainty around the assessment results. Some of the reasons for rejecting assessments relate to the diagnostics identified in Section 2.

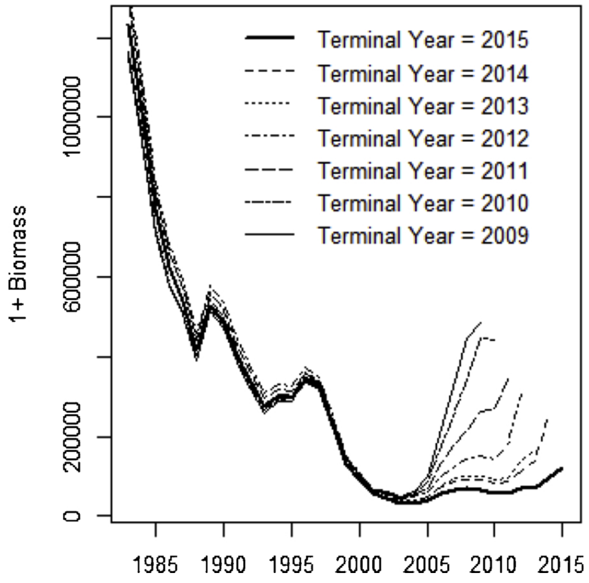

Fig. 1. Retrospective pattern in the 2015 assessment of Pacific mackerel (Pacific Fishery Management Council (PFMC, 2015b).

- Poor retrospective patterns. Fig. 1 shows the retrospective pattern for the biomass of animals of age 1 and older for Pacific mackerel (*Scomber japonicus*) off the USA west coast. Fig. 1 shows very clearly that the addition of new data leads to less and less optimistic results. Rejection due to retrospective patterns is common in Europe (Table 2) and the New England region of the USA (Table 1). Many factors have been identified as possible causes for retrospective patterns (e.g., Brooks and Legault, 2016; Legault and Chair, 2009; Hurtado et al., 2015; Szuwalski et al., 2018). In the USA New England region, these reasons include mis-reporting of catches, and changes in natural mortality caused by increasing predator populations. Not all assessments that exhibit strong retrospective patterns are necessarily rejected; some involve a rho-adjustment for stock status determination and provision of catch advice (Brooks and Legault, 2016).
- Extreme sensitivity to changes to the specifications of the assessment. While it should be expected that the outcomes of an assessment change given changes to model structure, fixed parameters, and data weighting, extreme changes (particularly those that cannot

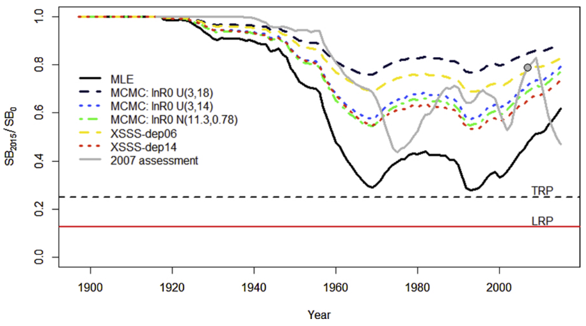

Fig. 2. Stock status time series for arrowtooth flounder off the US west coast across all potential base case models and treatments compared to the 2007 assessment. The point indicates where the 2007 assessment ended. Time series beyond that point are projected values. Target (TRP) and limit (LRP) reference points are indicated by the horizontal lines. Source: Cope (2015).

be explained easily) can lead to rejection of an assessment. For example, Fig. 2 shows the sensitivity of the relative biomass trajectory from a data-moderate Bayesian assessment (which included indices of abundance, but no composition data) of arrowtooth flounder (*Atheresthes stomias*) off the west coast of the USA to the choices for priors and the method for sampling from a Bayesian posterior. Several of the assessments in Tables 1 and 2 were rejected due to 'extreme sensitivity', but in general, how much sensitivity is too much was not defined in the peer-review report.

- Conflicts in the data on which the assessment is based. For example, the trends in the northern and southern regions for black sea bass (*Centropristis striata*) in the USA Mid-Atlantic were conflicting, but the assessment was based on a model that attempted to model the populations from both areas within a structure that assumed that the population was distributed homogeneously across the entire region. Fig. 3 shows the residuals for the assessment of blacknose shark (*Carcharhinus acronotus*). While the population model for this stock is able to mimic some of the data series fairly well (e.g., 'PC Gillnet Adult') there are clear residual patterns for at least two of the relative abundance series.

Although many of the rejections in Tables 1 and 2 follow directly from diagnostic statistics, there are an almost equal number of rejections due to factors that could be considered more subjective. For example:

- Unreasonable parameter estimates and model outputs. This problem includes unrealistically high harvest rates given the fishery involved (e.g., kelp greenling off the Pacific coast) or parameter estimates that seem implausible given auxiliary information. For example, the assessment of Pacific sanddab (*Citharichthys sordidus*) was rejected because the swept area biomass estimates (from fishery-independent sources) were 4–22 times the model estimates of biomass (Pacific Fishery Management Council (PFMC, 2013). Furthermore, in Europe, the assessment for sole (*Solea solea*) in Division 7d was downgraded from Category 1 to Category 3 because of an unrealistically large increase in the plus group (due to the way the XSA assessment estimates the plus group) that led to an almost doubling of catch advice (ICES, 2019c, d).
- An inability to determine the scale of fishing mortality and biomass. This occurs when the assessment is able to resolve trends in abundance (often expressed as biomass or fishing mortality relative to reference points) fairly robustly, but a wide range of absolute biomass levels fit the data equally well. Fig. 4 shows a likelihood profile for the logarithm of unfished biomass for stripetail rockfish (*Sebastes saxicola*) off the USA west coast, which indicates a 95 % confidence interval from ~7 upwards. Although it is not possible to estimate current biomass in this case, it can be concluded that the stock is likely above its target reference point and this information informed management decision making. An inability to determine 'scale' is usually associated with uninformative data or stocks that have not been fished intensively enough to detect a fishing signal in the monitoring data if there is one (e.g., Thorson and Cope, 2017).
- Uncertainty in stock structure and biological parameters. Assessments of black sea bass in the USA Mid-Atlantic and toothfish (*Dissostichus eleginoides*) off south Chile, were rejected due to a considerable uncertainty about population structure, among other concerns. Also, and depending on the model structure, some assessments such as that for anchovy (*Engraulis ringens*) off north Chile are rejected due to uncertainty in estimates of growth and age.

There are also case-specific reasons for rejection. The most common of these is that there is simply not enough information on which to base an assessment. It was, however, often difficult to determine from the peer-review and assessment reports what this meant. Other, somewhat unique, reasons for rejection include problems with model convergence (e.g., queen snapper, *Etelis oculatus*) and the use of an inappropriate model (e.g., black sea bass; the Atlantic Large Coastal Shark Complex). Another type of error, which can lead to a rejected assessment (or in the case of ICES, for an inter-benchmark process) is errors in how a model is coded and incorrect implementation of how the model is to be specified (e.g., turbot and saithe in the North Sea (ICES, 2018a, 2019e).

4. What happens when a stock assessment is rejected?

It is common for jurisdictions to have 'fall-back' positions so that it is possible to provide a scientific basis for the management advice even when an assessment is rejected. However, the design of that fall-back position varies substantially among jurisdictions.

The following approaches have been applied in cases when a 'fall -back' position is needed.

1 The simplest fall-back position is to continue with previous catch limit advice or regulations - even if they were based on a similar analysis to what was rejected. This has occurred for Atlantic mackerel (*Scomber scombrus*) in the U.S. Mid-Atlantic region (Mid-Atlantic Fishery Management Council (MAFMC, 2010; Transboundary Resources Assessment Committee (TRAC, 2010).

2 Basing management advice on the last agreed model but with additional data. Assessments for the USA North Pacific Fishery Management Council need to include a model run in which the assessment is based on the last agreed assessment but with updated input data (North Pacific Fishery Management Council (NPFMC, 2016). This means that even if all newer models are rejected, there is always a fall-back on which to base management advice. The peer-review panel that reviewed the assessment of Pacific mackerel in 2007 adopted a similar approach by endorsing an assessment conducted using the method used in the last accepted assessment and which did not have the undesirable behaviour of the application in the assessment report submitted for review (Pacific Fishery Management Council (PFMC, 2007).

3 'Downgrading' the assessment to a simpler assessment method. For example, all benchmark assessments to be used by the USA New England and Mid-Atlantic Fishery Management Councils must develop a 'plan B,' along with the proposed assessment in case the proposed assessment is rejected. The 'plan B' assessments are index-based, easy to compute, and theoretically require little review once agreed upon (Northeast Fisheries Science Center (NEFSC, 2017). This 'plan B' approach was developed to define roles, responsibilities and process in cases when assessment working groups or review panels deem that a stock assessment is insufficient or inappropriate, and empirical approaches are required to provide management advice. These simple 'plan B' approaches typically rely on trends in the survey data to change the management advice (similar to many

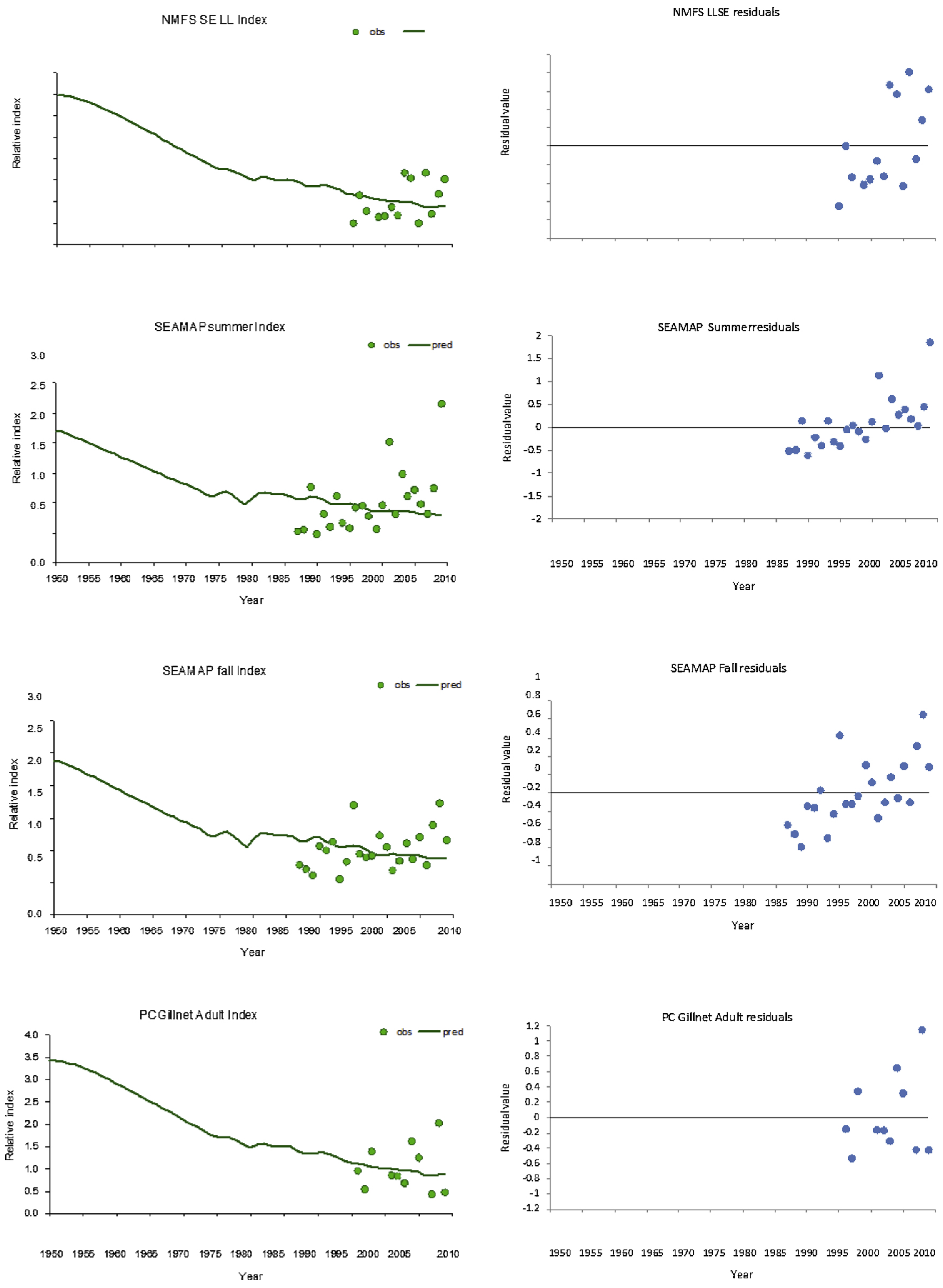

Fig. 3. Predicted fits to indices (left) and residual plots (right) for the base run for blacknose shark (source: SEDAR (2011b)).

data-limited approaches). Similarly, if problems with a stock assessment cannot be resolved at ICES, the stock assessment is 'downgraded' from Category 1 or 2 (stocks with quantitative assessments; stocks with analytical assessments and forecasts that are only treated qualitatively) to Category 3 (stocks for which survey-based assessments indicate trends) or 4 (stocks for which only reliable catch data are available and where catches can be used to estimate MSY; ICES, 2018c). The basis for catch advice when stocks are 'downgraded' may change from being based on moving the stock towards achieving MSY to ensuring a precautionary approach is applied (e.g., application of a "two over three" rule in combination with an uncertainty cap and precautionary buffer; ICES, 2012). A similar approach is taken in Japan and Australia where, if a model-based assessment is rejected, an empirical harvest control rule based on catch and CPUE data is used to provide advice on target catches (H. Okamura, Fisheries Research Agency, pers. comm; Little et al.,

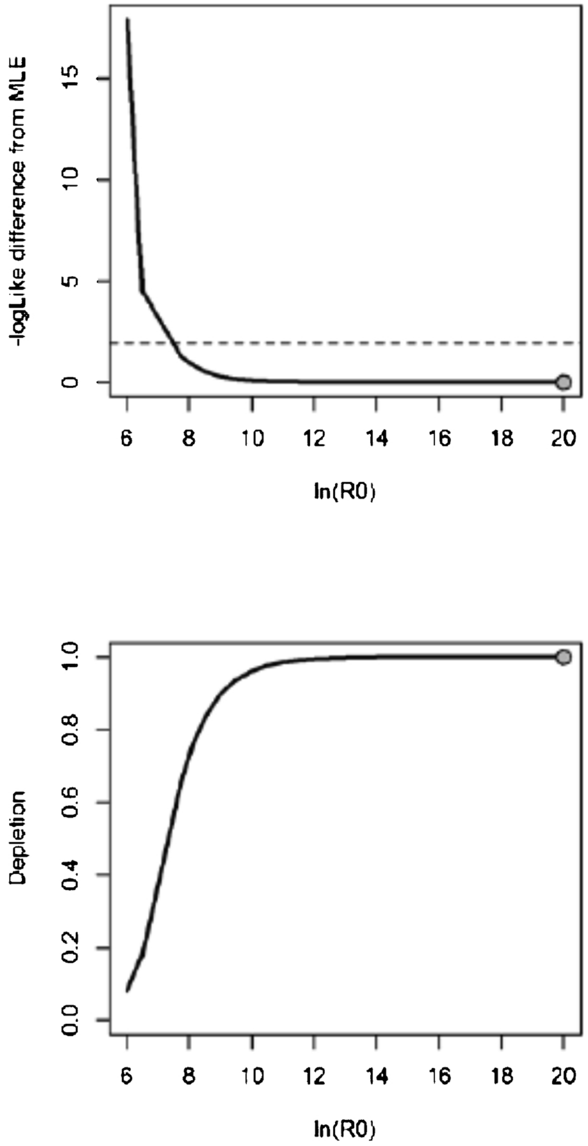

Fig. 4. Likelihood profile (upper panel) and the relationship between logarithm of unfished recruitment and current relative population size (lower panel) for stripetail rockfish (source: Cope et al., 2013).

2011). This type of approach was used by the Scientific and Statistical Committee (SSC) of the USA South Atlantic Fishery Management Council in the case of gray triggerfish when a data-limited approach (DCAC; MacCall, 2009) was applied after the model-based assessment was rejected.

4 Working to improve the assessment before a decision needs to be finalized. For example, for Atlantic surfclam, a subcommittee of the SSC of the USA Mid-Atlantic Council met with the analysts to evaluate several additional analyses that were not in, but supported, the assessment. Given the additional analyses, the assessment results were ultimately used for management. A similar process (referred to as the 'mop up' panel) is implemented for the USA Pacific Fishery Management Council.

In some cases, the peer-review body may conclude that even though the assessment is unacceptable from a model diagnostics viewpoint, it still provides the best information on which to base management recommendations. This occurred for the 2015 assessment of Pacific mackerel, which exhibited a strong retrospective pattern (Fig. 1). The peer-review body eventually decided to base management advice on the model configuration suggested by the analysts even given the concerns, noting '*While recognizing the substantial issues that remain in determining the scale of the stock, the SSC endorses the STAT-preferred assessment model as the best available scientific information for management of Pacific mackerel.*' (Pacific Fishery Management Council (PFMC, 2015c). The buffer between the Overfishing Level (OFL; the catch for the next year corresponding to a fishing mortality of F_{MSY} or a proxy thereof) and the Acceptable Biological Catch (the OFL reduced by the extent of scientific uncertainty) for this stock was more than doubled given the uncertainty associated with the assessment.

Finally, it is not uncommon for a stock assessment to be 'partially rejected'. Assessments are used for several purposes, and an assessment can be rejected for one purpose but not for others. In particular, assessments may be insufficient as the basis for assessing biomass and stock status but sufficient to support management decision analysis. Treating assessments as 'component analyses', where some parts may be useful for management while others are not, is a formal approach in the USA (Methot, 2019). For example, the assessment of Pacific sanddab off the US west coast was rejected for use as the basis for setting an OFL, but not for concluding that the stock was not overfished. This was because the primary uncertainty related to the ability of the assessment to resolve the scale of the population, with relatively small changes to the specifications of the assessment leading to marked change in biomass in absolute, but not relative, terms. The OFL was eventually set using the data-limited catch-only method Depletion-Based Stock Reduction Analysis (DB-SRA; Dick and MacCall, 2011). Similarly, the estimates of biomass and fishing mortality in the assessment of Atlantic surfclam in the USA Mid-Atlantic were deemed highly uncertain (including by the assessment analysts), but the results were sufficient to conclude that overfishing was not occurring, and the stock was not overfished (Mid-Atlantic Fishery Management Council (MAFMC, 2017). The SSC of the USA Mid-Atlantic Fishery Management Council was able to set an ABC based primarily on expert judgement (Mid-Atlantic Fishery Management Council (MAFMC, 2017). The results of the GADGET assessment of anchovy in ICES region 9a south (Atlantic Iberian water) were not used directly but rather as input to a data-limited (ICES category 3) assessment (Anon, 2019), an approach used for a number of stocks in ICES (i.e., inputs to the "two over three" rule being model-based instead of empirically-based; ICES, 2012).

It is clear from the above summary that some jurisdictions have a quite formal (and well-specified) process to handle cases in which assessments are rejected. At the other extreme, the expert judgement of the peer-review panel may be used to identify a sustainable catch limit. Expert judgement may be supplemented by simple analyses (such as the average catch over the period the fishery was considered stable or assuming that the biomass equals a recent average from a survey). The approaches above differ in terms of whether the management advice is more precautionary if a fall-back position is adopted (for example within the US system by increasing the buffer between the Overfishing Level and the Acceptable Biological Catch). It is not clear how effective these methods are in terms of achieving management goals, and we therefore do not provide recommendations for 'best' practice. Rather, we recommend that the closed-loop simulations (aka MSE) be undertaken to explore the advantages and disadvantages of each approach. ICES (2013) used simulations to compare the performance of ICES Categories 1–4 (with 1 = data-rich, with increasing data limitation for 2, 3, etc.) and found that the data-limited categories were not necessarily more precautionary.

5. Discussion

This paper highlights that the types of diagnostics that can be included in a stock assessment are well known and commonly reported and evaluated. These diagnostics also often form the basis for the decision whether to accept or reject an assessment. However, many diagnostics do not have specified thresholds that can be used to make objective decisions regarding the acceptance or rejection of an assessment.

Increasingly, most stock assessments report the same types of diagnostics. This is probably a consequence of peer-review processes that involve reviewers external to the region in which the assessed stock is found[2] . Involvement in a review panel also has benefits for assessments beyond that being reviewed because reviewers may encounter methods they are not familiar with, leading to 'convergent evolution' of diagnostics. The use of common diagnostics is also enhanced by attempts to identify best practice guidelines for stock assessments, such as by the workshops run by the Center for the Advancement of Population Assessment Methodology (CAPAM). The use of software packages to conduct stock assessments increases the likelihood that code is available to produce these diagnostics (which can be quite complicated) and provides a common reporting format. This is the case for stock assessments based on Stock Synthesis (Methot and Wetzel, 2013) for which the R package *r4ss* is used extensively. SAM (Nielsen and Berg, 2014; Berg and Nielsen, 2016) has an associated R package that runs the assessment and produces diagnostics. These packages have large user bases, making it cost effective to invest in developing software to automatically produce new diagnostics.

Terms of Reference established for conducting assessments and reporting their results implicitly provide a loose and reviewing 'schema' for the rejection process. However, those Terms of Reference do not provide hard and fast rules for rejection; the decision is ultimately left to individual peer-review groups. In addition, care should be taken so as not to develop Terms of Reference that set the requirements for accepting an assessment "too high" or "too low".

Whether an assessment is accepted or rejected may ultimately depend on the composition of the review group. In fact, it would be expected that two review panels provided with the same information could draw different conclusions regarding the acceptability of the same assessment. Increased use of a formal system for rejecting assessments would help (but not eliminate) this problem because peer-review always has a subjective component given the background and expertise of the reviewers. Various jurisdictions have adopted approaches to minimize the potential problem of subjectivity in the peer-review process. For example, the USA North Pacific Fishery Management Council reviews all crab assessments, and groundfish assessments for the Gulf of Alaska and Bering Sea and Aleutian Islands using the same groups, while the assessments for the USA Mid-Atlantic and New England regions are often reviewed in groups of similar stocks. The USA Pacific Fishery Management Council enhances consistency in its review process by having a common (external) reviewer among its groundfish assessments within each assessment cycle and by having the assessment review meetings chaired by a member of its SSC. Nevertheless, inconsistencies in process can occur. For example, the 2014 assessment of the Main Hawaiian Islands Deep 7 Bottomfish Complex was rejected by the peer-review process for the Western Pacific Fishery Management Council based on the quality of input data, even though the same data were used in previous (and later) assessments.

The peer-review processes surrounding model-based stock assessments can work with the analysts to identify an assessment that is as acceptable as possible. This is the case in South Africa and at the International Pacific Halibut Commission (IPHC) where an assessment can be refined over multiple meetings. For example, the IPHC assessment is presented to an independent scientific review board in June and September of each year, with an aim of the June review to identify potential improvements to the assessment, and a full independent external review done every third year. A similar process is applied by the USA North Pacific Fishery Management Council and was common for stocks managed by the Australian federal government in the past. However, in the majority of jurisdictions, the expectation of the review is to decide whether the proposed assessment satisfies the requirements for 'best available' science at the time.

Failing the peer-review process does not necessarily mean that no management advice can be given because some of the advice can be based on a 'rejected' assessment. For example, a stock assessment that is highly uncertain can still be used for management purposes, particularly when the management decision accounts for the extent of uncertainty when selecting how precautionary the management actions should be given the best estimates from the assessment. Moreover, some jurisdictions work with the analysts post-review to enable a 'rejected' assessment to be accepted for use in management. For example, within ICES, if an assessment is considered problematic during the usual annual assessment process and usually about one per regional assessment working group is 'problematic' each year, rather than rejecting the assessment, an 'inter-benchmark' process may be undertaken. The rationale and criteria for an inter-benchmark are not very well documented and are based on the perceptions of the assessment experts. The inter-benchmark generally involves investigating one issue, usually the performance of the model and assumptions, but sometimes the rejection or inclusion of more data sets. As with benchmark assessments, the approach and conclusions of an inter-benchmark assessment must be externally reviewed but the peer-review process is abbreviated compared to the standard process of reviewing benchmark assessments (Mark Dickey-Collas, ICES, pers comm).

A similar approach has been taken for groundfish off the USA west coast where assessments that are rejected for use in management during a first peer-review are sent to the 'mop-up' peer-review panel. As is the case with ICES, the 'mop-up' panel review is focused on a relatively small number of issues with no reviewers independent of PFMC assessment and review process. For example, there were several issues of concern with the assessment of Pacific ocean perch (*Sebastes alutus*) in 2017 (Pacific Fishery Management Council (PFMC, 2017b) for which estimated stock status was very different compared to the previous full assessment in 2011 (45–96 % vs ~20 % relative stock status). The main driver for the wide range of potential relative stock status values was the value assumed for the steepness of the stock-recruitment relationship. There was no information in the data for Pacific ocean perch about this parameter (according to likelihood profiles) and the estimated spawning output from the assessment model was highly sensitive to the assumed steepness value. In the end, the peer-review body (the SSC of the Pacific Fishery Management Council) balanced the concerns and fixed steepness in the base model at the value resulting in the mean end-year spawning output when profiling across steepness values ranging from 0.25–0.95, assuming these values are all equally likely (Pacific Fishery Management Council (PFMC, 2017c). This result contrasts with the Pacific sanddab example above where the assessment was rejected by the same review body because the likelihood profile led to an unreasonable result (for catchability), and therefore there was no set of reasonable values over which to integrate.

5.1. Thoughts on recommended practices

There will always be situations in which a proposed assessment fails to satisfy the standards for a scientifically defensible analysis, either because the data are insufficient to estimate the key parameters or because the model is incorrectly specified, or knowledge about the biology or ecology of the fish stock or marine system has changed and new understanding has not been included in the assumptions or

[2] Such as the USA Center for Independent Experts (CIE), which was established in 1998 as a national peer-review program. The CIE was operated by the University of Miami until 2007, at which point Northern Taiga Ventures, Inc. was contracted by the National Marine Fisheries Service to administer this program, which ensures that the appropriate expertise is identified among reviewers, that the reviews are appropriate and of high quality, and to make sure the reviewers follow conflict of interest guidelines. The CIE has completed hundreds of reviews, with many reviewers based outside the USA.

structure of the assessment model. It is unlikely that there will ever be a set of rules that determine whether a stock assessment, particularly a complex stock assessment based on many assumptions and data sets, should be accepted or rejected that can be applied automatically (and it is not clear that such rules would even be desirable). However, this paper provides some ideas regarding how the current process can be made less subjective:

- Increased use of external peer-reviewers where this is not already the case, to enhance consistency in terms of which diagnostics are reported in assessments across agencies and regions. In cases where multiple reviews are conducted within a short period of time, including at least one reviewer in all review panels would enhance consistency.
- More explicit specification of what constitutes unacceptable model outcomes / behaviour in Terms of Reference for stock assessments.
- More holistic evaluation across model diagnostics rather than basing acceptance / rejection decisions on a single diagnostic.
- Encouragement to peer reviewers (e.g., through peer review terms of reference) to evaluate the relative utility of the components of a stock assessment rather than perform an "all-or-nothing" evaluation as is typically the case for scientific journal article reviews. Because of the operational nature of stock assessments, a review should be conducted to determine what aspects are appropriate to inform management, as opposed to a simple reject/accept approach.
- Identification of metrics to summarize diagnostic plots. The thresholds defined by Hurtado et al. (2015) or Brooks and Legault (2016) for retrospective patterns is one example of such metrics. Similarly, defining a standard set of sensitivity tests to allow the extent to which an assessment is sensitive would help assess the relative sensitivity / stability of assessments, while tests already exist to determine whether the fit to a data set is mis-specified. 'Failing' such metrics (and thresholds) would not necessarily lead to rejection, but they would provide a common basis for evaluation, and increase the consistency of the review process. Carvalho et al. (2017) evaluated several commonly used diagnostic tools and methods. That analysis was based on one assessment method (Stock Synthesis) and a few violations of assumptions. Continuation and extension of this work to more assessment methods and a greater number of problems will help with the development of thresholds and best practices for individual diagnostics.
- Development of standards for how model output is produced and saved so that common software tools for running and summarizing diagnostics can be applied; at present there are usually far more diagnostics reported for assessments based on software packages (such as Stock Synthesis, CASAL (Bull et al., 2005; Doonan et al., 2016), and MULTIFAN (Fournier et al., 1998) than for assessments based on bespoke methods, likely due to the 'cost' of developing the software for diagnostics when that software will only be used by one, or a small number, of analysts.
- Documenting recommendations from a review panel to improve a stock assessment and responses to these from the lead scientist in subsequent assessments would increase the transparency of the review process and evolution of the stock assessment modelling efforts.
- Creation of easily accessible and machine-readable stock assessment results would facilitate transfer of knowledge among distantly located scientists and speed the formulation of best practices

Many jurisdictions have fall-back positions when assessments are rejected, and where this is not the case, development and testing of such positions should be considered a top priority. The fall-back position is ideally an opportunity for further work in collaboration with the peer-reviewers to identify an assessment that does not have the problems associated with the proposed and rejected assessment. However, that is not always feasible. Any fall-back approaches would have levels of uncertainty greater than that associated with the originally proposed assessment, and some allowance for this seems warranted when managers decide the level of precaution to apply when setting management regulations such as catch limits, particularly if the fall-back position is simply to 'roll over' the current management arrangements.

Acknowledgements

We thank our colleagues Kevin Hill and Paul Crone (NOAA, SWFSC); David Sampson (OSU); Hiroshi Okamura (Fisheries Research Agency, Japan); and Kyle Shertzer (NOAA; SEFSC) for their input. Malcolm Haddon (CSIRO Retired), the editor, and two anonymous reviewers are thanked for their comments on an earlier version of this paper. This publication is [partially] funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement NA15OAR4320063, Contribution No 2019-1041.

References

Anon, 2007. AFMA Management Position Paper: Recommendations for Total Allowable Catches for SESSF Quota Species for the 2008–09 Fishing Year. Australian Fisheries Management Authority, 15 Lancaster Place, Majura Park, ACT 2609, Australia.

Anon, 2012. Minutes of the Bering Sea Aleutian Islands Groundfish Plan Team. https://www.npfmc.org/wp-content/PDFdocuments/membership/PlanTeam/Groundfish/JGPTminutesNov12.pdf.

Anon, 2013. Southern and Eastern Scalefish and Shark Fishery Slope Resource Assessment Group (SLOPERAG). Australian Fisheries Management Authority, Australia Box 7501 Canberra Business Center, ACT 2610.

Anon, 2019. Anchovy (Engraulis encrasicolus) in division 9.a (Atlantic Iberian waters). ICES Advice on Fishing Opportunities, Catch, and Effort Bay of Biscay and the Iberian Coast Ecoregion. . http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2019/2019/ane.27.9a_2018.pdf.

Berg, C.W., Nielsen, A., 2016. Accounting for correlated observations in an age-based state-space stock assessment model. ICES J. Mar. Sci. 73, 1788–1797.

Brodziak, J., Yau, A., O'Malley, J., Andrews, A., Humphreys, R., DeMartini, E., Pan, M., Parke, M., Fletcher, E., 2014. Stock assessment update for the main Hawaiian Islands deep 7 bottomfish complex through 2013 with projected annual catch limits through 2016. NOAA Tech. Memo. U.S. Dep. Commer. NOAA-TM-NMFS-PIFSC-42, 61 p.

Brooks, E.N., Legault, C.M., 2016. Retrospective forecasting—evaluating performance of stock projections for New England groundfish stocks. Can. J. Fish. Aquat. Sci. 73, 935–950.

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., Smith, M.H., Bian, R., Fu, D., 2005. CASAL (C++ Algorithmic Stock Assessment Laboratory): CASAL User manualv2.30-2012/03/21. NIWA Technical Report 127. .

Carvalho, F., Punt, A.E., Chang, Y.-J., Maunder, M.N., Piner, K.R., 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. 192, 28–40.

Cope, J., 2015. The 2015 Stock Assessment of Arrowtooth Flounder (*Atheresthes Stomias*) in California, Oregon, and Washington Waters. https://www.pcouncil.org/wp-content/uploads/2015/05/D8_Att5_ATF_2015_data-mod_FULL-E-Only_JUN2015BB.pdf.

Cope, J., Dick, E.J., MacCall, A., Monk, M., Soper, B., Wetzel, C., 2013. Data-moderate Stock Assessments for Brown, China, Copper, Sharpchin, Stripetail, and Yellowtail Rockfishes and English and Rex Soles in 2013. http://www.pcouncil.org/wp-content/uploads/Data-Moderate_Assessments_2013_FINAL_160116.pdf.

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

Deroba, J.J., Butterworth, D.S., Methot, R.D.Jr., De Oliveira, J.A.A., Fernandez, C., Nielsen, A., Cadrin, S.X., Dickey-Collas, M., Legault, C.M., Ianelli, J., Valero, J.L., Needle, C.L., O'Malley, J.M., Chang, Y.-J., Thompson, G.G., Canales, C., Swain, D.P., Miller, D.C.M., Hintzen, N.T., Bertignac, M., Ibaibarriaga, L., Silva, A., Murta, A., Kell, L.T., de Moor, C.L., Parma, A.M., Dichmont, C.M., Restrepo, V.R., Ye, Y., Jardim, E., Spencer, P.D., Hanselman, D.H., Blaylock, J., Mood, M., Hulson, P.-J.F., 2015. Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods. ICES J. Mar. Sci. 72, 19–30.

Doonan, I., Lange, K., Dunn, A., Rasmussen, S., Marsh, C., 2016. Casal2: new Zealand's integrated population modelling tool. Fish. Res. 183, 408–505.

Fournier, D.A., Hampton, J., Sibert, J.R., 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, Thunnus alalunga. Can. J. Fish. Aquat. Sci. 55, 2105–2116.

Francis, R.I.C.C., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124–1138.

Hillary, R.M., Preece, A.L., Davies, C.R., Kurota, H., Sakai, O., Itoh, T., Parma, A.M., Butterworth, D.S., Ianelli, J., Branch, T.A., 2016. A scientific alternative to moratoria for rebuilding depleted international tuna stocks. Fish. Res. 17, 469–482.

Hurtado Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J., Johnson, K.F., Licandeo, R.R., McGillaird, C.R., Monnahan, C.C., Muradian, M.L., Ono, K., Vert-pre, K.A., Whitten, A.R., Punt, A.E., 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES J. Mar. Sci. 72, 99–110.

ICES, 2012. ICES Implementation of Advice for Data-limited Stocks in 2012 in Its 2012 Advice. ICES CM 2012/ACOM 68. 42 pp. .

ICES, 2013. Report of the Workshop on the Development of Quantitative Assessment Methodologies Based on LIFE-history Traits, Exploitation Characteristics, and Other Key Parameters for Data-limited Stocks (WKLIFE III), 28 October–1 November 2013. ICES CM 2013/ACOM:35. 98 pp. Copenhagen, Denmark. http://ices.dk/sites/pub/Publication%20Reports/Expert%20Group%20Report/acom/2013/WKLIFE3/Report%20WKILFE%20III.pdf.

ICES, 2017a. Herring (Clupea Harengus) in Subareas 1, 2, and 5, and in Divisions 4.a and 14.a, Norwegian Spring-spawning Herring (the Northeast Atlantic and the Arctic Ocean). . http://ices.dk/sites/pub/Publication%20Reports/Advice/2017/2017/her.27.1-24a514a.pdf.

ICES, 2017b. Report of the Benchmark Workshop on Pelagic Stocks, 6–10 February 2017. ICES CM 2017/ACOM:35. Lisbon, Portugal. http://www.ices.dk/sites/pub/Publication%20Reports/Expert%20Group%20Report/acom/2017/WKPELA/01%20WKPELA%202017%20Report.pdf.

ICES, 2018a. Report of the InterBenchmark Protocol for Turbot in the North Sea 2018 (IBP-Turbot). ICES IBPTurbot Report 2018 30–31 July, 2018. Ijmuiden, the Netherlands. ICES CM 2018/ACOM:50. . http://ices.dk/sites/pub/Publication%20Reports/Expert%20Group%20Report/acom/2018/IBPTurbot/IBPTurbot_Report_2018.pdf.

ICES, 2018b. ICES Advice on Fishing Opportunities, Catch, and Effort Faroes, Celtic Seas, and Oceanic Northeast Atlantic Ecoregions. . http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2018/2018/rng.27.5b6712b.pdf.

ICES, 2018c. Introduction to Advice. . https://www.ices.dk/sites/pub/Publication%20Reports/Advice/2018/2018/Introduction_to_advice_2018.pdf).

ICES, 2019a. Workshop on North Sea Stocks Management Strategy Evaluation (WKNSMSE). ICES Scientific Reports. 1:12. 378 pp. https://doi.org/10.17895/ices.pub.5090.

ICES, 2019b. Interbenchmark Workshop on the Assessment of Northeast Atlantic Mackerel (IBPNEAMac). ICES Scientific Reports. 1:5. 71 pp. https://doi.org/10.17895/ices.pub.4985.

ICES, 2019c. Inter-benchmark Protocol for Sole in the Eastern English Channel (IBPsol7d). ICES Scientific Reports. 1:75. 88 pp. https://doi.org/10.17895/ices.pub.5631.

ICES, 2019d. Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak (WGNSSK). ICES Scientific Reports. 1:7. 1271 pp. https://doi.org/10.17895/ices.pub.5402.

ICES, 2019e. Report of the Interbenchmark Protocol on North Sea Saithe (IBPNSsaithe). ICES Scientific Reports. VOL 1:ISS 1. 65 pp. https://doi.org/10.17895/ices.pub.4890.

ICES, 2019f. 22 Deep Pelagic *Sebastes mentella*. . https://www.ices.dk/sites/pub/Publication%20Reports/Expert%20Group%20Report/Fisheries%20Resources%20Steering%20Group/2019/NWWG/24%20NWWG%20Report%202019_Sec%2022_Deep%20pelagic%20Sebastes%20mentella.pdf.

International Whaling Commission (IWC), 2012. The revised management procedure (RMP) for Baleen Whales. J. Cetacean Res. Manage 13 (Suppl), 485–494.

Johnston, S.J., Butterworth, D.S., 2005. Evolution of operational management procedures for the South African west coast rock lobster (*Jasus lalandii*) fishery. NZ J. Mar. Freshw. Res. 39, 687–702.

Little, L.R., Wayte, S.E., Tuck, G.N., Smith, A.D.M., Klaer, N., Haddon, M., Punt, A.E., Thomson, R., Day, J., Fuller, M., 2011. Development and evaluation of a cpue-based harvest control rule for the southern and eastern scalefish and shark fishery of Australia. ICES J. Mar. Sci. 68, 1699–1705.

Legault, C.M., Chair, 2009. Report of the Retrospective Working Group, January 14–16, 2008. Northeast Fish Sci Cent Ref Doc. 09–01; 30 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026. US Dept. Commer., Woods Hole, Massachusetts. http://www.nefsc.noaa.gov/nefsc/publications.

Lynch, P.D., Methot, R.D., Link, J.S. (Eds.), 2018. Implementing a Next Generation Stock Assessment Enterprise. An Update to the NOAA Fisheries Stock Assessment Improvement Plan. U.S. Dep. Commer.. https://doi.org/10.7755/TMSPO.183. NOAA Tech. Memo. NMFS-F/SPO-183, 127 p.

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

Maunder, M.N., Crone, P.R., Valero, J.L., Semmens, B.X., 2014. Selectivity: theory, estimation, and application in fishery stock assessment models. Fish. Res. 158, 1–4.

Maunder, M.N., Crone, P.R., Punt, A.E., Valero, J.L., Semmens, B.X., 2016. Growth: theory, estimation, and application in fishery stock assessment models. Fish. Res. 180, 1–3.

Maunder, M.N., Crone, P.R., Punt, A.E., Valero, J.L., Semmens, B.X., 2017. Data conflict and weighting, likelihood functions and process error. Fish. Res. 192, 1–4.

McAllister, M.K., Ianelli, J.N., 1997. Bayesian stock assessment using catch-age data and the sampling/importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54, 284–300.

Methot, R.D., 2019. NOAA Fisheries Framework for Determining That Stock Status Determinations and Catch Specifications Are Based on Best Scientific Information. available (https://www.fisheries.noaa.gov/webdam/download/90600446). .

Methot, R.D., Taylor, I.G., 2011. Adjusting for bias due to variability of estimated recruitments in fishery assessment models. Can. J. Fish. Aquat. Sci. 68, 1744–1760.

Methot, R.D., Wetzel, C.R., 2013. Stock Synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99.

Mid-Atlantic Fishery Management Council (MAFMC), 2010. Report of May 2010 Meeting of the MAFMC Scientific and Statistical Committee. . https://static1.squarespace.com/static/511cdc7fe4b00307a2628ac6/t/5165e423e4b035d7482d9bfe/1365632035521/SSC_Report_11-12_May_+2010.pdf.

Mid-Atlantic Fishery Management Council (MAFMC), 2017. Report of the May 2017 SSC Meeting. . https://static1.squarespace.com/static/511cdc7fe4b00307a2628ac6/t/5938001659cc686cc1578f97/1496842263226/01_May+2017+SSC+Report.pdf.

MAFMC, 2012. Mid-Atlantic Fishery Management Council (MAFMC). . https://static1.squarespace.com/static/511cdc7fe4b00307a2628ac6/t/5165e509e4b00ae130ccb6cb/1365632265884/SSC_Report_30_Jul_2012.pdf.

Mohn, R., 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56, 473–488.

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. 158, 96–101.

Northeast Fisheries Science Center (NEFSC), 2011. 52nd Northeast Regional Stock Assessment Workshop (52nd SAW): Assessment Report. . https://www.nefsc.noaa.gov/publications/crd/crd1117/crd1117.pdf.

Northeast Fisheries Science Center (NEFSC), 2015. Operational Assessment of 20 Northeast Groundfish Stocks, Updated Through 2014. Northeast Fish Sci Cent Ref Doc. 15–24. US Dept Commer. https://www.nefsc.noaa.gov/publications/crd/crd1524/.

Northeast Fisheries Science Center (NEFSC), 2017. Overfishing Levels (OFLs) and Acceptable Biological Catch (ABC) Recommendations for Groundfish Stocks for Fishing Years 2018-2020. http://s3.amazonaws.com/nefmc.org/2017-Fall-NRCC-Briefing-Binder_171114_100136.pdf.

North Pacific Fishery Management Council (NPFMC), 2016. A Guide to the Preparation of Alaska Groundfish SAFE Reports. . https://www.npfmc.org/wp-content/PDFdocuments/membership/PlanTeam/Groundfish/SAFE_guidelines_2016_July_25.pdf.

Pacific Fishery Management Council (PFMC), 2005a. Kelp Greenling STAR Panel Report. . https://www.pcouncil.org/wp-content/uploads/Kelp_Greenling_STAR.pdf.

Pacific Fishery Management Council (PFMC), 2005b. Vermilion Rockfish STAR Panel Report. . https://www.pcouncil.org/wp-content/uploads/Vermilion_STAR_panel_report.pdf.

Pacific Fishery Management Council (PFMC), 2007. Pacific Mackerel Stock Assessment (STAR) Panel Meeting. https://www.pcouncil.org/wp-content/uploads/APP2_Mackerel_Assessment_2007-08.pdf.

Pacific Fishery Management Council (PFMC), 2013. Pacific Sanddab Stock Assessment Review (STAR) Panel Report. . http://www.pcouncil.org/wp-content/uploads/Sanddab_2013_STAR.pdf.

Pacific Fishery Management Council (PFMC), 2015a. Summary Minutes Scientific and Statistical Committee. ftp://ftp.pcouncil.org/pub/SSC_minutes/2015_June_SSC_Minutes.pdf.

Pacific Fishery Management Council (PFMC), 2015b. Scientific and Statistical Committee Report on Pacific Mackerel Assessment and Management Measures. . http://www.pcouncil.org/wp-content/uploads/2015/06/G2b_Sup_SSC_Rpt_JUN2015BB.pdf.

Pacific Fishery Management Council (PFMC), 2015c. Pacific Mackerel Stock Assessment (STAR) Panel Meeting Report. . http://www.pcouncil.org/wp-content/uploads/2015/05/G2_Att1_STAR_Rpt_JUN2015BBpdf.pdf.

Pacific Fishery Management Council (PFMC), 2017a. Yellowtail Rockfish Stock Assessment (STAR) Panel Meeting Report. . https://www.pcouncil.org/wp-content/uploads/2018/01/YT_STAR_Panel_Report_Final_July10-14_2017.pdf.

Pacific Fishery Management Council (PFMC), 2017b. Pacific Ocean Perch Stock Assessment Review (STAR) Panel Report. . https://www.pcouncil.org/wp-content/uploads/2018/01/POP_STAR_Panel_Report-FINAL_June2017.pdf.

Pacific Fishery Management Council (PFMC), 2017c. Minutes Scientific and Statistical Committee. file:///C:/Users/pun009/Downloads/2017_Nov_SSC_Minutes.pdf.

Pacific Fishery Management Council (PFMC), 2019. Terms of Reference for the Groundfish and Coast Pelagic Species Stock Assessment Review Process for 2019–2010. https://www.pcouncil.org/wp-content/uploads/2019/04/Stock_Assessment_ToR_REVISED_2019-20_APR2019_Final-2.pdf.

Plummer, M., Best, N., Cowles, K., Vines, K., 2006. CODA: convergence diagnosis and output analysis for MCMC. R News 6, 7–11.

Punt, A.E., 2017. Some insights into data weighting in integrated stock assessments. Fish. Res. 192, 52–65.

Punt, A.E., Deng, R.A., Siddeek, M.S.M., Buckworth, R.C., Vanek, V., 2017. Data weighting for tagging data in integrated size-structured models. Fish. Res. 192, 94–102.

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

Rademeyer, R.A., Plagányi, É.E., Butterworth, D.S., 2007. Tips and tricks in designing management procedures. ICES J. Mar. Sci. 64, 618–625.

Rademeyer, R.A., Butterworth, D.S., Plagányi, É.E., 2008. A history of recent bases for management and the development of a species-combined Operational Management Procedure for the South African hake resource. Afr. J. Mar. Sci. 30, 291–310.

Sharma, R., Maunder, M.N., Babcock, E., Punt, A.E., 2019. Recruitment: theory, estimation, and application in fishery stock assessment models. Fish. Res. 217, 1–4.

Southeast Data, Assessment, and Review (SEDAR), 2006. Large Coast Shark Complex, Blacktip and Sandbar Shark. https://sedarweb.org/docs/sar/Final_LCS_SAR.pdf.

Southeast Data, Assessment, and Review (SEDAR), 2007. Caribbean Queen Conch. . http://sedarweb.org/docs/sar/S14SAR3%20Queen%20Conch%20Report.pdf.

Southeast Data, Assessment, and Review (SEDAR), 2011a. U.S. Caribbean Queen Snapper. https://sedarweb.org/docs/sar/S26_Queen_SAR.pdf.

Southeast Data, Assessment, and Review (SEDAR), 2011b. HMS Gulf of Mexico Blacknose Shark. http://sedarweb.org/docs/sar/GoM_Blacknose_SAR.pdf.

Southeast Data, Assessment, and Review (SEDAR), 2014. SEDAR 37 The 2013 Stock Assessment Report for Hogfish in the South Atlantic and Gulf of Mexico. . http://sedarweb.org/docs/sar/SEDAR37_Hogfish_SAR.pdf.

Southeast Data, Assessment, and Review (SEDAR), 2016a. SEDAR 47 Stock Assessment Report. Southeastern U.S. Goliath Grouper. http://sedarweb.org/docs/sar/S47_Final_SAR.pdf.

Southeast Data, Assessment, and Review (SEDAR), 2016b. SEDAR 41 Stock Assessment Report. South Atlantic Gray Triggerfish. . http://sedarweb.org/sedar-41.

SUBPESCA/CCT-CD/2, 2013. Acta de Sesión N°2. Comité Científico Técnico de Crustaceos Demersales: 14 p. http://www.subpesca.cl/portal/616/articles-82134_documento.pdf.

SUBPESCA/CCT-RDZSA/4, 2016. Acta de Sesión N°4 – 2016. Comité Científico Técnico Recursos Demersales Zona Sur Austral: 11 p. http://www.subpesca.cl/portal/616/articles-95118_documento.pdf.

SUBPESCA/CCT-PP/6, 2017a. Acta de Reunión N°6–2017. Comité Científico Técnico Pesquerías de Pelágicos Pequeños: 24 p. http://www.subpesca.cl/portal/616/articles-98715_documento.pdf.

SUBPESCA/CCT-RDAP/4, 2017b. Acta de Sesión N°4. Comité Científico Técnico Recursos Demersales de Aguas Profundas: 15 p. http://www.subpesca.cl/portal/616/articles-98683_documento.pdf.

Sullivan, P.J., Haist, V., Klaer, N., Nilesen, A., 2016. Summary Report of the 62th Northeast Regional Stock Assessment Review Committee (SARC 62). . https://www.nefsc.noaa.gov/saw/saw62/sarc62_panel_summary_report.pdf.

Szuwalski, C.S., Ianelli, J.N., Punt, A.E., 2018. Reducing retrospective patterns in stock assessment and impacts on management performance. ICES J. Mar. Sci. 75, 596–609.

Taylor, I.G., Stewart, I.J., Hicks, A.C., Garrison, T.M., Punt, A.E., Wallace, J.R., Wetzel, C.R., Thorson, J.T., Takeuchi, Y., Ono, K., Monnahan, C.C., Stawitz, C.C., A'mar, Z.T., Whitten, A.R., Johnson, K.F., Emmet, R.L., Anderson, S.C., Lambert, G.I., Stachura, M.M., Cooper, A.B., Stephens, A., Klaer, N.L., McGilliard, C.R., Iwasaki, W.M., Doering, K., Havron, A.M., 2019. R4ss. https://github.com/r4ss/.

Thorson, J.T., Cope, J.M., 2017. Uniform, uninformed or misinformed?: the lingering challenge of minimally informative priors in data-limited Bayesian stock assessments. Fish. Res. 194, 164–172.

Thorson, J.T., Johnson, K.F., Methot, R.D., Taylor, I.G., 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fish. Res. 192, 84–93.

Transboundary Resources Assessment Committee (TRAC), 2010. Atlantic Mackerel in the Northwest Atlantic. Status Report 2010/01. . http://www.bio.gc.ca/info/intercol/trac-cert/documents/reports/TSR_2010_01_E.pdf.

Transboundary Resources Assessment Committee (TRAC), 2014. In: Proceedings of the Transboundary Resources Assessment Committee for Georges Bank Yellowtail Flounder Diagnostic and Empirical Approach Benchmark Report of Meeting Held 14-18 April 2014. TRAC Proceedings 2014/01. . https://www.nefsc.noaa.gov/saw/trac/TSR_2014_03_E_revised.pdf.

Wang, S.P., Maunder, M.N., Piner, K.R., Aires-da-Silva, A., Lee, H.H., 2014. Evaluation of virgin recruitment profiling as a diagnostic for selectivity curve structure in integrated stock assessment models. Fish. Res. 158, 158–164.