THE THIRD WORKSHOP ON GUIDELINES FOR MANAGEMENT STRATEGY EVALUATIONS (WKGMSE3)

VOLUME 2 | ISSUE 116

ICES SCIENTIFIC REPORTS

RAPPORTS
SCIENTIFIQUES DU CIEM

ICES INTERNATIONAL COUNCIL FOR THE EXPLORATION OF THE SEA
CIEM CONSEIL INTERNATIONAL POUR L'EXPLORATION DE LA MER

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H.C. Andersens Boulevard 44-46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

ISSN number: 2618-1371 I

ICES Scientific Reports

Volume 2 | Issue 116

THE THIRD WORKSHOP ON GUIDELINES FOR MANAGEMENT STRATEGY EVALUATIONS (WKGMSE3)

Recommended format for purpose of citation:

ICES. 2020. The third Workshop on Guidelines for Management Strategy Evaluations (WKGMSE3). ICES Scientific Reports. 2:116. 112 pp. http://doi.org/10.17895/ices.pub.7627

Editors

José De Oliveira

Authors[1]

Valerio Bartolino • Benoit Berges • Höskuldur Björnsson • Mollie Elisabeth Brooks • Doug Butterworth • Andrew Campbell • Massimiliano Cardinale • Tom Carruthers • Santiago Cerviño • Galina Chernega • Harriet Cole • Carryn de Moor • José De Oliveira • Jonathan Deroba • David Die • Nicholas Duprey • Gavin Fay • Simon Fischer • Dorleta Garcia • Kyle Gillespie • Daisuke Goto • Michaël Gras • Michelle Greenlaw • Stefanie Haase • Alex Hanke • Einar Hjörleifsson • Daniel Howell • Laurence Kell • Alexander Kempf • Toshi Kitakado • Christoph Konrad • Allen R. Kronlund • Gwladys Lambert • Polina Levontin • Mackenzie Mazur • Tanja Miethe • David Miller • Iago Mosqueira • Virginia Noble • Alessandro Orio • Ana Parma • Martin Pastoors • Alfonso Perez-Rodriguez • Sara Pipernos • Maris Plikšs • Claus Reedtz Sparrevohn • Norbert Rohlf • Andrea Ross-Gillespie • Sonia Sanchez • Rishi Sharma • Margaret Siple • Laura Solinger • Henrik Sparholt • Michael Spence • Marc Taylor • Robert Thorpe • Ash Wilson • Henning Winker

[1] 3 February 2021: Marc Taylor added to list of authors

ICES International Council for the Exploration of the Sea
CIEM Conseil International pour l'Exploration de la Mer

Contents

i Executive summary

This workshop is the third in a series of workshops on guidelines for developing Management Strategy Evaluations (MSEs) within ICES, and was intended to explore some of the issues that arose out of workshops that actually developed MSEs for a range of ICES stocks since the second MSE guidelines workshop was held in early 2019. It is intended that results and conclusions herein be used to update the guidelines for conducting MSEs within ICES further.

In addition to reviewing existing work, new analyses were prepared especially for this workshop under TORs c (risk and uncertainty), d (more efficient tuning in MSEs) and e (shortcut vs. full MSEs). TOR a covered reference points, and this workshop proposed a framework for calculating reference points from simulation models used in MSEs when an MSE is conducted for a stock. TOR b considered how to handle alternative operating models when reporting results from MSEs, and the workshop made two proposals for how this could be done based on approaches used in other fora where MSEs are developed routinely. Under TOR c, some ideas were put forward based on comparing risk to an unfished scenario such that the inclusion of additional (realistic) uncertainty was not unduly penalised, and could be used to deal with the situation (e.g. for short-lived species) where risk in an unfished scenario was already close to or greater than 5%. There were several proposals for increasing the efficiency of tuning in MSEs under TOR d, particularly when computation time was a factor (e.g. when full MSEs have relatively complex analytical assessments imbedded in the management procedure being evaluated). These were based both on statistical techniques, and a method for identifying the Pareto-optimal solutions that focuses on trade-offs among competing objectives. TOR e (shortcut vs. full MSEs) was the most contentious, but nevertheless useful because it made out the way harvest control rule evaluations (and more recently full MSEs) conducted in ICES compared to MSE approaches elsewhere more clearly evident, and highlighted alternative interpretations of the shortcut method. There was discussion of the pros and cons of the full and shortcut methods, including alternative views of how some of the shortcomings in each could be addressed.

Several recommendations are made, including a dedicated workshop on reference points, consideration of more flexible MSE approaches (such as developing empirical management procedures), improving communication between scientists, managers and stakeholders when developing MSEs, and a further MSE guidelines workshop that includes consideration of when performance of the HCR/management strategy is not as intended, under both the full and shortcut MSE approaches, as indicated by the results from the simulations.

ii Expert group information

Expert group name	The third Workshop on Guidelines for Management Strategy Evaluations (WKGMSE3)
Expert group cycle	Annual
Year cycle started	2020
Reporting year in cycle	1/1
Chair	José De Oliveira, UK
Meeting venue and dates	26–30 October 2020, online meeting, (60 participants)

1 Introduction

This third workshop on Management Strategy Evaluation (MSE), WKGMSE3, was intended to explore in greater detail issues that were uncovered during the work of WKNSMSE (ICES, 2019b) that could not be further explored at the time, given workload and time constraints, so that these could be used to update the guidelines developed during WKGMSE2 (ICES, 2019a). It picks up on some of the recommendations from WKNSMSE, including how to extract reference points from MSEs, how to deal with alternative operating models, how to define risk, and explore methods for finding optimised management strategies. It also compares shortcut and full MSE approaches. The Terms of Reference (TOR) for this workshop, including scientific justification, are provided in Annex 2.

The meeting was held by WebEx during 26–30 October 2020, and was well-attended (over 50 participants, though not all full time, spread across several time-zones and both hemispheres; Annex 1). The meeting was organised into 5 sessions, each connected to a TOR, but with additional sessions summarising work from other relevant recent ICES workshops (WKMSEMAC, WKREBUILD and WKRPChange), and an open session for topics not covered by the TORs (although there was not much time for discussion on these topics). This report is also organised by TOR (Sections 2–6) with each section accompanied by the summaries of the presentations and discussions under each respective TOR; conclusions agreed within the plenary for the TOR are provided at the end of each section. Section 7 provides some reflections on the use of MSE in the Northeast Atlantic, Section 8 describes recent ICES workshops with relevance to the TORs of this workshop and a brief summary of discussion points, and Section 9 provides a brief description of the presentations made during the open session. Section 10 lists recommendations. Annexes 3–7 provide more detail of the analyses prepared for the workshop, while Annexes 8 and 9 contain background information for Sections 6.5 and 7.1, respectively.

It was clear during discussions of several of the TORs (in particular, TOR e comparing the full and shortcut approaches; Section 6) that there was a need to clarify terminology. In order to help with this, an extract of relevant terms is provided in Table 1.0.1 below, taken from the glossary that was developed by a joint tuna RFMO Management Strategy Evaluation Working Group that met for the second time in Seattle, 13–15 June 2018 (https://www.tuna-org.org/mse.htm). This glossary was developed to encourage a consistent use of terms associated with harvest strategies, management procedures and MSE, and was developed from a range of sources and a range of MSE practitioners with broad experience in the field (Anon., 2018).

Table 1.0.1. Selected terms and definitions extracted from the "Glossary of terms for harvest strategies, management procedures and management strategy evaluation" (https://www.tuna-org.org/Documents/MSEGlossary_tRFMO_MSEWG2018.pdf; also Anon., 2018).

Term	Definition	Abbreviation
Harvest Strategy	Some combination of monitoring, assessment, harvest control rule and management action designed to meet the stated objectives of a fishery. Sometimes referred to as a Management Strategy (see below). A fully specified harvest strategy that has been simulation tested for performance and adequate robustness to uncertainties is often referred to as a Management Procedure.	HS
Management Procedure	A management procedure has the same components as a harvest strategy. The distinction is that each component of a Management Procedure is ***formally specified***, and ***the combination*** of monitoring data, analysis method, harvest control rule and management measure has been simulation tested to demonstrate adequately robust performance in the face of plausible uncertainties about stock and fishery dynamics.	MP
Management Strategy	Synonymous with harvest strategy. (But note that this is also used with a broader meaning in a range of other contexts.)	
Management Strategy Evaluation	A process whereby the performances of alternative harvest strategies are tested and compared using stochastic simulations of stock and fishery dynamics against a set of performance statistics developed to quantify the attainment of management objectives.	MSE
Operating Model(s)	A mathematical–statistical model (usually models) used to describe the fishery dynamics in simulation trials, including the specifications for generating simulated resource monitoring data when projecting forward in time. Multiple models will usually be considered to reflect the uncertainties about the dynamics of the resource and fishery.	OM(s)

It is worth noting some further terminology standards used in this document:

- We have dispensed with the terms "full-feedback" and "closed-loop" to describe application of a full MSE, which includes an analytical assessment model in the MP. This is because the terms "feedback" and "closed-loop" are terms associated with MSE in general, and not just a full MSE (i.e. they also apply to MSEs that use a shortcut approach or empirical MPs).
- We have adopted the term "Prob3" to describe the risk definition that ICES uses to evaluate whether management strategies are precautionary (see Section 4.2 of the WKGMSE2 report; ICES, 2019a); this risk measure is alternatively also called "risk3" and described as the maximum probability that SSB is below B_{lim}, where the maximum is taken over a pre-specified number of years.
- We have distinguished between "risk level" or simply "risk" (e.g. Prob3 = 0.05), and "risk threshold" (e.g. B_{lim}, as per definition of Prob3).
- We have generally used the term management strategy as per Table 1.0.1, but "management plan" has been used in the past to essentially mean the same thing in the ICES context.

2 TOR a: Reference points

Develop guidelines for when and how reference points should be extracted from an MSE when one is conducted.

TOR a deals with extracting reference points from MSEs when they are conducted, including the time-frame to be used; it came about because of discrepancies between reference points from the standard ICES approach (EqSim), and the MSEs conducted as part of WKNSMSE (ICES, 2019b).

2.1 ICES reference point framework

Presentation by David Miller

This presentation summarised the current ICES reference point framework, a detailed description of which can be found at ICES (2017a), with a more general description found in ICES (2019c).

Summary of discussion

It was noted that there may be some circularity in the way F_{MSY} is calculated: MSY $B_{trigger}$ follows the calculation of F_{MSY}, but is then used to calculate $F_{P.05}$, which may cap the F_{MSY} calculation. There was some discussion about absolute vs. relative reference points, the latter in the sense that reference points are recalculated with every update, instead of being kept constant. ICES does not typically update reference points, unless there is a benchmark or perhaps inter-benchmark assessment. Current MSE guidelines (ICES, 2019a) suggest recalculating reference points for each OM so that in the MSE context, reference points are internally consistent with the OM. A clear distinction should be made between reference points that are used for management advice (i.e. those that appear in ICES advice sheets) and reference points that are inherent properties of the OM.

There was some discussion about the different frameworks used for estimating reference points. The ICES guidelines for calculating reference points call for stochastic simulations of the stock and fishery to determine the impact of different rates of fishing mortality on the equilibrium (stationarity in the long term) catch and biomass of the stock. ICES typically uses EqSim (ICES, 2014a; 2014b), but an MSE framework, which also includes a stochastic simulation model, could also be used (ICES, 2019b; 2020a; Section 2.3). An MSE simulation model is, arguably, a more comprehensive basis for calculating reference points because it explicitly and more directly accounts for different sources of uncertainty. However, this latter way of calculating reference points is not really a full "MSE" procedure. It is simply using the carefully parameterised simulation model developed for an MSE to calculate the reference points instead of using the generic EqSim shortcut stochastic simulation method. Care should be taken when describing the calculation of reference points so as not to mix concepts and cause confusion; using different terminology for control parameters of an MSE would help avoid such confusion.

2.2 MSE experiences: North Sea whiting

Presentation by Tanja Miethe

Following ICES guidelines, MSY reference points such as F_{MSY} (F which delivers the long term sustainable yield), $F_{P.05}$ (an upper F limit that is considered precautionary for management strategies and MSY rules) and MSY $B_{trigger}$ are determined using EqSim software at benchmarks and checked again at inter-benchmarks (ICES, 2014a; 2014b). MSEs are run occasionally either as part of a benchmark, inter-benchmark or following a special request.

Considering the example of North Sea whiting, it can be shown that reference points estimated using EqSim are not always precautionary when tested in an MSE. For North Sea whiting, MSEs were run in 2016, at the Inter-benchmark (IBP; ICES, 2016a) to update natural mortality estimates in the assessment model, and again in 2019 during WKNSMSE following an EU-Norway request (ICES, 2019b). In each of these MSEs, discrepancies in performance of the HCRs in the standard ICES approach (EqSim) and the MSEs were observed. Such discrepancies are to be expected in general, because the assumptions underlying the two different simulations are likely to be different (in terms of observational error, time frame used, etc.), and one could not therefore expect that any two approaches necessarily produce the same results.

North Sea whiting is caught in a mixed fishery and considered bycatch. The stock has relatively high natural mortality estimates, which together with recruitment drive the stock dynamics (ICES, 2019d). At WKROUND 2013 (ICES, 2013a), it was concluded that there was no reliable estimate of F_{MSY} for this stock; instead, advice was given following the EU-Norway Management Plan with $F_{mgt} = 0.15$. At the IBP (Inter-Benchmark Protocol workshop) in 2016, it was estimated using EqSim that at a constant target fishing mortality, $F_{P.05}$ was 0.12. However, in the MSE, a constant target fishing mortality of neither 0.15, 0.12 nor 0.1 was sufficiently low to keep SSB above B_{lim} with at least 95% probability at the end of the projection period. Following this, an MSY approach was adopted for this stock. At the 2018 benchmark, F_{MSY} was updated again (capped by $F_{P.05}$) using EqSim (ICES, 2018a). However, at WKNSMSE 2019, this new value of F_{MSY} was found not to be precautionary when used as F_{target} in the HCR of the MSE (ICES, 2019b). Due to a lack of appropriate guidelines, EqSim reference points continued to be used and were not updated following the MSE.

In both cases, even though based on different lengths of data time series and assessment models (XSA in 2016, SAM in 2019), the EqSim estimates for F_{MSY} were not precautionary when considered as an F_{target} value in the HCR tested in an MSE. There are differences in these two approaches. EqSim can be considered a shortcut approach as it does not include running the actual assessment and forecast in the projection period. EqSim was run with a projection period of 200 years, while the MSEs were run with 20 years; this can have a strong impact on performance statistics, because for this stock Prob3 was shown to decrease slowly over time, potentially due to autocorrelation in the recruitment time series (ICES, 2019b).

To avoid large discrepancies in the estimates due to switching approaches, any approach to determine reference points needs to be practical enough to be run not only at benchmarks, but also at inter-benchmarks, when time, availability of resources, and expertise are limited. Furthermore, guidelines on how to deal with different outcomes with regard to precaution in reference point estimation and MSEs are needed.

Summary of discussion

A question was raised about whether whiting should be treated as a short-lived species with an escapement-type HCR, given that M is high and F low when compared to M. Although it is clear

that M is dominant and drives the dynamics, the assessment itself still has a number of ages (up to 8+, which is greater than for North Sea cod). There was some discussion about the role of reference points within ICES, and how useful they are, but the reference point framework remains an important component of ICES advice for forecasting catch opportunities, determining stock status, and informing indicators of regional environmental status. Discussions about the use of the MSE framework for extracting reference points have been transferred to Section 2.3, as it is more relevant there.

2.3 A potential framework for calculating reference point from simulation models used in MSEs

Presentation by José De Oliveira

Currently, there are no guidelines within ICES for how to extract reference points from an MSE framework when one is conducted. ICES currently relies on a standardised EqSim software for calculating reference points, which accounts for assessment and advice error (derived from the historical performance of assessments and forecasts), stochastic variation in recruitment (including a mix of different possible stock-recruit functions), and variability in biological and fishery quantities (e.g. mean weights at age, maturity, natural mortality and selectivity patterns). It does not currently explicitly account for other sources of assessment error (only the point estimates from an assessment are used, e.g. for the stock-recruit pairs), and density-dependent changes in underlying biological processes. It also does not account for bias in the way assessment and advice error is currently formulated. An MSE framework, on the other hand, is built to explicitly account for various sources of uncertainty and bias, and could arguably better account for these compared to the current version of EqSim; therefore, it might be preferable for estimating reference points. Table 2.3.1 provides an example comparison between the two frameworks (as used during WKMSEMAC; ICES 2020a).

Table 2.3.1. A comparison of EqSim and an MSE frameworks, as used in the WKMSEMAC process (ICES, 2020a). Note the MSE framework is not limited to just the full approach (it could also include a shortcut approach).

	EqSim	MSE framework
Assessment/advice Error	2 parameter (F_{cv}, F_{phi}) function derived from historic performance of assessment and forecast, adding auto-correlated (but unbiased) noise to the target fishing mortality.	Full feedback approach. Includes an assessment with consistent bias in estimates of SSB/FBar such that the realized F is always lower than the target fishing mortality.
Recruitment residuals	Process error generated around SR. Identical starting population for all iterations. Auto-correlation in residuals estimated and included in predicted residuals.	SR models assigned randomly to the 1000 populations derived from the variance-covariance matrix. Auto-correlated random deviations derived from ARIMA fit to log residuals for each iteration.
Simulation/Stat period	200 years/ 50 years	40 years/ 5 years

WKMSEMAC (ICES, 2020a) concluded that the MSE framework represents a more appropriate methodology than the generic EqSim tool for the estimation of reference points. In particular, they noted that the MSE framework:

- includes the actual assessment and forecast process and can therefore more appropriately handle the related uncertainties;
- is a consistent approach with that used for the evaluation of the long term management strategy.

A framework is proposed here for extracting reference points from an MSE simulation model, when an MSE has been conducted for a stock. It follows closely the current guidelines for estimating reference points using EqSim (ICES, 2017a), but with some adaptations, as schematically shown in Figure 2.3.1. One difference is the proposal to extract B_{pa} from the distribution of SSBs when fishing at an F that leads to Prob3 = 0.05 (Figure 2.3.1), which accounts for both stochastic recruitment and assessment uncertainty, and is consistent with how B_{pa} is defined, instead of being calculated using $B_{lim} \times e^{1.645 \times \sigma}$, where σ is based only on the assessment uncertainty in SSB in the terminal year (σ is the estimated standard deviation of ln(SSB) in the final assessment year; ICES, 2017a). The framework uses the (base case) OM conditioned on the current expert group assessment (according to the stock annex), and the same assessment as the MP (full approach) or characterising the assessment behaviour on the basis of the same assessment (shortcut approach). How to deal with alternative OMs was not considered in the proposal, because the proposed framework does not treat the process of extracting reference points as actually conducting an MSE (see below).

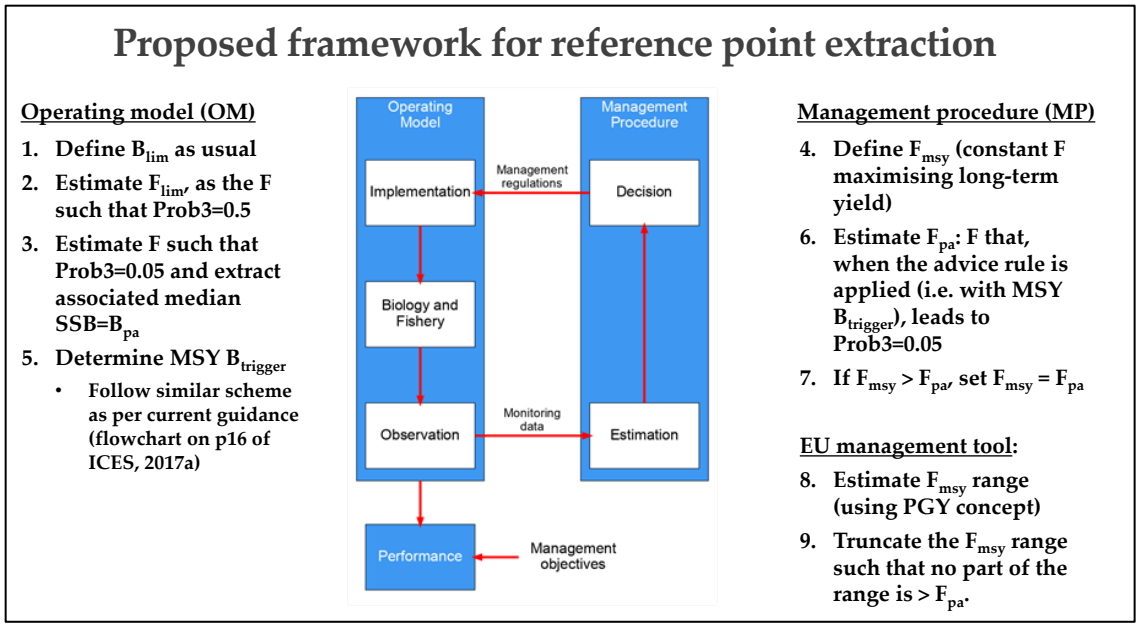

Figure 2.3.1. Schematic of proposed framework for reference point extraction when an MSE is conducted, with steps 1-9, in order. The scheme from the current guidance can be found in ICES (2017a). The advice rule is the ICES advice rule, which is a hockey-stick with breakpoint at F = F_{MSY} and SSB = MSY $B_{trigger}$ (ICES, 2019c). PGY is the "pretty good yield" concept (Rindorf *et al.*, 2017), and further information about calculating F_{MSY} ranges can be found in ICES (2016b). Although the full MSE schematic is shown, this framework will also be appropriate for the shortcut (middle plot of Figure 6.0.1).

One important point to note about the proposed framework is that the process itself does not fulfil the requirements of an MSE, so should not be labelled an MSE. This is because, for example,

MSY $B_{trigger}$ estimated within the OM is passed directly to the MP, and this would not be permitted within an MSE (anything passed to the MP has to go through the observation model in a full MSE, or through the emulator in a shortcut MSE).

The issue of number of projection years and number of replicates for the framework was raised, given that these are different for EqSim compared to an MSE framework (Table 2.3.1). The number of replicates is discussed elsewhere in this report (see Section 4.2). Running projections too far into the future was also questioned (see Section 4.2), so it is suggested that the same time frame as used for the MSE be used in the proposed framework; however, this would need further consideration in a dedicated workshop on reference points (see Recommendations; Section 10).

Summary of discussion

There was some concern about whether the proposed framework would replace EqSim as a general framework for estimating reference points, as this would not be practical because MSEs are not developed for every stock. The proposal simply provides guidelines for how to extract reference points from a self-consistent MSE simulation model when one is conducted. The proposed framework is less computationally onerous than an MSE because it does not require a grid search, but may suffer from the lack of standardised software, because MSEs are usually customised for the stock concerned (standardised software is a notable advantage of EqSim). Combining MSE updates with benchmark updates, and using the MSE framework for extracting reference points may be one solution for how this could work, but carries with it concerns about available time and human resources. The difference between the suite of ICES reference points estimated by the proposed framework and control parameters of HCRs that are evaluated in an MSE to be used for catch opportunity advice should be clearly communicated so that they are not confused (i.e. the standard names used for ICES reference points should not be duplicated for MP control parameters).

2.4 Conclusions

The group discussed the estimation of ICES reference points (B_{lim}, B_{pa}, MSY $B_{trigger}$, F_{lim}, F_{pa} and F_{MSY}) when an MSE has been carried out. Currently, EqSim is generally used to estimate ICES reference points using historical assessment error and autocorrelation (ICES, 2014b). This may result in ICES reference points being different to the control parameters of the MP, for example when the F_{MSY} from EqSim is compared to the F_{target} intended to maximise yield from an MSE (WKNSMSE report; ICES, 2019b). Furthermore, the MSE may conclude that the ICES F_{MSY} is not precautionary (e.g. the whiting and herring examples in the WKNSMSE report; ICES, 2019b; although in these cases, the time series of data used in EqSim compared to the MSE may have differed).

There should be a clear distinction between the reference points in the OM (used to evaluate performance against management objectives; e.g. F_{MSY}) and control parameters in the MP (used in a harvest control rule to trigger action; e.g. F_{target} and $B_{trigger}$).

The group discussed whether an MSE simulation framework that has been used for an MSE evaluation for a stock and associated fishery could be used to assess ICES reference points, instead of using EqSim. It could be argued that, in this context, an MSE simulation framework offers opportunities for a more comprehensive accounting of uncertainty, and in particular accounts for bias when estimating F_{MSY} and associated reference points (the F_{MSY} range and F_{pa}). A proposal was put forward to adapt the MSE code to mimic the estimation process as used in

EqSim (see Section 2.3). This will bring the process for updating reference points closer to (although still not identical to) the MSE that produced the original evaluation.

While this seems like a potential path to follow, and given the fact that the estimation of reference points requires special consideration in any case, WKGMSE3 recommends that the guidelines for estimating reference points in either the context of a benchmark or management strategy evaluation, be considered in a dedicated workshop (Section 10).

3 TOR b: Alternative operating models

Develop guidelines for how to treat the results of alternative operating models. Currently, these have been used as robustness tests for "optimised" management strategies.

A fundamental principle of MSE is to have a range of operating models (other than one based on the current assessment) to cover various important sources of uncertainty (Table 1.0.1), but how results from these are weighted and/or combined is not always clear. Currently, they have been used to check robustness of optimised management strategies. TOR b explores how to handle alternative operating models.

3.1 Operating model design in tuna RFMOs

Presentation by Rishi Sharma

Abstract from Sharma *et al*. (2020):

The five Regional Fishery Management Organizations dedicated to tunas (tRFMOs) are all either developing or implementing Management Strategy Evaluations (MSEs) to provide advice for the stocks under their competencies. Providing a comparative overview will help tRFMOs to learn from one another and to collaborate on common solutions and may also help to more clearly define the challenges of building decision support tools in contexts of large scientific uncertainty and where management requires cooperation across multiple stakeholders characterized by unequal power and divergent interests. For example, our overview showed that in most cases, a grid-based design with an emphasis on structural uncertainty has been adopted. However, uncertainties such as sampling errors and non-stationarity of important ecological processes, which are of potentially equal significance for demonstrating robustness of management procedures, were not considered. This paper identifies key issues for operating model (OM) design that challenges the tRFMOs, compares how these challenges are being met, summarizes what lessons have been learned and suggests a way forward. Although the current approach of using assessment models as the basis for OM design is a reasonable starting point, improvements should be made to the conditioning of OMs, especially with respect to enabling the inclusion of other important processes and uncertainties that are difficult to account for in stock assessments but that can crucially affect the robustness of advice. Attempts should also be made to improve documentation and communication of uncertainties that are included and those that are excluded from consideration in the process.

Summary of discussion

Within ICES MSEs, the estimator has often been a replicate of the assessment used to condition the Operating Model (OM); although it may be desirable to have a simpler estimator (see Section 7.2), the current management system does not, on the whole, accommodate this "simpler" method (involving the use of empirical estimators) as it tends to take a "best assessment" approach (see Section 7.1). Communicating MSE results back to managers is a complex task, and there is no single best way of doing it; furthermore, conclusions from simulations may depend on stocks, areas and time period (e.g. results may show it is preferable to have one larger cut in catches or several smaller ones over a longer period). For CCSBT (Commission for the Conversation of Southern Bluefin Tuna), tuning criteria are closely developed with managers. Generally,

it is easier to communicate choices between alternative MPs if the management objectives are clearly specified.

Implementation error is often stock/fishery specific, and there is a need to account for what is realistic for a particular stock/fishery (i.e. to consider not just IUU, but also bycatches, banking and borrowing, etc.).

How did different case studies within tRFMOs decide which OMs were important when narrowing down from over 1440 models to just a handful? This was done on a case specific basis, and came down to factors such as the effects of sample size, natural mortality, etc. A key principle is to narrow it down to what is driving the stock. Three points are important: (i) the MP may be biased, so tune it; (ii) if there are no good reference points, try to go for trends; and (iii) generation cycle is important – if the population doubles in two years, this has an impact. With MSE, it is desirable to find simple rules and feedback controls that work. What tends to be important in this context is the nature of the time series.

3.2 Conclusions

Although TOR b did not cover the choice of OMs, there was some discussion around the need to select a plausible set of OMs, and the basis for doing so. The first consideration was whether the hypothesis is sufficiently important to be included, where importance refers to either or both of "considered likely to reflect the actual situation (i.e. more plausible)" and "would have a large impact/influence on management advice if it did in fact reflect the actual situation". This should be followed by a demonstration of goodness of fit (e.g. through runs tests and other model diagnostics), as OMs must be consistent with the data available, and finally by cross-validation (against observations, demonstrating prediction skill). Once OMs have been identified for inclusion in an MSE, these may be split into a Reference Set of the more plausible OMs, and a Robustness Set of less plausible or less important OMs (particularly ones that were nevertheless potentially more influential), using the above criteria. For each case, decisions about plausibility will have to be made by the workshop conducting the evaluation, and should be clearly explained and agreed.

A robustness test may indicate where further research is needed to resolve any problem areas. It could also be used as a way to bound the applicability of evaluations (if the situation it describes actually happens, then there may be a need to review the approach prior to the next application of the management strategy). Robustness tests could also have a lower bar for MPs to pass in order to be considered 'acceptable' compared to the Reference Set, but any changes to acceptability criteria need to be agreed with managers or other relevant stakeholders.

Two rather different approaches were considered for reporting MSE outputs from the Reference Set of OMs, which can be conveniently called the "IWC" (International Whaling Commission) and "CCSBT" (Commission for the Conservation of Southern Bluefin Tuna) approaches, because they have been implemented by these two groups and.

"IWC" approach:

The "IWC" approach considers a manageable set of OMs (a few, and not more than about 50 at most). Performance statistics from these OMs are evaluated separately for each OM, rather than as a single performance measure derived from integrating performance statistics across all OMs. Under this approach, MPs must, for example, pass the risk level (5% for Prob3 in the case of ICES) for all OMs in the Reference Set; each OM in the Reference Set is effectively given equal weight.

Pros

- The performance of each MP under each alternative OM is evident; in particular, poor performance under one OM cannot be hidden (while it could become overlooked in an integration-based, "CCSBT" type approach).
- Relative weighting of OMs is not essential, potentially saving much discussion/negotiation time.

Cons

- Can only include a limited set of OMs.
- Outcomes can be heavily dependent on the decision of whether or not to include an OM in the Reference Set based on subjective "plausibility" criteria.

"CCSBT" approach:

The "CCSBT" approach typically considers a greater number of OMs (often in the 100s, though it can be applied to smaller numbers too) and allocates relative weighting to the OMs in the Reference Set. Typically, the Reference Set (or "grid") of OMs arises from a full cross evaluation of different levels along what are considered to be the major axes of uncertainty, with weighting based on fit likelihood and/or expert judgement. In this approach, performance statistics are integrated over the full Reference Set (e.g. by combining distributions across all the OMs).

Pros

- Can evaluate each MP under a wider range of uncertainties (OMs).
- Facilitates easier comparison between MPs as there is effectively a single set of integrated performance statistics for each MP.

Cons

- Requires (preferably) relative weighting for each OM.
- Integration over results could hide unacceptable performance due to a particular source of uncertainty (represented by an OM or a subset of OMs).

There are other examples where both these approaches address structural uncertainty. The Indian Ocean Tuna Commission (IOTC, 2019) is combining models representing different combinations of assumptions about stock biology and fishery dynamics. Reference Sets are comprised of dozens to hundreds of OMs covering large ranges of uncertainty in stock productivity, variability, and current status. Results from OMs are combined with equal weights once they satisfy some criteria for acceptance. MPs are tested for all the OMs and performance statistics are computed by integrating across the whole set.

Evaluations based on a reference OM and a set of robustness tests has been carried out for a number of stocks by different Expert Working Groups of the European Commission Scientific, Technical and Economic Committee for Fisheries. For example, STECF (2019) conducted an

evaluation of MPs for demersal stocks in the Adriatic. A single base case OM, including estimation uncertainty, was used to select the harvest control rule parameters, while a number of robustness OMs were then employed to test its behaviour under different possible conditions.

It seems more appropriate, at least initially, to begin with a smaller number of OMs offering a simpler and less computationally onerous way forward for ICES.

4 TOR c: Risk and uncertainty

Explore the relationship between estimated risk and assumed levels of uncertainty included in the MSE. Risk and uncertainty are closely related, and including more uncertainty affects the estimated level of risk from the MSE. Apart from uncertainty, consideration should also be given to:

i. The number of replicates and length of projection period used in the MSE;
ii. The stationarity of MSE projections, from which risk metrics are calculated;
iii. The risk metric itself (e.g. several definitions are given in the WKGMSE report of 2019).

TOR c covers issues related to the definition of risk, and in particular whether there is some way of benchmarking risk in relation to the amount of uncertainty incorporated in an MSE, and what to do in the presence of non-stationary MSE projections.

4.1 Alternative views of risk

Presentation by Carryn de Moor

Most fisheries are managed on the basis of a (perhaps implicit) chosen level of acceptable risk (usually expressed as probability). "Risk", in the context of managing fisheries, can be considered as the probability of not achieving a desired management goal. In some settings, a policy decision of an acceptable risk probability directly influences the Harvest Control Rule (e.g. the preferred level of risk aversion to overfishing is used by the Pacific Fishery Management Council to set the acceptable biological catch at a level below the overfishing limit).

In the context of MSE, however, risk is evaluated by means of a performance statistic used to determine if a candidate Management Procedure (MP) satisfies pre-specified objectives. The definition of risk consists of both a "threshold" against which one would judge a result, and the tolerance "level" or probability that the threshold is exceeded. For example, risk could be the probability of the simulated biomass under a given MP not reaching the target biomass such as B_{MSY} (the threshold) within a pre-specified time frame. Within an ICES context, the recommended risk definition (Prob3) relates to the maximum probability that spawner biomass (SSB) falls below B_{lim} (the threshold) over a pre-specified number of simulated years (WKGMSE: ICES, 2013b; WGKMSE2: ICES, 2019a). The policy choice for a pre-specified maximum acceptable risk probability is 0.05 for ICES stocks.

In contrast, for South African stocks, risk has been defined in a species-specific manner and has not always been explicitly considered in MSEs. For example, in some MSEs the selection amongst alternative candidate MPs has focussed primarily on simulated median catch levels or achievement of recovery targets. In these cases, the risk of the population dropping below an unacceptable level was considered on a relative basis when comparing leading candidate MPs, as part of the suite of performance statistics. In the small pelagic fishery, where a threshold similar to B_{lim} is used, the acceptable risk level (probability) is calculated by considering what the risk would naturally be in an unfished scenario. Changes in the Operating Models (OMs) used to simulation test one MP from another, such as changes in the rate of natural mortality or variability about the stock recruitment relationship, can substantially affect the perceived risk to the population in an unfished scenario. Similarly, the inclusion of additional uncertainty within an OM would result in wider confidence intervals, and thus a higher risk even in an unfished scenario. The simulated distribution of biomass under an MP will spread towards lower biomass levels compared to that for an unfished scenario. For example, in Figure 4.1.1, the distribution of the

biomass at the end of the projection period under alternative MPs shifts further to the left of the unfished distribution as the risk probability increases. This "leftward shift" is used to determine the acceptable level of risk to be used to tune a new MP for the South African small pelagic fishery (de Moor and Butterworth, in prep). The ratio of the 20th percentile of the end-year biomass distribution under an MP to that under the unfished scenario is kept constant over time from one selected MP to the next. This methodology has automated the process where the absence of a policy decision on the acceptable level of risk could have resulted in stalemate discussions on acceptable risk levels. The acceptable level of risk may therefore increase or decrease from one accepted MP to the next, depending on the underlying uncertainties which affect both the unfished and fished scenarios.

This method implicitly considers the multiplicative difference between the acceptable level of risk under the MP to that under an unfished scenario. When risk under an unfished scenario is very low, however, it may be more appropriate to consider risk additively. For example, within ICES, perhaps an MP could be said to be precautionary if Prob3 is ≤5% more than that simulated under an unfished scenario. This would ensure that the inclusion of additional (realistic/plausible) uncertainty – which is encouraged in an MSE – is not unduly 'penalised'.

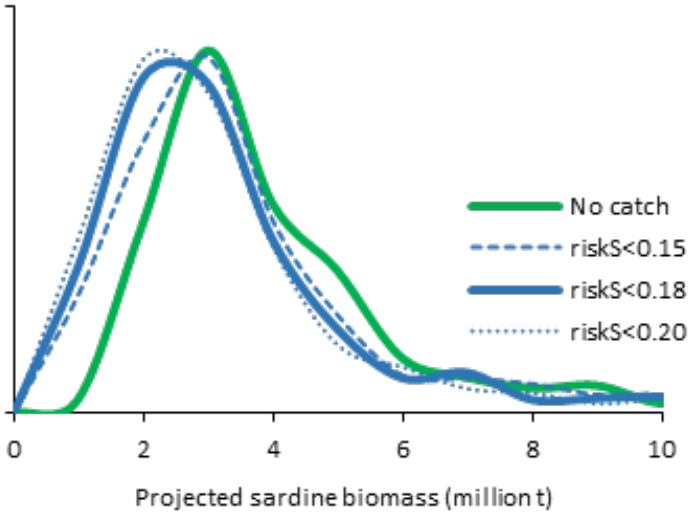

Figure 4.1.1. The simulated distribution of South African sardine biomass at the end of a projection period under a no future catch scenario compared to catch scenarios tuned to different levels of risk. In this example, risk was defined as the probability that the projected sardine biomass falls below the average 1991 to 1994 biomass at least once during the projection period of 20 years.

Summary of discussion

There was a brief description of risk in northeast USA, were key risk metrics are the probability that $F > F_{MSY}$ and probability that $SSB < 0.5B_{MSY}$. Currently, not many MSEs are used to inform management in the USA, and it is more of an academic exercise. There are examples of MSEs of Atlantic herring that include uncertainty about the relationships with other species, such as tuna growth being related to herring condition, links being explored between herring biomass and the productivity of common terns, and between dogfish natural mortality and herring abundance. In these cases, risk can be associated with meeting a threshold linked to a predator, for example. Therefore, risk can be defined quite differently in different parts of the world.

Risk can be affected not only by the uncertainty within an OM, but also by the range of OMs considered, which in turn is affected by questions of plausibility of alternative OMs, and how this is dealt with when calculating performance statistics (including risk). Selection of OMs should be such that they are balanced (i.e. in respect of OMs that are both more and less optimistic about the current status and productivity of the resource). Once results are integrated across a range of OMs, care must be taken with the use and interpretation of variance measures (including probabilities). The difficulty is that the results for these measures will depend on the number of OMs considered, because the greater that number, the larger any variance value will tend to be. This raises particular difficulties when, for example, tuning a management strategy to a particular risk value – risk evaluated over which set of OMs?

It is important to do an adequate number of robustness tests to avoid too-frequent updates of the MP, but there should be reasonable limits on this. It is also important to consider metrics other than risk, such as how quickly a management strategy is able to recover a resource once it gets into trouble (such as when there is a series of low recruitment). The arbitrary nature of the 5% risk level used by ICES was also discussed; there was a suggestion that it may have been based on early work by Bergh and Butterworth (1987), who argued that a reasonable approach to manage fisheries may be to choose a level of risk similar to that considered acceptable in other sectors of the economy, and initially suggested somewhere in the vicinity of 10%. This was a level also used by Beddington and Cooke (1983), although Butterworth and Bergh (1993) later argued for a common international convention for its choice (something we have not yet achieved almost 30 years later). Therefore, it is still not entirely clear where 5% came from, but it seems to have become institutionalised through repetition after 1998 when it was first presented to managers. Managers were recently requested to give their opinion about this level of risk (January 2019 MIRIA meeting, a Meeting between ICES and Recipients of ICES Advice), and answered that they were satisfied and had no alternatives to offer.

A review of the performance of management strategies that have been in place for a long time would help to give an overview of how successful these management strategies have been. Examples of such management strategies are Icelandic cod, North Sea herring and some South African stocks.

4.2 Effects of uncertainty in risk in MSEs

Presentation by Daisuke Goto (see Annex 3 for further details of the analysis)

Assessing the risk of overexploitation of fish stocks is critical to managing fisheries sustainably. The risk assessment, however, must be performed with imperfect knowledge of current stock status, sampling (observation) error, and environmentally driven year-to-year fluctuation in stock size (also known as "process error"). To determine how sensitive assessed risks are to varying levels of uncertainty, we conducted simulation experiments using the full management strategy evaluation (MSE) framework previously developed for North Sea saithe (*Pollachius virens*) as part of WKNSMSE (ICES, 2019b). This framework comprises an age-structured population model as an operating model (OM), two monitoring surveys, and a management procedure (MP). In the MP, a model-based harvest control rule (HCR) is implemented based on assessment of stock status using SAM (State-space Assessment Model; Nielsen and Berg, 2014). In this work, we analysed the risk (probability of spawner biomass dropping below a predefined threshold, B_{lim}, in the last 10 years of projection) by evaluating performance of the HCR set for saithe under increasing levels (10% to 100%) of six sources of uncertainty: two process errors (number-at-age and recruitment in the OM) and four observation errors (survey catchability, age-specific abundance index, biomass index, and catch). Overall, the HCR was robust to a moderate increase in

uncertainty levels tested, the risk level remained below 5% when uncertainty levels increased in number-at-age, recruitment, age-specific abundance index, and catch. By contrast, the performance of the HCR was sensitive to increasing levels in two of the observation errors, survey catchability and the biomass index, increasing the risk as much as 33-fold. To further test how sensitive the estimated risks of increasing levels of these two observation errors to simulation settings, we reran MSEs with varying numbers of replicates and projection periods. Analyses showed that the estimated risk becomes less stable and less accurate with elevated uncertainty levels, and it would require higher number of replicates and longer projection period to ensure its reliability. These findings underscore the importance of evaluating the reliability of assessed risks under varying uncertainty levels when conducting MSEs.

Summary of discussion

It was not clear why risk became more variable with increasing number of replicates, as shown in the presentation, but this effect has disappeared when recomputed using the resampling method with 10 000 replicates (see Annex 3, Figure A.3.3). Nevertheless, if risk changes in a non-smooth manner depending on the number of replicates, then the stability of the estimator should be investigated; furthermore, sufficient replicates should be used for the required level of precision (Annex 3, Figure A.3.3).

Running MSEs too far into the future (as was done when considering projection years (Annex 3, Figure A.3.4) was questioned, because managers are generally only interested in looking no further than 10–20 years into the future. Risk should in any case be considered on a relative basis (i.e. compared amongst alternative MPs). Results showed greater sensitivity to observation error compared to process error, but this was not surprising because observation error is generally a key driver in MSEs.

4.3 Conclusions

When considering risk, a distinction is needed between the risk metric (the probability of breeching a threshold), and the risk threshold (such as B_{lim}); the former is simply referred to as "risk" below, and the latter as the "risk threshold".

Performance statistics or summary metrics are a set of statistics used to evaluate the performance of Candidate MPs against specified pre-agreed management objectives, and the robustness of these MPs to uncertainties in resource and fishery dynamics of concern to stakeholders and managers (Section 3). These performance indicators are codified as properties of the system, e.g. the ratio of the realised catch to MSY, and the risk of the stock falling below a level where recruitment is impaired. They may be used to test the robustness of assumptions made in a stock assessment or within an MP, for example when B_{MSY} (based on exploitable biomass) or F_{MSY} (based on harvest rate) are used as part of the forecast for biomass dynamic models such as SPiCT (Pedersen and Berg, 2017). These may differ from the corresponding quantities in the OM (where, e.g. B_{MSY} is based on SSB, and F_{MSY} based on an instantaneous exploitation rate). There are two main ways to derive quantities to be used for performance statistics, namely: (i) using equilibrium assumptions (Sissenwine and Shepherd, 1987), or (ii) through stochastic simulation e.g. by projecting at $F = F_{MSY}$ or $F = 0$ (e.g. Carruthers *et al.*, 2016; de Moor *et al.*, 2011). The latter approach is preferable where environmental forcing or resonant cohort effects impact on productivity.

The group agreed a focus on relative performance statistics (i.e. relative to OMs) was preferable to absolute measures of performance statistics in terms of management objectives (e.g. recovery targets, risk, etc.). With regard to the risk threshold, such as B_{lim}, this measure is inherently linked

to the OM and should therefore be specific to an OM. It may therefore be useful to focus on the description of methodology to derive B_{lim}, e.g. SSB in a particular year or lowest historical SSB, rather than focus on its absolute value. However, it was also noted that it was important to work to the requirements of the management system in place, and that ICES and the requests from its clients currently require an assessment of absolute risk (Prob3≤5%) as part of the advice process.

In terms of risk itself, consideration should be given to setting the acceptable risk relative to that which would be achieved under a no fishing scenario. While some consideration has already been made towards this in the guidelines (e.g. short-lived stocks), there remain potential constraints of having an absolute risk. For example, if Prob3 for a short-lived stock is simulated to be 0.049 under a no future catch scenario, then the current guidelines allow for hardly any exploitation of this stock by requiring Prob3≤0.05 for any simulated MP. A way of addressing this could be to set the acceptable risk to an additive increase of that simulated under the no future catch scenario. As risk is linked to the level of uncertainty assumed in the OM, this method of considering the acceptable risk to be an additive increase of that simulated under the no future catch scenario could also help address the concern that the inclusion of additional uncertainty may 'penalise' the selection of MPs on the basis of satisfying an absolute probability. Hindcasting MSEs (see Section 6.3) can also provide a useful way of comparing realised risk (e.g. Prob3) under historical management when screening alternative MPs, and such approaches are encouraged. When considering future scenarios (such as regime shifts, as considered for mackerel: ICES, 2020a; and for Icelandic cod: ICES, 2010), care should be taken that reference points are appropriate to these future scenarios. Such scenarios could be handled as alternative OMs with associated different reference points, or by including relevant processes (e.g. density dependence in the OM).

5 TOR d: More efficient tuning in MSEs

Develop more efficient ways of conducting searches over a grid to the required level of precision. This is needed because of the high-performance computing requirements for full MSEs. This work could include investigating statistical properties that relate sample size to required precision, GAMs to interpolate over an incomplete grid, etc.

TOR d covers the practical problem of optimising management strategies under full MSEs when each cell of a grid over which the optimisation takes place takes a long time to run. This TOR covers more effective and efficient ways of conducting the optimisation (e.g. through statistical means, or by using methods such as genetic algorithms).

5.1 Statistical approach for more efficient grid searches

Presentation by Michael Spence (see Annex 4 for further details of the analysis)

Grid searches can be expensive, and running every grid value can be wasteful. By considering the outcome of the grid as uncertain and describing "your" beliefs about the grid based on previous runs, as well as knowledge about how the space behaves (e.g. one may expect the results of the grid search to be smooth), large areas of the grid space can be removed without exploring it. On many occasions, we may 'know' the outcome of a particular model run before it has been done. For example, if in a grid search all the surrounding points give very poor MSE results, then we would expect that this point would also give a bad result. Running the MSE only to confirm this is wasteful in terms of time and possibly money.

This idea is formalised using Bayesian statistics where one's beliefs about the outcome of an MSE is quantified and design experiment techniques are used to improve one's knowledge about the space. In this presentation, methods for doing this were described, and it was demonstrated, using a Gaussian process emulator with North Sea cod, that a good approximation to the full grid results can be accomplished with only 15% of clock time and with only 12% of the model runs.

Summary of discussion

The obvious candidates for solving problems similar to the grid search presented here are the geostatistical techniques used for oil exploration. The level of precision used for the search (two decimal places for F) was also questioned, given the considerable uncertainty included in the MSE, but this is the level ICES typically works with. The Gaussian process emulator used in this work can handle multiple dimensions, so can be extended to more complex problems than the one presented, and can also include multiple objectives. It could also be interesting to look at performance statistics other than those generally considered; this might help discriminate between control parameter combinations that show similar performance under a restricted set of performance statistics. If there were more than a single OM, then the procedure developed here could simply be run multiple times, although the approach can also work on ensemble models with their associated uncertainties.

The North Sea cod example used had a very smooth and well-behaved surface, which may not always be the case. However, the method can work with more difficult surfaces that have local minima using "history matching" techniques (adjustments to reproduce past behaviour in order

to have reasonable future predictions). The method also used Sobol sequences, which are quasi-random low-discrepancy sequences that attempt to fill the space of possibilities more evenly, resulting in faster convergence and more stable estimates. Furthermore, the iterative design used means that starting points don't matter too much.

5.2 A bootstrapping approach to streamline MSEs

Presentation by Iago Mosqueira (see Annex 5 for further details of the analysis)

Using bootstrap on the evaluations of MPs along a grid of HCR parameters allows the calculation of the precision of the quantile estimates, e.g. the 5% probability of $B < B_{lim}$. The number of iterations to be run can then be limited to those needed to obtain the required level of precision. It can also be used to limit the number of iterations in areas of the parameter grid that do not result in the desired performance. An example was presented based on the recent analysis of MPs for North Sea cod.

The algorithm aims at a predefined precision level, computed as the relative error of the Median Absolute Deviation (MAD). Iterations are carried out in steps until the target relative error, or a set maximum number, is reached. A selection is then made on continuing with the run only if the upper limit of the confidence interval includes the performance objective value. The tests conducted on the NS cod MSE indicated that the same solution as obtained in the original analysis could be found, and with the result determined to the desired level of precision (1% in this case), by computing only 30% of the simulation runs.

A draft version of the source code is available in Annex 5, and a complete method will be distributed as part of the mse package of the FLR platform (http://flr-project.org/mse).

Summary of discussion

The possible combination of this algorithm with other approaches was discussed. By quickly identifying areas of the parameter grid likely to return the chosen objective, and then applying the bootstrap method, a limited number of runs could be made that (i) focus on the useful parameter values and (ii) return results with the required precision level.

5.3 Pareto-optimal solutions using machine learning and support vector regression

Presentation by Laurence Kell

When tuning control parameters (i.e. hyper-parameters) in a management procedure, it is important to use an efficient and robust search strategy to sample potential parameter combinations, such as gain terms in an empirical MP, or values of F_{target} and $B_{trigger}$. There are two main approaches to find the best parameters: grid search which exhaustively considers parameter combinations, and random search which has been shown to be more efficient for hyper-parameter optimization than trials on a grid (Bergstra and Bengio, 2012).

Tuning a management procedure often focuses on finding the best possible objective value; however, in practice there are multiple objectives and trade-offs between them (Kell *et al*., 2019). An alternative is to use evolutionary techniques which have been found to be both practical and

efficient for such multi-objective problems, e.g. Non-Dominated Sorting Genetic Algorithm II (Zhihuan *et al*., 2010) and Multi-Objective Genetic Algorithms (Murata and Ishibuchi, 1995).

An example application was presented based on random search which used support vector regression (Smola and Schölkopf, 2004), a non-parametric machine learning technique, to model the relationship between the objectives and the control parameters. These relationships were then used to identify the Pareto-optimal solutions using a Genetic Algorithm (Whitley, 1994). These optimal solutions fall along the Pareto frontier that identifies situations where no objective can be improved without making at least one other objective worse.

An advantage of this method is that it focuses on trade-offs and does not assume that the "best" solution is known and promotes a dialogue between stakeholders.

Summary of discussion

A question was raised about whether machine learning techniques were better at finding minima compared to more traditional techniques. The approach presented was based on cross validation and involved random search, which is more able to look into all the "nooks and crannies"; this has proved to be quite successful. Key trade-offs among objectives are explored through the Pareto frontier, which provides a powerful tool to be used with stakeholders.

5.4 Conclusions

Due to the high computational requirements, before embarking on optimisation over a grid, one initial approach could be to agree with the requestors of advice what performance statistics are of most importance to them, the tolerance within which they need them calculated, and consequently (for example), how many different combinations of F_{target} and $B_{trigger}$ would be required.

Several approaches were presented to the group that offered various ways of improving the efficiency of identifying optimum parameters in MPs, based either on searching for the optimum combination of control parameters on a grid, or using random search techniques based on machine learning. More sophisticated approaches may be needed for MPs which include a wider number of control parameters with interactions between them (e.g. see Section 6.3).

Two statistical approaches were presented that led to more efficient grid searches involving computer-intensive methods. They were tested on the NS cod MSE from WKNSMSE (ICES, 2019b), which offered results for a full grid of 451 cells.

- The Gaussian process emulator (Section 5.1, Annex 4) found the solution with 56 evaluations instead of the full 451 grid search, and only ran an MSE simulation (one cell) when required (study sped up by 85%).
- The bootstrap method (Section 5.2, Annex 5), that allows approximation of the confidence intervals and relative errors around other estimates, found the solution with a reduction of the number of replicates required by 70%.

These improvements are context dependent (here within the context of the NS cod MSE), and more work may be needed in other contexts to better understand the behaviour of these methods. These two approaches could be also used in combination.

Different approaches to these statistical methods are the use of a shortcut MSE (if found to provide an adequate approximation to the full MSE) to explore the full grid (quick to run), or the use of fewer replicates to scope the grid initially. These are then followed by full MSEs in the focus areas identified.

The final method presented (Section 5.3) uses efficient machine learning techniques to model the relationship between objectives and control parameters, which could then be used to identify the Pareto-optimal solutions (with genetic algorithms) that focus on trade-offs among these competing objectives. The approach uses a random search, support vector regression and genetic algorithms, and can be used subsequent to having run the MSE, unlike the previous optimisation approaches which require the MSE to be rerun.

Machine learning (ML) has the potential to help society adapt to major global challenges, e.g. changing climate and natural resource management (Rolnick *et al.*, 2019). For example, when conducting MSE, the objective is to find robust feedback control rules that, despite uncertainty, still meet multiple management objectives. ML can help to find general rules and provide efficient methods to fine-tune rules on a case-specific basis (Fischer *et al.*, in press 2020). When the potential parameter space increases, however, i.e. when tuning MPs or considering a range of uncertainty, simulations can often take far longer than can be searched in a reasonable time. Additionally, the objectives are usually mutually exclusive, leading to a decision having to be made as to which objective is more important or optimising over a combination of the objectives. An alternative is to use Genetic Algorithms to identify the Pareto frontier and find the optimal parameter sets for all combinations of objectives. The Pareto frontier can then be used to identify the sets of optimal parameters that are 'best' for a given combination of objectives – thus allowing decisions to be made with more complete knowledge (Kell *et al.*, 2019).

6 TOR e: Shortcut versus full MSEs

Compare the shortcut and full MSE approaches, providing guidelines for use of the former as an approximation for the latter, if appropriate. Consideration should be given to MSE with alternative operating models (i.e. operating models not solely based on the currently-used assessment).

TOR e covers something that is topical for MSEs conducted within ICES, namely comparing the shortcut approach to carrying out computationally intensive full MSEs. Figure 6.0.1 illustrates the difference between the full and shortcut approach, and additionally includes the empirical approach for comparison. The main difference between the full and shortcut approaches is that the shortcut one replaces the estimation model with an estimation emulator, where typically numbers at age or biomass are taken directly from the operating model, but modified with noise before being used in the decision model. Empirical approaches bypass the estimation model and use observations directly in the decision rules. We have dispensed with the term "full-feedback" to describe the full approach to avoid ambiguity, because all of these methods rely on the principle of feedback-control.

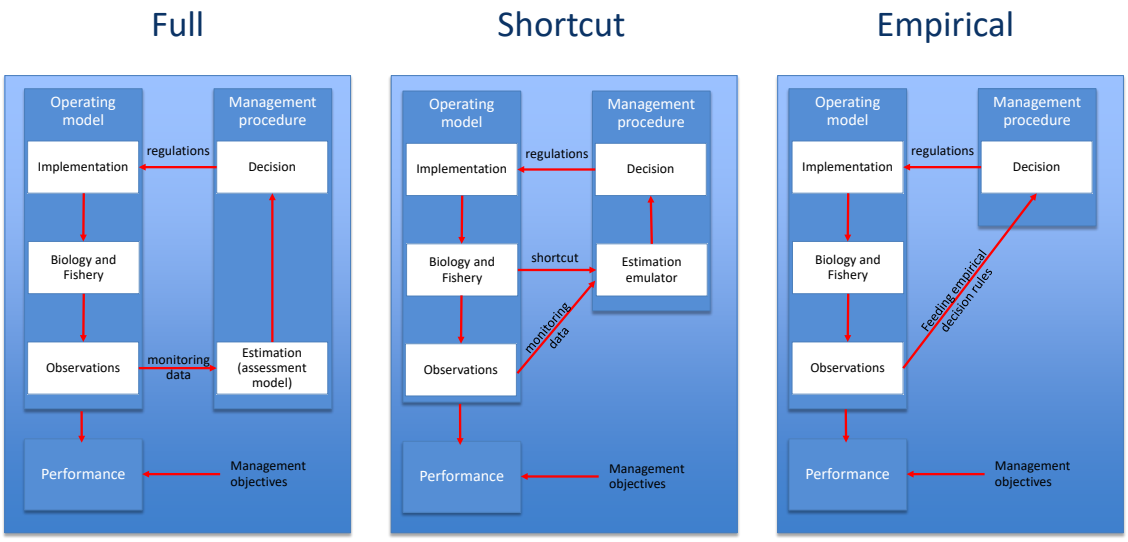

Figure 6.0.1. Schematic of different forms of MSE, modified from Punt *et al*. (2016).

It became clear in discussions during and after the meeting that there were two interpretations of the shortcut approach, and it is useful here to understand the difference between them in order to place the analyses presented in Sections 6.1–6.5 into an appropriate context.

The first interpretation of the shortcut approach is in the stricter sense of an MSE that will deliver an MP (Table 1.0.1). In this interpretation, the four components (monitoring data, assessment, harvest control rule and management action) are pre-specified [this was the interpretation of the full method as used by WKNSMSE (ICES, 2019b) and WKMSEMAC (ICES, 2020a); this is possible given the current ICES system of benchmarks, which deliver stock annexes that pre-specify monitoring data and the assessment method]. Here, the shortcut approach is viewed as an approximation to the full method (the shortcut emulator replaces the monitoring data and assessment components; Figure 6.0.1), and if it were to be used, there would need to be a demonstration that the shortcut approach provides an adequate approximation (see WKGMSE2 report and

references therein; ICES, 2019a). The approximation can be achieved in various ways, for example using assessment error (e.g. variance-covariance matrix of numbers and exploitation pattern at age; see Section 6.1 and Annex 6) or using an analytical retrospective analysis (see below, Section 6.2 and Annex 7); crucially, both are based on the accepted ICES benchmark assessment. In this approach, if any of the monitoring data or assessment method were to change (e.g. such as would require another benchmark or inter-benchmark), then the MSE would need to be rerun (unless it could be shown that the changes would have only a minor impact on results of the MSE).

The second interpretation of the shortcut approach is a rather looser interpretation of MSE, with a focus on Harvest Control Rule (HCR) evaluation, and no longer on a pre-specification of what the monitoring data and assessment method are. This approach strives for a more generic application of the HCR, which means that the approach is no longer viewed as an approximation (the lack of pre-specification means there is nothing against which to compare), but rather as broadly modelling changes seen historically and assuming that similar changes will occur in future. In the IWC context, where MSE was pioneered, this approach is more closely aligned to the NMP (New Management Procedure), rather than the RMP (Revised Management Procedure), which insists on pre-specification. This shortcut approach is typically characterised by considering historical retrospective patterns (i.e. not linked to any particular monitoring data set or assessment method, because these may have changed historically, and do so again in future) rather than the more narrowly defined analytical retrospective patterns (which are calculated from a single assessment model and the associated data). It should be noted that historical retros typically consider the current benchmark assessment as the baseline against which previous assessments (for which model and associated data may have changed over time) are compared; in contrast, analytical retros will keep the same model and data sources, and simply peel off one year of data at a time (the method typically used to calculate Mohn's rho). See Sections 6.3, 6.4 and 6.5 (and associated Annex 8) for examples of this second interpretation of the shortcut approach. A challenge for this version of shortcut approach is that if a generic HCR (i.e. without pre-specifying monitoring data and assessment) is to be tested, the testing (by definition) needs to encompass assessment uncertainty and uncertainty about the data to be used that is then sufficient to cover the expected range (which would require a much more comprehensive evaluation than a case-specific evaluation, such as that in the more narrowly-defined first interpretation of the shortcut approach).

6.1 Sprat MSE example

Presentation by Mollie Brooks (see Annex 6 for further details of the analysis)

An MSE was conducted on sprat in the North Sea as part of the 2018 benchmark for that stock. The methods developed there were used to compare full and shortcut MSEs with alternative operating models. General MSE software was not used because seasonality and the inclusion of a zero age-group are thought to be important aspects of the dynamics that need to be included in simulations and assessments of sprat. The decision rule for this stock is an escapement strategy where the goal is to set the TAC each year such that SSB in the following summer after the TAC is taken is expected to equal B_{MSY}. To do this, a short-term forecast solves for the F-multiplier that will achieve this escaped SSB, using estimated stock numbers (N), exploitation pattern (E), and biological parameters. Estimated N and E come from a SMS assessment in the full MSE, or in the shortcut version (i.e. the assessment emulator) by giving the true N and E multivariate log-normal errors. In this study, the assessment emulator simulated errors using the variance-covariance

matrix of log(N) and log(E) produced by the benchmark version of SMS, but see comments below about future improvements.

In a first version of the comparison, biological parameters were provided from the operating model to the assessment and short-term forecast without error. In that case, the shortcut and full versions produced quite similar results (performance statistics, worm plots, timelines with CI) even when changes were made to biological parameters in the operating model (and passed to the HCR without error), but this was not considered to be a useful comparison.

In a second iteration of the comparison, bias of plus or minus 10% was added to the natural mortality provided to the short-term forecast (in both shortcut and full versions) and to the SMS assessment in the full versions. In this comparison, when natural mortality was observed with error in the HCR, then the performance statistics of the full and shortcut versions did not match to a satisfactory degree (Figure 6.1.1)

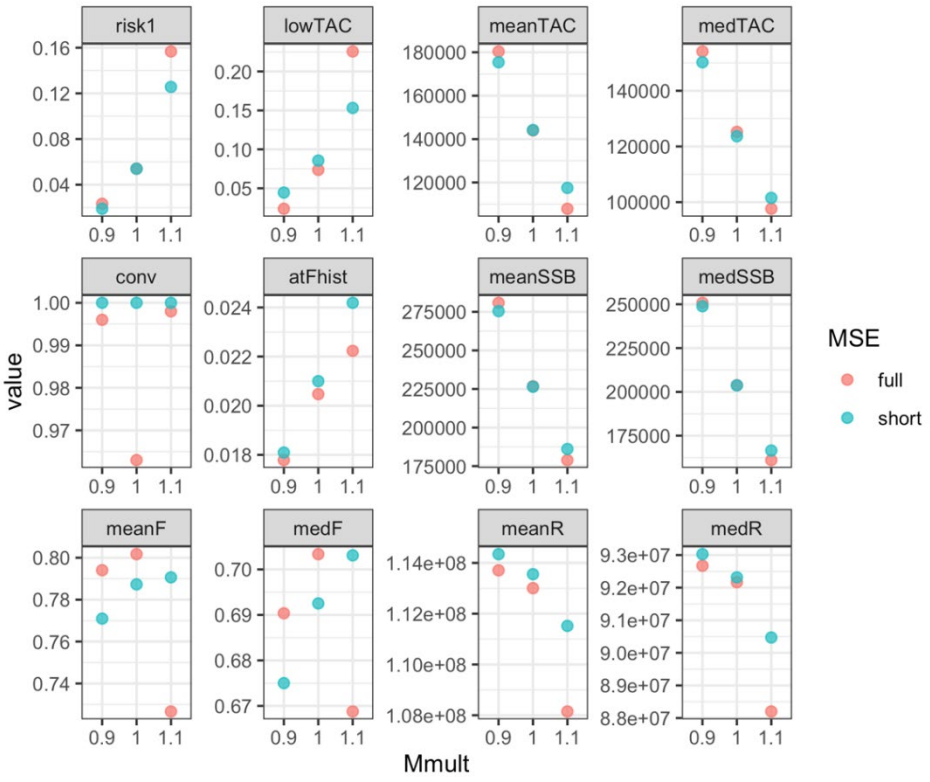

Figure 6.1.1. The x-axis represents bias in the observed natural mortality sent to the management procedure (90%, 100% or 110% of the true value used in the operating model). Each panel is a different performance statistic produced for the 2018 benchmark.

No attempt was made to pass biological parameters with error into the assessment emulator (i.e. variance-covariance matrix of lognormal error) in the shortcut version because they are not explicitly part of the emulator. In the future, it may be possible to improve the shortcut version's ability to match the full version if it used a variance-covariance matrix for log(N) and log(E) that comes from rerunning SMS with input that matches the input to the SMS run in the full version (i.e. input biological parameters with observation error).

Summary of discussion

Convergence of SMS did seem to be an issue in the full approach. Not all SMS runs converged. It was commented that the runs that are not converging could have a common structure, so that by excluding them, the MSE results become biased. What one would do in practice when an assessment is not converging is some intervention to make it converge. In a simulation context one cannot easily do that. That may mean that one is testing a procedure that is different from reality. Non-convergence may be a general concern for complex assessment models within the simulations, especially if there is some underlying property of the simulation dynamics which affects convergence, and lack of convergence is not randomly distributed among the simulations, which could lead to biased results. This may even happen in biomass dynamic models, if the OM is sufficiently flexible. Convergence should be monitored and a warning be raised if more than a very small number of runs do not converge.

In the presentation of MSE results, it is important to show a number of worm plots to be able to gauge how the feedback is working.

6.2 North Sea cod MSE example

Presentation by Simon Fischer (see Annex 7 for further details of the analysis)

Recently, ICES conducted a workshop to evaluate long-term management strategies for several North Sea fish stocks (WKNSMSE; ICES, 2019b). This evaluation was done with a full MSE and included an analytical stock assessment model (SAM) and a stochastic short-term forecast within the MSE feedback loop. For this meeting, the analysis of North Sea cod from WKNSMSE was repeated with a shortcut MSE approach, where the stock assessment was approximated. This allowed a direct comparison of the results from a full MSE to a shortcut MSE. For the baseline operating model (OM), the shortcut approach yielded similar results. However, there was a small shift in the optimised harvest control rule (HCR) parameterisation, and the shortcut suggested a parameterisation which was considered non-precautionary in the full MSE. The shortcut MSE approximation was crucially dependent on the level of stock assessment error, whereas autocorrelation in the assessment errors had a minor influence. The exploration of alternative OMs indicated a substantial bias in the evaluation of the HCR between the full MSE and the shortcut approach, related to the mismatch between the OM and assessment model. The conclusion from this exercise was that shortcut MSE approaches cannot entirely replace full MSEs, but they can complement one another, and a hybrid approach could be considered. In such a hybrid approach, a shortcut MSE can be useful for initial explorations of the search space, reducing the computational complexity. However, outcomes should be confirmed with a full MSE, which is particularly important for alternative OMs.

Summary of discussion

The method for how to derive uncertainty and bias estimates to be applied in a shortcut approach was questioned. There may be a large discrepancy between the uncertainty and bias that is derived from the analytical retro, and what has been seen happening in the working groups. The latter is due to changes in models, including or excluding data and fixing certain problems during working group meetings. There were different views about what was an appropriate characterisation of the approximation for the shortcut approach, with one view more closely aligned with the MP approach (and therefore relying on an analytical retro for this characterisation), while the other view did not put the same emphasis on specifying the monitoring data and assessment method used in the MP (and therefore relied on a historical retro reflecting changes in

monitoring data and assessment methodology over time). These alternative views reflect the different interpretations of the shortcut approach (see introduction to this section). However, care should be taken that comparisons between the full and shortcut approaches are made under the same conditions and assumptions (which was the framework adopted in this presentation, where the first interpretation of a shortcut approach was used, consistent with the full approach). It should also be noted that a basic rule for a full MSE is that if there are any such changes to monitoring data and assessment methodology as noted above, then the full MSE should be updated (although there may be cases where differences are considered minor, so a decision taken not to update the MSE).

When simulating future projections, the generation of future "data" needs to incorporate observation error, e.g. the error about the relationship between a survey index of abundance and the underlying resource biomass. It is important that this generation process produces realistic variation in these data. Often, for example, such errors are assumed to be independent from one year to the next, whereas in reality they are positively correlated; failure to take that into account will lead to over-optimistic inferences about how well the assessment method will perform in practice. However, in developing statistical models of errors (residuals) in fits to historical data to use for such projections, care must be taken not to confuse what is in fact a systematic pattern in recent residuals with positive autocorrelation. The latter will simply diminish the information content of data but will not otherwise bias results. However, if such a pattern is instead indicating model-misspecification, either the associated bias must be modelled explicitly when projecting forward, or the assessment model needs to be modified to remove that trend as the related consequences for inferences about assessment model performance could be quite different from those arising from autocorrelation.

6.3 Alternative HCRs for blue whiting using hindcasting

Presentation by Martin Pastoors

In 2019, the Pelagic Advisory Council (PELAC) commissioned a hindcast evaluation for blue whiting to assess the potential implications that different types of harvest control rules would have had given the observed dynamics of the stock. A simulation framework was developed in R using FLR (Kell *et al.*, 2007) designed to build simulation models representing alternative hypotheses about stock and fishery dynamics. An Operating Model (OM) was developed to run simulations of the stock under the different HCRs. The OM was conditioned on the current ICES stock assessment (ICES, 2019e). A Beverton and Holt stock recruitment relationship with a steepness of 0.9 was assumed so that simulated recruitments were similar to those observed historically but if the stock crashed, recruitment would also be impaired.

Two HCRs were implemented and simulation tested (HCR1: The Standard ICES MSY rule using an $F_{target} = F_{MSY} = 0.32$ and a $B_{trigger} = $ MSY $B_{trigger} = 2.25$ Mt, and HCR2: a two-tier approach with two levels of target fishing mortality and two biomass triggers). Both scenarios were executed with and without a stability mechanism of 20% down and 25% up when the stock is assessed to be above $B_{trigger} = $ MSY $B_{trigger}$. Simulations started in the initial year (2000) and then the stock was projected forward using either of the two alternative HCRs and with or without bounding the variability in TACs. Uncertainty in stock assessments was taken at 0.3, derived from the retrospective analysis of the SAM assessment.

Overall, the two-tier HCR2 performed similarly to the standard HCR1, the main difference being the additional level of safety provided by HCR2, which reduced F and catch at low biomass, i.e. in 2010–2015. Introducing bounds on the amount of change in TACs if the stock is above MSY $B_{trigger}$ did lead to stock collapses, as the large reductions in stock biomass seen were driven by

poor recruitment and the bounds resulted in F not being reduced quickly enough. In addition, the bounds prevented the TAC from being increased substantially as the stock recovered. A deterministic example of the workings of the TAC bounds showed that in 2010 the stock was still estimated above MSY $B_{trigger}$, and therefore the bound on TAC decrease applied. This resulted in a high fishing mortality for that year. The next year, the stock was below MSY $B_{trigger}$, so the bounds no longer applied and the TAC was reduced substantially. In 2012, the stock was again above MSY $B_{trigger}$ but because the bounds applied again, the catches remained low for a number of years. This demonstrates that the use of bounds in mitigating changes in TACs may have counter-intuitive and unwanted consequences.

The results of the hindcast analysis was welcomed by the stakeholder in the PELAC, because they found the results intuitive to understand. Hindcast exercises are easier to explain than general MSE results.

Summary of discussion

The issue of a precise definition of recovery was raised, which needs to be addressed when evaluating recovery plans using MSE.

Sablefish in Canada currently has an asymmetric meta rule, where if the management procedure indicates the catch limit should be reduced, then TAC is reduced to that level. However, if the MP indicates that the catch limit should increase, the magnitude of the increase must be greater than a threshold before the TAC is increased.

One may need a higher bound on the extent of TAC change on increases compared to decreases to deal with a downward ratchet effect (e.g. 100×1.1×0.9≠100). Recent Special Requests (e.g. ICES, 2019b) have dealt with this issue by requesting upward constraints of 25% and downward constraints of 20% (e.g. 100×1.25×0.8=100). However, this depends on the HCR being used, because the feedback control can take care of the ratchet effect; this is because the ratio of up vs. down is not necessarily 50:50. Furthermore, one should be careful with %-based stability constraints: if the TAC drops very low, different constraints would be required in that domain to allow for reasonable increases in the TAC as the stock recovers.

6.4 Using shortcut MSEs to evaluate horse mackerel rebuilding plans

Presentation by Martin Pastoors

The western horse mackerel stock has been hovering just above B_{lim} for a number of years. The PELAC has commissioned work to develop and evaluate a potential rebuilding plan for this stock. This resulted in HCR analyses based on two different assessment methods (SS3 and SAM) and two different HCR evaluation tools (EqSim-based and SAM HCR). The EqSim simulator is an extended version of the SimpSIM approach that was used for the blue whiting MSE in 2016 (ICES, 2016c). The code was further developed by Andrew Campbell and Martin Pastoors to improve standardisation, documentation and visualization of results. The EqSim simulator makes use of an Operating Model (OM) and a Management Procedure (MP). The SAM HCR forecast is a simple stochastic forecast with an HCR to evaluate management for fish stocks that need rebuilding in the short-term. The stochastic forecasts begin at the estimated current level of the stock, i.e. the assessment estimates currently used for tactical management advice, with consideration of the uncertainty in these estimates. Rebuilding is evaluated forward for a specified number of years and for different target fishing mortality values. Both HCR evaluation tools can

be considered a type of 'shortcut' with appropriate conditioning of the uncertainties in the assessment based on historical assessment and forecast CV and autocorrelation.

Three different types of harvest control rules were evaluated:

- Constant F strategy: fixed F_{target} independent of biomass level
- ICES Advice Rule: breakpoint at $B_{trigger}$ and linear decline in F to zero below $B_{trigger}$.
- Double Breakpoint rule: breakpoint at $B_{trigger}$ and straight decline in F to 20% of F_{target} at B_{lim}. Below B_{lim} continued fishing at F = 0.2 * F_{target}.

Given that the EqSim simulator with SS3 evaluation is closest to the ICES advisory practice, this was used as the basis for the preferred rebuilding plan by the PELAC. The PELAC-preferred options are:

- Target fishing mortality at $F_{target} = F_{MSY}$ = 0.074 (approximated by 0.075 in the simulations)
- B_{lim} at ICES B_{lim} (834 480 t)
- $B_{trigger}$ at ICES MSY $B_{trigger}$ (1 168 272 t)
- Double breakpoint rule with 20% constraint on interannual variability in TAC (IAV) above $B_{trigger}$
- Minimum F when stock is below B_{lim} at 20% of F_{MSY} = 0.015

The selected rebuilding plan had a 50% probability of rebuilding to B_{lim} by 2021 (similar to zero catch option) and a 50% probability of rebuilding to B_{pa} = MSY $B_{trigger}$ by 2024 (similar to the zero-catch option).

This work demonstrated that the uptake of the results of an MSE (in this case a rebuilding plan) is greatly facilitated by close interaction with stakeholders in a way that they understand the trade-offs that can be made.

Summary of discussion

No discussion points were raised with this topic.

6.5 Using Muppet to compare full and shortcut approaches

Presentation by Höskuldur Björnsson (see Annex 8 for further details of analyses)

Muppet is a stock assessment and prediction model written at the MFRI in Iceland. The model is written in ADMB and is designed for stock assessment, short-term prediction and management strategy evaluations using the shortcut method.

The model can be run both as a separable model (selection allowed to change at prespecified periods) or a VPA model. The VPA option is useful for testing CVs by age for surveys, year factors in surveys, etc., where a catch at age model would set one of the estimated standard deviations to zero, something that a normal VPA cannot do because it has already set the standard deviation of observation error in catch at age to zero. Therefore, the VPA option can be useful as an operating model.

The Muppet model was changed somewhat to include the possibility of using a "closed-loop" approach.

What is done is:

1. Run mcmc simulations of the assessment model saving every n^{th} iteration. From this step, a set of parameters for stochastic simulations is generated.

2. Stochastic simulations run with the parameters generated in step 1. Stochasticity in future recruitment and biological parameters added.

Muppet has, until now, been run by adding autocorrelated lognormal assessment error to the stock biomass estimated by the (operating) model and basing next year's advice (decision rule) on this perturbed estimate of stock biomass, in some cases predicted one year ahead. In the closed-loop approach, a new "harvest control rule" is added where the advice for each year is based on deterministic simulations using an F multiplier or Harvest Rate. The model has combined a deterministic short-term prognosis with the assessment since it was developed in 2002, and an assessment without prognosis is not allowed in the model.

For many Icelandic stocks, the biomass at the beginning of the assessment year is the basis for the advised TAC in the following year. Those decision rules are simpler to simulate because a short-term prediction is not needed in simulating the decision rule. Furthermore, changes in weight and maturity at age are automatically taken care of. A TAC developed from the biomass-based decision rules does not depend on the selection patterns for the fisheries, making the rules robust (although not completely insensitive) to relatively large change in such selection.

Two test cases, where the shortcut and full methods are compared, are presented in Annex 8: NEA mackerel and Icelandic cod. The assessment of Icelandic cod is relatively precise and based on a long time series of data (survey series of 36 and 25 years long, and reasonably well standardised), but the NEA mackerel assessment is relatively uncertain, with relatively short time series that could have trends included. The CV of the estimated spawning stock for Icelandic cod is 7% in 2020 and the CV of reference biomass, 5.5%. For mackerel, the estimated CV of the spawning stock in 2020 is 16% if tagging data are not included, but 12% if tagging data are included. The spawning stock forms a much larger proportion of the mackerel stock than of the cod stock. For both of those stocks, the uncertainty estimated by the assessment model appears to be too low and reduced over time in the full MSE, because tuning series generated on the basis of the likelihood function in the assessment alone may be too "well-behaved", and are continued in the future.

The most important result is that the shortcut and full methods lead to comparable results for the same operating model. There will always be some differences, but this is not precision science. Using an analytical retro to reflect the assessment error for the shortcut approach is a relatively precautionary measure, as the series on which retrospective runs are based are shorter than what the current assessment is based on. For the two stocks in Annex 8, the choice of operating model has a greater influence than use of the shortcut or full approaches. For the shortcut approach, the settings of the assessment error do not have a major effect on long term results, and there is no bias.

For Icelandic cod, estimated uncertainty in the assessment becomes too low in the future simulations. Combinations of three long, high-quality data series and low harvest rate imply that each stock assessment is based on many data points. It could be argued that the assessment has too much inertia because data that are 5–7 years old have considerable weight. Variability in M might be a factor; there are no indications of much variability in M for adult cod, but there must be some, and including that might lead to the quality of values deteriorating faster. An interesting observation supporting that M is close to constant for adult cod is that VPA runs for Icelandic cod estimate lowest CVs for ages 2, 6 and 7 in the March survey (similar results for the autumn survey). Low CV of age 2 indicates relatively low variability in M.

In the shortcut simulations, the short-term calculations must be appropriately formulated. In Muppet, the estimated stock biomass and TAC from the current assessment are input to the model. Iterations where the stock biomass is small start with overestimation (positive assessment error) and *vice versa*, increasing risk in the short term compared to not linking the start of

assessment error to the stock size in each iteration. A full analysis should automatically take care of this problem. One difference between the shortcut and full approaches in the work presented is that assessment error decreases gradually with time in the full MSE, where all sampling is assumed to continue unchanged for few decades.

Inertia in the assessment is demonstrated in the results in Figure A.8.12 (Annex 8), both from autocorrelation and bias. Many of the virtual future worlds are biased, and the one we live in is biased (retrospective bias is 6%). Therefore, bias in the empirical/analytical retrospective pattern based on 30 years is not unexpected. Bias in the retrospective pattern used to specify assessment error in the shortcut approach is not a problem, because it is taken into account by reducing target fishing mortality. But the average over all the future worlds is not biased. This relates back to the problem of whether bias in analytical or empirical retros should be included in the shortcut approach, a question that was answered above. It is not possible to prove that observed bias in retrospective pattern is just autocorrelation and analysts must act in accordance with that, taking it into account.

The observation model for Icelandic cod needs to be revisited, because the age disaggregated non-linearities are not sufficient to lead to sufficient contrast in the data (Annex 8, Figure A.8.11), and the relationship between uncertainty and stock numbers should be investigated further (low values include substantial sampling error, lognormal with the offset used in Muppet). Furthermore, temporal correlation in the surveys and correlation between surveys might have to be included if realistic values for uncertainty are to be obtained. One possible way to do this is to introduce temporal variations in M; this helps improve model fit, but the cost is that it can be difficult to fit operating model and assessment model together. The same considerations apply for mackerel.

In the shortcut approach, the observation model and assessment model are considered in combination when using the term "assessment error"; this is relatively easy to estimate if one considers that a converged assessment is true. On the other hand, separation of the assessment error into a wrong biological model (temporally varying M), a wrong observation model, and observation residuals, is difficult.

To summarize the results, the shortcut method is sufficient for stocks with long timeseries of age disaggregated data. A limited full approach can be useful to test settings regarding assessment error. Long series of age-disaggregated data do usually give a reasonable picture of what is happening with the stock, and important factors of the operating model, such as the stock-recruitment function (as it has limited effect on the historical assessment) and density dependent growth (which does not affect the historical assessment) can easily be tested in the shortcut approach (ICES 2010; Björnsson, 2013).

Where the full approach is useful is where operating and assessment models differ. Muppet is quite flexible in potentially having a different structure for the operating and assessment models. A limitation is that all data series included in the assessment model must be included in the operating model (the observation model can be different) so that future data for all series can be generated. Reducing the weight of data series in the operating model is a possibility for a series that is not intended to affect the operating model but will be used in the assessment model.

And finally, those that undertake management strategy evaluations have to look at the data. If the data do not tell indicate approximately what is happening in reality, then setting up an MSE will be challenging.

Summary of discussion

[Discussion not captured by rapporteur notes.]

6.6 Conclusions

ICES clients frequently request that the generalised HCRs which are used to provide TAC advice be tested to confirm that they respect certain requirements, such as being consistent with the ICES precautionary approach and capable of achieving catches corresponding to F_{MSY}. Several examples of such evaluations have been presented in this section of the report (see also Section 7).

With regard to evaluation tools, a distinction can be made between two different approaches: shortcut and full. Shortcut approaches refer to evaluation tools where the assessment and advice processes are emulated on the basis of past uncertainties and biases, and can be based on an analytical retro (i.e. using the current expert group assessment and associated data, according to the stock annex) or on a historical retro (i.e. comparing how advice has been produced in the past, including any changes to the assessment method, with the current expert group assessment). A key issue with the shortcut is that it requires a demonstration that the emulator is working as expected (i.e. that it adequately approximates the behaviour of the assessment and advice processes it is trying to emulate). This is important because the choice of stock and fishery monitoring data, the assessment method, and the harvest control rule interact in ways that cannot easily be predicted in advance. Full approaches are simulation tools where the assessment and forecast methods currently used by the expert group for advice are included as the estimator and advice generator in the simulation approach; the assumption is that the current assessment and forecast methods will continue to be used in the near-future for advice, and any changes to this (e.g. via a benchmark or inter-benchmark) would potentially require the full MSE to be rerun. In the full approach, the observation errors that are used to generate simulated inputs to the assessment model in the MP are based on the fits to historical data of the model used in the OM, and there is an underlying assumption that the once these observation errors are imposed, the assessment model will perform similarly on the simulated data to how it performs in reality, if reality was reflected by that OM. A key issue is that what is being tested is what will be implemented in practice.

There may be tuning of some control parameters of the HCR (such as F_{target} and $B_{trigger}$) in the simulation evaluation process; this is to exclude unacceptable combinations of values of those parameters, and to provide a basis for selection amongst other combinations based on trade-offs in performance (e.g. average catch vs TAC variability from year to year). There may even be tuning of the assessment method used in the MP itself, such as the choice of Bayes priors or size of process vs. observation error.

Both shortcut (ICES, 2005a; 2005b; 2015a; 2015b; 2016d) and full (ICES, 2017b; 2019b; 2020a) approaches have been applied in the past by ICES when evaluating HCRs and management strategies. As noted in WKGMSE2, historically most evaluations within ICES have been using shortcut approaches; however, as available computing power has increased, together with an improved understanding of the MSE process, there have been more full approach evaluations in recent years.

Analyses presented during the workshop (see e.g. Sections 6.1, 6.2 and 6.5) indicate that the shortcut and full approaches can lead to similar future SSB and F trajectories for the same values of control parameters (see A.7.2 and A.8.9 in Annexes 7 and 8, respectively). However, when considering performance statistics, some differences may arise in the combinations of control parameter values when tuned to achieve identical objectives, partially because of the different ways in which uncertainties and biases have been included. In the case of the NS cod example for the baseline OM (Section 6.2), the optimised combination under the shortcut approach was not precautionary under the full approach (i.e. Prob3>5%), when tested under the same conditions (i.e. accepting that the current assessment is used to characterise assessment behaviour for

both approaches). Differences were larger under alternative OMs, which are difficult to capture in the shortcut approach.

The pros and cons of the two approaches are listed below, along with some recommendations to deal with the cons.

Full method:

Pros:

- Tests the management procedure that will be implemented in practice.
- Conceptually straightforward to include multiple OMs.
- May generate plausible patterns in future assessments that were not seen in the assessment in its estimates for past years.

Cons:

- May still contain some approximations (e.g. GAM indices not being updated, some data that are "too difficult" to generate may be omitted), which means one may not be testing exactly what is being implemented.
- Computationally intensive (when including an analytical assessment in the MP), thereby limiting utility, especially for requests that require quick answers.
- Tendency to focus on the technical challenges of implementing the complex assessment model in the evaluation framework. Less time therefore spent examining HCR variants.
- Requires case-by-case code development, although generic code elements can be used. May be difficult to quality control all the code being used for the evaluation.
- Convergence of the assessment model in the simulation may be an issue.

Shortcut method:

Pros:

- Might be able to be standardized across multiple situations (e.g. standardized tools such as HCS [Skagen, 2015], EqSim/SimpSim [ICES, 2016d; Pastoors *et al*., 2020], Prost [Åsnes, 2005], etc.); there is the possibility to quality control the software.
- Can take historical performance of the management process into account (WG/advice system) under the second interpretation of shortcut.
- Quick to run, therefore more focus on alternative HCRs, or on the details of the stock dynamics and/or wider ecosystem interactions, becomes possible. Less resource intensive (human as well as computer) and can thus be run on more stocks where the resources for a full MSE may not be available.

Cons:

- Not testing the management procedure (i.e. advice process) that will be implemented in practice.
- Assessment behaviour (e.g. positive assessment errors which tend to be larger than negative assessment errors; lags; retrospective patterns in the assessment model outputs) and bias can be difficult to characterise in an emulator, which attempts to reflect the performance of a more complex method and data generation adequately, and requires testing to confirm this.
- Difficult to test MPs against alternative OMs because of the need to take account of how the information from the OM would end up in the MP (i.e. a more complex emulator, and therefore more difficult to develop, will be required).

Recommendations to deal with cons:

There was a divergence of views regarding the remedies to deal with the cons of the two approaches. Therefore, these divergent views are listed separately below as view 1 and view 2.

Full method

View 1:

Requesters of MSEs conducted through ICES expect that the monitoring data and assessment used to provide advice (as stipulated in the stock annex following a benchmark or inter-benchmark) will be used to parameterise the management strategy being evaluated if and when it is implemented. The full method acknowledges this as a given, and attempts to evaluate what will be implemented in practice, even if computationally onerous. What is evaluated in the full approach includes data collection schemes, the specific analyses applied to those data and the harvest control rules used to determine management actions based on the results of those analyses; this follows MSE best practice in testing exactly or very closely what is planned to be implemented, and is widely considered to be the most appropriate way to assess the consequences of uncertainty for achieving management goals (Punt *et al.*, 2016). A basic rule of the full approach is that if data inputs to the assessment, or the assessment itself, change in any way (such as would necessitate an update of the stock annex), then the full MSE would need to be updated. It is important to set up the observation error model carefully so that it generates data with similar statistical properties to those seen in the past (see ICES, 2019a, for guidelines on this), otherwise the full approach is not being implemented properly.

The full approach should therefore ensure that what is being implemented is what will be tested, and if this is not possible, then this should be made clear and alternatives proposed, such as:

- Testing (and then implementing) empirical MPs
- Testing (and then implementing) model-based MPs with simpler estimators

These alternatives may require changes to the management process (see Section 7).

View 2:

Requesters of evaluation of HCRs conducted through ICES expect that the parameters of the HCR are being calculated in line with the current understanding of stock dynamics and reference points. This requires an appropriate conditioning of uncertainty and bias. Requesters have – until very recently – never specified that they required a full analysis which included the (complex) assessment model as part of the simulation loop (see Section 7.1).

The full method does not capture the annual human intervention leading to adjustment of assessment models that may occur within expert groups, and also does not capture changes that occur over time, including during benchmarks and inter-benchmarks. The observation error model may also not adequately generate data with similar statistical properties as seen in the past. This could, however, be addressed by carrying out hindcast analysis using the full approach.

The major challenges with the full method within the ICES context (i.e. HCR evaluations with relatively complex assessment models) relate to the resource intensive nature of the simulations, the tendency to focus on the technical challenges of implementing the complex assessment model in the evaluation framework, and the case-by-case code development. This could be addressed by using full simulations more for strategic evaluations of generic management approaches and

less for operational advice generation. It was agreed that developing full methods with simpler assessment methods or empirical MPs would be advantageous.

Shortcut method

View 1:

The shortcut approach (middle plot of Figure 6.0.1) attempts to emulate the assessment and advice process (in the left-most plot of Figure 6.0.1, this refers to the estimation model, which may include a forecast), and does this by characterising assessment behaviour through analytical retrospective analyses and/or analyses of estimation error and comparing these to the expert group assessment currently in use (first interpretation of shortcut), or through historical retrospective analyses (second interpretation of shortcut). There are two main challenges with this approach. The first challenge is that it is harder to demonstrate that the shortcut method provides an adequate emulation of the intended process than the full method, which includes the stock and fishery monitoring data, assessment method and HCR intended for application. It is imperative that there be a demonstration that the shortcut emulator adequately captures the behaviour of the process it is attempting to approximate. Punt *et al*. (2016) note that although a failure to simulate application of the actual assessment method allows a broader set of hypotheses to be explored quickly, the risk is that the actual error distribution associated with assessments does not match that assumed, and hence the values calculated for the performance statistics are incorrect. In the extreme, the resultant relative ranking of MPs may be rendered incorrect, as was demonstrated in the sprat and North Sea cod example (Sections 6.1 and 6.2) and by ICES (2008b). Punt *et al*. (2016) suggest that the justification for using an approximation to an MP may be examined by running a few simulations for the full MP and the approximation, and comparing the results to ascertain whether the approximation is adequate; this is part of the current ICES guidelines for MSE (ICES, 2019a). The second challenge is dealing with alternative OMs. The current shortcut approach of characterising the behaviour of the estimator through retrospective analyses relies on comparisons with the expert group assessment in current use. The difficulty then is in characterising the behaviour of the estimator under alternative OMs, where it is no longer possible to make use of retrospective analyses. These two challenges lead to the following recommendations:

- The shortcut approach can be considered valid only if it has been demonstrated to provide an adequate approximation to the process it is trying to emulate. Therefore, develop techniques (e.g. machine learning) to demonstrate that the shortcut emulator produces estimates of stock status for use in the HCR that are of very similar size and error structure as those produced by the estimator during a corresponding full MSE (whichever interpretation of shortcut is used).
- Given alternative OMs, the shortcut emulator needs to be checked to confirm that it provides an adequate approximation to behaviour were the current management approach to be applied to each. The emulator cannot simply mimic the OM, as that is not its purpose; it should be emulating the estimator that will be used in practice, and how this estimator reacts to alternative OMs.

An alternative would be that one focusses rather on evaluating the performance of simpler MPs, as outlined in Section 7.2.

View 2:

The shortcut approach (second interpretation, where it is considered a method for carrying out more generic HCR evaluations) is a valid method in its own right, has been widely used by ICES

in the past to evaluate management strategies (see Section 7.1), and should not be viewed as somehow inferior to the full MSE approach. Rather it is a case of where one chooses to focus limited resources – this shortcut approach allows for evaluations to be conducted with less resources, and/or allows more focus on aspects other than the assessment model (e.g. alternate HCRs or more realistic dynamics in the OM). However, there is room for improvement, as reflected in the following recommendations:

- Focus on standardised methods and tools to generate parameters to mimic working group and/or assessment behaviour and bias.
- In order to deal with alternative OMs, the emulator should not simply mimic the OM, as that is not its purpose. The emulator should be emulating how the estimator would perform under alternative OMs. This could be based on generic studies of shortcut and full methods that provide insights into the types of biases that may be expected in certain situations.
- Develop generic techniques to demonstrate under which conditions and assumptions a shortcut emulator produces estimates of stock status that are of the same order and error structure as the estimator it is trying to emulate.

Possible way to deal with divergent views:

Having divergent views, as expressed above, without trying to resolve them in some way does not move the group forward. A possible resolution may be in the form of a recommendation to investigate when performance of an HCR/management strategy is not as intended (as indicated by the results from the simulations) under either the full or shortcut methods (including alternative interpretations of the latter) in MSEs that include a range of alternative operating models.

7 The use of MSE in the NE Atlantic

This section provides some reflections authored by different members of the Working Group on the use of MSE in the ICES context, both in the past and a potentially different way to use it in the future.

7.1 Some reflections on MSE in the context of the ICES advisory process

The use of MSE-type approaches in the context of the ICES advisory process commenced in the late 1990s and early 2000s when ICES started giving advice on the precautionary nature of harvest control rules (HCRs; Patterson *et al.*, 1997; ICES, 2000; 2005a). The scientific advisors (ICES) and the clients of the scientific advice (fishery managers) have gradually become educated in and accustomed to an HCR evaluation process, which in most cases resulted in a binary conclusion (an HCR was either in accordance with or not in accordance with the precautionary approach). In those early years, there was, in most cases, only a limited connection between the evaluation of an HCR and the stock assessment methods that were used in providing the annual advice.

Over time, the process of carrying out HCR evaluations has developed in terms of scope and transparency. Instead of only coming up with binary results, the HCR evaluations have become explorations of potential F_{target} and $B_{trigger}$ values conditional on a number of constraints, such as Interannual Variability in TAC (IAV) and banking and borrowing (ICES, 2008a; 2012; 2014c; Skagen and Miller, 2013). This is when the F_{target}–$B_{trigger}$ tables were introduced as a mechanism to summarise the results of the evaluations. However, in the requests to ICES up until 2017, the clients did not mention the assessment method to be used in the evaluation or in the subsequent process of generating annual advice (nevertheless, it should be noted that since the introduction of the benchmark system in 2009, there has always been an understanding among the clients that the advice will be generated from the benchmark assessment, and this is what they expect to be simulation tested). Clients were asking for an evaluation of the parameters of a harvest control rule (and whether it was in conformity with the precautionary approach); they were arguably not asking for an MSE in the stricter sense as implied by the IWC or tuna commission, where MSE is seen as testing pre-specified components (Table 1.0.1), including their robustness to uncertainty. However, the requests to ICES have become more and more elaborate as the clients and stakeholders have become 'educated' in the language of management strategy evaluations. As an example, the requests for HCR/management strategy evaluations for NEA mackerel from 2007-2020 are in Annex 9. Furthermore, the requests for evaluations have also become more frequent. Often, when either the assessment method and/or reference points were changed in a benchmark process, there would be a request to evaluate the harvest control rule in the context of the new assessment or reference points. HCR evaluations have mostly been perceived as a technical exercise that one needs to go through in order for an HCR to be 'vetted' by the scientists, and once that is done, one can simply apply it on the basis of the annual advice that is produced by ICES using the best possible assessment method. This is consistent with the ICES approach of trying to provide the best possible assessment for each stock and for every year, going through a benchmark process if some issues are detected with a particular assessment result. The advisory system in ICES is supporting and promoting the concept of 'best possible assessment'

through a continued emphasis on improving the data and assessment methods for providing scientific advice.

In the 'best assessment' paradigm, the role of HCR evaluations is mainly to fulfil the requirements of the precautionary approach and the MSY approach (ICES, 2019c) so that HCRs can be implemented that can be expected to be in line with those objectives. In order to do the evaluations, one needs to be able to mimic the capacity within the advisory system to come up with annual advice, but when conducting HCR evaluations (as opposed to MSEs that deliver MPs; Table 1.0.1) there is no need to actually run the preferred assessment model as part of the evaluation process. As long as the uncertainties and biases that have been seen in the assessment process in the past are adequately represented in any future simulations and that the general approach to the assessment stays the same, this is sufficient for the purpose of the evaluation. The best assessment model will then be run to calculate the annual values in relation to the agreed HCR.

In contrast to the 'best assessment paradigm', one could use the 'MSE paradigm' to describe the approach where management strategy evaluations are essentially about simulation-testing prespecified components of the management strategy, and the robustness of this management strategy to prevailing uncertainties (Rademeyer *et al*., 2007; Punt *et al*., 2016). In that approach, it does not make much sense to invest a lot of time and energy in trying to find the absolute best assessment approach and to apply that every year to all the stocks. Rather, it seems more sensible to look for a robust assessment and management approach that would work under a wide range of plausible scenarios of how the fisheries system works. It also does not really make sense to limit a full MSE to one with a complex OM that is derived from a complex assessment method and then apply that same complex assessment method as the basis of management decisions in the MP. What would potentially be far more useful than this (and have comparable performance) is a management strategy that is based on much simpler models or indicators on which to manage the fishery on a year by year basis.

After WKGMSE2 (ICES, 2019a) and the applications in WKNSMSE (ICES, 2019b) and WKMACMSE (ICES, 2020a), a hybrid system between the best assessment paradigm and the MSE paradigm has developed. That is a full MSE approach in which both the baseline operating model and the management procedure are based on the same, highly complex assessment model. In practice, this means that a lot of time and effort is spent on getting the assessment model to run as part of the management procedure. We are facing serious challenges in the amount of time, and computing capacity required to run the evaluations, but also in understanding the behaviour of the assessment model itself, given the data that we are feeding it from the OM. Because the assessment models are rather complex, it is not clear whether the behaviour that we observe in the MP are simply artefacts from the way the model has been parameterised or that it is some real behaviour that may be anticipated in the management process (Figure 7.1.1). The hybrid MSE approach is arguably not necessarily better or more scientifically sound than the simpler HCR evaluation methods that have been applied in the past (2000–2018). It is simply different.

If one would want the 'pure' MSE paradigm that is more feasible to implement considered and accepted in the ICES context of providing advice to management (see Section 7.2), a demonstration of the implications (benefits, drawbacks) of such an approach to the management of stocks would be required. This could entail a clear demonstration of the reduction in the amount of effort spent on the annual assessment and advice process while maintaining robustness against uncertainties. A fully worked out scenario would be required. So instead of mainly looking at the trade-off between F_{target} and $B_{trigger}$, one would be looking at the robustness of the management system to potential unknowns in the real system. By developing such a demonstration project, in consultation with managers, stakeholders and scientists, we could raise the awareness of the benefits of redirecting the current focus on very detailed annual advice (with many digits behind

the comma) towards simpler management approaches. The examples from the management approach used in Iceland could provide inspiration.

On the other hand, if we also aim to directly improve the current management and advice system, we could improve the benchmark system by including full MSE approaches for testing whether changes in the assessment methodology could be an improvement over the current methods and to improve the process of quantifying uncertainty and bias in the assessment and advice process of the past. So far, the evaluations using shortcut methods have underplayed the importance of uncertainty and bias, and this could be remedied by doing some analyses on assessments and their behaviour.

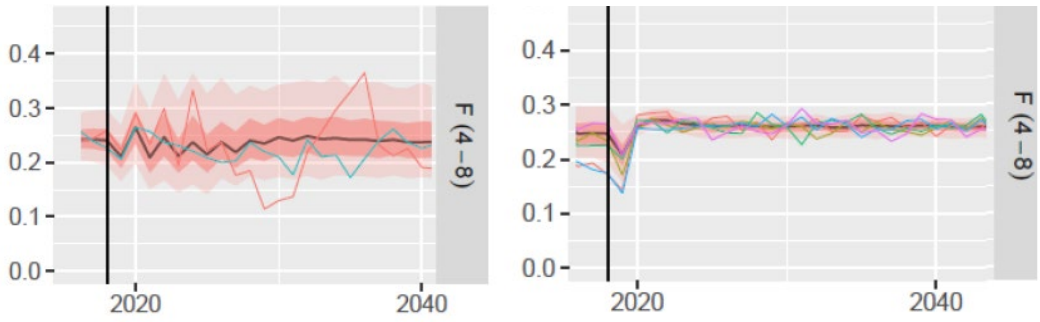

Figure 7.1.1. Simulations conducted including SAM as part of the MP (left) and excluding SAM by using a shortcut (right). Taken from the WKMSEMAC report (ICES, 2020a).

7.2 Some reflections on a potentially different way to use MSE in the management context of the NE Atlantic

The simulation testing framework that has developed over time in ICES, though termed MSE, has come to differ considerably from what is applied elsewhere in the world under the designation of "MSE" or the "Management Procedure (MP) approach" (see Section 7.1). The original development of that approach took place primarily in the IWC Scientific Committee in the 1980s, motivated by the realisation (implicit also for the ICES framework) that it was essential that methods being used to compute recommendations (such as TACs) for managing fisheries be simulation tested to check that they would behave in practice as intended, which could depend (*inter alia*) on the quality of the input data. As with typical fisheries assessments, this MP approach usually used those complex assessments to provide the testing framework (the Operating Models, OMs), but importantly did NOT require that the same assessment process was used in the algorithm applied to compute the management recommendation (the MP itself). Instead, this MP approach admitted considerably greater flexibility, including the following:

- The basis for selection amongst alternative MPs was purely their relative performances in meeting quantified forms of the management objectives set for the fishery.
- The calculation process used by the MP to determine the management recommendation did not need to be the same as the assessments (OMs) used to provide the testing framework.

- The values of the control parameters used for the harvest control rule component of the MPs did not need to be related to standard reference points (e.g. F_{MSY}, MSY $B_{trigger}$) for the assessments/OMs[1].

Of key importance was that particular stress was laid on checking that the performance of any MP was robust to the endemic uncertainties about the dynamics of the resource and fishery concerned, i.e. the MP needed to be shown by simulation to perform adequately not only if the best assessment reflected the underlying dynamics exactly, but also for other plausible alternatives (OMs; see Section 3). This simulation process is useful for evaluating the likely consequences of management choices, but is essential for eliminating MPs that are unlikely to perform well in actual application. Poor performers can be rejected from further consideration on the basis that they fail to meet objectives in simulation. Although a procedure that performs well in simulation is not guaranteed to have the same performance when applied to the real fishery, robustness to uncertain stock and fishery dynamics represented by OMs is gained through filtering out those MPs that do not adequately meet objectives in simulation.

Most implementations of this form of MSE/MP approach do not use complex assessment models as components of such MPs, but rather either use considerably simpler assessment models or work directly from the resource monitoring data themselves (e.g. abundances indices from surveys or CPUE) which are used as input to such "empirical" MPs.

Particular advantages that this MP approach offers over the present ICES framework include that it side-steps the problems associated with simulation testing complex assessment methods, such as non-convergence occurring for some of the replicates and/or extremely onerous computational time burdens. Of special importance is the matter of stakeholder buy-in. When, for example, the scientific advice is to change a TAC, stakeholders want to understand the reasons why before accepting this. That can be difficult to achieve when a complex method is used which operates as a "black-box" in generating an output value that can be difficult to explain. Buy-in is much more readily achieved for the simpler methods employed by many MPs, for which the reasons for the direction of any TAC change is typically easily explained to laypersons.

[1] However, performance statistics and/or associated targets would often be specified in terms of such reference points for the OM concerned. For example, a general objective of recovering the resource to an abundance capable of yielding maximum catch sustainably might be operationalised as seeking the median of the distribution of projected B/B_{MSY} in 20 years' time being close to 1 for each OM under consideration. Note that, for example, B_{MSY} is OM-specific, i.e. its estimated value differs depending on the OM being considered due to differences in assumptions and data applied in each OM.

8 ICES Workshops relevant to TORs

The outcomes of three recent ICES workshops that have relevance to WKGMSE3 were reported to the meeting, and edited executive summaries with accompanying discussions summarised below.

8.1 WKREBUILD (TORs a, c, e)

Edited executive summary of WKREBUILD report (ICES, 2020b) (presented by Martin Pastoors):

The Workshop on guidelines and methods for the evaluation of rebuilding plans (WKREBUILD) chaired by Vanessa Trijoulet (Denmark) and Martin Pastoors (Netherlands) met from 24 to 28 February 2020. The workshop attracted 27 participants from the US, Canada, Europe and FAO.

When stocks are estimated to be below B_{lim} and there is no perceived possibility of rebuilding above B_{lim} within the time-frame of a short-term forecast, ICES has regularly recommended zero catch in combination with the development of a rebuilding plan.

A review was carried out of the international experience regarding the development, evaluation and implementation of rebuilding plans for fisheries management in the Northeast Atlantic and in other fora around the world. In the Northeast Atlantic, rebuilding plans have been implemented in the past (e.g. the cod recovery plans of the early 2000s) but ICES has played a limited role in evaluating the performance of such recovery plans and does not currently have the tools or criteria to evaluate such plans. Recently, when a rebuilding plan for herring in 6.a and 7.bc was submitted to ICES for evaluation, ICES refrained from providing such an evaluation. In the US and Canadian approaches, the legal framework determines the triggering and required elements of re-building plans. Such a legal imperative does not exist in the Northeast Atlantic. Nevertheless, the US and Canadian experiences provided useful elements that could be included in establishing the ICES approach to rebuilding plans.

Several case studies were presented on potential tools for the evaluation of rebuilding plans. Particular attention was given to evaluating options for harvest control rules of such a plan. The tools focused mostly on short- to medium-term explorations of the probability of achieving rebuilding. Because rebuilding plan evaluations need to be ready and available at short-notice when required, it was concluded that relatively standardized tools (i.e. packages or compiled code) to carry out such evaluations would be preferable over custom-made evaluation tools. In addition, certain modelling considerations were highlighted as important, such as realistic assumptions of productivity, uncertainty, bias in assessments, and implementation error, and estimating the probability of achieving a rebuilding of stocks.

Criteria for the acceptability of rebuilding plans will require an agreed Limit Reference Point (LRP) for initiating a rebuilding plan, definition of targets for fishing mortality or stock biomass, time-frames, and the acceptable probabilities of whether the rebuilding targets have been achieved. All of these should take into account realistic levels of uncertainty and being consistent with international best (scientific) practices. Although it was recognized that B_{lim} would be the most likely candidate LRP triggering a rebuilding plan, the current approach in ICES for the determination of B_{lim} was questioned during the workshop, because it requires a subjective classification of the stock-recruitment pairs into different types. In other regions, the LRP is often set as a certain proportion of the SSB at maximum sustainable yield (B_{MSY}), e.g. 40% B_{MSY}. If changes in productivity have been experienced in recent years, and these are taken into account when

estimating MSY reference points, the proportion of B_{MSY} approach would likely lead to greater changes in the estimated value of LRP than the current ICES procedures used to estimate B_{lim}, which rely on stock-recruitment pairs or a definition of the lowest observed biomass (B_{loss}). This could have a large impact on the rebuilding target for stocks that experience changes in productivity regimes. Some concerns were raised regarding the often small distance between B_{lim} and MSY $B_{trigger}$ reference points for ICES stocks in comparison to the distance between a trigger and limit in other jurisdictions. MSY $B_{trigger}$ could therefore represent a late trigger to start decreasing fishing mortality when SSB is decreasing. The workshop recommended a future workshop on the revision of the procedure to estimate reference points within the ICES framework.

An estimate of the minimum time (T_{MIN}) by which rebuilding may be expected to be achieved, could be calculated by assuming zero catch and should be used as a baseline for comparison with other rebuilding scenarios. The maximum time for rebuilding in the US and New Zealand is set to $T_{MAX} = 2 \times T_{MIN}$ or to T_{MIN} plus one generation time (average length of time between when an individual is born and the birth of its offspring; NRC, 2014). While the workshop did not arrive at an overall agreement on a default value for T_{MAX}, it was suggested that $T_{MAX} = 2 \times T_{MIN}$ could be explored as a potential bounding on the rebuilding period, even though this should be subject to scientific analysis of potential effects on the stock in question.

The workshop generated a guidance table summarizing the best practices for evaluation of rebuilding plans against the potential criteria of acceptability. The guidance table includes elements such as estimation of reference points, time-frames for rebuilding, rebuilding targets, handling uncertainties and bias, probability of achieving rebuilding targets, and visualizing results. The workshop recommended that a follow-up workshop (WKREBUILD2) be organized for testing the guidelines with actual test cases, with the aim of defining more specific criteria and guidelines, i.e. learning by doing.

Some of the elements that were discussed in the workshop, but that have not (yet) entered the guidelines for evaluation of rebuilding plans, are socio-economic trade-offs (e.g. between fast and slow rebuilding), mixed fisheries aspects (e.g. unavoidable bycatch due to mixed fisheries), and elements in rebuilding plans other than the HCR part (e.g. monitoring to improve the knowledge base).

Most of the discussion at WKREBUILD was centred on stocks with analytical assessments (Categories 1 and 2). Identifying when a data limited stock is in need of rebuilding (or has rebuilt), and how to evaluate rebuilding plan options for such stocks, would likely require a separate process.

Summary of discussion

Rebuilding plans could form part of longer-term management strategies in the form of a rebuilding phase, but focus for rebuilding plans, or a rebuilding phase of a management strategy, should be short-term and encompass issues such as time to recovery. Canada has not yet formally established timelines. However, the interval T_{MIN} to $2*T_{MIN}$ has been adopted in New Zealand. T_{MIN} at least has the advantage of embedding understanding of stock dynamics and reflects current depletion.

8.2 WKRPChange (TOR a)

Summary of WKRPChange meeting (presented by Daniel Howell):

WKRPChange was held by WebEx 21–24 September 2020, chaired by Anna Rindorf, Jeremy Collie, and Daniel Howell. The meeting was tasked with examining how ICES handles the estimation of target and limit reference points in the face of changing environmental conditions. In particular, the meeting was asked to review the robustness of the current ICES procedures and to suggest specific improvements that could be made, especially relating to changes in stock productivity arising from environmental conditions, species interactions, and density-dependent effects. Part of the work involved reviewing the basis of the ICES reference points, and contrasting the ICES procedures with those in the USA and Canada, and part on providing specific guidance for future reference point estimation within ICES. The meeting also highlighted the recent work at WKIRISH6 (ICES, 2020c), which gave scope to "fine tune" the F_{target} to account for small changes in environmental drivers without requiring full re-estimation of the reference points.

One common approach to changing environmental conditions is to truncate data series. WKRPChange agreed that this may be necessary in some cases, but several studies were presented showing that the estimation of reference points becomes unreliable (both noisy and potentially biased) as the time series is reduced, and therefore recommending that modelling the specific process involved is generally a better approach than truncation. The meeting noted several examples where reference points within current ICES management are allowed to vary (F in the case of NEA cod, B_{lim} in the case of Iberian Sardine) according the prevailing conditions. WKRPChange noted that this was only required if conditions were expected to change significantly over the lifespan of the reference points, and that where it was implemented the status determination (the "traffic lights") should be made accordingly.

The key recommendation of WKRPChange is to support the conclusions in WKGMSE2 (ICES, 2019a), that for each stock a scoping exercise should be undertaken to identify any key drivers, and where there is good evidence for ecosystem-driven changes in stock productivity, that process should be accounted for in setting reference points. The meeting highlighted that reference points have a finite lifespan, generally related to the benchmark cycle, and the estimation of the reference point should take into account processes likely to be important over that lifespan. WKRChange noted that many ICES stocks are managed by Harvest Controls Rules which are evaluated through an MSE process. In this case, there is considerable scope for including such environmentally driven processes in the Operating Model. However, many stocks are managed through the standard ICES HCR and reference points derived through the EqSim program. There is therefore a specific recommendation that density dependence be incorporated into EqSim, to allow for more realism to be included in the estimation of reference points where the evidence indicates that this is important.

Summary of discussion

It was argued that B_{lim} should be derived from biology (e.g. breakpoint of a segmented regression, or similar from other stock-recruit curves) instead of being based on, e.g. some fixed fraction of B_0. The parameters of stock-recruit curves are notoriously difficult to estimate, and often little is gained from a single stock-recruit fit, but meta-analysis and the use of distributions as a Bayesian prior could provide a useful starting point from which meaningful updates could occur. A strong recommendation from WKRPChange was to avoid truncating time-series wherever possible when estimating reference points (any such truncation should be strongly justified).

Furthermore, focus should be on reference points in a changing environment rather than on the reference point estimation *per se*.

8.3 WKMSEMAC (TORs a, b, c, d, e)

Edited executive summary of WKMSEMAC report (ICES, 2020a) (presented by Andrew Campbell):

WKMSEMAC (Workshop on Management Strategy Evaluation of Mackerel) took place during the period January to July 2020 with a kick-off meeting at ICES HQ, Copenhagen from 7 to 9 January followed by a number of WebEx meetings. A second scheduled physical workshop was cancelled following the COVID-19 pandemic, necessitating completion of the work by correspondence. The workshop was chaired by Andrew Campbell (Ireland) and was attended by 26 participants including three reviewers. The purpose of the workshop was to evaluate a long-term management strategy for northeast Atlantic mackerel, following a request from the EU, Norway, and the Faroe Islands (see Annex 9).

In line with the request, the approach adopted was to include an assessment and forecast in a full MSE simulation. This approach was computationally challenging such that the simulations had to be executed in a high-performance computing environment. In comparison with earlier iterations of this evaluation (which adopted a shortcut approach), this requirement limited the number of scenarios that could be evaluated and led to significant resource pressure. However, sufficient scenarios were explored to identify a range of harvest control rule parameters (F_{target} and $B_{trigger}$) that are both precautionary in the long term (risk to B_{lim} of < 5%) and maximize long term yield for management strategies both with and without stabilisation measures. All scenarios considered were precautionary in the short term, due to the current high stock size.

Strategies incorporating stabilisation measures were evaluated for a subset of the base case (without stabilisation). Measures included TAC change limitation, 5% constant banking, and an extreme bank and borrow scenario (10% alternating bank/borrow), and were tested both in isolation and in combination. In terms of long-term yield and risk, none of the stabilisation measures had a notable impact compared to the base case. Limiting the interannual change in TAC mitigated extreme changes in TAC, particularly for HCRs with relatively low $B_{trigger}$ values. Median interannual variability (IAV) in TAC is only marginally reduced compared to the base case for which the IAV was generally lower than the limitation imposed by the stabilisation measure.

Simulation of future recruitment within the HCR evaluations was based on assessment estimates of recruitment from a contemporary period (1998 onwards) which were strongly influenced by the recruitment index. These were considered to be reflective of the current situation and resulted in an increased perception of stock productivity compared to previous MSE evaluations. An evaluation with an alternative operating model based on recruitment estimated from the abundance of older (fully selected) fish indicated that although the maximum yield was associated with different HCR parameter values, all combinations considered precautionary under the base case remained so under this alternative.

Due to time constraints, it was not possible to fully explore robustness to alternative operating models with respect to natural mortality and density-dependent weight as included in the request. The results were contingent on both recent data and assessment performance such that should the perception of recruitment, biological characteristics, or exploitation change, re-evaluation of the management strategies tested should be considered.

Reference points were reviewed and updated values proposed based on evaluation with the MSE framework software.

Summary of discussion

Although concern was expressed about possible biases of the estimation model (SAM) in the MSE, this was not surprising, and should not be of concern, because the role of the MSE is precisely about testing how well the method performs; what is key when running the MSE is that the behaviour of the estimation model is as it would be when implemented in practice so that its performance can be appropriately evaluated. Nevertheless, there were some behavioural aspects of the SAM estimation model when running simulations that could not be fully explained (see Figure 7.1.1).

On reflection, the use of a hybrid approach, involving both shortcut and full methods, would have been beneficial. Furthermore, time may have been better spent focussing on uncertainty and robustness to OMs, instead of worrying about gaining a high amount of precision for the HCR solutions, particularly in the context of OMs that may reflect a large amount of uncertainty.

The MSE process was useful for guidelines, and guidelines should be pragmatic. However, the aspect of computational complexity is not the main problem; meetings are more expensive, and trying to process scenarios is more important. The main issues are the definitions of the terms validation, plausibility and realism. The models cannot be validated in fisheries science because the reality is unknown and can only be estimated with some observations. As an example, the oscillatory behaviour of the estimation model in the full approach (left plot of Figure 7.1.1) is interesting and either really important for the advice (if correct) or the model is set up incorrectly. The shortcut had a totally different outcome in the past compared to the current full approach, and this needs to be resolved.

Given the known historical trend (decade or so) of catches exceeding scientific advice for this stock, what steps can/should be taken to run scenarios with and without implementation error (i.e. to test scenarios for robustness)? Do ICES have guidelines on how best to take into account 'appropriate' levels of implementation error when it is known to take place, and if this trend of TAC overage is considered to have a high likelihood of continuing (whether the source is a TAC overshoot or bycatches/discards on top of catch under the European Union's Landing Obligation policy, etc.)? In the context of the mackerel, requesters specifically did not want this type of implementation error included; however, it is still possible to include it when considering alternative operating models, and current guidelines recommend it (ICES, 2019a).

The mackerel MSE exercise showed that there were issues in the MSE process, with timing, computing, resources, etc. and ICES may need to think about that. Considering the number of stocks in ICES, is it possible to do full MSEs for all of them? Full MSEs are only feasible for a limited number of stocks; an alternative is to consider simpler MPs, but this would need stakeholder buy-in (see Section 7).

SPM as an OM (part of WKMSEMAC report: ICES, 2020a; presented by Henrik Sparholt):

A surplus production model (SPM) as OM was presented as an alternative to the more complex OMs that typically take age and/or length structure into account. The argument for doing this was that an SPM, by default, has the advantage that all density dependent factors are included, while in the typical OMs, only density dependence in recruitment is included. The most realistic SPM for a mackerel OM was found to be one that was based on the mis-reported corrected catch and corresponding assessment done by WGWIDE in 2013 (ICES, 2013c). From the meta-analysis by Thorson *et al*. (2012), the form of the SPM was obtained (that of Perciformes fish to which mackerel belongs). The F_{MSY} value from the F_{MSY}-project (Sparholt *et al*., 2020) was used. The carrying capacity K was obtained by fitting to observed Surplus Production (SP) (catch + change in

stock biomass in a given year) from the time series of catch and total biomass from the assessment. More details of this approach can be found in ICES (2020a).

Summary of discussion

Age-based surplus production models are possible, which would allow more flexibility (e.g. when management strategies being tested require more structure, such as age). Biomass dynamic models can work well, but need to be tested first with more complex models. Furthermore, it is not the production function *per se* that is important, but rather process error (e.g. Walters *et al*., 2008; Botsford *et al*., 2014). It is important to capture real dynamics accurately. There is a need to be careful about B_{lim}, i.e. where does production fall off in a production curve? There is no indication of where this happens in a biomass dynamic model (e.g. through recruitment). Furthermore, in simulations, you could go below biomass levels for which no data exists, and as the production curve is not based on data; such declines are very slow in biomass dynamic models, and there could be a need for more pessimistic OMs to ensure management strategies are robust to these.

There is a history of large misreporting for NEA mackerel, and if this were included, the production curve could look entirely different. The simulation requires a more complex OM: age data are completely ignored in the biomass dynamic model, although it is parameterised using age-based models. It may be possible to mimic the behaviour of the biomass dynamic model with an age-based model coupled with a Ricker curve, but it was argued that such an approach would lack accounting of density dependence in more than just the stock-recruit curve. However, process error is driven by complex processes and cannot be easily modelled.

9 Open Session

The open session on the final day of the meeting allowed an opportunity for topics not directly covered by the TORs for this meeting to be presented. These are summarised below. Unfortunately, due to time constraints, there were no substantive discussions on these topics.

9.1 Density-dependence in operating models

Presentation by Henrik Sparholt

The standard assessment approaches used in the Northeast Atlantic area for estimating F_{MSY} do not account for density-dependent (DD) processes other than recruitment. Density can also affect growth, maturation, and natural mortality; failing to account for these processes in stock assessments can result in biased estimates of F_{MSY} (see Lorenzen, 2016, for a review; Sparholt *et al*., 2020). As populations rebuild, interactions such as predation and food competition strengthen, leading to higher mortality and slower growth, basic elements of ecosystem dynamics that determine ecosystem carrying capacity and density-dependent mechanisms.

Figure 9.1.1 below shows the effect of missing some of the DD factors when calculating F_{MSY} for North Sea cod (from Sparholt *et al*., 2019). When only DD in recruitment is included F_{MSY} is estimated to be 0.20. When DD in growth is added, F_{MSY} is estimated to be 0.30. When furthermore, DD in cannibalism is included, F_{MSY} is estimated to 0.70. In this case, DD in maturity only increased F_{MSY} slightly so seems not needed. This high F_{MSY} estimated when including all DD factors is in line with a recent MSE for this stock presented in WKNSMSE (ICES, 2019b).

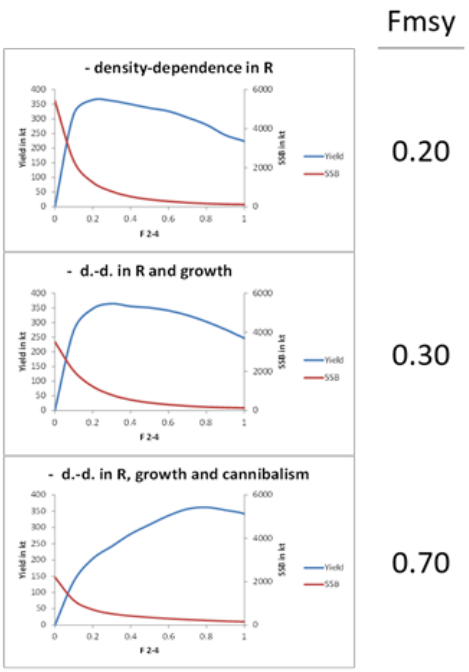

Figure 9.1.1. North Sea cod. Yield and SSB vs F for three scenarios of stock dynamics. Top panel only density dependence in R per SSB (the S-R model included). Middle panel DD in growth added. Bottom panel DD in growth and cannibalism added.

Density dependence is of course equally important for MSEs as for F_{MSY} calculations and it seems important that they are included in the OMs used in MSEs.

A new approach for estimating the fishing mortality benchmark F_{MSY} is proposed by Sparholt *et al*. (2020). The approach includes all the density-dependent (DD) factors. The analysis considers 53 data-rich fish stocks in the Northeast Atlantic. The new F_{MSY} values are estimated from an ensemble of data sources, including all DD factors: 1) applying traditional surplus production models on time-series of historical stock sizes, fishing mortalities, and catches from the current annual assessments; 2) dynamic pool model (e.g. age-structured models) estimation for stocks where data on density-dependent growth, maturity, and mortality are available; 3) extracts from multispecies and ecosystem literature for stocks where well-tested estimates are available; 4) the "Great Experiment" where fishing pressure on the demersal stocks in the Northeast Atlantic slowly increased for half a century; and 5) linking F_{MSY} to life history parameters. The new F_{MSY} values are substantially higher (average equal to 0.38 year^{-1}) than the current F_{MSY} values (average equal to 0.26 year^{-1}) estimated in stock assessments and used by management, The new F_{MSY} values are similar to the fishing pressure in the 1960s, and about 30% lower than the fishing pressure in 1970–2000, when the stocks generally were regarded as over-fished.

The new F_{MSY} values are 50% higher than current F_{MSY} values. Because the current F_{MSY} values are based on calculations only including DD in recruitment (with only two exceptions – Northeast Arctic cod and haddock), this probably means that DD in recruitment contributes 2/3 of the F_{MSY} value, and DD in growth, maturity and natural mortality thus contribute the rest, i.e. 1/3 of the F_{MSY} value.

The total catch in the Northeast Atlantic has been around 13 to 15 million tonnes per year between 1970 and 2000 (Figure 9.1.2). It is only about 9-10 million t in recent years. F has gone down significantly since the overfishing in 1970 to 2000, but catches have not improved. The study by Sparholt *et al*. (2020) indicates that the current F level is substantially lower than F_{MSY} and therefore at least partly, explains the low catch in recent years.

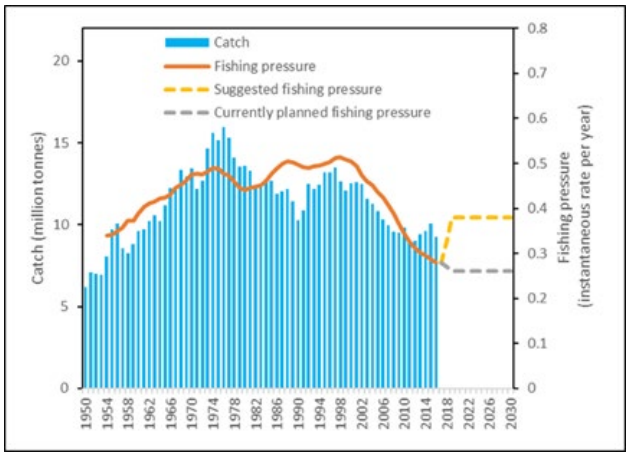

Figure 9.1.2. Catch by year in the Northeast Atlantic (FAO Area 27. From ICES database (http://www.ices.dk/marine-data/dataset-collections/Pages/Fish-catch-and-stock-assessment.aspx) except unreported catch (discards and IUU catch) which is from the "Sea Around Us"-database (http://www.seaaroundus.org/). Average fishing (5-year running means) for 53 data-rich Northeast Atlantic fish stocks. Until 2017, the values are historic values based on actual catches. From 2018 and onward, it is forecasts. The "Currently planned fishing pressure" curve is the development forecasted if the current F_{MSY} values are used in management, and the "Suggested fishing pressure" curve is the forecasted fishing pressure if the new F_{MSY} values suggested in the study by Sparholt *et al*. (2020) are used.

9.2 Investigating sampling levels and associated risk for sea bass using MSE

Presentation by Gwladys Lambert

The European sea bass is widely distributed in the Northeast Atlantic shelf waters, with the northern stock unit covering the North Sea, Channel, Celtic Sea and Irish Sea. Over the past 10 years, the northern stock has declined rapidly due to a combination of poor recruitment and high fishing mortality, leading to tighter management measures for both commercial and recreational fisheries. Recreational catches of seabass are a large proportion of total removals, representing at least 25% of the total catch. However, recreational data are limited with only a single estimate from 2012 currently available and informing the Stock Synthesis (SS) assessment. Routine surveys are being developed in several countries that exploit the stock, but estimation of recreational fisheries catches is a challenge. Hence, there is a need to understand how uncertainty and bias in recreational fisheries catches impact the assessment, the advice and ultimately the status of the stock. Here, the risk associated with varying levels of uncertainty in estimates of recreational catches was assessed using a closed loop simulation framework developed for SS. This allowed the comparison of the performance of the assessment and the current management rules under a range of different recreational data quality scenarios expressed in terms of precision and bias. Performance of management was measured using indicators including long-term trends in catch and stock biomass or the risk of falling below some reference points. The outcomes from the model were used to test the robustness of the current assessment to recreational data uncertainty and will be used to inform future regional sampling programmes for recreational fisheries. The work is still in development and challenges need to be addressed, notably with regards to setting up the operating model(s) (OM) and covering an appropriate level of uncertainty in the replicates, as well as using the accepted SS assessment model in the management procedure (MP), which slows the simulations but may also be too complex to fit in a loop.

9.3 Shiny app for MSE results

Presentation by Tom Carruthers

A Shiny App 'SLICK' was presented that allows users to upload MSE results and present these informatively and concisely. The interactivity of such Apps allows a wide range of MSE stakeholders to gain intuition about the interplay between operating models, management procedures and performance metrics. Figure 9.3.1 shows an example of SLICK results for multiple performance metrics across various management procedures.

While still in development, SLICK will be temporarily hosted at:
http://142.103.48.20:3838/SLICK/

Users can download the R package from GitHub repo:
https://github.com/tcarruth/SLICK

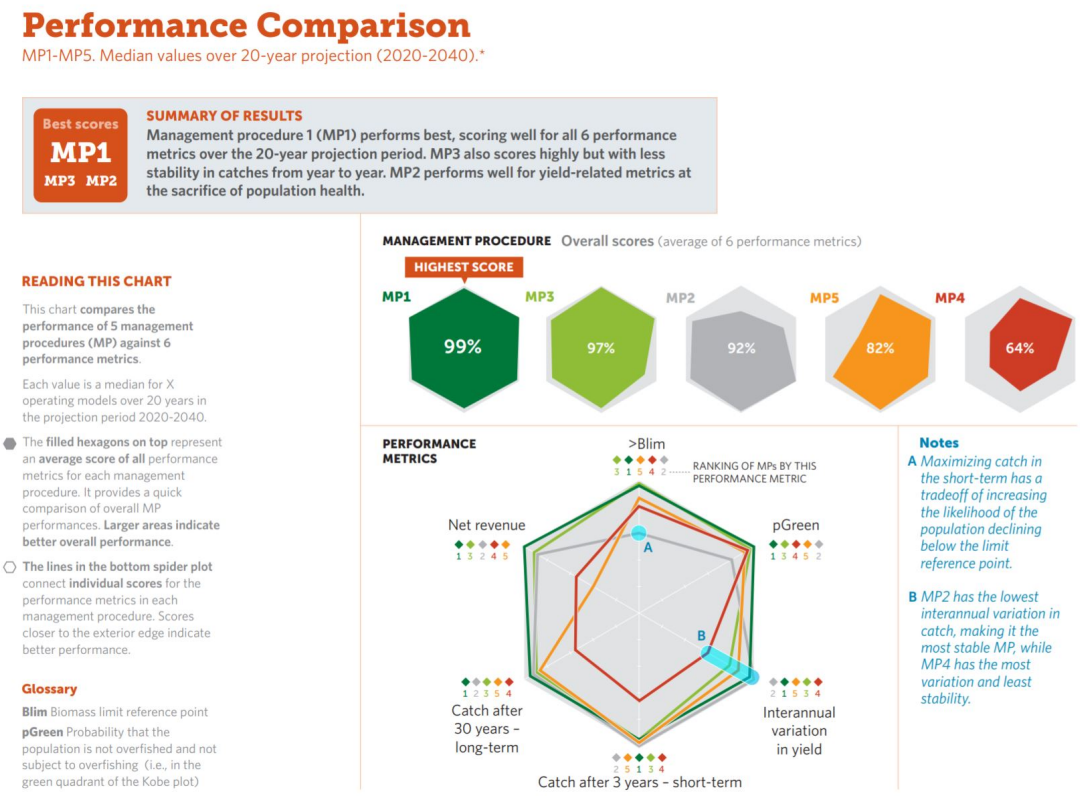

Figure 9.3.1. An example of SLICK results for multiple performance metrics across various management procedures. SLICK is supported by the Ocean Foundation, the design of the results figures is by 5W interactive.

10 Recommendations

1. ICES should reconsider the basis for calculating B_{lim} as this forms the cornerstone of the ICES reference point framework and underpins MSE evaluations; B_{pa} should also be considered given the proposal for an alternative way to calculate it (see Section 2.3). This should form part of a wider review of reference points in general, in a dedicated workshop, that takes into account international experiences on reference points, including rebuilding plans. Such a workshop should also make comparisons of reference points coming from different frameworks, such as EqSim and the proposed framework for reference point extraction from MSEs (see Section 2.3). [ACOM]. Suggested TORs for this workshop are:
    i. Re-evaluate the basis for calculating B_{lim} for ICES stocks (which forms the cornerstone of the ICES reference point framework) and associated reference points, accounting for previous work of ICES: SGPRP (ICES, 2003) and WKMSYREF (ICES, 2014a; 2014b; 2016b; Rindorf *et al*., 2017), the outcomes of more recent meetings: WKREBUILD (ICES, 2020b) and WKRPChange (held in 2020), as well as international developments.
    ii. Compare reference points derived using different frameworks, such as EqSim and the proposed framework for reference point extraction from MSEs (Section 2.3).
2. Comments were offered that the simulation testing process ICES has currently in place (to address requests from its clients) differs from the considerably more flexible approach also termed "MSE" (or the "Management Procedure approach") that is being implemented elsewhere in the world (such as empirical management strategies that are tuned by MSE to meet specific management objectives). Moving towards that other approach would bypass some of the problems being experienced in implementing the current ICES testing process, and also offers some other advantages for the overall management advice process. It is strongly recommended that this matter be accorded further consideration by ICES and its clients when developing and testing management strategies within an MSE framework. [ACOM]
3. Communication between managers, scientists and stakeholders is a key part of the MSE process (Miller *et al*., 2018), particularly when a request to develop and test management strategies is being drafted. In the ICES context, this has sometimes led to unsatisfactory outcomes because of the difficulty in getting feedback on a request (such as at WKNSMSE, when banking and borrowing scenarios were tested in combination with TAC stability scenarios, as requested, and not separately, which would have made interpretation of results clearer; ICES, 2019b). It is strongly recommended that managers and stakeholders play a more active role throughout the MSE process, and not just at the start and end of the process. [ICES Secretariat; ACOM]
4. There were strongly divergent views about the merits of full versus shortcut approaches, which could not be resolved from first principles. A possible resolution is a recommendation to be included in a further workshop (WKGMSE4) to investigate when performance of an HCR/management strategy is not as intended (as indicated by the results from the simulations) under either the full or shortcut methods (including alternative interpretations of the latter) in MSEs that include a range of alternative operating models. [ACOM]

11 References

Anon. 2018. Glossary of terms for harvest strategies, management procedures and management strategy evaluation, https://www.tuna-org.org/Documents/MSEGlossary_tRFMO_MSEWG2018.pdf.

Åsnes, M.N. 2005. Prost User Guide. ICES Arctic Fisheries Working Group, Murmansk, Russia 19-28 April 2005. WK2.

Beddington, J.R. and Cooke, J.G. 1983. The potential yield of fish stocks. F.A.O. Fish. Biol. Tech. Pap. 242: 1-47.

Bergh, M.O. and Butterworth, D.S. 1987. Towards rational harvesting of the South African anchovy considering survey imprecision and recruitment variability. *In* A.I.L. Payne, J.A. Gulland and K.H. Brink (ed.). The Benguela and Comparable Ecosystems. S. Afr. J. mar. Sci., 5: 937-951.

Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13(1), pp.281-305.

Björnsson, H. 2013. Report of the evaluation of the Icelandic haddock management plan, ICES CM 2013/ACOM:59. 47pp. http://www.ices.dk/sites/pub/Publication%20Reports/Expert%20Group%20Report/acom/2013/ADHOC/IntroAndHad.pdf.

Botsford, L.W., Holland, M.D., Field, J.C., and Hastings, A. 2014. Cohort resonance: a significant component of fluctuations in recruitment, egg production, and catch of fished populations. ICES Journal of Marine Science, 71: 2158–2170.

Butterworth, D.S. and Bergh, M.O. 1993, The Development of a management procedure for the South African anchovy resource. In Risk Evaluation and Biological Reference Points for Fisheries Management. Smith, S.J., Hunt, J.J. and D. Rivard (Eds). Can. Spec. Publ. Fish. Aquat. Sci. 120: 83–99.

Carruthers, T.R., Kell, L.T., Butterworth, D.D., Maunder, M.N., Geromont, H.F., Walters, C., McAllister, M.K., Hillary, R., Levontin, P., Kitakado, T. and Davies, C.R., 2016. Performance review of simple management procedures. ICES Journal of Marine Science, 73(2), pp.464–482.

de Moor, C.L., Butterworth, D.S. and De Oliveira, J.A.A., 2011. Is the management procedure approach equipped to handle short-lived pelagic species with their boom and bust dynamics? The case of the South African fishery for sardine and anchovy. ICES Journal of Marine Science, 68(10): 2075–2085.

de Moor, C.L. and Butterworth, D.S. In prep. How should risk be quantified for short-lived, highly variable species? Some concepts developed from the case of the South African small pelagics fishery.

Fasiolo, M., Wood, S.N., Zaffran, M., Nedellec, R. and Goude, Y. 2020. Fast Calibrated Additive Quantile Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2020.1725521.

Fischer, S.H., De Oliveira, J.A.A., Mumford, J.D. and Kell, L.T. In press 2020. Using a genetic algorithm to optimise a data-limited catch rule. ICES Journal of Marine Science.

Garud, S.S, Karimi, I.A. and Kraft, M. 2017. Design of computer experiments: A review. Computers & Chemical Engineering, 106: 71-95. https://doi.org/10.1016/j.compchemeng.2017.05.010.

Holden, P.B., Edwards, N.R., Garthwaite, P.H. and Wilkinson, R.D. 2015. Emulation and interpretation of high-dimensional climate model outputs. Journal of Applied Statistics, 42 (9): 2038-2055. https://doi.org/10.1080/02664763.2015.1016412.

ICES. 2000. Request to ICES on Norwegian spring-spawning herring. Section 3.1.10. ICES Cooperative Research Report No. 242, Part 1: 87-91. https://doi.org/10.17895/ices.pub.5375.

ICES. 2003. Report of the Study Group on Precautionary Reference Points for Advice on Fishery Management (SGPRP), 24–26 February 2003, ICES Headquarters, Copenhagen, Denmark. ICES CM 2003/ACFM:15. 85 pp.

ICES. 2005a. Report of the Ad hoc group on long term management (AGLTA), Copenhagen, 12-13 April 2005. ICES C.M. 2005/ACFM: 25.

ICES. 2005b. Report of the Herring assessment working group for the area south of 62, Copenhagen, 8-17 March 2005. ICES C.M. 2005/ACFM:16.

ICES. 2008a. Report of the Workshop on Herring Management Plans (WKHMP), 4-8 February 2008. ICES C.M. 2008 / ACOM:27.

ICES. 2008b. Report of the Working Group on Methods of Fish Stock Assessments (WGMG), 7–16 October 2008, Woods Hole, USA. ICES CM 2008/RMC:03, 147 pp.

ICES. 2010. Report of the Ad hoc Group on Icelandic Cod HCR Evaluation (AGICOD), 24-26 November 2009 ICES, Copenhagen, Denmark. ICES CM 2009 \ACOM:56. 89 pp. https://doi.org/10.17895/ices.pub.5279.

ICES. 2012. Report of the Workshop for revision of the North Sea herring Long Term Management Plan (WKHELP). IJmuiden, 3-4 September 2012 and Copenhagen, 1-2 October 2013. ICES C.M. 2012 / ACOM:72.

ICES. 2013a. Report of the Benchmark Workshop on Roundfish Stocks (WKROUND). 4-8 February 2013, Marine Laboratory, Aberdeen, UK. ICES CM 2013/ ACOM:47, 213 pp.

ICES. 2013b. Report of the Workshop on Guidelines for Management Strategy Evaluations (WKGMSE) , 21 - 23 January 2013, ICES HQ, Copenhagen, Denmark.

ICES. 2013c. Report of the Working Group on Widely Distributed Stocks (WGWIDE), 27 August - 2 September 2013, ICES Headquarters, Copenhagen, Denmark. ICES CM 2013/ACOM:15. 950 pp.

ICES. 2014a. Report of the Workshop to consider reference points for all stocks (WKMSYREF2), 8-10 January 2014, ICES Headquarters, Copenhagen, Denmark. ICES CM 2014/ACOM:47. 91 pp.

ICES. 2014b. Report of the Joint ICES-MYFISH Workshop to consider the basis for FMSY ranges for all stocks (WKMSYREF3), 17–21 November 2014, Charlottenlund, Denmark. ICES CM 2014/ACOM:64. 147 pp.

ICES. 2014c. Report of the EU Workshop on the NEA Mackerel Long-term Management Plan (WKMACLTMP), 24–27 June and 17–19 November 2014, Copenhagen, Denmark. ICES CM 2014/ACOM:63. 120 pp.

ICES. 2014d. Report of the Benchmark Workshop on Pelagic Stocks (WKPELA), 17–21 February 2014, Copenhagen, Denmark. ICES CM 2014/ACOM:43.

ICES 2015a. 9.2.3.1 EU, Norway, and the Faroe Islands request to ICES to evaluate a multi-annual management strategy for mackerel (*Scomber scombrus*) in the Northeast Atlantic. In Report of the ICES Advisory Committee, 2015.

ICES 2015b. 9.2.3.2 EU and Norway request to evaluate the proposed Long-Term Management Strategy for herring (*Clupea harengus*) in the North Sea and the Division IIIa herring TAC-setting procedure. In Report of the ICES Advisory Committee, 2015.

ICES. 2016a. Report of the Inter-benchmark protocol for Whiting in the North Sea (IBP Whiting), By correspondence, March 2016. ICES IBP Whiting Report 2016. ICES CM 2016/ACOM: 48, 119 pp.

ICES. 2016b. Report of the Workshop to consider FMSY ranges for stocks in ICES categories I and 2 in Western Waters (WKMSYREF4), 13-16 October 2015, Brest, France. ICES CM 2015/ACOM:58. 187 pp.

ICES. 2016c. NEAFC request to ICES to evaluate a long-term management strategy for the fisheries on the blue whiting (*Micromesistius poutassou*) stock. ICES Advice 2016, section 9.4.2, special request.

ICES. 2016d. Report of the Workshop on Blue Whiting Long Term Management Strategy Evaluation (WKBWMS), 30 August 2016, Copenhagen. ICES C.M. 2016 / ACOM: 53.

ICES. 2017a. ICES fisheries management reference points for category 1 and 2 stocks. ICES Advice Technical Guidelines, ICES Advice 2017, Book 12, Section 12.4.3.1. Published 20 January 2017. https://doi.org/10.17895/ices.pub.3036.

ICES. 2017b. EU, Norway, and the Faroe Islands request concerning long-term management strategy for mackerel in the Northeast Atlantic. In Report of the ICES Advisory Committee, 2017. ICES Special Request Advice, Ecoregions in the Northeast Atlantic and Arctic Ocean, sr.2017.19. Published 29 September 2017: 14 pp. https://doi.org/10.17895/ices.pub.3031.

ICES. 2017c. Report of the Benchmark Workshop on North Sea Stocks (WKNSEA), 14–18 March 2016, Copenhagen, Denmark. ICES CM 2016/ACOM:37. 698 pp.

ICES. 2017d. Report of the Benchmark Workshop on Widely Distributed Stocks (WKWIDE), 30 January-3 February 2017, Copenhagen, Denmark. ICES CM 2017/ACOM:36. 196 pp.

ICES. 2018a. Report of the Benchmark Workshop on North Sea Stocks (WKNSEA 2018), 5–9 February 2018, Copenhagen, Denmark. ICES CM 2018/ACOM:33, 634pp.

ICES. 2018b. Report of the Working Group on the in the North Sea and Skagerrak Assessment of Demersal Stocks (WGNSSK), 24 April - 3 May 2018, Oostende, Belgium. ICES CM 2018/ACOM:22: 1250 pp.

ICES. 2019a. Workshop on Guidelines for Management Strategy Evaluations (WKGMSE2). ICES Scientific Reports. 1:33. 162 pp. http://doi.org/10.17895/ices.pub.5331.

ICES. 2019b. Workshop on North Sea stocks management strategy evaluation (WKNSMSE). ICES Scientific Reports. 1:12. 378 pp. http://doi.org/10.17895/ices.pub.5090.

ICES. 2019c. Advice basis. In Report of the ICES Advisory Committee, 2019. ICES Advice 2019, section 1.2. https://doi.org/10.17895/ices.advice.5757.

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

ICES. 2019e. Working Group on Widely Distributed Stocks (WGWIDE). ICES Scientific Reports. 1:36. 948 pp. http://doi.org/10.17895/ices.pub.5574.

ICES. 2019f. 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. 2019g. Workshop on the management strategy evaluation of the reference point, F_{cap}, for Sprat in Division 3.a and Subarea 4 (WKspratMSE). 11-12 December 2018. ICES HQ, Copenhagen, Denmark. ICES CM 2018/ACOM:69. 35 pp

ICES. 2019h. EU and Norway request concerning the long-term management strategy of cod, saithe, and whiting, and of North Sea autumn-spawning herring. In Report of the ICES Advisory Committee, 2019. ICES Advice 2019, sr.2019.06. 24 pp. https://doi.org/10.17895/ices.advice.4895.

ICES. 2020a. Workshop on Management Strategy Evaluation of Mackerel (WKMSEMAC). ICES Scientific Reports. 2:74. 175 pp. http://doi.org/10.17895/ices.pub.7445.

ICES. 2020b. Workshop on guidelines and methods for the evaluation of rebuilding plans (WKREBUILD). ICES Scientific Reports. 2:55. 79 pp. http://doi.org/10.17895/ices.pub.6085.

ICES. 2020c. Workshop on an Ecosystem Based Approach to Fishery Management for the Irish Sea (WKIrish6; outputs from 2019 meeting). ICES Scientific Reports. 2:4. 32 pp. http://doi.org/10.17895/ices.pub.5551.

IOTC. 2019. Report of the 10th Session of the IOTC Working Party on Methods. Pasaia, Spain 17-19 October 2019. IOTC–2019–WPM10–R[E]: 36 pp.

Kell, L.T., Mosqueira, I., Grosjean, P., Fromentin, J-M., Garcia, D., Hillary, R., Jardim, E., Mardle, S., Pastoors, M.A., Poos, J.J., Scott, F., and Scott, R.D. 2007. FLR: an open-source framework for the evaluation and development of management strategies. ICES Journal of Marine Science, 64: 640–646. https://academic.oup.com/icesjms/article/64/4/640/640024.

Kell, A.J.M., Forshaw, M., and McGough A.S. 2019. Optimising energy and overhead for large parameter space simulations. 2019. Tenth International Green and Sustainable Computing Conference (IGSC), Alexandria, VA, USA, 2019, pp. 1-8, doi: 10.1109/IGSC48788.2019.8957205.

Kennedy, M.C. and O'Hagan, A. 2001, Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63: 425-464. https://doi.org/10.1111/1467-9868.00294

Lorenzen, K. 2016. Toward a new paradigm for growth modeling in fisheries stock assessments: Embracing plasticity and its consequences. Fisheries Research, 180: 4–22. https://doi.org/10.1016/j.fishres.2016.01.006.

Miller, S.K., Anganuzzi, A., Butterworth, D.S., Davies, C.R., Donovan, G.P., Nickson, A., Rademeyer, R.A. and Restrepo, V. 2018. Improving communication: the key to more effective MSE processes. Can. J. Fish. Aquat. Sci. 76: 643-656. dx.doi.org/10.1139/cjfas-2018-0134.

Murata, T. and Ishibuchi, H. (1995). MOGA: Multi-objective genetic algorithms. (November):289–294

National Research Council [NRC]. 2014. Evaluating the effectiveness of fish stock rebuilding plans in the United States. National Academies Press 155 p. http://www.nap.edu/catalog.php?record_ id=18488.

Nielsen, A. and Berg, C.W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research, 158: 96–101. http://dx.doi.org/10.1016/j.fishres.2014.01.014.

Oakley, J. and O'Hagan, A. 2002. Bayesian inference for the uncertainty distribution of computer model outputs, Biometrika, 89 (4): 769–784. https://doi.org/10.1093/biomet/89.4.769

Pastoors, M.A., Campbell, A., Trijoulet, V., Skagen, D., Gras, M., Lambert, G., Sparrevohn, C. R. and Mackinson, S. 2020. Western Horse Mackerel Technical Focus Group on Harvest Control Rule Evaluations, 2020. Pelagic Advisory Council: 43 pp. https://www.pelagic-ac.org/media/pdf/1920PAC90%20to%20COM%20Annex%20I%20report%20on%20Western%20Horse%20Mackerel.pdf.

Patterson, K.R., Skagen, D., Pastoors, M.A. and Lassen, H. 1997. Harvest control laws for North Sea herring. Working Document for the ICES ACFM November Meeting 1997.

Pedersen, M.W. and Berg, C.W. 2017. A stochastic surplus production model in continuous time. Fish and Fisheries, 18: 226–243.

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

Rindorf, A., Cardinale, M., Shephard, S., De Oliveira, J.A.A., Hjorleifsson, E., Kempf, A., Luzenczyk, A., Millar, C., Miller, D.C.M., Needle, C.L., Simmonds, J. and Vinther, M. 2017. Fishing for MSY: using "pretty good yield" ranges without impairing recruitment. ICES Journal of Marine Science, 74: 525–534. https://doi.org/10.1093/icesjms/fsw111.

Rademeyer, R.A., Plagányi, E.E. and Butterworth, D.S. 2007. Tips and tricks in designing management procedures. ICES Journal of Marine Science, 64: 618-625.

Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A. and Luccioni, A. 2019. Tackling climate change with machine learning. arXiv preprint arXiv:1906.05433.

Sharma, R., Levontin, P., Kitakado, T., Kell, L., Mosqueira, I., Kimoto, A., Scott, R., Minte-Vera, C., De Bruyn, P. Ye, Y., Kleineberg, J., Walton, J.L., Miller, S. and Magnusson, A. 2020. Operating model design in tuna Regional Fishery Management Organizations: Current practice, issues and implications. Fish and Fisheries, 21: 940–961. https://doi.org/10.1111/faf.12480.

Sissenwine, M.P. and Shepherd, J.G., 1987. An alternative perspective on recruitment overfishing and biological reference points. Canadian Journal of Fisheries and Aquatic Sciences, 44(4), pp.913-918.

Skagen, D. and Miller, D.C.M. 2013. Blue Whiting HCR Evaluations, Summer/Autumn 2013, Copenhagen, Denmark. ICES CM 2013/ACOM:76. 60 pp.

Skagen, D.W. 2015. HCS program for simulating harvest control rules. Program description and instructions for users. Version HCS 15_1. August 2015. [Obtainable at www.dwsk.net.]

Sparholt, H., Bogstad, B., Christensen, V., Collie, J., Gemert, R.v., Hilborn, R., Horbowy, J., How-ell, D., Melnychuk, M.C., Pedersen, S.A., Sparrevohn, C.R., Stefansson, G., and Steingrund, P. 2019. Report of the 3rd working group meeting on optimization of fishing pressure in the Northeast Atlantic, Rhode

Island March 2018. NORDIC WORKING PAPERS http://dx.doi.org/10.6027/NA2019-906 NA2019:902, ISSN 2311-0562. www.norden.org/en/publication/report-3rd-working-group-meeting-optimization-fishing-pressure-northeast-atlantic-rhode.

Sparholt, H., Bogstad, B., Christensen, V., Collie, J., van Gemert, R., Hilborn, R., Horbowy, J., Howell, D., Melnychuk, M.C., Pedersen, S.A., Sparrevohn, C.R., Stefansson, G., and Steingrund, P. 2020. Estimating Fmsy from an ensemble of data sources to account for density dependence in Northeast Atlantic fish stocks. ICES Journal of Marine Science, doi:10.1093/icesjms/fsaa175. Published online 15 October 2020.

Spence, M.A., Alliji, K., Bannister, H.J., Walker, N.D. and Muench, A. 2020. Fish should not be in isolation: Calculating maximum sustainable yield using an ensemble model, arXiv:2005.02001. https://arxiv.org/abs/2005.02001

Smola, A.J. and Schölkopf, B. 2004. A tutorial on support vector regression. Statistics and computing, 14(3):199–222.

STECF. 2019. Multiannual Plan for the fisheries exploiting demersal stocks in the Adriatic Sea (STECF-19-02). Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-04009-5, doi:10.2760/026674, JRC116731.

Thorson, J.T., Cope, J.M., Branch, T.A., and Jensen, O.P. 2012. Spawning biomass reference points for exploited marine fishes, incorporating taxonomic and body size information. Canadian Journal of Fisheries and Aquatic Sciences, 69: 1556–1568.

Vernon, I., Goldstein, M. and Bower, R. 2014. Galaxy Formation: Bayesian History Matching for the Observable Universe. Statistical Science, 29 (1): 81–90. www.jstor.org/stable/43288453.

Walters, C.J., Hilborn, R. and Christensen, V. 2008. Surplus production dynamics in declining and recovering fish populations. Can. J. Fish. Aquat. Sci. 65: 2536–2551.

Whitley, D. 1994. A genetic algorithm tutorial. Statistics and computing, 4(2):65–85.

Wiedenmann, J., Wilberg, M.J., Sylvia, A., and Miller, T.J. 2015. Autocorrelated error in stock assessment estimates: Implications for management strategy evaluation. Fisheries Research, 172: 325–334. Elsevier B.V. http://dx.doi.org/10.1016/j.fishres.2015.07.037.

Zhihuan, L., Yinhong, L. and Xianzhong, D., 2010. Non-dominated sorting genetic algorithm-II for robust multi-objective optimal reactive power dispatch. IET generation, transmission & distribution, 4(9), pp.1000-1008.

Annex 1: List of participants

Name	Institute	Country (of institute)	Email
Alessandro Orio	SLU - Swedish University of Agricultural Sciences	Sweden	alessandro.orio@slu.se
Alex Hanke	Fisheries and Oceans Canada	Canada	alex.hanke@dfo-mpo.gc.ca
Alexander Kempf	Thünen-Institute of Sea Fisheries	Germany	Alexander.kempf@thuenen.de
Alfonso Perez-Rodriguez	IMR - Institute of Marine Research	Norway	alfonso.perez-rodriguez@hi.no
Allen R. Kronlund	Interface Fisheries Consulting	Canada	interfacefisheries@gmail.com
Ana Parma	CONICET - Centro Nacional Patagónico	Argentina	anaparma@gmail.com
Andrea Ross-Gillespie	University of Cape Town	South Africa	andrea.ross-gillespie@uct.ac.za
Andrew Campbell	Marine Institute	Ireland	andrew.campbell@marine.ie
Ash Wilson	Pew Charitable Trusts	United Kingdom	awilson@pewtrusts.org
Benoit Berges	WUR - Wageningen University & Research	Netherlands	benoit.berges@wur.nl
Carryn de Moor	University of Cape Town	South Africa	carryn.demoor@uct.ac.za
Christoph Konrad	JRC - Joint Research Centre	Other	Christoph.KONRAD@ec.europa.eu
Claus Reedtz Sparrevohn	Danish Pelagic Producers' Organisation	Denmark	crs@pelagisk.dk
Colin Millar (ICES secretariat)	ICES	Other	colin.millar@ices.dk
Daisuke Goto	IMR - Institute of Marine Research	Norway	daisuke.goto@hi.no
Daniel Howell	IMR - Institute of Marine Research	Norway	daniel.howell@hi.no
David Die	Rosenstiel School of Marine and Atmospheric Sciences	United States	ddie@rsmas.miami.edu
David Miller (ICES secretariat)	ICES	Other	david.miller@ices.dk
Dorleta Garcia	AZTI-Tecnalia	Spain	dgarcia@azti.es
Doug Butterworth	University of Cape Town	South Africa	doug.butterworth@uct.ac.za
Einar Hjörleifsson	Marine and Freshwater Research Institute	Iceland	einar.hjorleifsson@hafogvatn.is
Ernesto Jardim	Marine Stewardship Council	Other	ernesto.jardim@msc.org
Galina Chernega	Atlantic branch of Russian Federal Research Institute of Fisheries and Oceanography	Russian Federation	chernega@atlant.baltnet.ru
Gavin Fay	University of Massachusetts Dartmouth	United States	gfay@umassd.edu

Name	Institute	Country (of institute)	Email
Gwladys Lambert	CEFAS - Centre for Environment, Fisheries and Aquaculture Science	United Kingdom	gwladys.lambert@cefas.co.uk
Harriet Cole	Marine Scotland Science	United Kingdom	H.Cole@MARLAB.AC.UK
Henning Winker	JRC - Joint Research Centre	Other	Henning.WINKER@ec.europa.eu
Henrik Sparholt	University of Copenhagen	Denmark	henrik.sparholt@gmail.com
Höskuldur Björnsson	Marine and Freshwater Research Institute	Iceland	hoskuldur.bjornsson@hafogvatn.is
Iago Mosqueira	WUR - Wageningen University & Research	Netherlands	iago.mosqueira@wur.nl
Jette Fredslund (ICES secretariat)	ICES	Other	jette.fredslund@ices.dk
Jonathan Deroba	NOAA - National Oceanic and Atmospheric Administration	United States	Jonathan.Deroba@noaa.gov
José De Oliveira (chair)	CEFAS - Centre for Environment, Fisheries and Aquaculture Science	United Kingdom	jose.deoliveira@cefas.co.uk
Kyle Gillespie	Fisheries and Oceans Canada	Canada	Kyle.Gillespie@dfo-mpo.gc.ca
Laura Solinger	University of Southern Mississippi	United States	Laura.Solinger@usm.edu
Laurence Kell	SEA++	United Kingdom	laurie@seaplusplus.co.uk
Mackenzie Mazur	University of Maine	United States	mmazur@Gmri.org
Marc Taylor	Thünen Institute of Sea Fisheries	Germany	marc.taylor@thuenen.de
Margaret Siple	University of Washington	United States	mcsiple@gmail.com
Māris Plikšs	BIOR - Institute of Food Safety Animal Health and Environment	Latvia	Maris.Plikss@bior.lv
Martin Pastoors	Pelagic Freezer-Trawler Association	Netherlands	mpastoors@pelagicfish.eu
Massimiliano Cardinale	SLU - Swedish University of Agricultural Sciences	Sweden	massimiliano.cardinale@slu.se
Michael Gras	JRC - Joint Research Centre	Other	michael.gras@ec.europa.eu
Michael Spence	CEFAS - Centre for Environment, Fisheries and Aquaculture Science	United Kingdom	michael.spence@cefas.co.uk
Michelle Greenlaw	Fisheries and Oceans Canada	Canada	Michelle.Greenlaw@dfo-mpo.gc.ca
Mollie Elisabeth Brooks	DTU Aqua - National Institute of Aquatic Resources	Denmark	molbr@aqua.dtu.dk
Nicholas Duprey	Fisheries and Oceans Canada	Canada	Nicholas.Duprey@dfo-mpo.gc.ca
Norbert Rohlf	Thünen-Institute of Sea Fisheries	Germany	norbert.rohlf@thuenen.de
Polina Levontin	Imperial College London	United Kingdom	polina.levontin02@imperial.ac.uk

Name	Institute	Country (of institute)	Email
Rishi Sharma	FAO - The Food and Agriculture Organization of the United Nations	Other	Rishi.Sharma@fao.org
Robert Thorpe	CEFAS - Centre for Environment, Fisheries and Aquaculture Science	United Kingdom	robert.thorpe@cefas.co.uk
Santiago Cerviño	IEO - The Spanish Institute of Oceanography	Spain	santiago.cervino@ieo.es
Sara Pipernos	The Ocean Foundation	United States	spipernos@oceanfdn.org
Simon Fischer	CEFAS - Centre for Environment, Fisheries and Aquaculture Science	United Kingdom	simon.fischer@cefas.co.uk
Sonia Sanchez	AZTI-Tecnalia	Spain	ssanchez@azti.es
Stefanie Haase	Thünen Institute of Baltic Sea Fisheries	Germany	stefanie.haase@thuenen.de
Tanja Miethe	Marine Scotland Science	United Kingdom	t.miethe@marlab.ac.uk
Tom Carruthers	University of British Columbia	Canada	t.carruthers@oceans.ubc.ca
Toshi Kitakado	Tokyo University of Marine Science and Technology	Japan	kitakado@kaiyodai.ac.jp
Valerio Bartolino	SLU - Swedish University of Agricultural Sciences	Sweden	valerio.bartolino@slu.se
Virginia Noble	Fisheries and Oceans Canada	Canada	Virginia.Noble@dfo-mpo.gc.ca

Annex 2: Resolutions for WKGMSE3

2019/2/FRSG36: **The third Workshop on guidelines for management strategy evaluations (WKGMSE3)**, chaired by José De Oliveira*, United Kingdom, will be established and will meet by correspondence and online meetings 26-30 October 2020 to:

a) Develop guidelines for when and how reference points should be extracted from an MSE when one is conducted.

b) Develop guidelines for how to treat the results of alternative operating models. Currently, these have been used as robustness tests for "optimised" management strategies.

c) Explore the relationship between estimated risk and assumed levels of uncertainty included in the MSE. Risk and uncertainty are closely related, and including more uncertainty affects the estimated level of risk from the MSE. Apart from uncertainty, consideration should also be given to:

  i) The number of replicates and length of projection period used in the MSE;

  ii) The stationarity of MSE projections, from which risk metrics are calculated;

  iii) The risk metric itself (e.g. several definitions are given in the WKGMSE report of 2019).

d) Develop more efficient ways of conducting searches over a grid to the required level of precision be investigated. This is needed because of the high-performance computing requirements for full MSEs. This work could include investigating statistical properties that relate sample size to required precision, GAMs to interpolate over an incomplete grid, etc.

e) Compare the shortcut and full MSE approaches, providing guidelines for use of the former as an approximation for the latter, if appropriate. Consideration should be given to MSE with alternative operating models (i.e. operating models not solely based on the currently-used assessment).

WKGMSE3 will report by 27 November 2020 for the attention of ACOM and FRSG.

Supporting information

Priority	This workshop picks up on some of the recommendations from WKNSMSE, covering aspects that the latter could not fully explore. They cover extracting reference points from MSEs, defining risk, and methods for finding optimised management strategies. An additional TOR has been added to compare shortcut and full MSE approaches.
Scientific justification	This workshop explores in greater detail issues that were uncovered during the work of WKNSMSE, and that could be further explored at the time given workload and time constraints. TOR (a) deals with extracting reference points from MSEs when they are conducted, including the time-frame to be used; it came about because of discrepancies between reference points from the standard ICES approach (EqSim), and the MSEs conducted as part of WKNSMSE. A fundamental principle of MSE is to have a range of operating models (other than one based on the current assessment) to cover uncertainty, but how results from these are weighted and/or combined is not always clear. Currently, they have been used to check robustness of optimised management strategies. TOR (b) explores how to handle alternative operating models. TOR (c) covers issues related to the definition of risk, and in particular whether there is some way of benchmarking risk in relation to the amount of uncertainty incorporated in an MSE, and what to do in the presence of non-stationary MSE projections. TOR (d) covers the practical problem of optimising management strategies under full MSEs when each cell of a grid over which the optimisation takes place takes a long time to run. This TOR covers more effective and efficient ways of conducting the optimisation (e.g. through statistical means, or by using methods such as genetic algorithms). Finally, TOR (e) covers something that is topical for MSEs conducted within ICES, namely comparing the short-cut approach to carrying out computer-intensive full MSEs.
Resource requirements	
Participants	
Secretariat facilities	
Financial	
Linkages to advisory committees	
Linkages to other committees or groups	This work follows on from WKNSMSE and is closely linked to WKGMSE.
Linkages to other organizations	

Annex 3: Effects of uncertainty on risk assessments in management strategy evaluation (TOR c)

Daisuke Goto

Management strategy evaluation for North Sea saithe

As part of TOR c, we explored the relationships between uncertainty levels and risk in management strategy evaluations (MSEs). We analysed uncertainty effects in risk by performing simulation experiments using the MSE framework previously developed for North Sea saithe (*Pollachius virens*) as part of WKNSMSE (ICES, 2019b) as a case study. The framework consists of sub-models that simulate 1) 'true' population and harvest dynamics at sea and observations through monitoring surveys (an operating model or OM), and 2) management processes including assessments based on observations from the surveys and subsequent decision-making (a management procedure or MP) (Figure A.3.1a). We conditioned the OM on the 2018 assessment for the stock (Figure A.3.1b, ICES 2018b) and projected 21-year (2018–2038) forecasts. We ran all simulations in R using the FLR (Fisheries Library in R) mse package (https://github.com/flr/mse) (ICES, 2019b). Data, models, and codes to run the saithe MSE are available on GitHub (https://github.com/ices-taf/wk_WKNSMSE_pok.27.3a46).

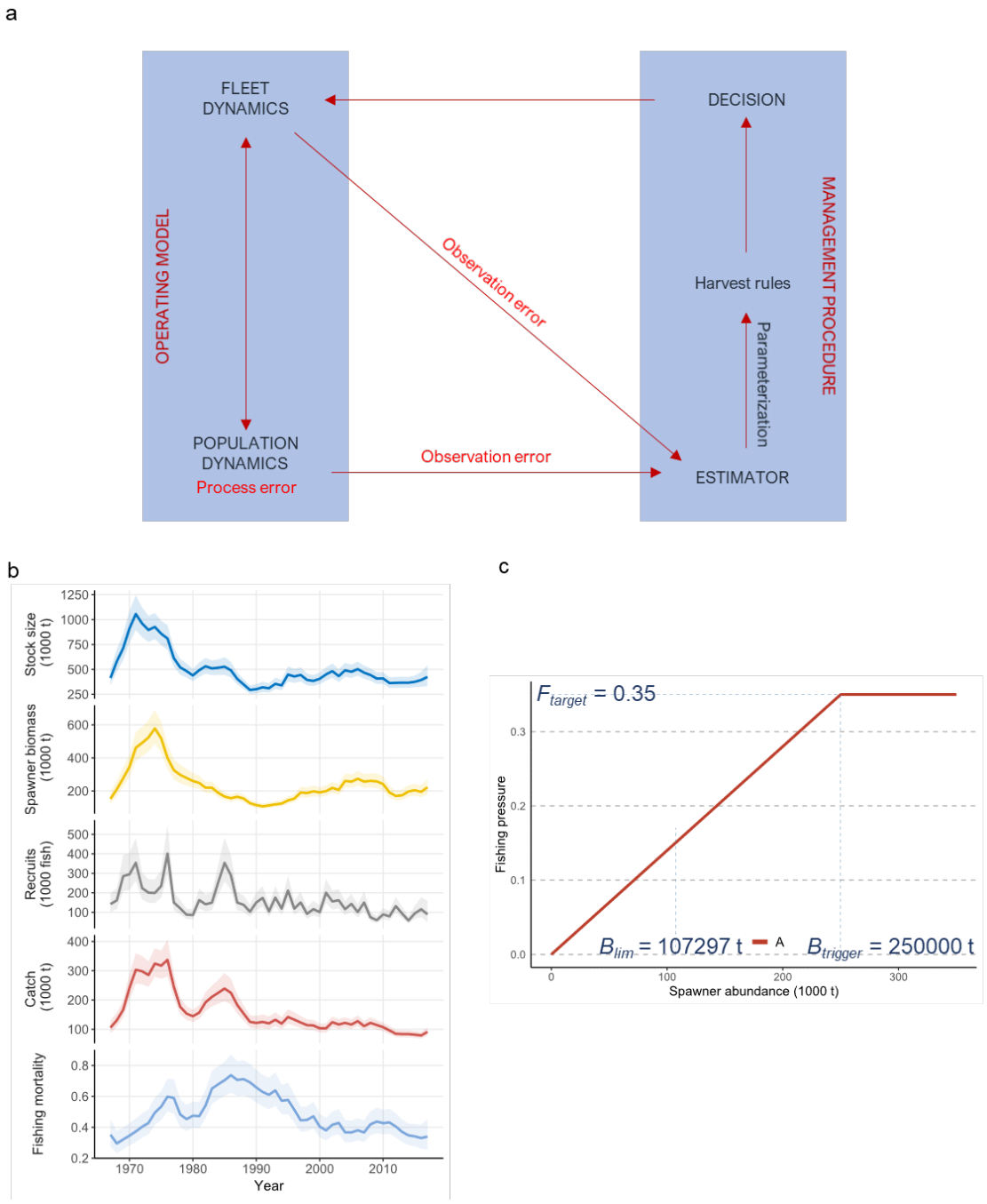

Figure A.3.1. Management strategy evaluation (MSE) framework developed in WKNSMSE (ICES, 2019b) and adopted for WKGMSE3 (a), the 2018 stock assessment (b), and harvest control rule (c) for North Sea saithe.

Operating model (OM)

We used a stochastic, age-structured population model that accounts for environmental stochasticity. The data sources, survey methods, and model structure have been extensively documented in ICES (2017c) and ICES (2019f). Briefly, we parameterized the model with 51-year estimates of age-specific masses (g), maturity rates, and natural mortality rates (0.2 year^{-1} for all ages and years). Then, we fitted it to time series data of commercial catch (age-aggregate biomass of German, French, and Norwegian trawlers in 2000–2017, t) and age-specific (ages 3–8) abundance indices (IBTS-Q3 in 1992–2017) (ICES, 2018b). We simulated density-dependent regulation

of recruitment with a segmented regression. We parameterized the spawner–recruit model by fitting it to the 1998–2017 data. To account for environmental stochasticity in density-dependency of recruitment, we first used a kernel density function to smooth the resulting distribution of residuals from the fitted regression. Then, we resampled residuals (with replacement) from the distribution and applied to model outputs to generate recruits every simulation year; this process was repeated independently for each replicate.

Preliminary analyses showed little evidence of temporal autocorrelation in recruitment. To account for process uncertainty, we generated 1000 realisations of stochastic populations using the variance-covariance matrix of estimable parameters (age-specific numbers and fishing mortality rates) taken from the 2018 assessment. We derived a set of mean age-specific masses, maturity rates, and gear selectivity by randomly selecting a year (with replacement) from the 2008–2017 data; this process was repeated independently for each replicate every simulation year.

We simulated future annual monitoring of the population and harvest by adding observation error to survey indices and age-specific catch computed from the population OM, assuming that the model is fixed (life-history parameters such as maturity rates are time-invariant). We simulated deviances to the observed survey index (IBTS-Q3) using the variance-covariance matrix for the survey index to account for observation error correlated between ages. Observation error is included on age-specific abundance indices as multiplicative lognormal error. We simulated uncertainty in reported catch by computing an exploitable biomass index generated from the population OM.

Management procedure (MP)

In the MP, the current status is assessed annually by fitting an estimation model (EM) to the time series data passed on from the observation model (survey and catch indices) before provision of catch advice in May. We used the State-space Assessment Model (Nielsen and Berg, 2014) as an EM and harvest control rule set for saithe ("HCR-A", Figure A.3.1c); model settings and forecast assumptions are fully described in ICES (2019f). Under this HCR, the following year's catch target is derived from a target exploitation rate (F_{target}) and stock abundance (t) when the estimated spawning stock biomass (SSB) in the current assessment year remains above a fixed threshold ($B_{trigger}$) (Figure A.3.1c). When the SSB falls below $B_{trigger}$, exploitation rate is adjusted to F_{target} scaled to the proportion of SSB relative to $B_{trigger}$ (Figure A.3.1c).

Uncertainty scenarios and risk assessment

To explore the relationship between uncertainty levels and risk, we performed MSEs under scenarios of varying levels of uncertainty. We increased uncertainty levels in six sources of uncertainty, two process errors (standard deviation in number-at-age and recruitment in the OM) and four observation errors (survey catchability, covariance in age-specific abundance index, standard deviation in biomass index, and standard deviation in catch), by 10% to 100% (1.1x to 2.0x) from estimated values in the 2018 assessment (Table A.3.1). We then computed SSB and risk in the last 10 years of projection period from 1000 realizations of annual assessments to evaluate performance of the HCR applied. We computed two types of risk metrics (Prob1 and Prob3) defined in ICES (2019a), the mean and maximum probabilities of SSB falling below a limit threshold, B_{lim} (respectively; ICES, 2019a). We estimated B_{lim} using the EqSim R package (https://github.com/ices-tools-prod/msy); B_{lim} for saithe is set to 107 297 t (ICES, 2019b).

Table A.3.1. Estimated uncertainty levels of six select sources included in the North Sea saithe MSE

Source	Median	Range
SD in number-at-age	2.72	1.53-8.58
SD in recruitment	2.74	0.83-61.7
SD in survey catchability	6.90×10^{-6}	5.38×10^{-6}-8.92×10^{-6}
SD in exploitable biomass index	0.10	0.04-0.20
CV in age-specific survey index (IBTS-Q3)	2.77	1.04-6.88
SD in catch	2.72	1.39-23.4

Relationships between uncertainty levels and risk

Simulations showed that the HCR set for saithe was robust to increased uncertainty levels (both Prob1 and Prob3 remained below 5%) in all the sources tested except for exploitable biomass index and survey catchability (Figure A.3.2). The HCR was highly sensitive to increased uncertainty levels in exploitable biomass index and survey catchability (Figure A.3.2); in particular, the HCR became non-precautionary if the uncertainty level in catchability increased by more than 20% from the estimated value (Figure A.3.2). Furthermore, in addition to getting less precise estimates of stock size from the EM, increased uncertainty levels also modified the temporal dynamics. For example, elevated uncertainty levels in number-at-age in the OM initially led to underestimation of stock size, which, in turn, allowed the stock to reach higher biomass (by as much as 12%) by the end of projection period, whereas elevated uncertainty levels in catchability led to persistent overestimation of stock size.

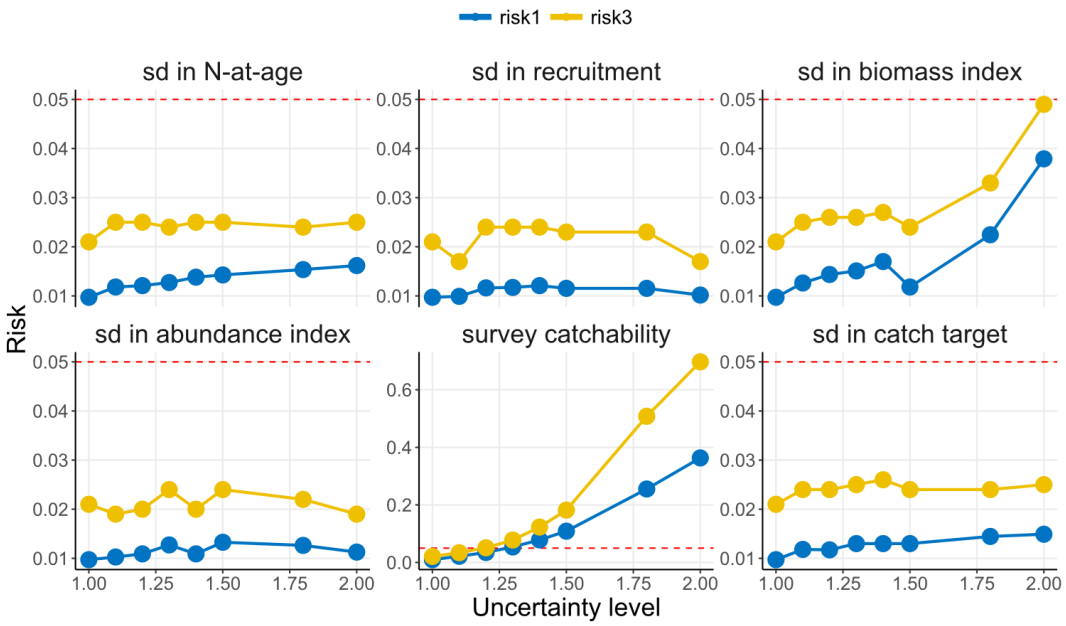

Figure A.3.2. Relationship between uncertainty levels and risk (risk1 = Prob1 and risk3 = Prob3).

Stability in risk with varying numbers of replicates

To explore how elevated levels of uncertainty influence the reliability of risk, we performed additional MSEs with varying numbers (1000 to 10 000) of replicates for the baseline and two

scenarios with Prob3 exceeding 5%; 2.0x exploitable biomass index and 1.2x survey catchability. Overall, risk1 was more stable than Prob3; Prob3 was also overestimated when the number of replicates was less than 2000 (Figure A.3.3). Furthermore, compared with the baseline scenario, when the uncertainty levels were elevated, Prob3 stabilized only with more than 5000 replicates (Figure A.3.3).

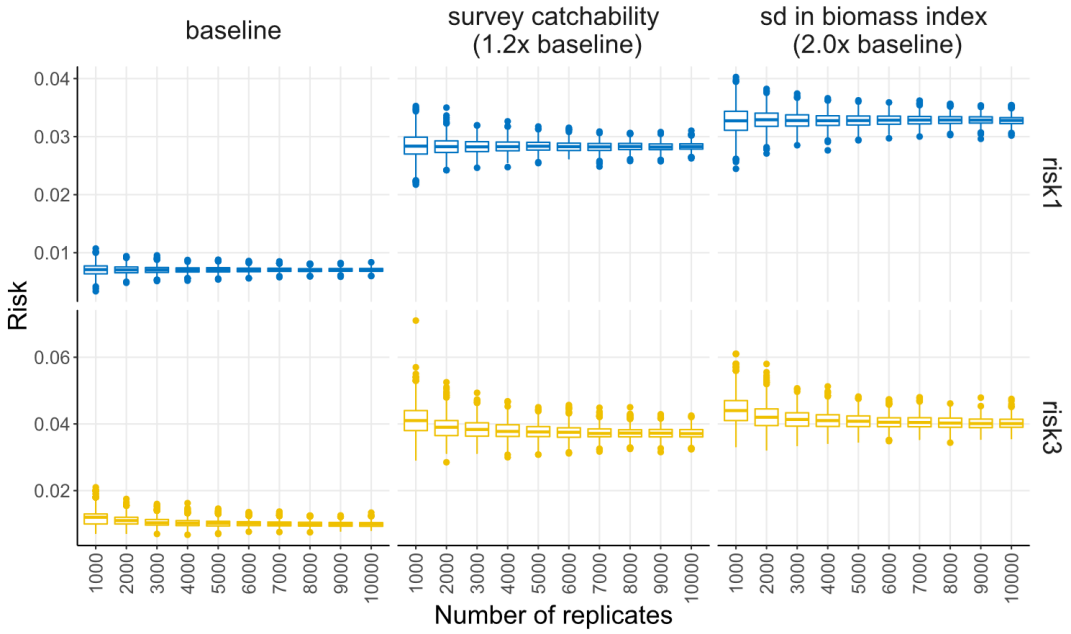

Figure A.3.3. Stability in risk with varying numbers of replicates. The risks (risk1 = Prob1 and risk3 = Prob3) were computed by re-sampling with replacement from simulations with 10 000 replicates. In the box and whisker plots, the thick horizontal line, the edges of the box, the whiskers, and the circles indicate the median, the 25th and 75th percentiles, the 2.5th and 97.5th percentiles, and outliers, respectively.

Stability in risk with varying lengths of projection period

To further explore how elevated levels of uncertainty influence the stability of risk, we also performed MSEs with varying lengths (10 to 100 years) of projection period for the same three scenarios above. Under the baseline scenario, both the risk metrics stabilized by year 50 (Figure A.3.4). By contrast, when uncertainty levels were elevated, the risks remained variable even after 100 years; Prob1 continued declining, whereas Prob3 oscillated (Figure A.3.4).

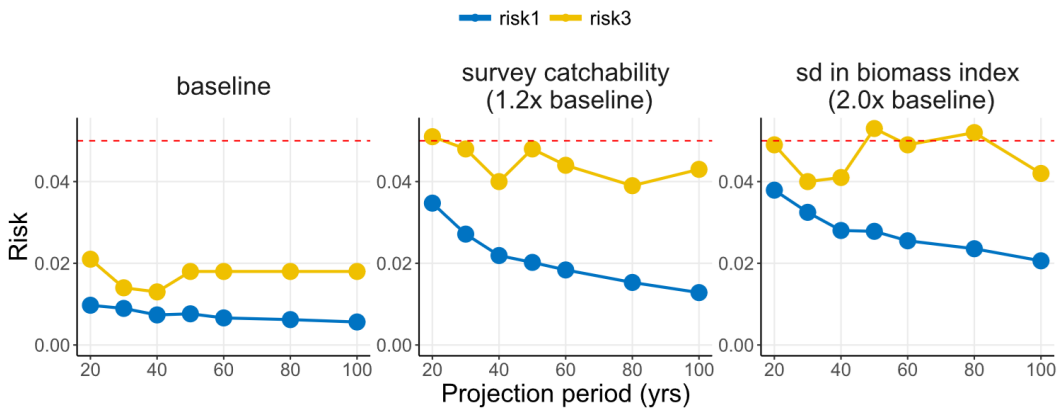

Figure A.3.4. Stability in risk with varying lengths of projection period (risk1 = Prob1 and risk3 = Prob3).

Conclusions

Analyses based on model simulations show that estimated risk strongly depends not only levels of uncertainty but also sources of uncertainty included in MSEs. In the case of North Sea saithe, the HCR set for the stock is robust to a moderate increase in both process errors and observation errors. However, when uncertainty levels become progressively severe, estimated risk may respond nonlinearly, depending on sources of uncertainty. Furthermore, when uncertainty levels are elevated, estimated stock size becomes more variable and less precise. As a result, risk may be overestimated. Although increasing the number of replicates and length of projection period can improve the stability of risk to some extent, these changes in simulation settings also increase the computation cost of performing MSEs. High computation cost is one of the major issues in performing full MSEs. Conducting exploratory simulations with varying numbers of replicates and lengths of projection period is recommend to ensure the reliability of estimated risk under estimated or assumed levels of uncertainty.

Annex 4: Statistical approach for more efficient grid searches involving computer-intensive methods (TOR d)

Michael Spence

Grid searches can be expensive and running every grid value can be wasteful. By considering the outcome of the grid as uncertain and describing "your" beliefs about the grid based on previous runs, as well as knowledge about how the space is characterised, e.g. one may expect the results of the grid search to be smooth, and large areas of the space can be removed without exploring it.

Starting with Kennedy and O'Hagan (2001), this approach has been used in many other fields including cosmology (Vernon at al., 2014), climate science (Holden *et al*., 2015), and more recently in fisheries (Spence *et al*., 2020).

Here, I demonstrate these approaches to the grid search problem to search for the optimal combination of F_{tgrt} and $B_{trigger}$ for North Sea cod that satisfies the criteria:

1. Maximises the median long-term yield
2. Keeps risk of SSB being below B_{lim} to be < 0.05

This work was previously done with a grid search with $F_{trgt} \in \{0.1, 0.11, \ldots, 0.5\}$ and $B_{trigger} \in \{110000, 120000, \ldots, 210000\}$. As this work was for demonstrative purposes, I will use these runs; however, I do not recommend the use a regular search grid. For a review on the design of computer experiments, see Garud *et al*. (2017). The grid search took 451 MSE evaluations, with a single MSE evaluation taking 2 hours. Furthermore, it was possible to do 20 evaluations in parallel using high-performance computing (HPC). Therefore, the full grid search would take a little under 48 hours in computer time

To describe my beliefs about the outcome of an MSE for a particular F_{tgrt} and $B_{trigger}$ value, I used a Gaussian process. I built two independent Gaussian processes for the median long-term yield and the risk that SSB was below B_{lim}. For more details, see Oakley and O'Hagan (2002).

My aim was to learn about the grid and to exclude combinations of F_{tgrt} and $B_{trigger}$ that I was 99.99% or greater confident did not satisfy the criteria. This took an initial run of 20 combinations of F_{tgrt} and $B_{trigger}$ (Figure A.4.1). Figure A.4.2 and Figure A.4.3 show the marginal outcome of the first round runs for the median long-term yield and the risk that the SSB will fall below B_{lim}, respectively. Using quantile regression, generalised additive models (Fasiolo *et al*. 2020), and a Gaussian process, I was able to calculate my beliefs about the median value for the whole space (Figure A.4.4 and Figure A.4.5, respectively). Using the Gaussian process, I am able to calculate my beliefs, expressed as a probability, that each combination of F_{tgrt} and $B_{trigger}$ will lead to a higher median long-term yield than the best one currently found (Figure A.4.6) and that the risk that SSB would be below B_{lim} is less than 0.05 (Figure A.4.7). Excluding the area where I believed the risk that SSB would be below B_{lim} was greater than 0.05 (with probability at least 0.9999), Figure A.4.8 shows my beliefs that the median long-term catch will be larger than my current best estimate, and that the risk of SSB falling below B_{lim} will be larger than 0.05.

I excluded the dark blue region and repeated the process with 20 more model runs. After repeating the process I got Figure A.4.9, opening up a region of space that was previously rejected.

After rejecting the space after round 2, we were left with just 16 potential F_{tgrt} and $B_{trigger}$ combinations. We ran all of these points. We found the best value had F_{trgt} to be 0.38 and $B_{trigger}$ to be 170000. We did this in 56 MSEs in 3 rounds, thus taking 6 hours. See Figure A.4.10 for the searches.

In this demonstration, I got the same answer as the grid search but with only 15% of the model runs. The Appendix to this Annex provides the R code used for the analysis.

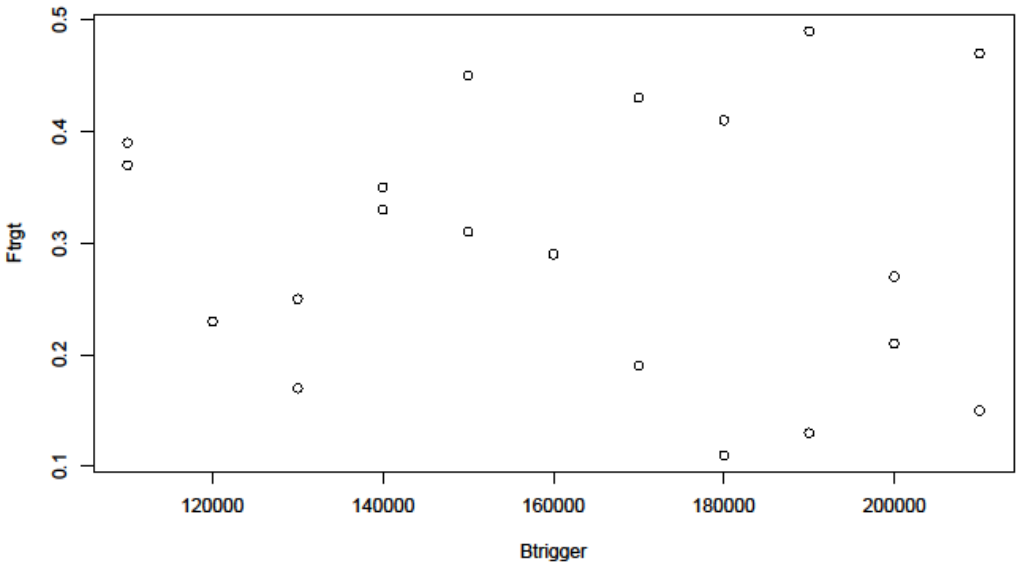

Figure A.4.1: The first-round design points.

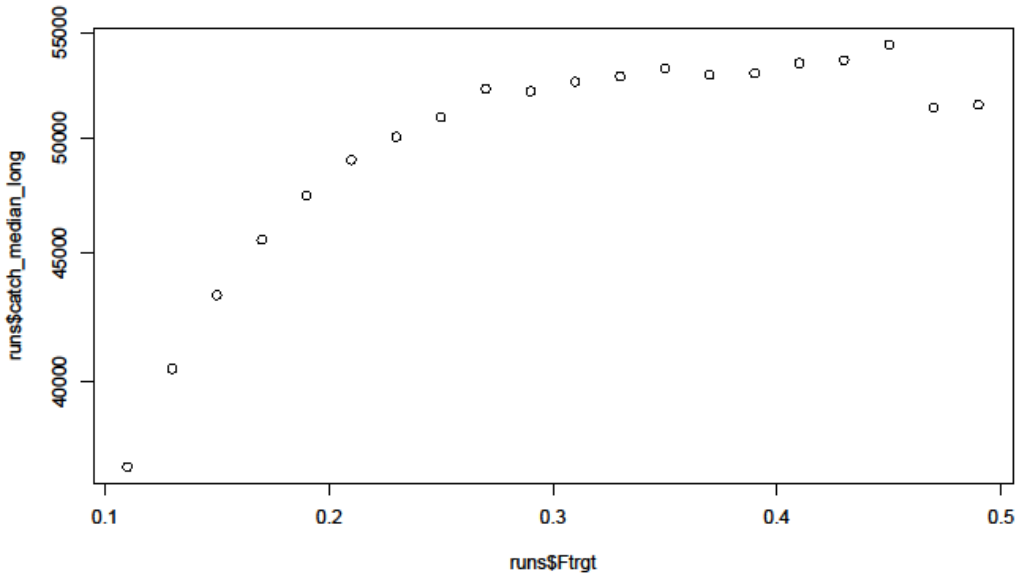

Figure A.4.2: The median long-term catch from the first round.

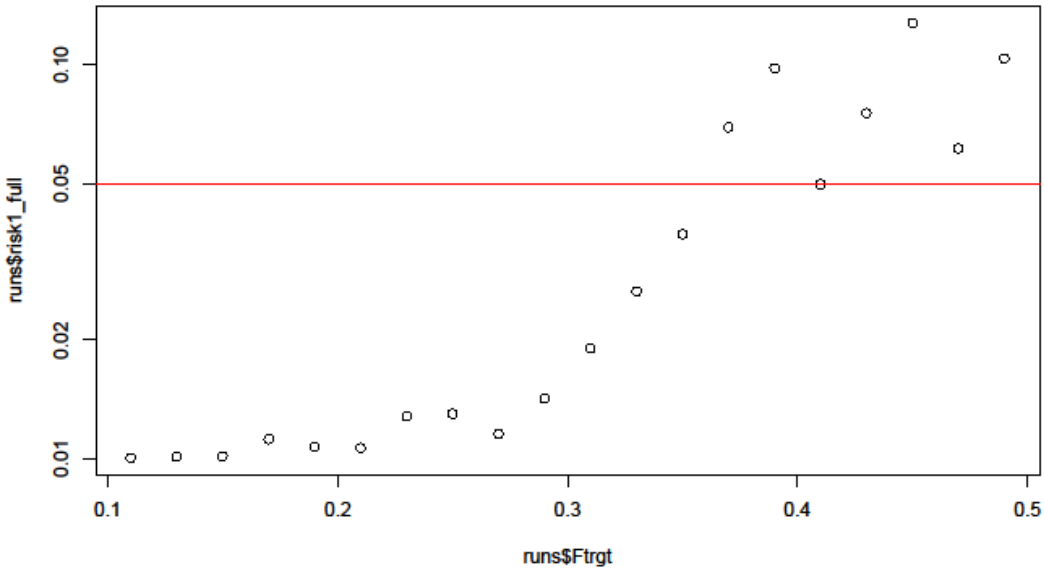

Figure A.4.3: The risk of SSB being below B_{lim} for the runs in the first round. The red line is a risk of 0.05.

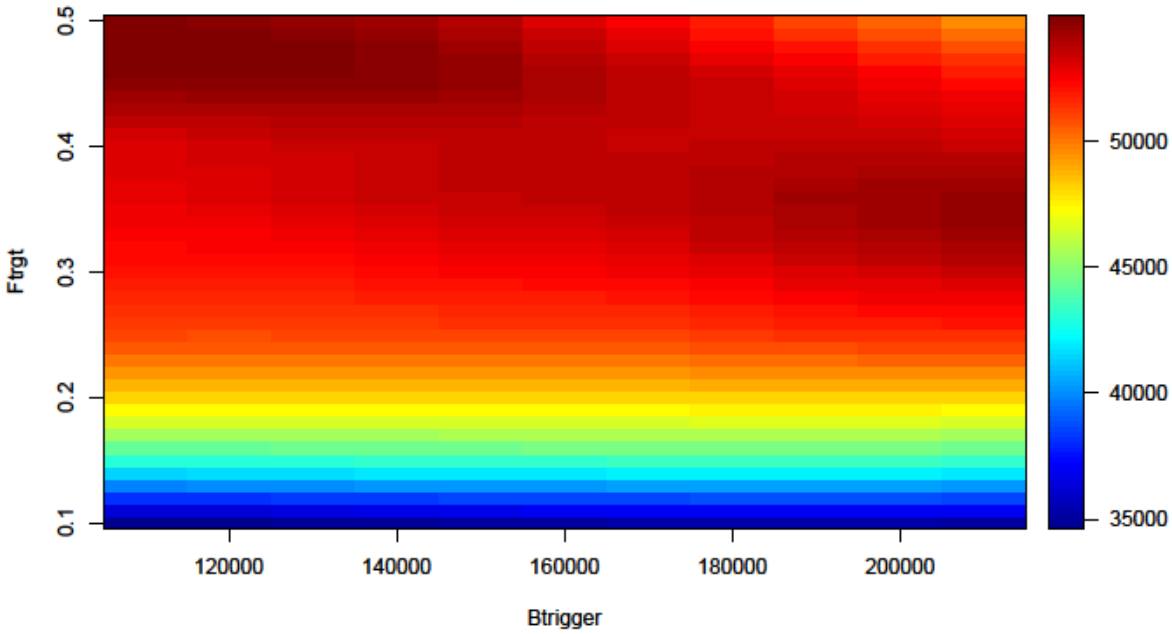

Figure A.4.4: After round 1, my median beliefs of the median long-term catch.

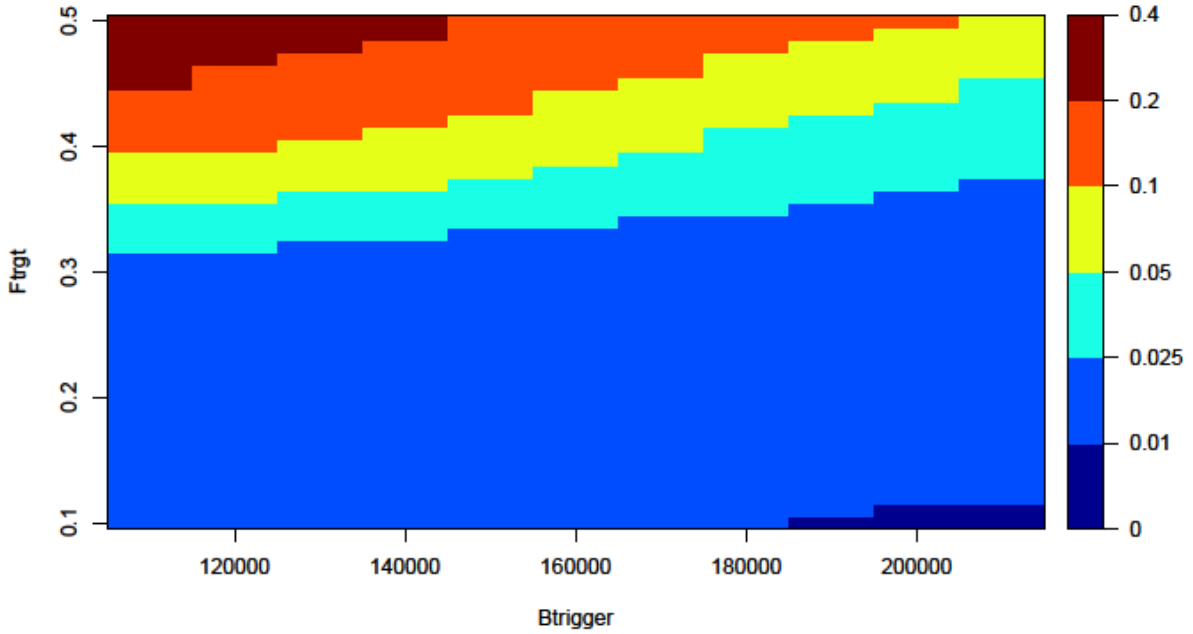

Figure A.4.5: After round 1, my median beliefs of the risk of SSB being below B_{lim}.

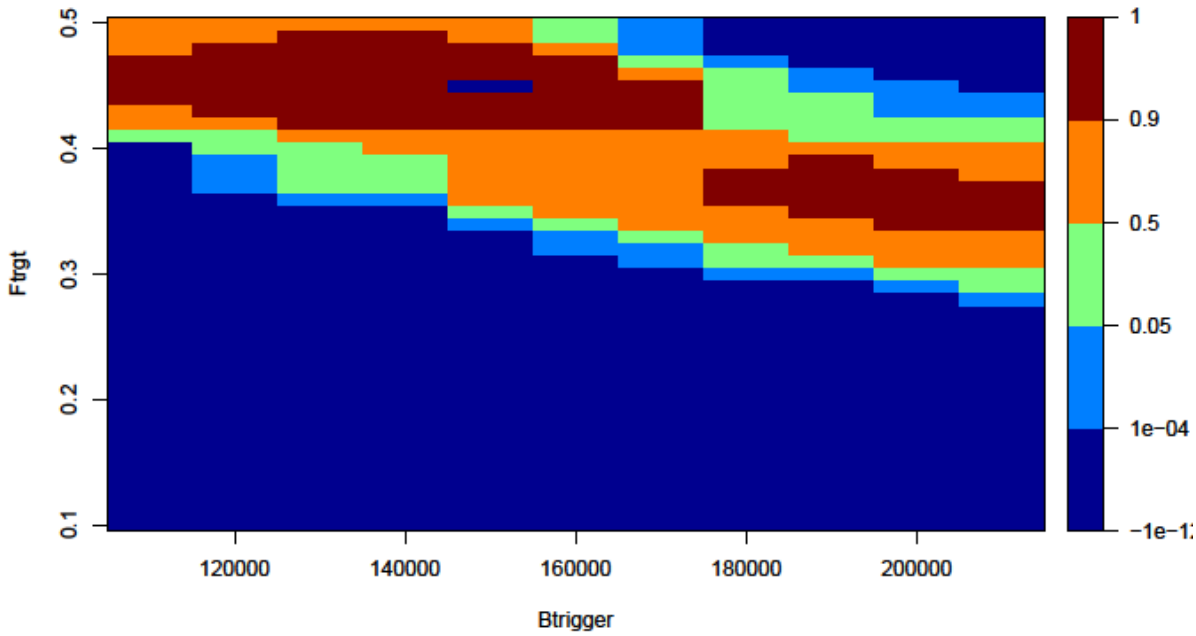

Figure A.4.6: My belief that the median long-term catch will be larger than my current best after round 1.

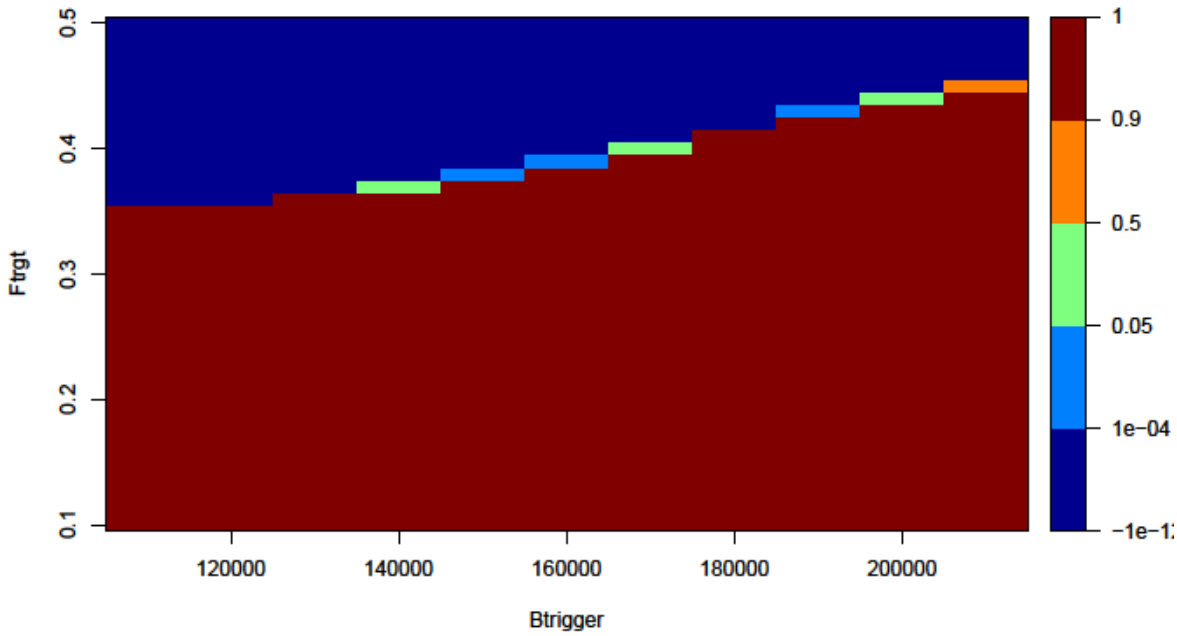

Figure A.4.7: My belief that the risk of SSB falling below B_{lim} will be larger than 0.05 after round 1.

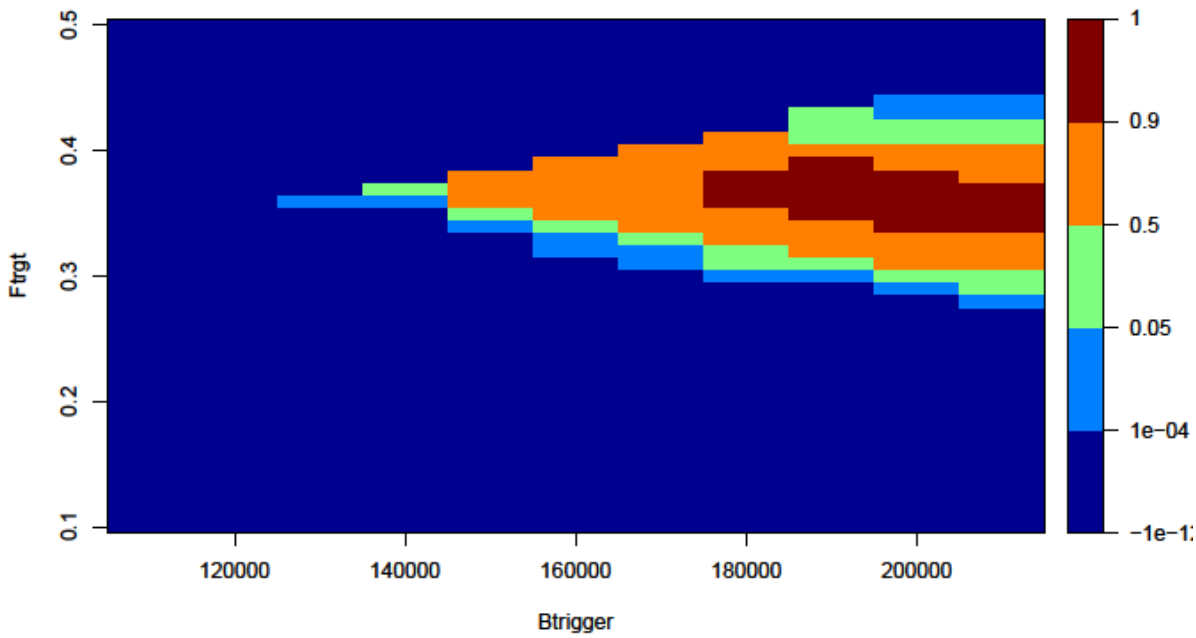

Figure A.4.8: My belief that the median long-term catch will be larger than my current best and that the risk of SSB falling below B_{lim} will be larger than 0.05 after round 1.

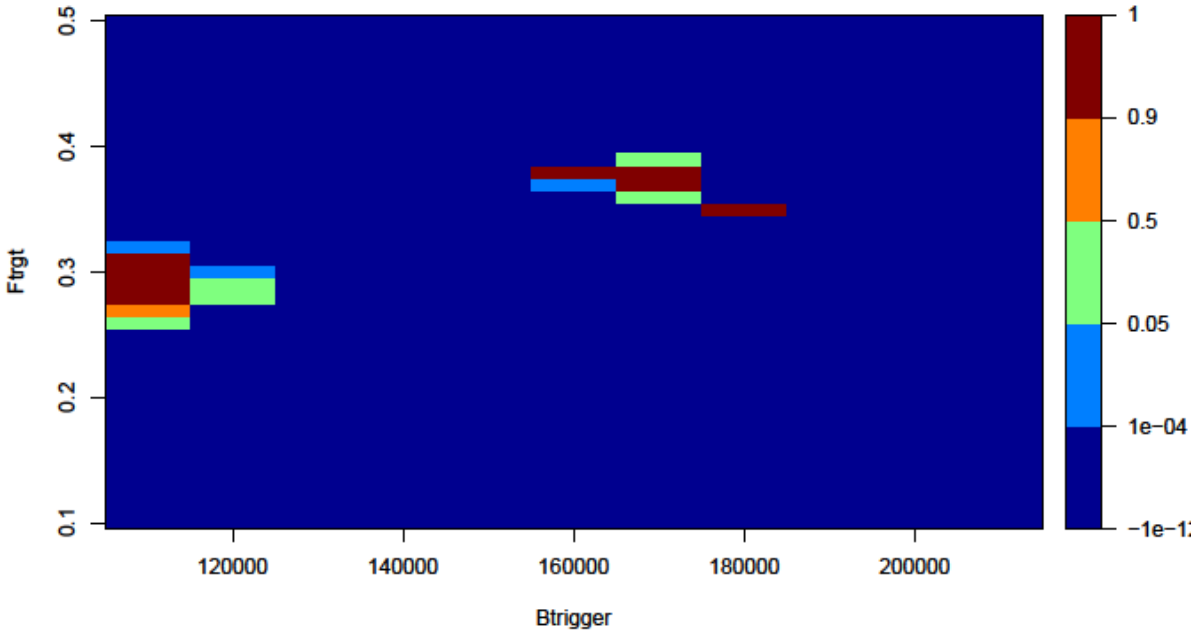

Figure A.4.9: My belief that the median long-term catch will be larger than my current best and that the risk of SSB falling below B_{lim} will be larger than 0.05 after round 2.

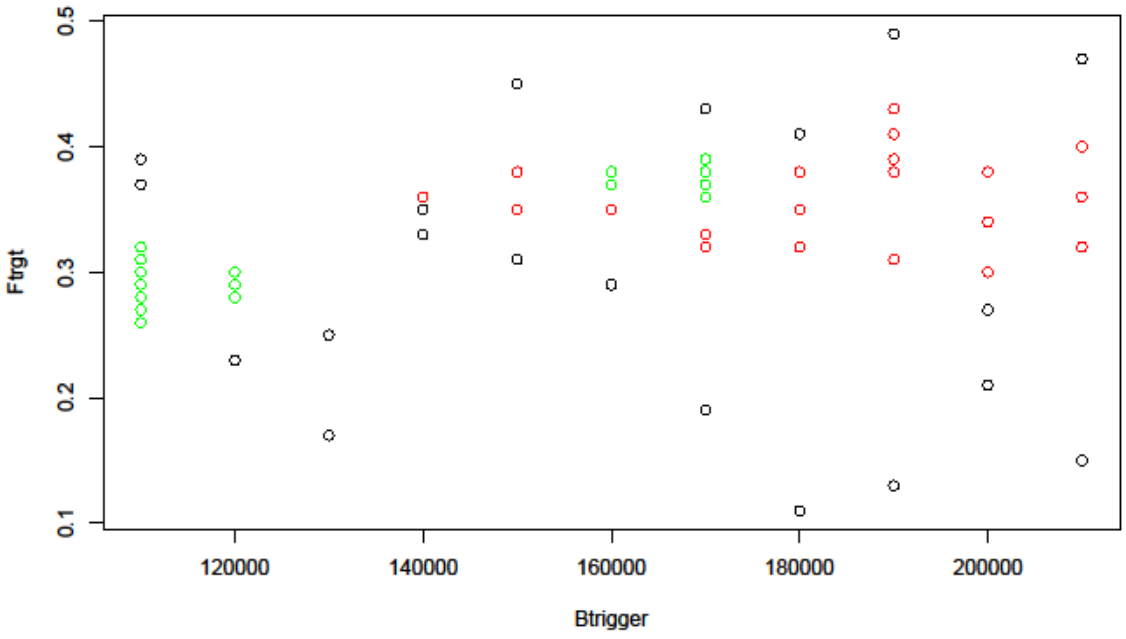

Figure A.4.10: All runs of the MSE. Black is the first round, red the second and green the third.

Appendix to Annex 4

The R code used to do the work. Please note that this work was done very quickly and this code should not be used to repeat the work. For more code see www.mucm.ac.uk. I recommend the uses of the BACCO package rather than DiceKrigging to do the Gaussian process. The main difference is that the Gaussian process should be fitted in a Bayesian framework, especially as we are interested in the extreme tails of the distribution, whereas DiceKrigging fits by maximum likelihood. The reason for using DiceKrigging as opposed to the full Bayesian approach was due to my previous experience of using Gaussian process emulators and the short turnaround of the demonstration for WGMSE3 (this work was done the day before the meeting). Previously I was interested in more probable regions of the space and thus fitting by maximum likelihood was sufficient for such problems.

```
#### Code for Gaussian process emulator
### Code is for demonstration process - NB more time needs to be spent on the Gaussian Process. I recommend the BACCO package but have used the DiceKriging package here

`stats_full_HCR-A` <- readRDS("C:/Users/MS23/OneDrive - CEFAS/ICES working groups/WKGMSE 2020/Playing about/stats_full_HCR-A.rds")
dat <- `stats_full_HCR-A`

plot(dat$Ftrgt,log(dat$catch_median_long),log="y")
plot(dat$Btrigger,log(dat$catch_median_long),log="y")
plot(dat$Ftrgt,dat$risk1_full,log="y")
abline(h=0.05,col="red")
plot(dat$Btrigger,dat$risk1_full,log="y")
abline(h=0.05,col="red")

## 1 cell 2 hours -- optimised
## 451 cells = 902 hours

## for speed lets use gams and DiceKriging although other techniques maybe useful
####
length(unique(dat$Ftrgt)) ## 41 values
length(unique(dat$Btrigger)) ## 11 Btrigger

set.seed(14)
## for now lets do a random sample although a space filling algorithm would be better
round1 <- data.frame(Ftrgt=sample(unique(dat$Ftrgt)[seq(2,40,2)]),Btrigger=sample(rep(unique(dat$Btrigger),2),size = 20))
plot(round1[,2:1])
## collect runs
library(dplyr)
runs <- left_join(round1,dat)
##
plot(runs$Ftrgt,runs$catch_median_long,log="y")
plot(runs$Ftrgt,runs$risk1_full,log="y")
abline(h=0.05,col="red")

gridd <- dat[,c("Ftrgt","Btrigger")]
## order is messed up
gridd<- gridd[order(gridd$Ftrgt,gridd$Btrigger),]

library(qgam)
library(DiceKriging)
### build the emulator for median catch
qgams_cat <- qgam(log(catch_median_long)~s(Ftrgt),data=runs,qu=0.5) ## quantile regression for robustness
res_cat <- log(runs$catch_median_long) - predict(qgams_cat)
gp_cat <- km(~1,design=runs[,c("Ftrgt","Btrigger")],estim.method="LOO",response = res_cat) ### need a better way of doing GPs -- see www.mucm.ac.uk maybe
### build the emulator for risk
qgams_risk <- qgam(log(risk1_full)~s(Ftrgt),data=runs,qu=0.5) ## quantile regression for robustness
res_risk <- log(runs$risk1_full) - predict(qgams_risk)
gp_risk <- km(~1,design=runs[,c("Ftrgt","Btrigger")],estim.method="LOO",response = res_risk) ### need a better way of doing GPs

### now lets check to see where we want to search next -- first lets look at risk
pred_risk1_q <- predict(qgams_risk,newdata=gridd)
pred_risk1_g <- predict(gp_risk,newdata=gridd,type="SK")
```

```
## median risk
med_risk1 <- exp(pred_risk1_g$mean + pred_risk1_q)
## plot it
library(plot3D)
image2D(ma-
trix(med_risk1,nrow=11),breaks=c(0,0.01,0.025,0.05,0.1,0.2,0.4),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrig-
ger)),xlab="Btrigger",ylab="Ftrgt")
## now lets look at the probability that the risk is more than
prisk1 <- pnorm(log(0.05),pred_risk1_g$mean + pred_risk1_q,pred_risk1_g$sd+1e-12)
image2D(matrix(prisk1,nrow=11),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrigger)),xlab="Btrig-
ger",ylab="Ftrgt",breaks=c(-1e-12,0.0001,0.05,0.5,0.9,1))

### now the catch
pred_cat1_q <- predict(qgams_cat,newdata=gridd)
pred_cat1_g <- predict(gp_cat,newdata=gridd,type="SK")
## median risk
med_cat1 <- exp(pred_cat1_g$mean + pred_cat1_q)
image2D(matrix(med_cat1,nrow=11),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrigger)),xlab="Btrig-
ger",ylab="Ftrgt")
max1 <- max(runs$catch_median_long[runs$risk1_full < 0.05])### the max so far -- maybe chose something else, maybe
we want to know more about the space around the maximum
pcat1 <- ifelse(pred_cat1_g$sd==0,as.numeric(abs((pred_cat1_g$mean + pred_cat1_q) - log(max1)) > 1e-
5),pnorm(log(max1),pred_cat1_g$mean + pred_cat1_q,pred_cat1_g$sd+1e-12))

image2D(matrix(1-pcat1,nrow=11),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrigger)),xlab="Btrig-
ger",ylab="Ftrgt",breaks=c(-1e-12,0.0001,0.05,0.5,0.9,1))

### find the ones that could be larger

possible <- (apply(cbind((1-pcat1) , prisk1),1,min) >  1e-04)
## plot the possible ones
image2D(matrix(possible * (1-pcat1),nrow=11),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrigger)),xlab="Btrig-
ger",ylab="Ftrgt",breaks=c(-1e-12,0.0001,0.05,0.5,0.9,1))

pot_points <- gridd[possible,]
p_later<-any(pot_points$Ftrgt==0.38 & pot_points$Btrigger==170000)
best_so_far<- runs[runs$risk1_full < 0.05,][which.max(runs$catch_median_long[runs$risk1_full < 0.05]),c("Ftrgt","Btrig-
ger")]
pot_points <- pot_points[-which(pot_points$Ftrgt==best_so_far$Ftrgt & pot_points$Btrigger==best_so_far$Btrigger),]

#### round2 points -- doing them randomly another 20 -- need to improve this search algorithm
nums <- sample(nrow(pot_points),20)

round2 <- pot_points[nums,]
plot(round1[,2:1])
points(round2[,2:1],col="red")
runs <- rbind(runs,left_join(round2,dat))

plot(runs$Ftrgt,runs$catch_median_long,log="y")
plot(runs$Ftrgt,runs$risk1_full,log="y")
abline(h=0.05,col="red")

### update our beliefs -- NB not refitting -- should refit really!
res_cat1 <- log(runs$catch_median_long) - predict(qgams_cat,newdata=runs)
gp_cat1 <- km(~1,design=runs[,c("Ftrgt","Btrigger")],estim.method="LOO",response = res_cat1,coef.trend =
gp_cat@trend.coef,coef.var = gp_cat@covariance@sd2,coef.cov = gp_cat@covariance@range.val)
res_risk1 <- log(runs$risk1_full) - predict(qgams_risk,newdata=runs)
gp_risk1 <- km(~1,design=runs[,c("Ftrgt","Btrigger")],estim.method="LOO",response = res_risk1,coef.trend =
gp_risk@trend.coef,coef.var = gp_risk@covariance@sd2,coef.cov = gp_risk@covariance@range.val)

pred_risk2_q <- predict(qgams_risk,newdata=gridd)
pred_risk2_g <- predict(gp_risk1,newdata=gridd,type="SK")
## median risk
med_risk2 <- exp(pred_risk2_g$mean + pred_risk2_q)
## plot it
image2D(ma-
trix(med_risk2,nrow=11),breaks=c(0,0.01,0.025,0.05,0.1,0.2,0.4),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrig-
ger)),xlab="Btrigger",ylab="Ftrgt")
## now lets look at the probability that the risk is more than
prisk2 <- pnorm(log(0.05),pred_risk2_g$mean + pred_risk2_q,pred_risk2_g$sd+1e-12)
image2D(matrix(prisk2,nrow=11),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrigger)),xlab="Btrig-
ger",ylab="Ftrgt",breaks=c(-1e-12,0.0001,0.05,0.5,0.9,1))

pred_cat2_q <- predict(qgams_cat,newdata=gridd)
pred_cat2_g <- predict(gp_cat1,newdata=gridd,type="SK")
```

```
## median catch
med_cat2 <- exp(pred_cat2_g$mean + pred_cat2_q)
image2D(matrix(med_cat2,nrow=11),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrigger)),xlab="Btrig-
ger",ylab="Ftrgt")
max2 <- max(runs$catch_median_long[runs$risk1_full < 0.05])### the max so far
#pcat2 <- pnorm(log(max2),pred_cat2_g$mean + pred_cat2_q,pred_cat2_g$sd+1e-12)

pcat2 <- ifelse(pred_cat2_g$sd==0,as.numeric(abs((pred_cat2_g$mean + pred_cat2_q) - log(max2)) > 1e-
5),pnorm(log(max2),pred_cat2_g$mean + pred_cat2_q,pred_cat2_g$sd+1e-12))

image2D(matrix(1-pcat2,nrow=11),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrigger)),xlab="Btrig-
ger",ylab="Ftrgt",breaks=c(-1e-12,0.0001,0.05,0.5,0.9,1))

## only this many left
possible1 <- (apply(cbind((1-pcat2) , prisk2),1,min) >  1e-04)
## just the possible ones
image2D(matrix(possible1*(1-pcat2),nrow=11),y=sort(unique(dat$Ftrgt)),x=sort(unique(dat$Btrigger)),xlab="Btrig-
ger",ylab="Ftrgt",breaks=c(-1e-12,0.0001,0.05,0.5,0.9,1))

pot_points <- gridd[possible1,]
p_later2<-any(pot_points$Ftrgt==0.38 & pot_points$Btrigger==170000)
## remove the best
best_so_far<- runs[runs$risk1_full < 0.05,][which.max(runs$catch_median_long[runs$risk1_full < 0.05]),c("Ftrgt","Btrig-
ger")]
pot_points <- pot_points[-which(pot_points$Ftrgt==best_so_far$Ftrgt & pot_points$Btrigger==best_so_far$Btrigger),]

################ round 3
### do the rest
plot(round1[,2:1])
points(round2[,2:1],col="red")
#points(round3[,2:1],col="blue")
points(pot_points[,2:1],col="green")
runs <- rbind(runs,left_join(pot_points,dat))
#### check we have it all
# solution
MSY<- runs[runs$risk1_full < 0.05,][which.max(runs$catch_median_long[runs$risk1_full < 0.05]),]
print(MSY[,1:2])
p_later
p_later2
nrow(pot_points)
```

Annex 5: Development of a bootstrapping approach to streamline management strategy evaluations (TOR d)

Iago Mosqueira

The method implemented here relies on the ability of the bootstrap (Efron, 1979) to provide a measure of the precision of an estimate or calculation. This procedure is especially useful when dealing with the precision of extreme quantiles, as is the case with the 5% risk probability of SSB < B_{lim}. The bootstrapping procedure resamples (with replacement) the time series of a given output parameter and returns confidence intervals (CI) around the yearly quantiles of its distribution (e.g. 5% quantile of SSB). In addition, the relative error of the quantile across years is returned (Median Absolute Deviation (MAD) over the median), which can be used as an overall precision metric, useful in the definition of stoppage rules for conducting additional MSE iterations.

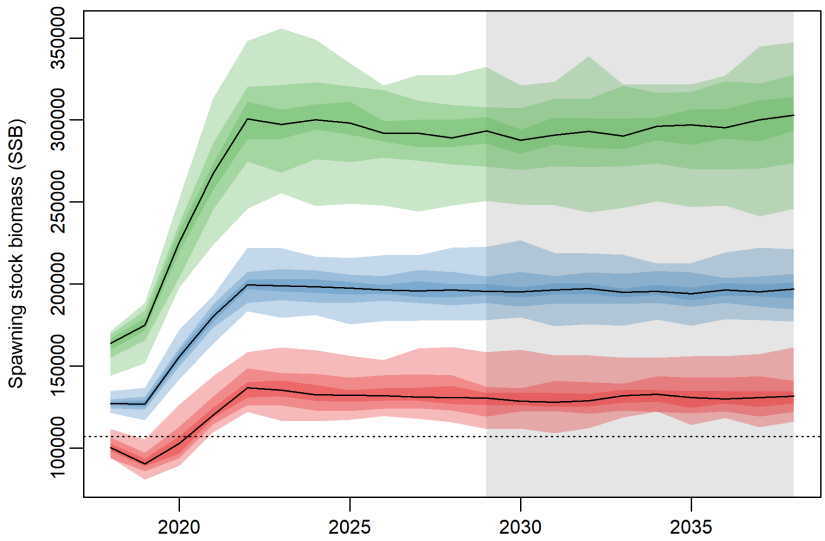

Figure A.5.1. Precision of the estimate of the 5, 50 and 95% quantiles, in red, blue and green, for SSB over the projection period. Black line is the "true" value, from the 1000 iterations, while the shades show the relative error with a larger number of iterations.

The following example demonstrates how the method might be used in a more efficient grid search for HCR parameters that both comply with the threshold of risk to SSB < B_{lim} and maximize catch. For a given combination of HCR parameters (e.g. $B_{trigger}$ and F_{target}), the relative error of the quantile estimate (5%), is computed for an increasing number of iterations, until its value is smaller than some pre-defined value, or a maximum number of iterations is reached. If the upper CI of the desired quantile is then found to be less than the desired target (e.g. B_{lim}), this grid cell is not considered further. Those cells where this is not true are ranked according the secondary

objective (catch). An extra number of iterations is then carried out for the top 75% of them, with a minimum of 30.

The relative error metric is expected to decrease with increasing number of iterations, and the relationship is roughly linear in log-scale. Using progressively smaller subsets of the completed iterations, the relationship for a given quantile can be derived and used to predict the number of total iterations required for a given level of precision. The code below implements the bootstrap calculation and then plots the relationship between the number of iterations carried out and the precision level for a series of distribution quantiles.

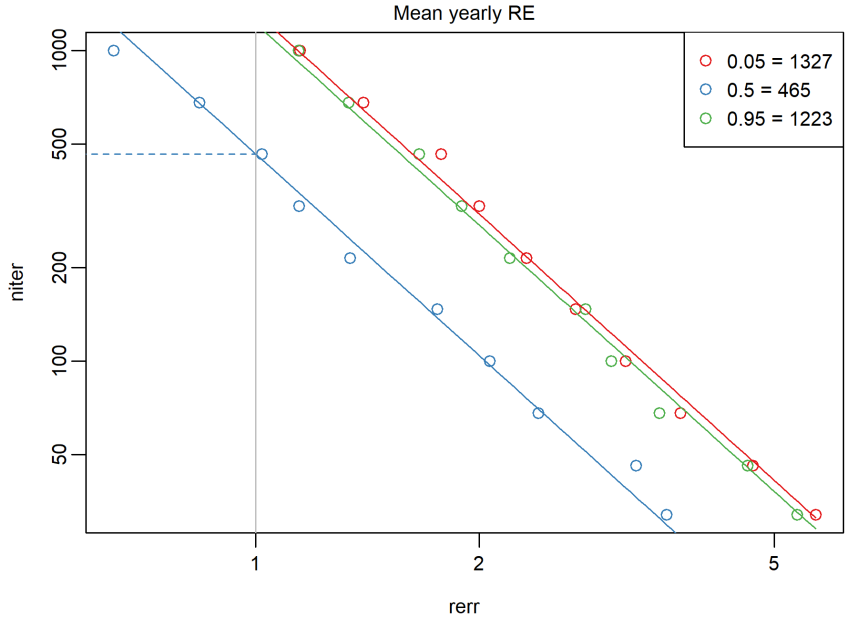

Figure A.5.2. Log-log plot of mean bootstrapped relative error (MAD) versus number of iterations by quantile. Predicted linear regressions are shown by solid lines, and predicted number of iterations needed for the target error level are shown by dashed lines.

Appendix to Annex 5

Code for bootstrap approach to streamline MSE

```
#' Calculate MSE quantile estimation error via bootstrapping and predict
#' number of iterations needed for a given precision.
#'
#' @description Calculates bootstrapped estimates of quantiles (median).
#'   Estimation error (median absolute deviation (MAD)) and confidence
#'   intervals (CI) are also estimated. The summary statistic of
#'   relative error (rerr=(MAD/est)*100) is also returned. When estimates are
#'   performed over various levels of iterations (`niter`), a linear
#'   regression is fit to predict the number of iterations needed to achieve
#'   a given level of error (`rerrTarget`).
#'
#' @param X matrix. Statistic of interest across years (columns) and
#'   iterations (rows)
#' @param nboot numeric. Number of bootstraps.
#' @param quant vector. Quantile probabilities
#'   (Default: \code{quant = c(0.95,0.75,0.5,0.25,0.05)}).
#' @param niter vector. Cumulative number of iterations to perform bootstrapped
#'   estimates (Default: \code{niter = exp(seq(log(nrow(X))*0.5, log(nrow(X)),
#'   length.out = 5))}).
#' @param ci numeric. Confidence interval level to use for
#'   lower (\code{cilow}) and upper (\code{ciup}) limits of bootrapped
#'   quantiles (Default: `ci = 0.95`).
#' @param aggfun string. Name of function to apply to quantile relative error
#'   across years (Default: \code{aggfun = "max"}). Use of "max" is
#'   consistent with 'Prob3', which is the maximum annual probability of SSB
#'   dropping below Blim. 'Prob2' would be the average annual probability
#'   of SSB dropping below Blim (i.e. \code{aggfun = "mean"}). Statistic is
#'   applied to each level of `niter``.`
#' @param rerrTarg numeric. Desired relative error (as a percent; Default: 1)
#'   for which to predict the required number of iterations. Only applicable if
#'   three or more levels of `niter` are bootstrapped, allowing for the fitting
#'   (and prediction) of relative error as a function of number of iterations.
#' @param verbose logical. Should progress be printed
#'   (Default: code\{verbose = TRUE})
#'
#'
#' @return list. Contains bootstrapped estimates of quantiles (`quant`)
#'   (estimate: `est`)
#' @examples
#' library(FLCore)
#' data(ple4)
#' X <- t(rlnorm(400, log(ssb(ple4)), 0.4)[drop=TRUE])
#' qbssb <- qboot(X, nboot=100, aggfun="mean")

qboot <- function(X, nboot = 30, quant = c(0.95,0.75,0.5,0.25,0.05),
  niter = exp(seq(log(nrow(X))*0.5, log(nrow(X)), length.out = 5)),
  ci = 0.95, aggfun = "mean", rerrTarg = 1, verbose = TRUE) {

  # add some tests of X dims
  X <- as.array(X)

  # empty matrix for bootstrapped quantiles
  qn <- array(NaN, dim=c(ncol(X), length(quant), nboot),
    dimnames = list(dimnames(X)[[2]], quant=quant, boot=seq(nboot)))
  # empty matrix for std.err and mean results
  est <- mad <- cilow <- ciup <- array(NaN, dim=c(ncol(X), length(quant), length(niter)),
    dimnames = list(dimnames(X)[[2]], quant=quant, niter=niter))

  if(verbose & length(niter) > 1)
    pb <- txtProgressBar(min = 1, max = length(niter), style = 3)

  for(i in seq(length(niter))) {

    # determine the quantiles of sub-sampled iterations (bootstrapping)
    for(n in seq(nboot)) {
      if(ncol(X)==1){
        qn[,,n] <- quantile(X[sample(nrow(X), niter[i], replace = TRUE),],
          prob = quant)
      } else {
        qn[,,n] <- t(apply(X[sample(nrow(X), niter[i], replace = TRUE),],
```

```
        MARGIN = 2, FUN = quantile, prob = quant))
    }
   }

   # determine the CIs of bootstrapped quantiles
   Q <- apply(qn, MARGIN = 1:2, FUN = quantile, prob = c((1-ci)/2, 0.5, ci+(1-ci)/2))
   cilow[,,i] <- Q[1,,]
   est[,,i] <- Q[2,,]
   ciup[,,i] <- Q[3,,]

   # median absolute deviation of bootstrapped quantiles
   mad[,,i] <- apply(qn, MARGIN = 1:2, FUN = function(x){
    median(abs(x - median(x)))
   })
   if(verbose & length(niter) > 1) setTxtProgressBar(pb, i)
  }
  if(verbose & length(niter) > 1) close(pb)

  # relative error of MAD
  rerr <- mad / est * 100

  # fit log-log linear regression; lm(log(niter) ~ log(rerr) + quant)
  if(length(niter) >= 3){
   df <- expand.grid(dimnames(rerr))
   df$rerr <- c(rerr)
   agg <- aggregate(rerr ~ quant + niter, data = df, FUN = aggfun)
   agg$niter <- as.numeric(as.character(agg$niter))
   fit <- lm(log(niter) ~ log(rerr) + quant, agg)
   newdat <- data.frame(rerr = rerrTarg, quant = factor(quant))
   newdat$niter <- exp(predict(fit, newdata = newdat))
  }else{
   if(verbose){
    print("length(niter) < 3; No fitting of model conducted: log(niter)~log(rerr)+quant")
   }
   agg <- NULL
   fit = NULL
   newdat = NULL
  }

  # return results
  ret <- list(
   est = est, cilow = cilow, ciup = ciup, mad = mad, rerr = rerr,
   fit = fit, fitpred = newdat,
   nboot = nboot, quant = quant, niter = niter, agg = agg,
   aggfun = aggfun, ci = ci, ciLevs = c((1-ci)/2, ci+(1-ci)/2), rerrTarg = rerrTarg)
  class(ret) <- "qboot"
  return(ret)
}

#' Plot results of qboot
#'
#' @param obj Result of running qboot, a list of class 'qboot'
#' @param pch Symbol to use in plot, defaults to 19
#' @param col Colors to apply to each quantile
#' @param ...
#'
#' @return plot of qboot results
#' @export
#'
#' @examples
#' library(FLCore)
#' data(ple4)
#' X <- t(rlnorm(400, log(ssb(ple4)), 0.4)[drop=TRUE])
#' qbssb <- qboot(X, nboot=100, aggfun="mean")
#' plot(qbssb)

plot.qboot <- function(obj, pch = 19, col = NULL, ...){
  if(missing(col)) col <- seq(obj$quant)

  if(is.null(obj$fit)){
   stop("Nothing to plot. No fitting of model conducted: log(niter)~log(rerr)+quant")
  }

  newdat <- expand.grid(
   rerr = seq(min(obj$agg$rerr), max(obj$agg$rerr), length.out = 100),
```

```
  quant = unique(obj$agg$quant))
newdat$niter <- exp(predict(obj$fit, newdata = newdat))

plot(niter ~ rerr, obj$agg, log = "xy", pch = pch, col = col,
  xlab="Relative error", ylab="No. iterations", ...)
legend("topright", legend = paste(obj$quant, "=", round(obj$fitpred$niter)),
  pch = pch, col = col, ...)

for(i in seq(levels(newdat$quant))){
  lines(niter ~ rerr, data = newdat,
    subset = quant == levels(newdat$quant)[i],
    col = col[i], ...)
}
abline(v = obj$rerrTarg, col = 8)
for(i in seq(levels(newdat$quant))){
  hit <- which(factor(obj$fitpred$quant) == levels(newdat$quant)[i])
  segments(x0 = 1e-6, x1 = obj$rerrTarg,
    y0 = obj$fitpred$niter[hit], y1 = obj$fitpred$niter[hit],
    col = col[i], lty = 2)
}
```

Annex 6: Sprat MSE: full vs shortcut (TOR e)

Mollie Brooks

Introduction

As part of the sprat benchmark in 2018, WKspratMSE performed MSEs on North Sea sprat (ICES, 2019g). The MSEs were conditioned using the assessment chosen during the benchmark. The assessment model is SMS with a quarterly time-step, tracking age groups 0, 1, 2, and 3+. For details see ICES WKspratMSE Report (ICES, 2019g). The MSE framework used in WKspratMSE was modified for this WK by adding observation error to biological parameters; previously (including a first iteration presented to the group), they were assumed to be observed without error due to intensive sampling of this data rich stock.

All MSEs were run for 35 years. Each had 1000 simulation trials which varied in their true exploitation patter (E) and true biological parameters as described in ICES WKspratMSE Report 2018. Here, we used $F_{cap} = 0.69$ in the harvest control rule on top of the escapement strategy because that was determined to be precautionary in the past. Performance statistics were calculated in the same way for all MSEs, as defined in ICES WKspratMSE Report 2018, i.e. we did not change B_{lim} depending on the OM.

Observation simulator

The observation simulator produces observations of the catch, surveys, and biological parameters. All 3 surveys use the catchability estimates from the most recent benchmark. All OMs included observation error on the catch and surveys, parameterized based on the SMS fit from the benchmark. For the baseline MSE, it was assumed that biological parameters are observed without error, but see section on alternative operating models below. As shown in Figures A.6.1 and A.6.2, observed biological parameters are passed to both the assessment model (SMS) and short term forecast in the full MSE, while they are only passed to the short term forecast in the shortcut MSE.

Assessment emulator

The same assessment emulator was used in all MSEs. It is based on the variance covariance matrix of log(N) and log(E) from the assessment chosen at the benchmark. That matrix is used to generate multivariate normal errors with mean zero to add error to the true values of log(N) and log(E) from the OM.

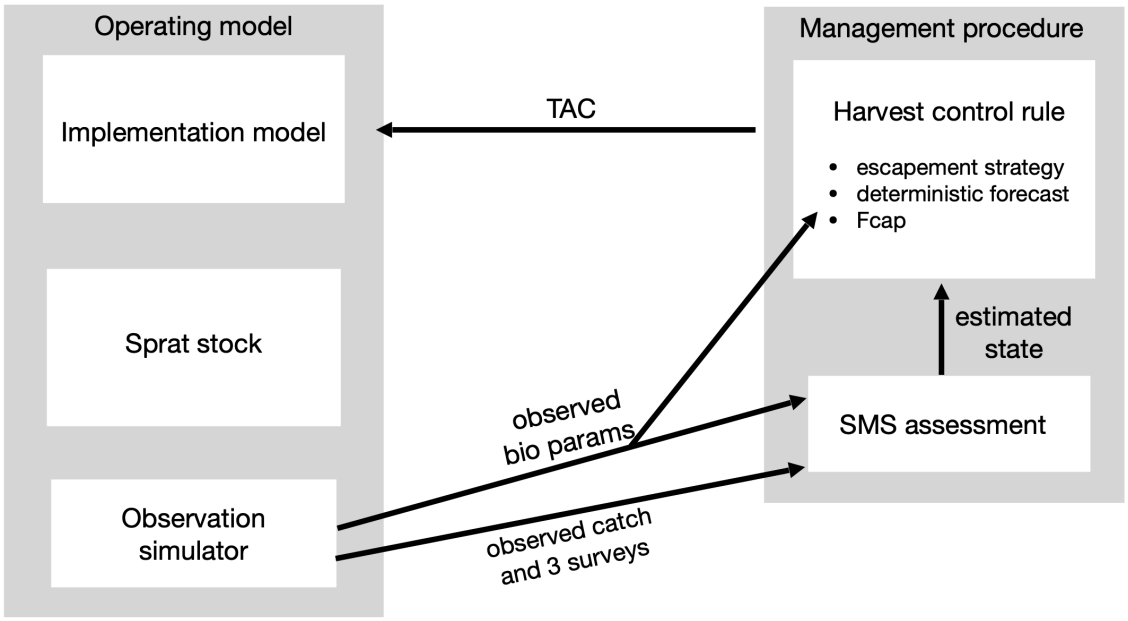

Figure A.6.1 Full MSE for North Sea sprat conceptual drawing. The "estimated state" contains seasonal and age structured N and exploitation pattern.

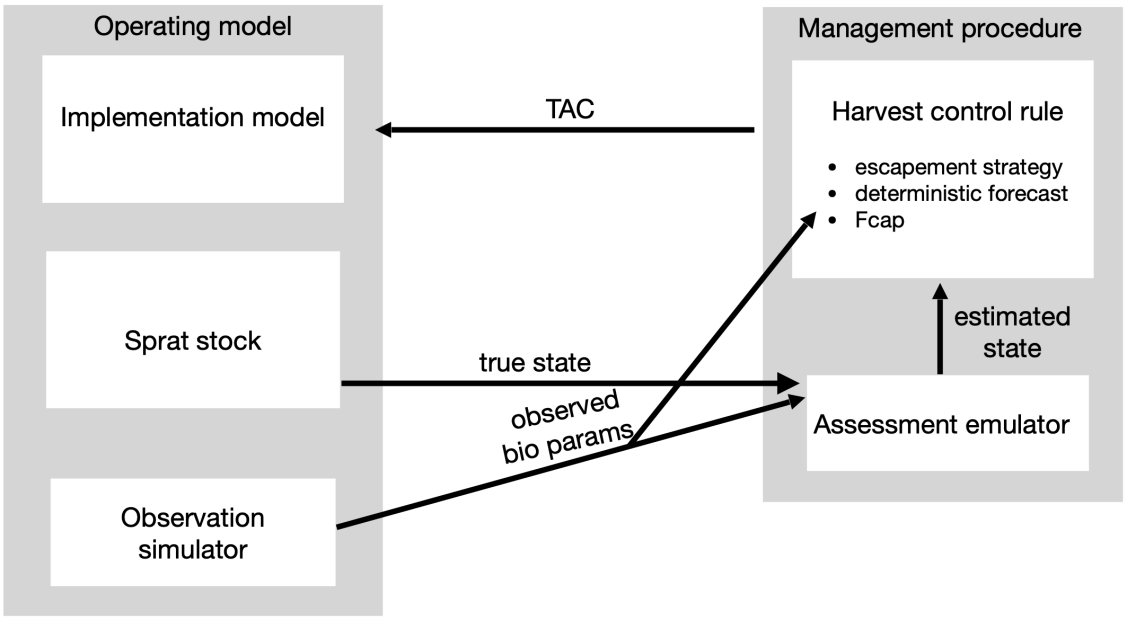

Figure A.6.2 Shortcut MSE for North Sea sprat conceptual drawing. The "true state" and "estimated state" contain seasonal and age structured N and exploitation pattern.

Alternative operating models

In an extended comparison, OMs were conditioned to have either biased natural mortality or weight at age compared to the baseline conditioning while the baseline values were still used in the MP. This simulates a possible scenario where biological parameters change in the real system, but the change is not observed and therefore the old values are used in the MP. In OM_M1 natural mortality was multiplied by 0.9. In OM_M2 natural mortality was multiplied by 1.1. This results in -/+ 10% bias in the observed natural mortality. In OM_Wslow, weight of ages 0, 1, 2,

and 3+ were multiplied by 0.5, 0,5, 0.75, and 1.0 respectively, thus slowing down weight gain of earlier ages.

Results

Figures A.6.3–A.6.10 show that the full and short cut MSEs perform quite similarly when natural mortality is correctly specified in the MP. However, differences in the performance statistics are notable when mortality is higher in the OM than the MP. Further work could be done to try to improve the ability of the assessment emulator to match the properties of the SMS assessment. Perhaps the matrix used to produce lognormal error on N and E could come from a run of SMS with higher natural mortality input.

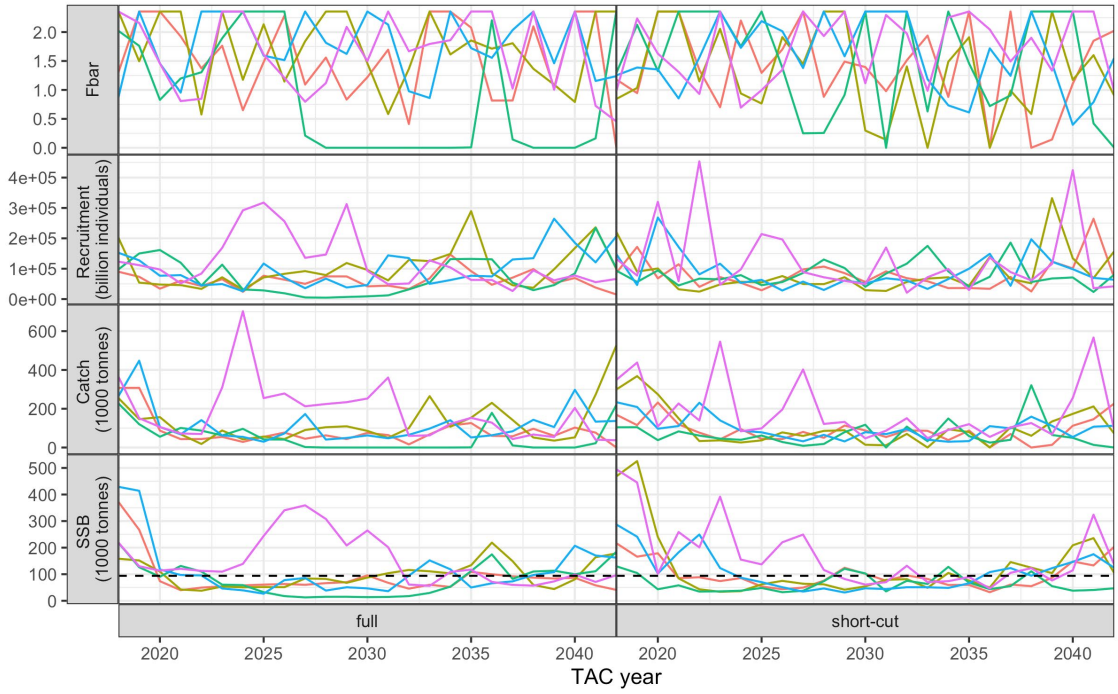

Figure A.6.3 Worm plots of 4 simulation trials in full (left) and shortcut (right) in the baseline MSE.

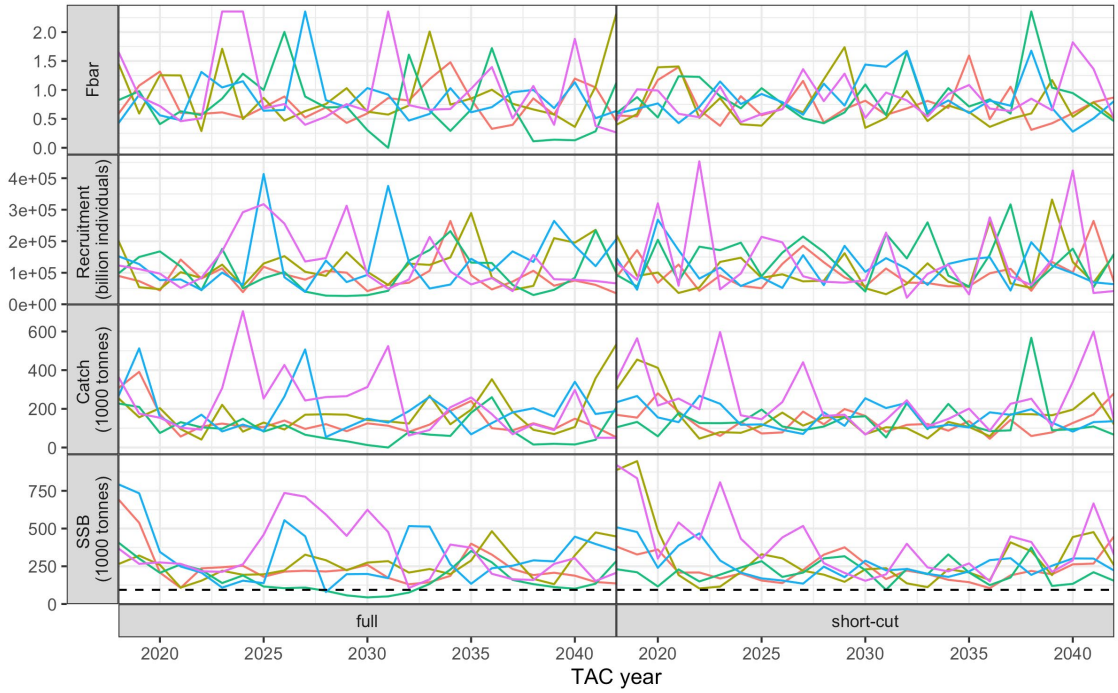

Figure A.6.4 Worm plots of 4 simulation trials in full (left) and shortcut (right) MSE runs with true natural mortality (in OM) 10% higher than the observed values (in MP).

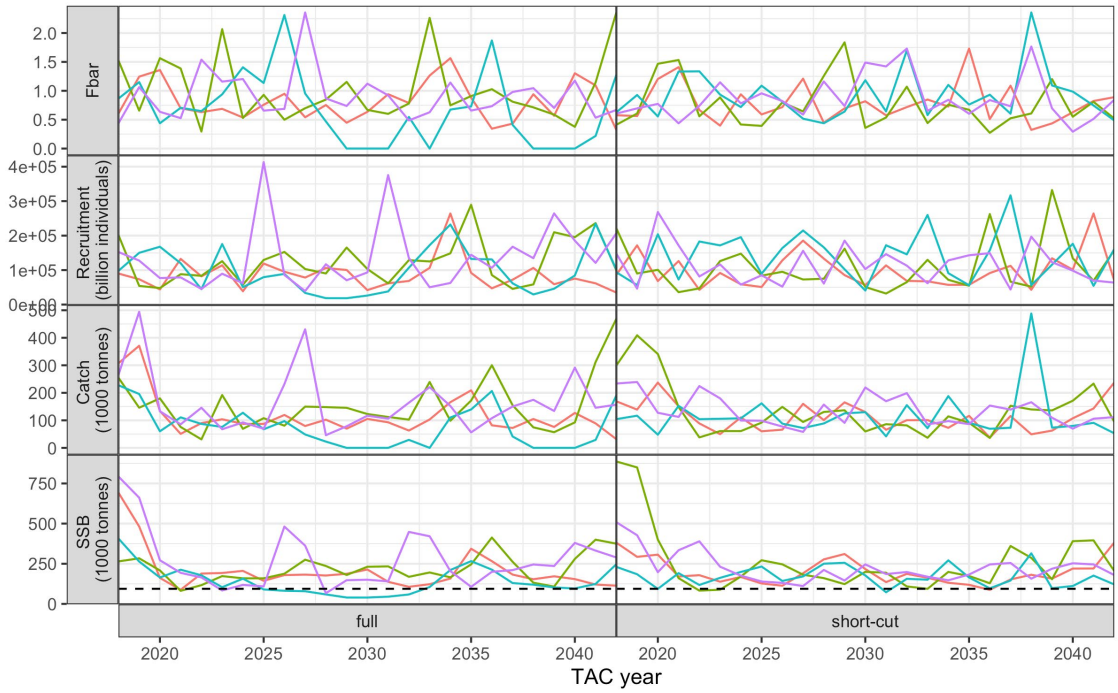

Figure A.6.5 Worm plots of 4 simulation trials is full (left) and shortcut (right) MSE runs with true weight at age (in OM) lower than the observed values (in MP).

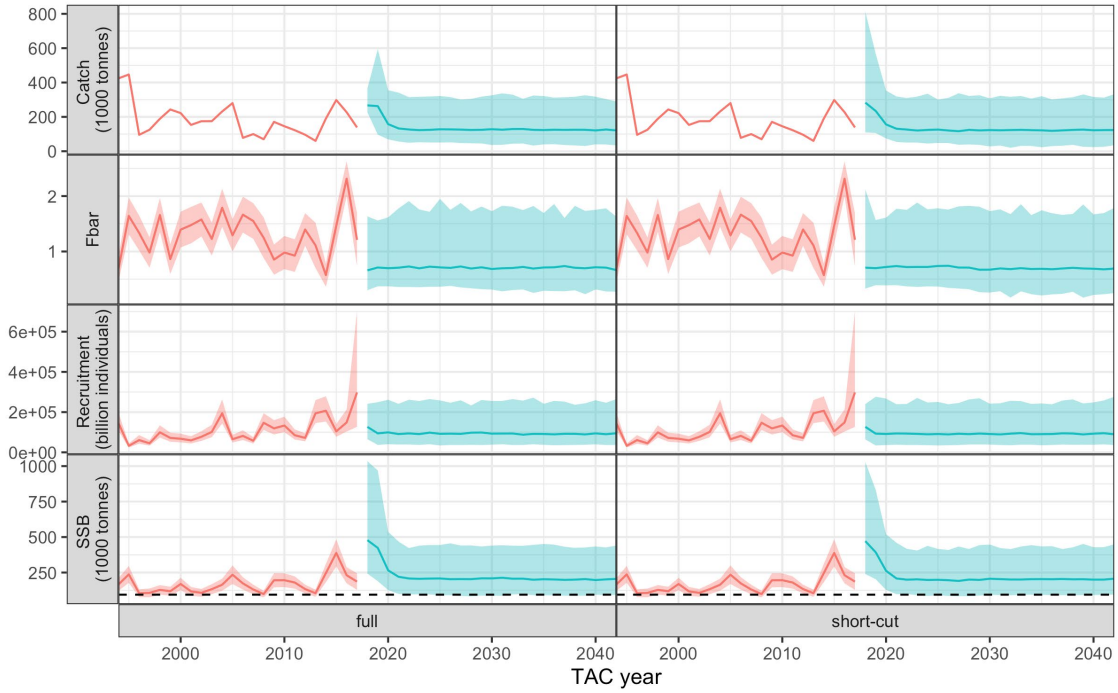

Figure A.6.6 Timelines with confidence intervals in full (left) and shortcut (right) in the baseline MSE. Red portions are estimated by SMS or observed as in the catch. Blue portions are simulation trials with median and 95% quantiles.

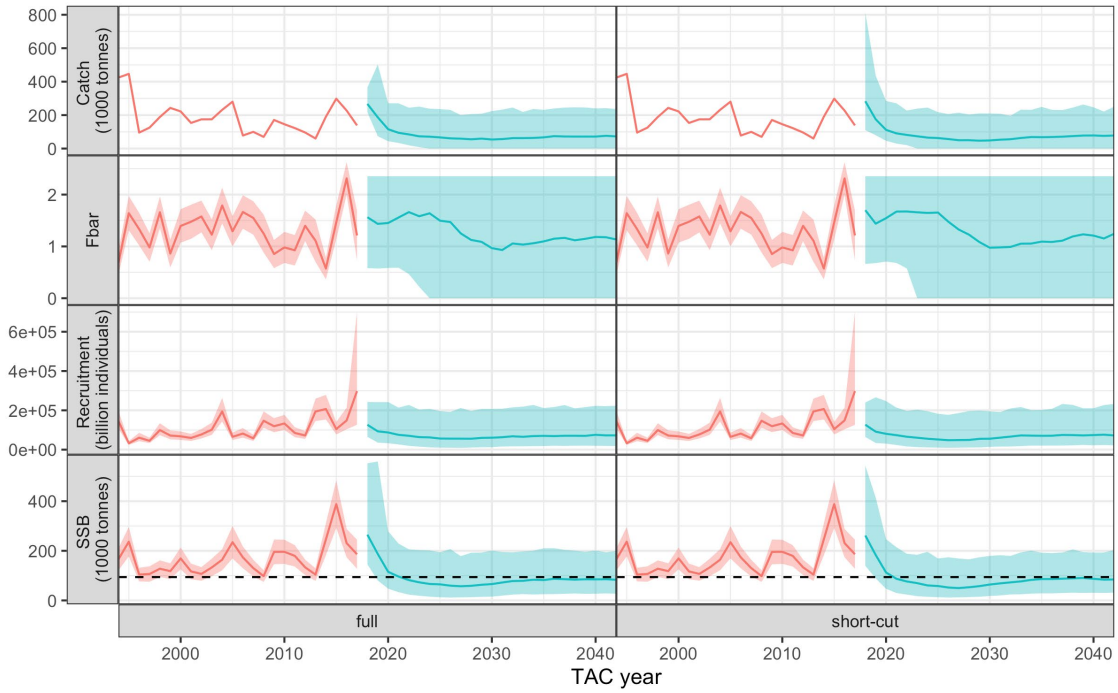

Figure A.6.7. Timelines with confidence intervals in full (left) and shortcut (right) MSE runs with true natural mortality (in OM) 10% higher than the observed values (in MP). Red portions are estimated by SMS or observed as in the catch. Blue portions are simulation trials with median and 95% quantiles.

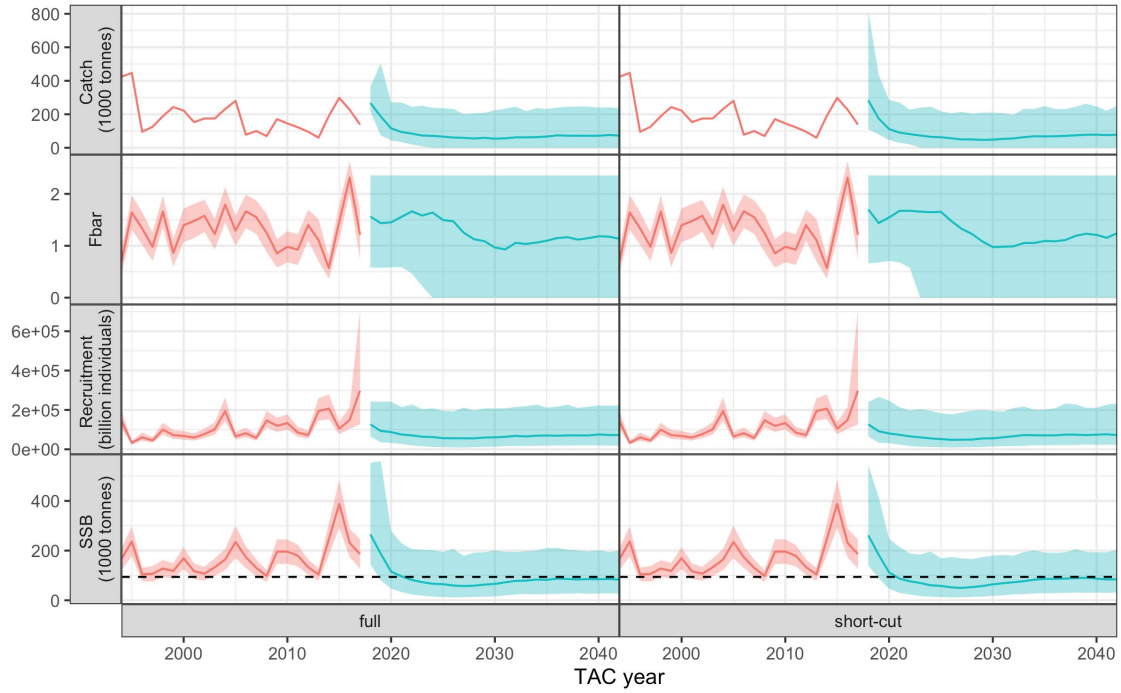

Figure A.6.8. Timelines with confidence intervals in full (left) and shortcut (right) MSE runs with true weight at age (in OM) lower than the observed values (in MP). Red portions are estimated by SMS or observed as in the catch. Blue portions are simulation trials with median and 95% quantiles.

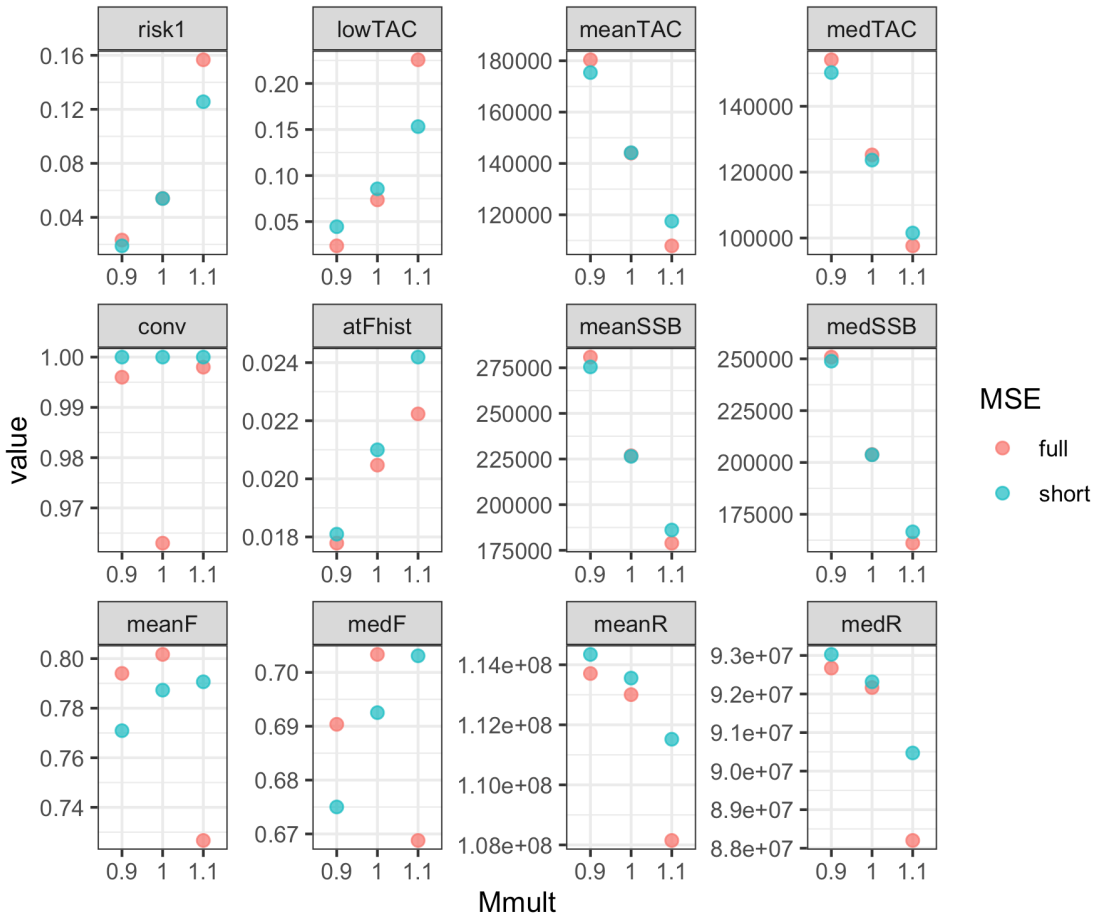

Figure A.6.9. Performance statistics in full (red) and shortcut (blue) MSE runs with natural mortality in the OM multiplied by Mmult, relative to the MP. The baseline is in the middle (where Mmult = 1) and sometimes the points overlap completely.

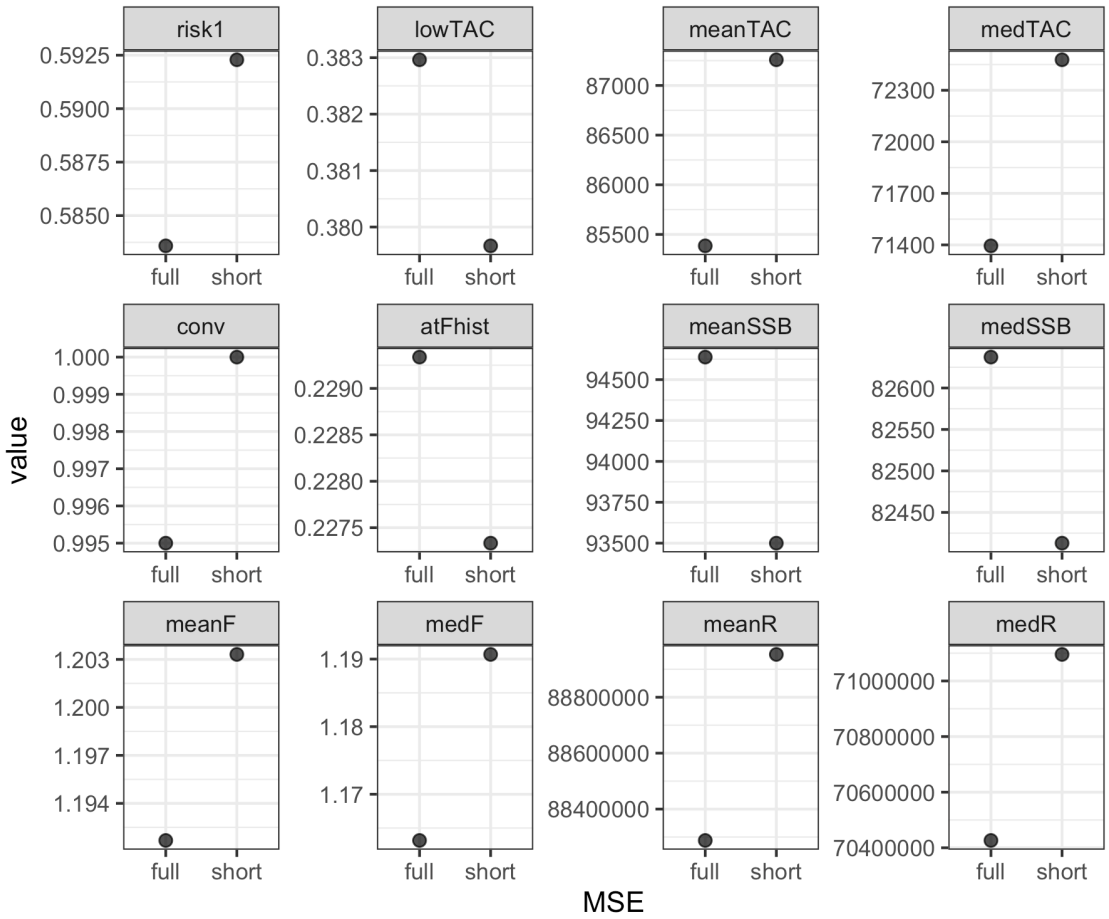

Figure A.6.10. Performance statistics in full and shortcut MSE runs where weight gain was slower in the OM compared to the MP.

Annex 7: North Sea cod MSE: full vs shortcut (TOR e)

Simon Fischer

Introduction

Recently, ICES conducted a workshop to evaluate long-term management strategies for several North Sea fish stocks (WKNSMSE; ICES, 2019b). This evaluation was done with a full MSE and included an analytical stock assessment model (SAM; Nielsen and Berg, 2014) and a stochastic short-term forecast within the MSE feedback loop. For this meeting (WKGMSE3), the analysis of North Sea cod from WKNSMSE was repeated with a shortcut MSE approach, and this allowed a direct comparison of the results from a full MSE with that of a shortcut MSE.

MSE background

The details of the full MSE are described in the WKNSMSE report (ICES, 2019b). The baseline operating model (OM) was conditioned on the 2018 SAM stock assessment of WGNSSK (ICES, 2018b) and included 1000 simulation replicates. The MSE was conducted using the Fisheries Library in R (FLR; Kell *et al.*, 2007) and made use of FLR's mse package (https://github.com/flr/mse). The projection period was set to 20 years, and the last 10 years (long-term) were used for the optimisation of a harvest control rule (HCR). Several HCRs were explored during WKNSMSE; however, for the shortcut comparison, only HCR A was considered. This HCR has a hockey-stick functional form (similar to the ICES MSY advice rule) with two tuneable parameters; F_{trgt} (the target fishing mortality) and $B_{trigger}$ (the SSB defining the breakpoint of the HCR where F is reduced when the SSB falls below this value). The evaluation of the SSB relative to $B_{trigger}$ is done for the advice year (the year for which the catch advice is given) and based on a short-term forecast. A second short-term forecast is required for the translation of the target F into a catch value.

Shortcut approach

The shortcut was based on the framework developed for the full MSE during WKNSMSE, and changes were only made where necessary (the code for the MSE is available on GitHub: full at https://git.io/JTX1F and shortcut at https://git.io/JTX1p; to access these links, access to the ICES TAF GitHub repositories is required). The management procedure (MP) received the same data which is passed to the stock assessment in a full MSE (weights at age, maturity, natural mortality, etc.; Figure 6.0.1). The assessment was approximated by assuming the perception of the stock can be emulated with an unbiased lag-1 autocorrelated lognormal error from the OM (Wiedenmann *et al.*, 2015). This error was included for the stock numbers at age. The same error was included for all ages so that the entire stock (e.g. SSB) was scaled accordingly; however, errors were different between simulation replicates and years. The quantification of the assessment emulator uncertainty was based on a 10-year analytical retrospective analysis of the stock assessment used to condition the OM. For this purpose, the terminal SSB estimates of the retro peels were compared to the final stock assessment (Figure A.7.1) and the standard deviation (SD) and autocorrelation calculated from the residuals. This resulted in SD = 0.11 and a lag-1 autocorrelation with $\varrho = 0.25$. The fishing mortality in the assessment emulator (required for the short-term

forecast) was calculated from the perceived stock and catch numbers at age using the Baranov catch equation. The stochastic short-term forecasts from the full MSE were approximated with deterministic forecasts and the values from the assessment emulator.

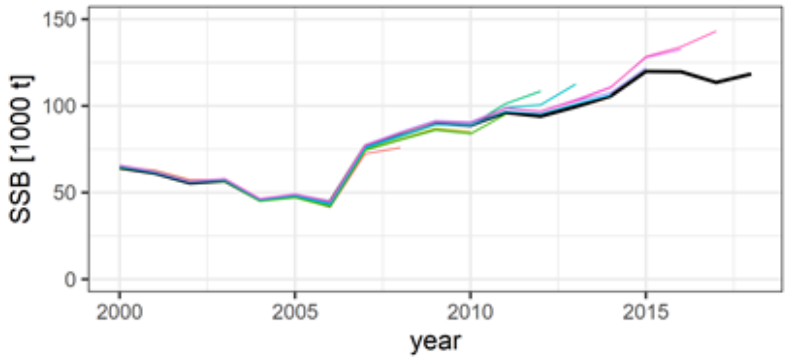

Figure A.7.1. Ten-year analytical retrospective analysis of the 2018 WGNSSK SAM stock assessment for North Sea cod (ICES, 2018b).

Baseline OM: comparison of full MSE and shortcut

Figure A.7.2 shows a comparison of the shortcut MSE projection to the full MSE for one HCR parameter combination. The trajectories, including the percentiles and individual simulation replicates, appeared similar.

Figure A.7.3 compares the results of a full grid search for the full and shortcut MSE. The outcome was similar; however, the HCR parameter combinations where risk 3 exceeded 5% is slightly shifted towards higher F_{trgt} values for the shortcut. Additionally, the optimum combination (the HCR parameterisation with the highest yield under the condition that risk does not exceed 5%) is shifted, and the shortcut suggests a parameterisation which was classified as non-precautionary in the full MSE.

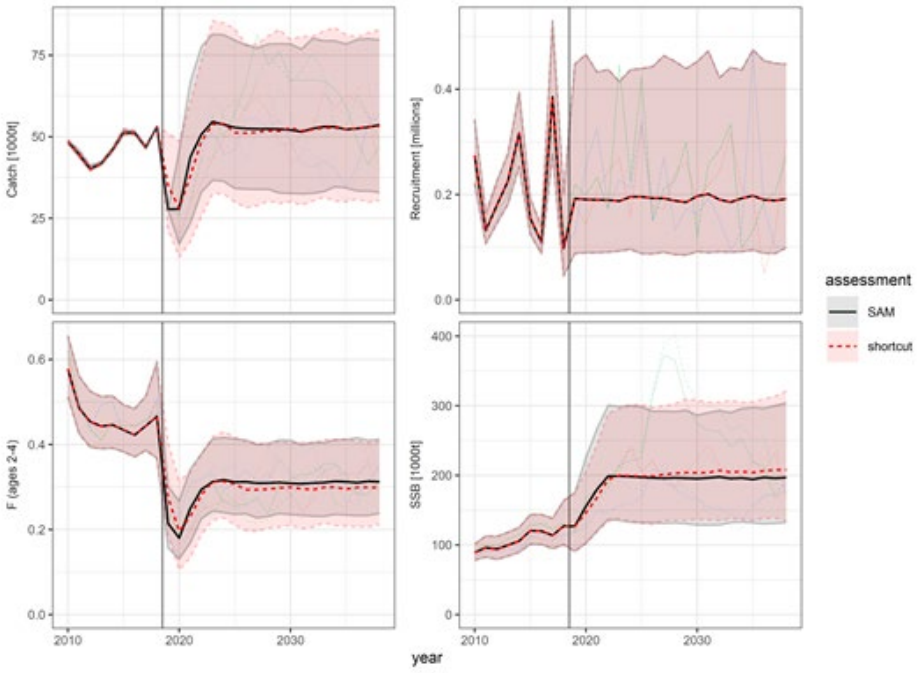

Figure A.7.2. Comparison of the North Sea cod MSE projection of the shortcut and full MSE for one HCR parameter combination (F_{trgt} = 0.31, $B_{trigger}$ = 150 000t). Shown are medians surrounded by 90% confidence intervals and three simulation replicates (coloured).

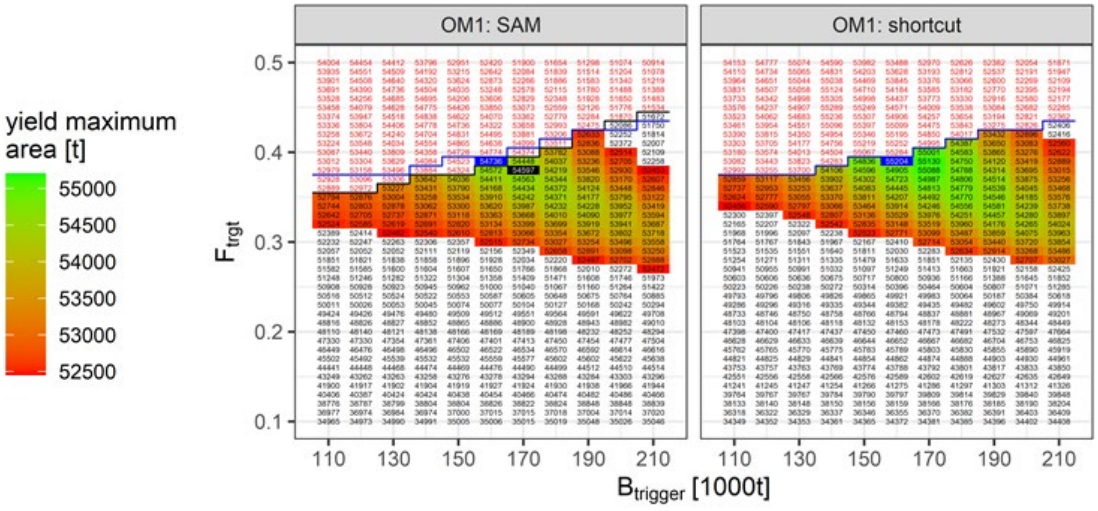

Figure A.7.3. Comparison of the grid search of the full (left) and shortcut MSE (right) for North Sea cod. Shown is the yield (catch) and text in red indicates that risk 3 exceeds 5%. The coloured cells indicate the optimum yield area where risk 3≤5%. The black rectangle shows the optimum for the full MSE, and the blue rectangle the optimum of the shortcut MSE; the black (full) and blue (shortcut) lines delineate where risk 3 exceeds 5%.

Robustness to assessment uncertainty estimates

The robustness of the shortcut MSE to the uncertainty assumptions was evaluated by testing various values for the uncertainty, expressed as SD (0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8, 1; default: 0.11) while keeping the default autocorrelation ϱ, and in turn keeping the default SD and exploring various ϱ values (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 0.99; default: 0.25), and repeating the shortcut grid search for all these values. The results of this analysis are shown in Figure A.7.4.

The levels of uncertainty in the assessment emulator had a big influence on the results. Smaller uncertainty shifted the threshold where risk 3 exceeded 5% towards higher F_{trgt} values and increased the steepness of the curve delineating the 5% risk 3 threshold (relative to $B_{trigger}$ values), and vice versa. The ϱ values had a minor impact on the risk 3 threshold and the location of the optimum.

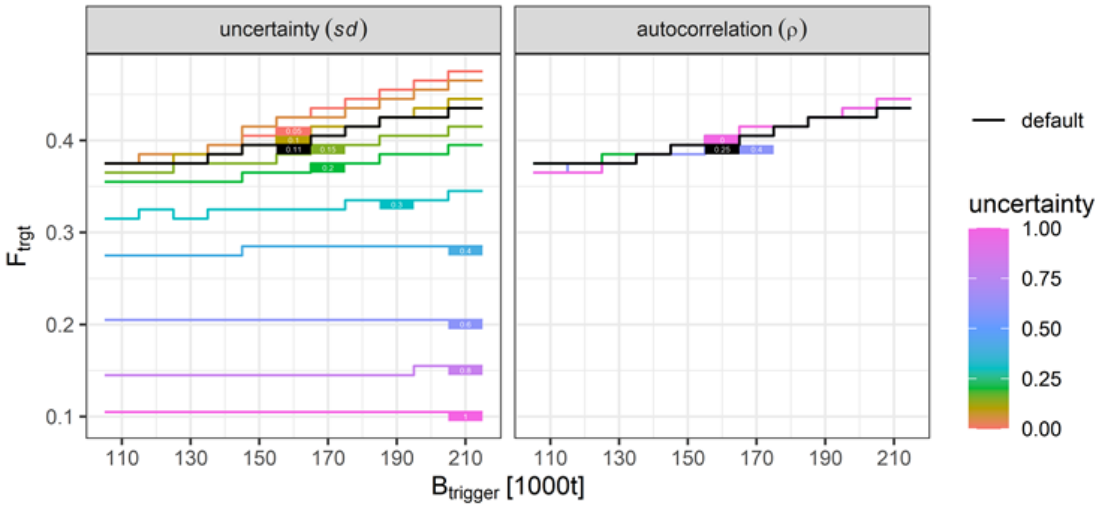

Figure A.7.4. Influence of the assessment emulator uncertainty on the results of the grid search for the North Sea cod shortcut MSE. Shown are the thresholds where risk 3 exceeds 5% for different values of SD (left) and the autocorrelation (right). The default (SD = 0.11 and ρ = 0.25) is shown in black. The coloured cells indicate the location of the optimum HCR parameterisation for each configuration.

Considering alternative OMs

WKNSMSE considered three alternative OMs for North Sea cod apart from the baseline OM (OM1): OM2 which assumes higher recruitment, OM3 which is based on a stock assessment that considers year effects in the survey indices, and OM4 which includes density dependence for natural mortality (cannibalism). In the full MSE, these alternative OMs provided a means of testing the robustness of the MP when the stock behaved differently compared to OM1, but the stock assessment in the MP was unchanged, i.e. there was a deliberate mismatch between the MP and the OM.

The WKNSMSE paradigm was that the selected optimised HCR should be robust to the selected plausible alternative OM assumption. WKNSMSE's conclusion was to reject the optimum control parameters from OM1, because this HCR parameterisation yielded a risk 3 for OM3 well above 5%. Instead, the default ICES MSY parameterisation (F_{trgt} = 0.31 and $B_{trigger}$ = 150 000t) was recommended, because this parameterisation was precautionary for all OM scenarios.

All three alternative OMs from WKNSMSE were recycled for the shortcut exploration, and the MP configuration from OM1 was implemented unchanged. Due to the reduced computational complexity of the shortcut MSE, full grids could be explored for these alternative OMs, and the results are shown in Figure A.7.5. The general behaviour of the alternative OMs relative to the baseline OM for the shortcut MSE was similar to the full MSE and OM3 (with survey year effects) was most restrictive. However, a detailed comparison of the full MSE vs. the shortcut for OM3 revealed a substantial bias (Figure A.7.6), and the 5% risk curve was shifted to higher F_{trgt} values in the shortcut MSE. This led to two important issues with the short approach: (1) the shift of a possible precautionary HCR parameterisation from OM1, because of OM3, would be less severe

compared to the full MSE (e.g., when comparing the shift of the 5% risk line from OM1 to OM3 for the range of $B_{trigger}$ from 160 000 t-180000t, the shift was 0.03 along the F_{trgt} axis for the shortcut: see Figure A.7.5; and 0.05 along the F_{trgt} axis for the full MSE: compare Figure A.7.3 "OM1: SAM" and Figure A.7.6 "OM3: SAM"), and (2) the suggested HCR parameterisation of the shortcut MSE OM3 was considered non-precautionary in the OM3 full MSE (top plot of Figure A.7.6).

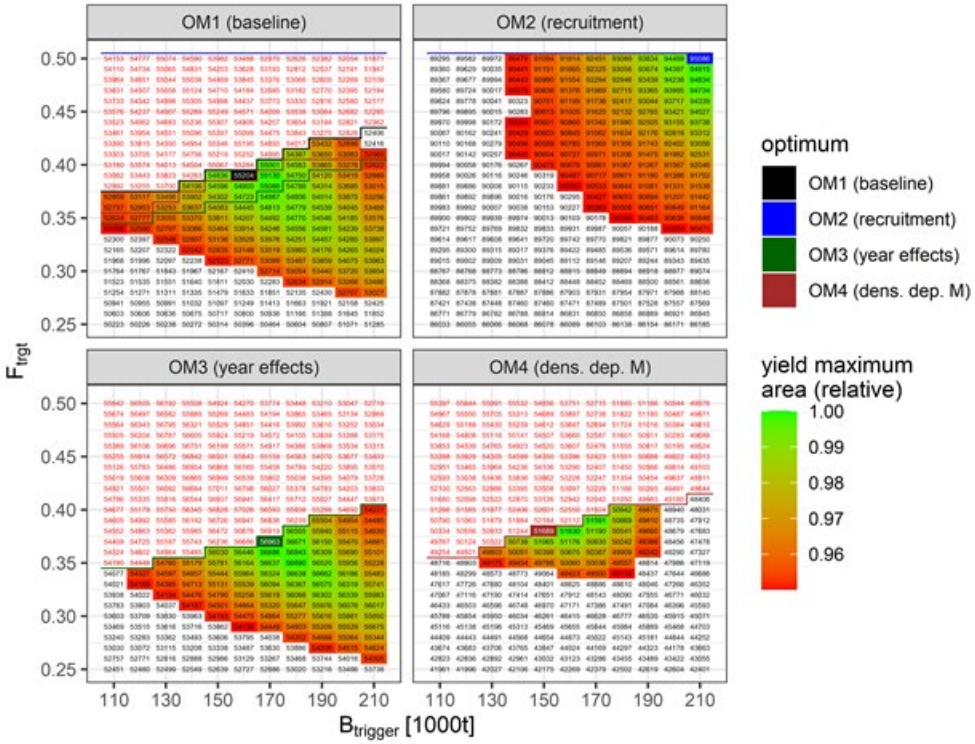

Figure A.7.5. Full grid search for alternative OMs in the shortcut MSE for North Sea cod.

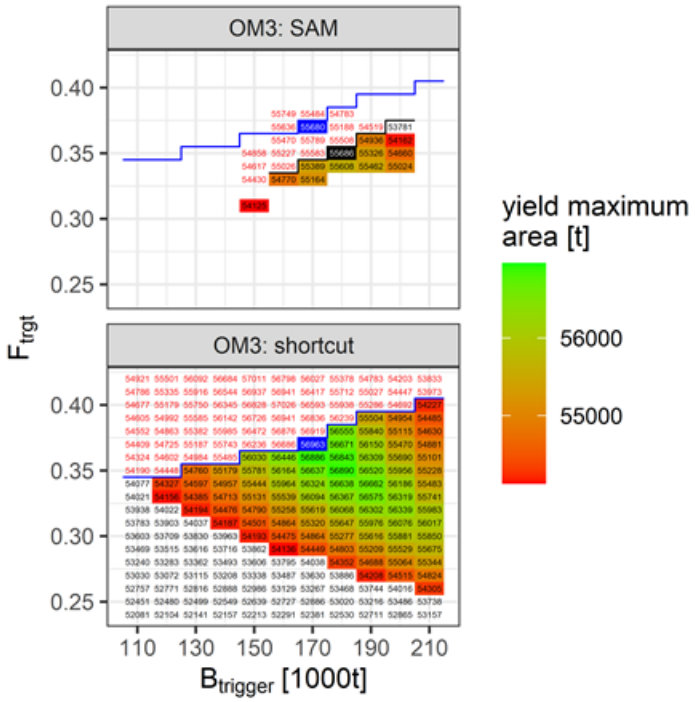

Figure A.7.6. Comparison of the full MSE (top) with the shortcut MSE (bottom) for OM3 (with survey year effects) for North Sea cod.

Discussion

The main benefit of the shortcut MSE was the reduction in computational complexity. The full grid for the full MSE for OM1 took around 20 000 CPU hours. This immense runtime required the utilisation of massive parallelisation techniques, a high-performance computing system, and associated costs for using such a system. The shortcut MSE was more than 500 times faster and the same computing time for running a single grid cell of the full MSE was enough to run a full grid of the shortcut MSE.

The shortcut approach allowed a quick and efficient exploration of the search space; however, the results were crucially dependent on the assumptions of assessment uncertainty.

The WKNSMSE North Sea cod MSE example was very "well behaved" in terms of one well-defined optimum and the search space surface being smooth without local optima. This situation facilitated the use of a shortcut approach, but this might not always be the case for other stocks or operating models, possibly impairing the applicability of shortcut approaches.

The bias between the full and shortcut MSE was small for the baseline OM. This bias, however, was much larger for alternative OMs. Relying solely on the shortcut approach for cod would have led to HCR parameterisations which were considered non-precautionary in the full MSE.

A shortcut MSE might not always be able to entirely mimic the assessment behaviour, which can have an important influence on the outcome of an MSE. If a stock assessment, or the MP in general, behaves in a specific way, e.g. it contains systematic bias, or cyclical behaviour, it is important to include this in an MSE. The WKNSMSE North Sea cod is an example of that. The stock assessment in the MP exhibited some overestimation of F in the early years of the MSE projection, when compared to the OM. This behaviour influenced the performance of the MP and the selection of an HCR parameterisation. Shortcut MSEs do not always include this (usually the

assessment error is assumed to be random without bias), and outcomes of a shortcut MSE, therefore, do not test what is implemented in reality.

Shortcut approaches cannot entirely replace full MSEs; however, they can complement MSE processes, particularly for early explorations. The best approach might be a hybrid of full, and shortcut considerations, where (1) the baseline OM is explored with a shortcut for a quick overview, (2) the robustness of the shortcut approach to uncertainty is checked, and (3) the outcome of the MSE (e.g. HCR parameter optimisation) is verified with a full MSE. Alternative OMs require more consideration because of possible large biases in shortcut approximations, and consequently should always be confirmed with full MSEs.

Annex 8: Comparing shortcut and full approaches using the Muppet model (TOR e)

Höskuldur Björnsson

Northeast Atlantic mackerel

The management strategy for NEA mackerel has changed since this section was written, but the comparison is based on $F_{4–9} = 0.22$ and no trigger.

The data for the deterministic simulations are based on estimated variances, correlations and catchabilities that are included in the set of stochastic parameters. The procedure is best described by an example from one year, for example 2030.

TAC for the year 2030 is based on the 2029 assessment that uses catch at age until 2028. Catch and survey data for what would be called the 2030 assessment are generated using this TAC for the year 2030 (based on the 2029 assessment). The predicted survey indices are generated using proportion of F and M before the surveys. For some surveys (egg mackerel) the index for 2030 will also have to be compiled from the TAC because catch is not available at the time of the assessment. Predicted catch at age for 2030 is also compiled based on the TAC, estimated stock size, estimated selection pattern, and proportion of F before the survey.

When the predicted catch and survey indices have been compiled, the "observed" values are compiled based on estimated observation variances and correlations. These observed values are then added to data files and used in a deterministic model run that is used to generate advice for the year 2031 using, as TAC constraint in 2030, the catch compiled from last year's closed loop assessment. Here it helps that standard deterministic Muppet runs can do a short-term prognosis with TAC constraint in the assessment year and fishing mortality or harvest rate for subsequent years. What the model requires from the deterministic run is only the TAC for the year 2031 (the year following the assessment year). Other values are kept track of to be able to look at the characteristics of the closed loop assessment. In the runs presented, future mean weight and maturity at age in the deterministic simulation is based on the average of last 5 years in the operating model, something that can be changed. Variability in mean weight at age is included but variability in maturity is not included in the stochastic simulations but is easy to add.

Variability in selection could also be added but the main effect of selection is when compiling the TAC so a selection pattern should really be specified when using the average F_{4-8} in management strategy. Estimation of recent selection pattern is, in any case, highly uncertain.

Results are presented in the figures below using $F_{target} = 0.22$ and no $B_{trigger}$. Variablity in the past comes from estimation of a multiplier on catches before 1998 using catch in numbers at age back to 1998. The data used in the assessment are catch at age, egg survey, pelagic survey, recruitment index and RFID tags. The RFID tags are based on tagging years 2011–2018, recapture years 2012–2019, and all ages except the plus group. Tagging mortality, tag loss, and dispersion are estimated. Steel tag data are not included, but they were included in some other runs. They do though not help much in estimating misreporting as misreporting and estimated tagging mortality are confounded (see discussion later).

Most of what is presented below is shown for a randomly selected iteration 10 (i.e. not selected because it looks "nice"!). The number of iterations was small (100–200), but subsequently one run is available with 500 iterations. When predicting 60 years ahead, the number of iterations is approximately 150 per 24 hours. This is then just for one fishing mortality and one $B_{trigger}$. Many different HCRs can be run in parallel. At the MFRI, there are something like 80–100 cores to do the calculations, often occupied by Gadget runs. Nevertheless, the number of possibilities that can be explored is rather limited. To be realistic, what we have here could be used to help in specifying the "assessment error" in simulations based on the "shortcut approach".

The Muppet model has until now been used in shortcut HCR simulations. The operating model has been derived from a run of an assessment model where the estimated SSB-recruitment relationship is the most important part, but has little effect on historical assessments. Only one stock-recruitment relationship can be used in each simulation, but autocorrelation of recruitment residuals can be an estimated parameter. The mean weight at age in the operating model is the average of last 3 years multiplied by a lognormal autocorrelated noise. Maturity at age in the operating model is fixed. Density dependent growth has been included, for example for Icelandic haddock and for NEA mackerel in 2017. Density dependent growth (DDG) will, however, not be included in this exercise; one reason is that estimation of DDG is confounded with misreporting.

The retrospective runs of spawning stock (Figures A.8.1) show some problems in the final years of each assessment. The main reason is probably that little is known about cohort size until at age 3 (the model has limited belief in the recruitment index that is estimated as a good index in the official assessment), and age 2–3 already have some weight in the spawning stock. Autocorrelation of recruitment in the operating model leads to worse retrospective patterns. The model cannot start with the recruitment index as nearly correct and later correct it with deviations in M as is done in the SAM assessment.

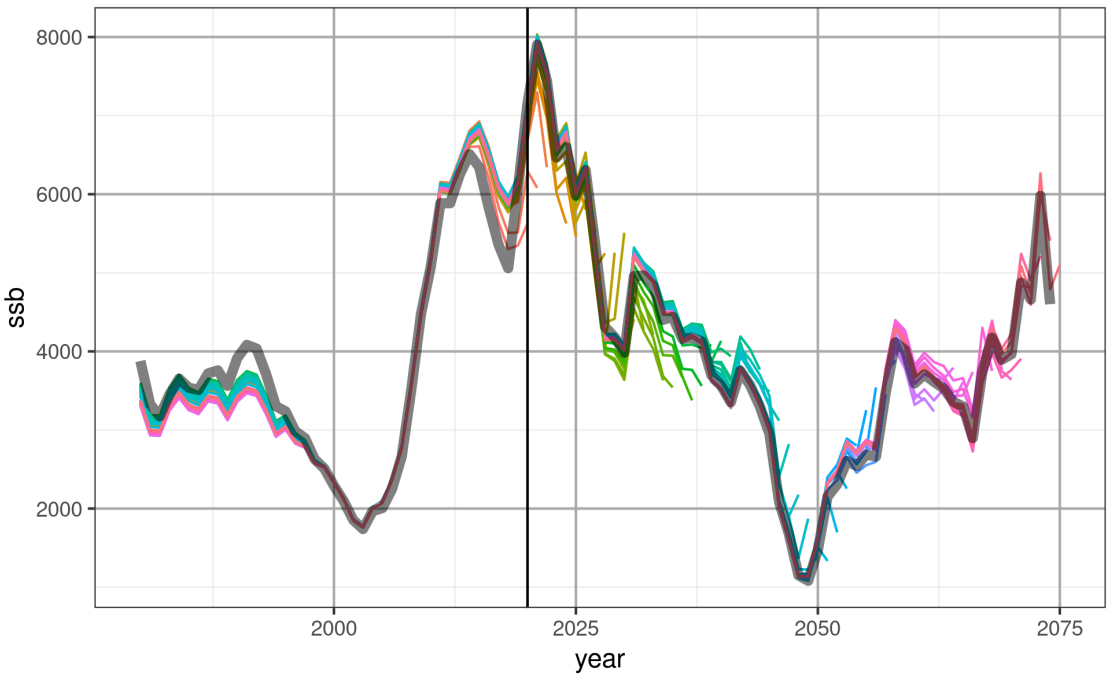

Figure A.8.1. Retrospective pattern of spawning stock ending year after assessment year for iteration 10. Generated tagging data not included in the full MSE. The wide line shows the converged assessment. SSB in thousand tonnes.

In the run presented in Figure A.8.2, F_{target} is always 0.22, so deviations from 0.22 are due to assessment error. ρ and σ were estimated based on 30 years (2045–2074). The average value of ρ was 0.56 but standard deviation of $\log(F)$ is 0.228. This is a rather high value of autocorrelation, but the standard deviation is moderate. For comparison, estimated CV of $\log(F)$ in 2021 is 0.3 and 0.26 in 2022.

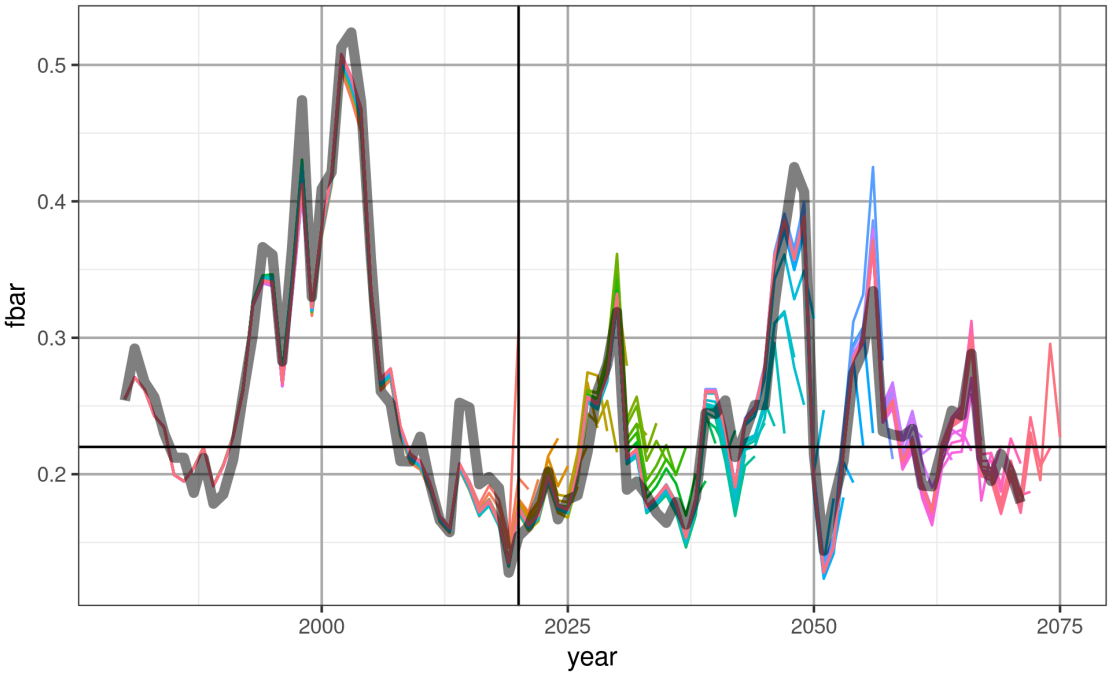

Figure A.8.2. Retrospective pattern of $F_{4\text{-}8}$ ending year after assessment year for iteration 10. The horizontal line shows the target $F_{4–8}$ = 0.22. 500 iterations. Generated tagging data not included in the full MSE. The wide line shows the real F i.e. from converged assessment.

Figure A.8.3 shows the distribution of main metrics for the real stock. The plots do not show increasing trend in F with time, while the same run done with 100 replicates showed a slight increase in F with time.

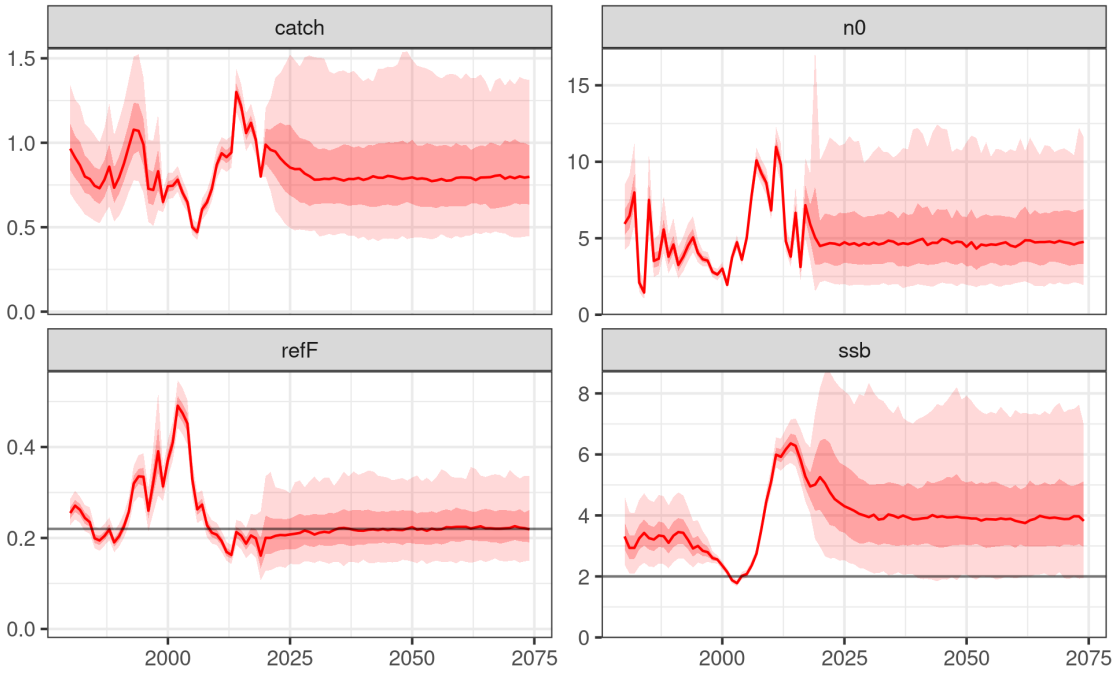

Figure A.8.3. Distribution of some of the main metrics of the "real stock", the shaded areas show 5th, 25th, 75th and 95th percentile. Generated tagging data not included in the full MSE; 500 iterations. SSB and catch in million tonnes, and n0 in billions.

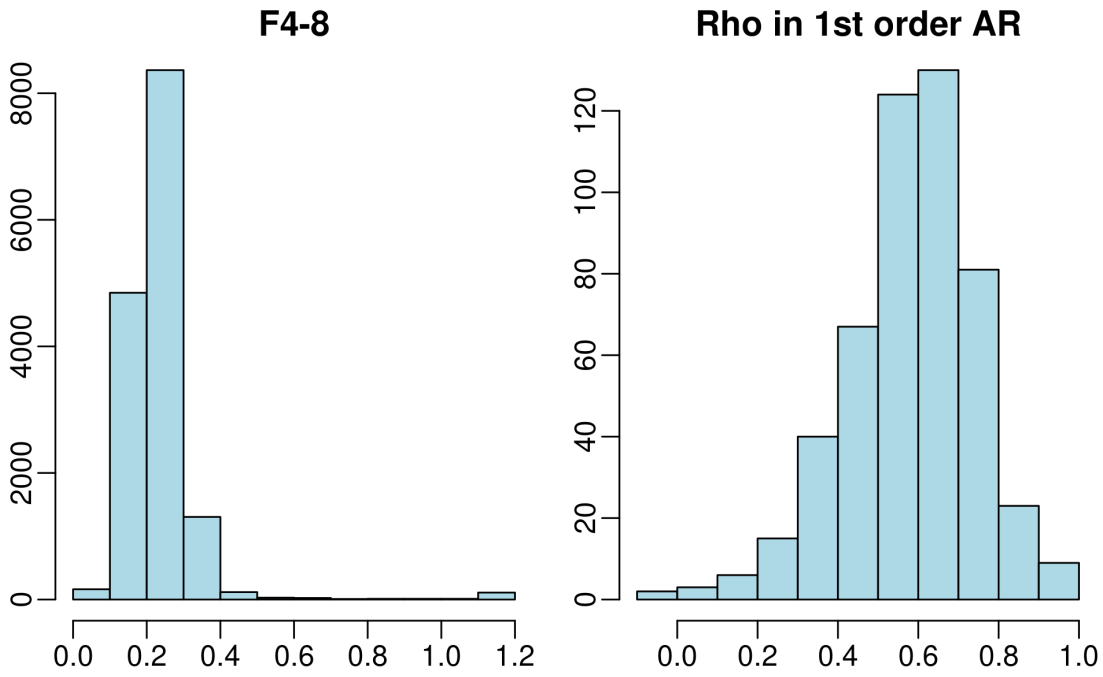

Figure A.8.4. Distribution of F4-8 2045–2070 and autocorrelation of the deviations. Tags not included in the full MSE.

One interesting question to look at is what the generated data look like. The plots have not yet been made for the tagging data (not clear how to do these yet). The plots for the surveys are all

shown as one total index per year. The pelagic survey is of course age disaggregated but considerable correlation between age groups exists, so the total index is descriptive.

The predicted recruitment index (Figure A.8.5) in iteration 10 is relatively variable. The recruitment index is one of the major differences between SAM and Muppet. In SAM, it is estimated to be ver4y precise ($CV \approx 0.16$), but in Muppet the estimated CV is 0.41. When compiled, this index is heavily smoothed by geostatistical analysis using the square root of numbers. Therefore, a model such as SAM with flexible M can decide to make it the truth, using variable M for later adjustment. One additional difference between Muppet and SAM is that the first estimate of a year class in Muppet is from a stock-recruitment model. This option exists in SAM but is not used in the adopted assessment.

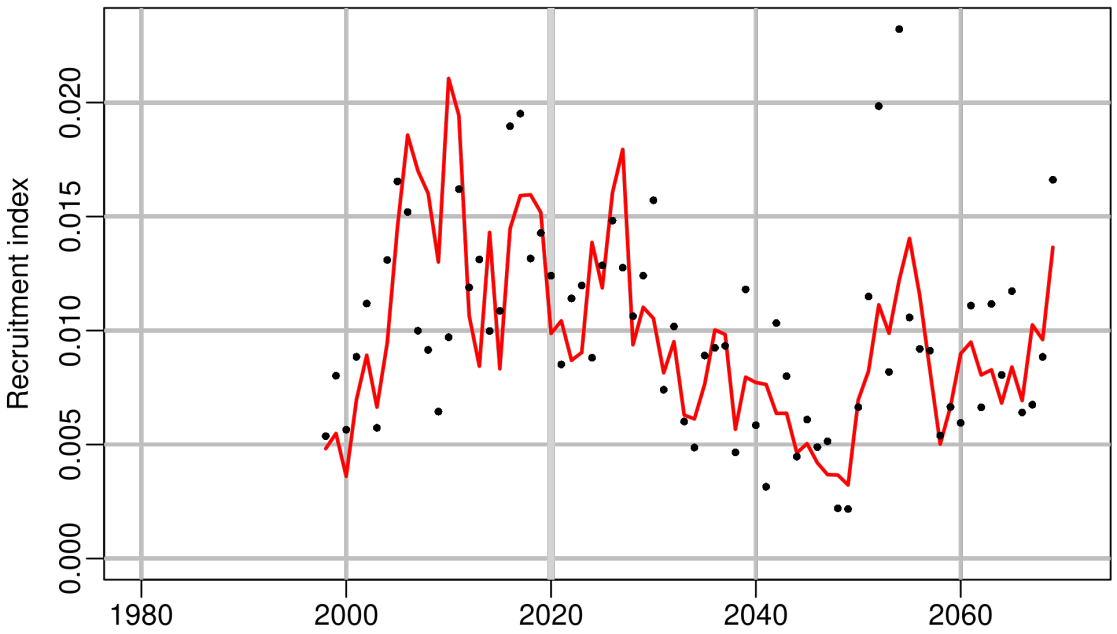

Figure A.8.5. Observed and predicted recruitment index one iteration.

The egg survey is assumed to be conducted every 3rd year (Figure A.8.6).

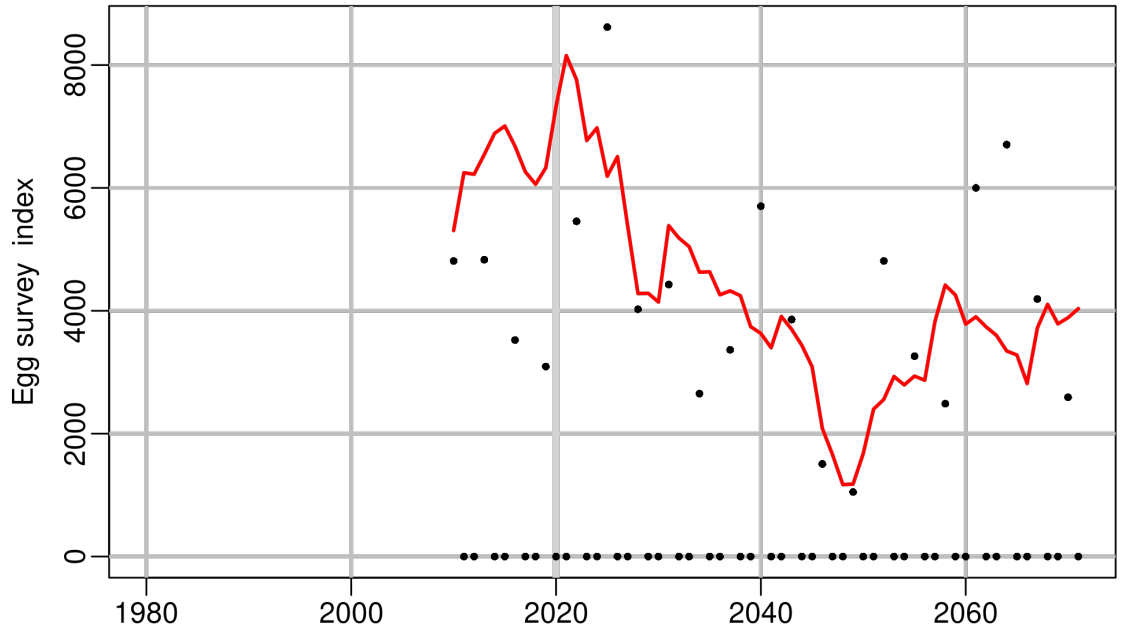

Figure A.8.6. Observed and predicted egg survey one iteration.

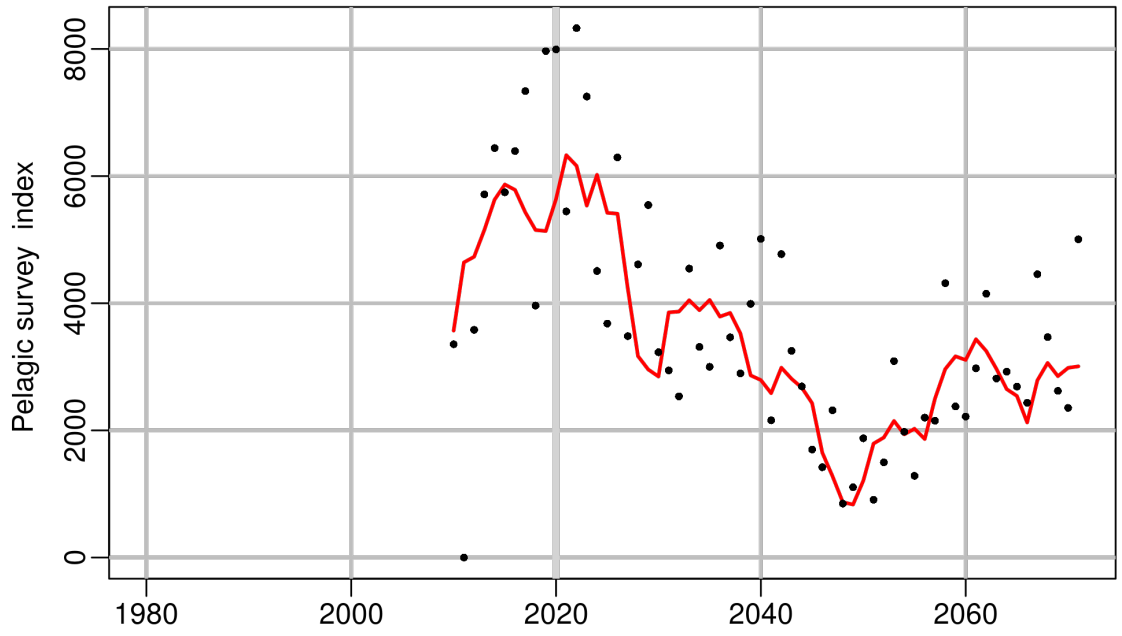

Figure A.8.7. Observed and predicted pelagic survey one iteration. The 2011 value is not included in the plot nor the assessment.

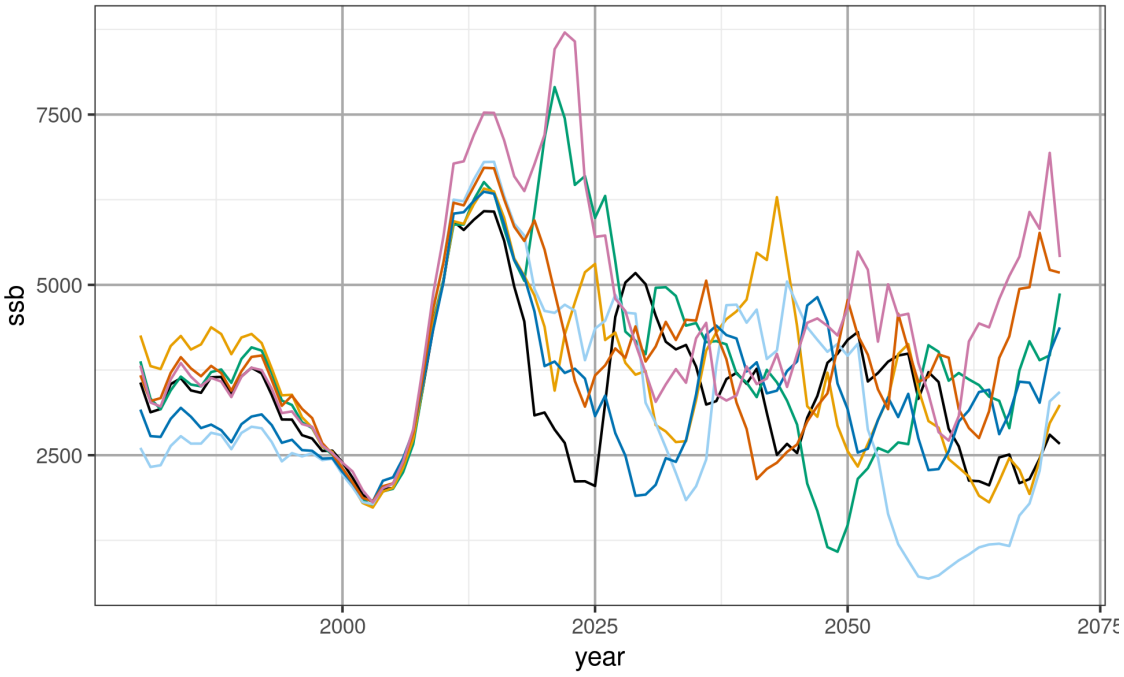

Figure A.8.8. Seven replicates of the spawning stock (thousand tonnes). Generated tag data not included in the full MSE.

All the model runs are in the estimation phase (operating model) using tagging data; tagging years are 2011–2018, recapture years 2012–2019, age groups 2–11, estimating tag loss as well as tagging mortality and dispersion. This work was presented in Ijmuiden in March 2019. The result of that work was that the 2011 and 2012 tagging data, where relatively small fish were tagged (not used in the official run), fitted best, most likely due to fewer problems with the age length keys that are used both at tagging and recapture.

A module generating tagging data was added, but needs some testing. It is based on a random negative binomial distribution using the estimated parameters, i.e. tagging mortality, tagloss and dispersion. The question will still be if some "additional problem" will have to be added; year factors in tagging mortality or detection, and the double use of age-length keys might be a problem. Tagging younger mackerel is probably a good idea for advice purpose, if they can be found.

The plan was to use generated RFID tagging data in some full MSE runs, but a mistake was detected in the 500 iterations run, so generated RFID data were ignored. The 100 iterations run is available using generated RFID data. Including the generated tagging data in the full MSE assessment did not lead to major changes compared to not including them. Median F after 2045 changed from 0.227 to 0.225, CV from 0.228 to 0.215, and autocorrelation from 0.56 to 0.55. Minor improvements but not significant.

The same could of course apply to the surveys, where temporal correlation are not implemented (correlation between age groups in the same year are implemented), but correlation in time do probably occur and can look like trend if the timeseries are short. Variability in M can lead to similar result as temporal correlations in surveys.

The results shown here give SSB_{05} as 1970 thousand tonnes and average catch 871 thousand tonnes based on the years 2045–2074. The estimated value of the hockey stick SSB breakpoint is around 2300 thousand tonnes so F_{msy} is little lower than 0.22 according to those results. $\rho_{rec} = 0.3$ in those simulations, but it should really be estimated.

The results shown here indicate that appropriate F_{target} is not far from 0.22, but considerably lower if implementation error is included.

Looking at retrospective patterns for the stock size at the beginning of the assessment year leads to $\sigma = 0.106$ and $\rho_{ass} = 0.66$ based on the years 2045–2074. A shortcut approach, where the statistical properties of the assessment error were based on those values, was performed. The main metrics are similar in the shortcut and full MSE methods (Figure A.8.9). The main difference is in the first years, but the shortcut starts with a fixed TAC, and for each replicate, the first value of the assessment error is based on the value $\frac{SSB_{mup,2020}}{SSB_{wg,2020}}$ where $SSB_{wg,2020}$ is the value of SSB from WGWIDE in 2020 (part of the input files in Muppet). Minor adjustment at the start would make the full and shortcut approaches nearly identical. To repeat, both simulations are based on exactly the same operating model.

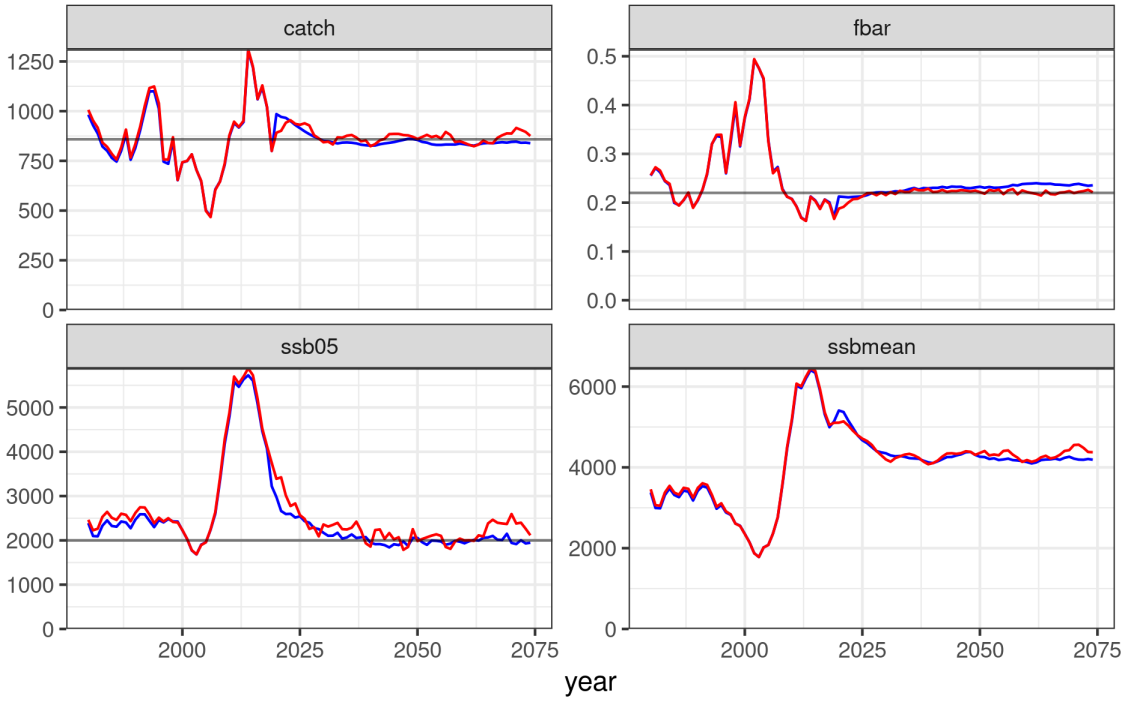

Figure A.8.9. Comparison of main metrics between shortcut (red) and full (blue) approaches when not using generated tagging data in the full MSE. Characteristics of assessment error in the shortcut derived from the full MSE. Horizonal lines represent average catch 1980–2018 for catch, 2 million tonnes for SSB05 and 0.22 for $F_{4\text{-}8}$. CV and ρ estimated from 500 runs. SSB and catch in thousand tonnes.

The full MSE simulations were run only based on one set of parameters, i.e. $F_{4-8} = 0.22$, $B_{trigger} = 0$, but to compile F_{msy} for the stock, where precautionary measures will first kick in, requires $B_{trigger} = B_{pa}$ for a stock like mackerel.

The results of those exercises are.

1. The shortcut approach leads to similar results as the full MSE approach based on the same operating model if statistical properties of the advice error are reasonable.
2. The full approach can be useful to get ideas about statistical properties of the advice error. Here we have a caveat that the observation models that we use in assessment models are not "nasty enough", and the generated data tend to be too well behaved.
3. For the mackerel, uncertainty in the operating model is an important issue. Potential misreporting in the past means that the stock might have been much larger before 1998, average productivity therefore much higher and density-dependent growth not an issue.
4. The Muppet model does not estimate a high level of misreporting in the past. This result is not based on much data (2 egg surveys). Really high misreporting in the past would lead to an increase in estimated survival from the steel tags that is already quite high.
5. Looking at different values of M, the model fits best to M considerably lower than 0.15 (<0.05). The reason is not clear, but the same problem has been observed for other stocks, for example Icelandic cod. There is a need to test the effect of using higher Ms in the assessment model than the operating model, something that can only be done with the full MSE approach, which is the most appropriate method for considering cases where the operating model and assessment model differ.
6. The Muppet model can estimate autocorrelation of recruitment in the assessment. This option should be used when the model is used as an operating model, but not when it is used as assessment model.

Icelandic cod

The assessment of Icelandic cod is based on catch in numbers 1955-2019, a 580 stations survey in March 1985–2020, and a 300 stations survey in October 1996–2019 (Figure A.8.10). Tuning is done by assuming that residuals in a year are multivariate normal (half year factor), but temporal correlation of survey q is not modelled. Nonlinearity is estimated for ages 1–5, but further work is needed on nonlinear relationships between survey indices and stock numbers, both in the assessment model and at a station level in the surveys.

Age groups modelled in the assessment are 1–14. Fish older than 8 used to be uncommon, with relatively few caught in the surveys. After a reduction in fishing effort in 2007, older age groups became a much larger component of the stock, and a signal started appearing for older age groups in the surveys. Now the assessment is tuned with ages 1–14, but the introduction of these older fish caused some changes in the assessment.

Observed and predicted survey indices for the surveys are shown in Figure A.8.10. Tuning is done on indices by age, but what is shown here is the sum of observed and predicted indices multiplied by weights at age. Observed survey biomass shows more contrast than predicted biomass, indicating a problem with the observation model in the assessment. The statement made earlier with mackerel that the observation model is not "nasty enough" applies here.

The full MSE assessment results are incredibly consistent, but residuals $R_y = log(\frac{B_{y,y}}{B_{2065,y}})$ are around 0.04 (average for all iterations) in the period 2030–2058. $B_{2065,y}$ in the equation means reference biomass 4+ in year y as estimated in the 2065 assessment

Figure A.8.12 demonstrates 3 metrics based for R_y in the years 2030–2060: 1st order autocorrelation, bias and standard deviation. Based on 30 years, different iterations show considerable bias, something that can be expected when the autocorrelation of the assessment error is so high.

In Figure A.8.12 average standard deviation of the residuals is only ≈ 0.03 while the average value is ≈ 0.04. The difference is explained by bias in different iterations.

The main problem in the evaluation of the management strategy for Icelandic cod is the stock-recruitment function in the operating model that does not have any effect on historical assessments, but is the single most important factor in simulations of the future.

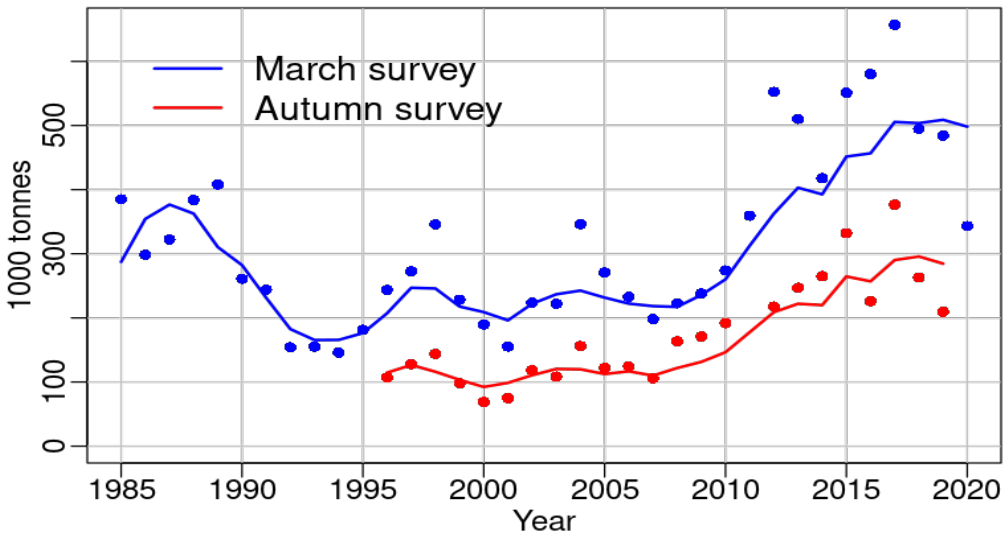

Figure A.8.10. Icelandic cod. Observed and predicted biomass from the surveys used for tuning.

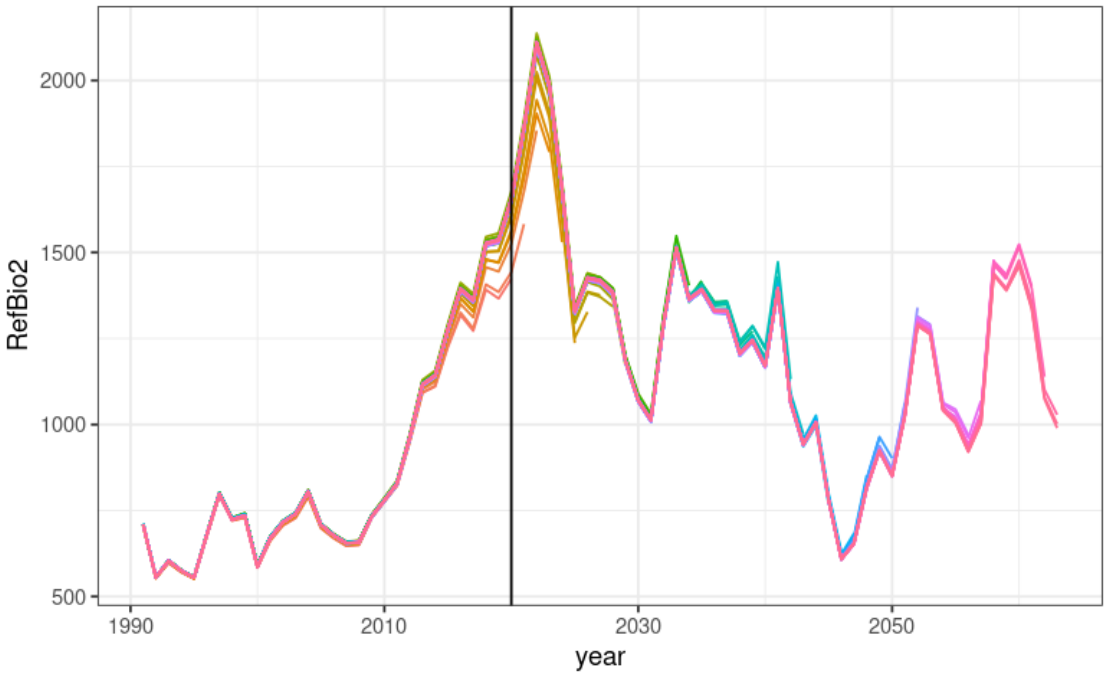

Figure A.8.11. Retrospective pattern of reference biomass (4+) for shortcut and full MSE simulations. The TAC is compiled from $TAC_{y+1} = \frac{TAC_y + 0.2 \times B_y}{2}$.

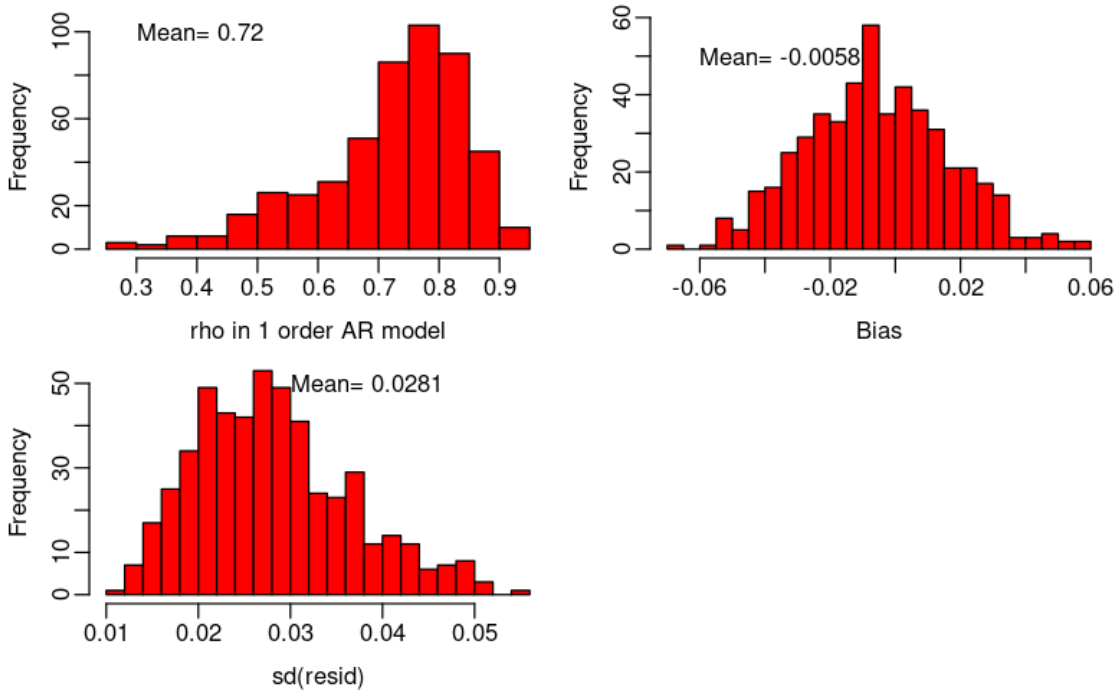

Figure A.8.12. Histogram of properties of assessment error for each of the 500 iterations based on the years 2030–2060. The averages over all the values are shown.

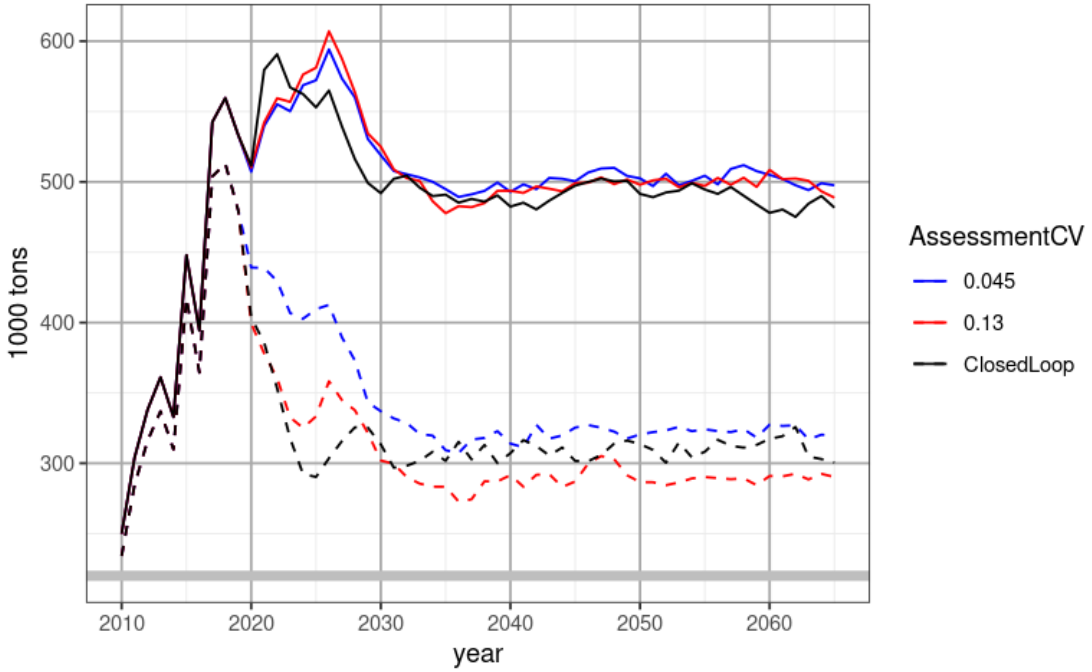

Figure A.8.13. Development of spawning stock for 2 shortcut and 1 full MSE simulation (termed "ClosedLoop"). $\rho_{ass} = 0.72$ in the shortcut simulations. Solid lines show medians, and dashed lines the 5th percentile. Horizontal thick grey lines is at $B_{trigger}$.

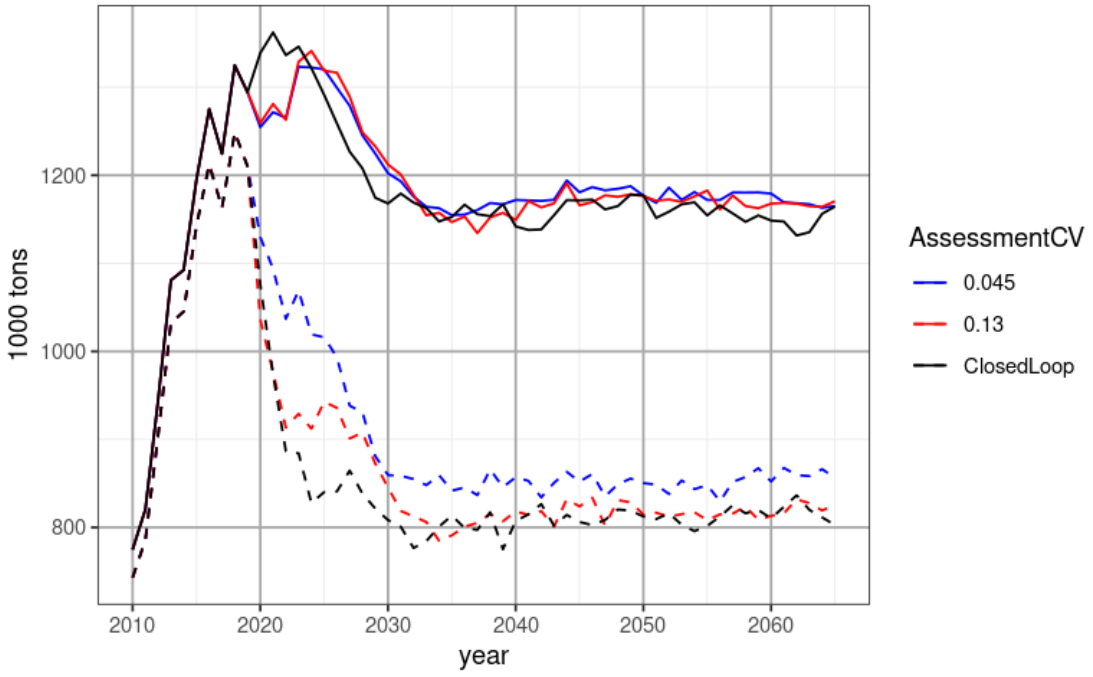

Figure A.8.14. Development of reference biomass for 2 shortcut and 1 full MSE simulation (termed "ClosedLoop"). $\rho_{ass} = 0.72$ in the shortcut simulations. Solid lines show medians, and dashed lines the 5th percentile.

How to set up operating models for stock with a "long" history of age disaggregated data

Many generations of catch in numbers at age and survey indices at age are available for a number of ICES stocks. Many of those stocks were heavily exploited earlier, but harvest rates have decreased in recent years, changing the stocks from being short lived to medium or long lived, thereby increasing generation time. For many of those stocks, which age-based assessment model is used does not matter much, as long as M is the same (SAM is then excluded as M is variable).

Best start for setting up an operating for those stocks is running a VPA, estimating CV of each age groups in surveys (something that can only be done in a VPA model) and correlation in surveys. Estimation of M can also be attempted but should be treated with care as it is confounded with correlations in the observation model.

A period of heavy exploitation can give good information about year-class strength; counting the numbers caught is enough to estimate year-class size, the exact value of M is just a minor disturbance.

Retrospective runs of the model give us an idea about assessment error that is just deviance from the VPA model. The deviance can either apply to biomass in the assessment year or F a year later (EqSim), the former is better if a short-term prediction using 1 year of TAC is included in the simulation model (HCS, Muppet) as it takes care of the amplification of error by the catch in the assessment year. The assessment error is a combination of M deviations and observation error. Models like SAM try to do the nearly impossible by splitting the assessment error between deviations in M and observation error. The estimated assessment error is most likely an overestimate if the tuning series become short in the first years of retrospective runs. Where retros go through period of high and low fishing mortality, the statistical properties of the assessment error change, something that full MSE models can be used to test.

For Muppet and many other models written in ADMB or TMB, mcmc runs can be generated saving replicates of the stock numbers in the final year (for EqSim) or replicates of SSB-recruit parameters (Muppet). Of the stock-recruitment parameters autocorrelation of the residuals is the most important one.

When this first step is finished, problems that have to be included in the operating model can be identified. A few examples are:

1. Do we see indications of nonlinear relationships between stock size and index, either for recruits or adult stock? Direct log-log comparisons are not enough as they do not handle low values correctly. Nonlinearity for the adult fish might have to be based on the sum of indices over many age groups.
2. Should we use a different M in the operating model and the assessment model?
3. Is there a trend in recruitment at the same time as there is a trend in fishing mortality? Are there indication of a wrong M? Part of M may be discards or even injuries from the fishing gear. M of pre-recruits works like a part of the SSB-recruitment function, but all ”human“ sources of M need to be identified.
4. Is there a change in recruitment at certain time point?
5. Is there any indication of density dependence in growth, as year and/or year-class effect?
6. Is there any indication of size-based selection? Constant selection over time could still mean variable selection at age. Not a problem if selection used for computing the TAC is included as part of HCR. For Icelandic haddock, selection was modelled as a function of stock weights at age in an age based model.

Of all these factors, bullet 2 can only be done in a full MSE simulation. Parameters of the stock-recruitment function, including trend and change in productivity at a specified time, can be estimated, and the estimation included as part of the operating model in shortcut or full MSE simulations. Different functional forms of the SSB-recruitment model can be tested; for example, the Ricker model is a good choice for approximating a Shaefer-type behaviour.

More complicated operating models (nonlinearities, correlations) can be put in the observation model with a simpler assessment model in the full MSE simulation. They can also be tested by retros to get the assessment error if that operating model is the same as assessment model. Longer time-series are required to get retros when temporal correlations are estimated.

Density dependent growth can be included, both in shortcut and full MSE simulations. They will lead to changes in selection by age. One way to solve this problem is by size-dependent selection and size-based HCR decision rules. Average F over a range of age groups is a poor HCR when selection by age varies. Those factors do not need a full MSE approach.

Indications of bias in analytical or empirical retros should preferably be corrected, but if that is not possible, the bias needs to be taken care of by reducing harvest rate.

For stocks where HCRs are based on analytical assessments, large number of operating models are often not required. A full MSE approach is also not necessarily required, but can be useful to look at specific problems. The management strategies for a number of Icelandic stocks have been tested by a shortcut approach. All of those stocks have long time-series of data. Cod, haddock, saithe, and herring have long series of age disaggregated data, and the HCRs have been evaluated by the Muppet model. Ling and tusk have long series of length data but a limited number of age data, so the Gadget model has been used.

Operating models of these stocks include number of features that vary from stock to stock. A decision on what is included is based on an investigation of the data.

- Cod: Number of SSB-recruit functions; change in productivity.
- Haddock: Size-based selection; density dependent growth, both year-class and year factors.
- Herring: Ichtyophonus epidemic, modelled as increased M in a few years occurring occasionally; the form of advice is to make the advice less dependent on the knowledge that epidemic is going on.

Annex 9: Special requests for MSEs for NEA mackerel: 2007–2020

Special Request 2007: EC request on evaluation of management plan for NEA mackerel

[http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2007/Special%20Requests/EC%20MP%20for%20NEA%20mackerel.pdf]

ICES is requested to identify multi-annual plans of the following form, and assuming that egg surveys of mackerel continue on a triennial basis:

1. The sum of the regulated catches for the combined stock of NEA mackerel (covering all areas where mackerel are caught) shall be set according to a fishing mortality of [A].
2. Notwithstanding paragraph 1 above, the sum of the regulated catches for the combined stock of mackerel shall no be altered by more than [B]% with respect to the sum of the regulated catches for the combined stock of the previous year.
3. Notwithstanding paragraph 1 and 2, in the event that the spawning stock size for mackerel shall be estimated at less than [C tonnes / appropriate model specified units], the sum of the regulated catches for the combined stock of mackerel, and other conservation measures as appropriate, shall be adapted to assure rebuilding of the spawning stock size to above [C] without incurring the restriction referred to in paragraph 2.

ICES is asked to identify combinations of values for A, B and C that would assure management of mackerel stock that would conform to the precautionary approach, i.e. a low risk of stock depletion, stable catches and sustained high yield. Values of A in the range 0.15 to 0.2 values of B in the range 5% to 20% and values of C above the present B_{pa} are of particular interest to managers, but ICES should explore other relevant scenarios on its own initiative as appropriate. ICES are also invited to suggest other approaches to multi-annual management of mackerel on its own initiative.

Special Request 2015: EU, Norway, and the Faroe Islands request to ICES to evaluate a multi-annual management strategy for mackerel (*Scomber scombrus*) in the Northeast Atlantic

[http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2015/Special_Requests/EU_Norway-Faroe_Islands_MAMS_for_mackerel.pdf]

In order for the Parties to develop a revised management plan for mackerel on which to base the appropriate fishing levels in the years 2015 to 2018, ICES is requested to:

1. Evaluate new biological reference points for the North East Atlantic mackerel stock based on the revised (WKPELA; ICES, 2014d) mackerel assessment method.
2. Evaluate the alternative fishing mortalities corresponding to F_{MSY}, 0.20, 0.25, 0.30 and 0.35 for appropriate age groups as defined by ICES.
3. Each alternative should be assessed in relation to how it performs with respect to stock development in the short, medium and the long term and the level of uncertainty in the stock assessment, inter annual TAC variability, long term yield, as well as in relation to the precautionary approach.
4. Each alternative shall be evaluated with an annual quota flexibility of 10%.
5. Each alternative shall also be assessed with a stability clause where the TAC shall not deviate by more than 20% from the TAC of the preceding year, but the F shall not deviate by more than 10% from the target F.

Special Request 2017: EU, Norway, and the Faroe Islands request concerning long-term management strategy for mackerel in the Northeast Atlantic

[http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2017/Special_requests/eu-fo-no.2017.19.pdf]

In order to revise the long-term management strategy, an evaluation of some alternative harvest control rules is needed. The Parties therefore ask ICES to evaluate the following harvest control rules:

Evaluate new fishing mortality reference points for the Northeast Atlantic mackerel stock based on ICES (2017d).

ICES is requested to update all the Tables given in its response to the EU, Norway and Faroe Islands request to ICES to evaluate a multi-annual management strategy for mackerel in the North East Atlantic (published 13 February 2015; ICES, 2015a), using:

- A range of $B_{trigger}$ from two to five million tonnes with an appropriate range of target Fs
- A harvest control rule with a fishing mortality equal to the target F when SSB is at or above $B_{trigger}$.
- In the case that the SSB is forecast to be less than $B_{trigger}$ at spawning time in the year for which the TAC is to be set, the TAC shall be fixed consistently with a fishing mortality that is given by: $F = F_{target}*SSB/B_{trigger}$

When updating the Tables referred to above, ICES should omit the constraint on F that had been evaluated in 2015.

All alternatives should be evaluated with and without a constraint on the inter-annual variation of TAC. When the rules would lead to a TAC, which deviates by more than 20% below or 25% above the TAC of the preceding year, the Parties shall fix a TAC that is respectively no more than 20% less or 25% more than the TAC of the preceding year. The TAC constraint shall not apply if the SSB at spawning time in the year for which the TAC is to be set is less or equal to $B_{trigger}$.

Evaluation and performance criteria

Each alternative shall be assessed in relation to how it performs in the short term (2018–2022), medium term (2023–2032) and long term (2033–2052) in relation to:

- Average SSB
- Average yield
- Indicator for year to year variability in SSB and yield
- Risk of SSB falling below B_{lim}
- Average mean weight for age groups 3-8 years in relation to long-term average mean weight

Evaluation of the management strategies shall be simulated with:

- both fixed weight-at-age and with density dependent weight-at-age.
- assessment uncertainty representing the present assessment model and input data. ICES is invited to use the values established by WKMSYREF4 (ICES, 2016b) as default if it is not possible to estimate present assessment uncertainty for NEA mackerel.

Special Request 2020: EU, Norway, and the Faroe Islands request on the long-term management strategies for Northeast Atlantic mackerel (full feedback approach)

[http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2020/Special_Requests/eufono.2020.07.pdf]

The European Union, Norway, and the Faroe Islands jointly request ICES to advise on the long-term management strategies on Northeast Atlantic Mackerel. A request is provided below.

ICES is requested to identify appropriate precautionary combinations in the Tables given in its response to the EU, Norway and the Faroe Islands request to ICES to evaluate a multi-annual management strategy for mackerel in the North East Atlantic (ICES, 2017b), using:

- A range of $B_{trigger}$ from two to five million tonnes with an appropriate range of target Fs
- A harvest control rule with a fishing mortality equal to the target F when SSB is at or above $B_{trigger}$
- In the case that the SSB is forecast to be less than $B_{trigger}$ at spawning time in the year for which the TAC is to be set, the TAC shall be fixed consistently with a fishing mortality that is given by: $F = F_{target}*SSB/B_{trigger}$

All alternatives should be evaluated with and without a constraint on the inter-annual variation of TAC. When the rules would lead to a TAC, which deviates by more than 20% below or 25% above the TAC of the preceding year, the Parties shall fix a TAC that is respectively no more than 20% less or 25% more than the TAC of the preceding year. The TAC constraint shall not apply if the SSB at spawning time in the year for which the TAC is to be set is less or equal to $B_{trigger}$.

The constraint mechanism shall be tested separately from and in combination with 10% banking and borrowing mechanism.

Evaluation and performance criteria

Each alternative shall be assessed in relation to how it performs in the short term (5 years), medium term (next 10 years) and long term (next 25 years) in relation to:

- Average SSB
- Average yield
- Indicator for year to year variability in SSB and yield
- Risk of SSB falling below B_{lim}

The approach should follow the same full feedback methodology that has been recently used to evaluate stocks in the North Sea (ICES, 2019h). The evaluation should be conducted to identify options that are robust to alternative operating models including but not limited to:

A. Investigating alternative plausible recruitment dynamics and scenarios,
B. Alternative natural mortality assumptions,
C. The potential impact of density dependent growth.

Following initial consideration of the request by ICES, the requesting parties confirmed that the strategy should also be evaluated with a banking and borrowing scheme representative of recent behaviour. The requesters furthermore confirmed that banking and borrowing should be suspended when SSB is below $B_{trigger}$, and that the implications of any future catch scenario that exceeds the advised catch should not be evaluated.