FINAL REPORT OF THE ICCAT SHORT-TERM CONTRACT: MODELLING APPROACHES: SUPPORT TO ICCAT NORTH ATLANTIC SWORDFISH MSE PROCESS

L.T. Kell[1] and P. Levontin[1]

SUMMARY

Management Strategy Evaluation was conducted for North Atlantic swordfish using an Operating Model (OM) conditioned using Stock Synthesis. A generic procedure for model validation and a shiny-app to visualise risk and uncertainty were developed. Residual runs test showed that the indices of abundance were in conflict, which may be due to model misspecification. Problems with the residuals from the fits to the indices also mean that it will be difficult to simulate psuedo data in the Observation Error Model to evaluate alternative Management Procedure. A hindcast (a forecast made retrospectively) identified that the assessment used to condition the OM has poor prediction skill. Although the OM itself does not have to predict the future state of the stock it should be representative of the main uncertainties in resource dynamics. A potential problem was that although the implied values of r and K were within plausible ranges the OM production function was highly skewed and hence B_{MSY} could be below the limit reference point (B_{lim}). This behaviour is mainly determined by parameters that are fixed (i.e. M and steepness), and has major implications for the assessment of the risk posed to the stock by harvesting.

RÉSUMÉ

L'évaluation de la stratégie de gestion a été réalisée pour l'espadon de l'Atlantique Nord à l'aide d'un modèle opérationnel conditionné par Stock Synthesis. Une procédure générique de validation de modèle et une application de type shiny permettant de visualiser les risques et l'incertitude ont été développées. Le test des scénarios résiduels a montré que les indices d'abondance étaient contradictoires, ce qui peut être dû à une spécification erronée du modèle. Des problèmes liés aux valeurs résiduelles des ajustements aux indices signifient également qu'il sera difficile de simuler des pseudo-données dans le modèle d'erreur d'observation pour évaluer une autre procédure de gestion. Une prévision rétrospective a fait apparaître que l'évaluation utilisée pour conditionner le modèle opérationnel avait une mauvaise capacité de prédiction. Bien que le modèle opérationnel en soi n'ait pas à prédire l'état futur du stock, il devrait être représentatif des principales incertitudes liées à la dynamique des ressources. Un problème potentiel était que, bien que les valeurs implicites de r et K se situaient dans des gammes plausibles, la fonction de production du modèle opérationnel était très asymétrique et que, par conséquent, la BPME pouvait être inférieure au point de référence limite (Blim). Ce comportement est principalement déterminé par des paramètres fixes (c'est-à-dire M et inclinaison) et a des implications majeures pour l'évaluation du risque que présente la capture pour le stock.

RESUMEN

Se llevó a cabo una evaluación de la estrategia de ordenación para el pez espada del Atlántico norte utilizando un modelo operativo (OM) condicionado con Stock Synthesis. Se desarrolló un procedimiento genérico para la validación del modelo y una aplicación shiny para visualizar el riesgo y la incertidumbre. Las pruebas de ensayos residuales mostraban que los índices de abundancia eran contradictorios, lo que puede deberse a la especificación errónea del modelo. Los problemas con los residuos a partir de los ajustes a los índices también significaban que será difícil simular pseudo datos en el modelo de error de observación para evaluar procedimientos de ordenación alternativos. Una simulación retrospectiva (una previsión realizada de forma retrospectiva) identificó que la evaluación utilizada para condicionar el OM tenía una escasa

[1] Centre for Environmental Policy, Imperial College London, London, United Kingdom.

capacidad predictiva. Aunque el OM en sí no tiene que predecir el estado futuro del stock, sí debería representar las principales incertidumbres en la dinámica del recurso. Un problema potencial fue que, aunque los valores implícitos de r y K se encuadraban en las gamas plausibles, la función de producción OM era muy asimétrica y, por tanto, BRMS podría hallarse por debajo del punto de referencia límite (Blim). Esta conducta está determinada por parámetros que son fijos (a saber, M e inclinación) y tiene importantes implicaciones para la evaluación del riesgo planteado al stock por la captura.

KEYWORDS

Hindcasting; Management Procedure; Management Strategy Evaluation; Operating Model; Production Function; Runs tests, Residuals; Risk; Swordfish; Uncertainty; Validation; Visualisation

Executive Summary

An initial Management Strategy Evaluation (MSE) was conducted for North Atlantic swordfish using an Operating Model (OM) conditioned using the Stock Synthesis base case from the most recent stock assessment. Several Management Procedures (MPs), based on those previously simulation-tested by the tuna RFMOs, were developed and evaluated. These include both model-based and model-free procedures. The uncertainty assumed about resource dynamics as represented by the Operating Model, and the performance of the Management Procedures can be examined using the publicly available shiny-app developed specifically for the North Atlantic Swordfish MSE in collaboration with graphic design experts.

All the code is available from a GitHub repository and the main documents, on the implementation of the MSE, are written in the form of reproducible reports using R with markdown, knitr and pandoc. This means that the work is reproducible in R. All SCRS papers were written using Google Docs to allow collaboration on multi-authored documents.

A wiki has also been developed for collaboration with the swordfish species group (SG), where a wiki is “the perpetual collective work of many authors, which anyone can edit”. It is an essential tool for the continued collaborative development of the MSE as it will allow all members of the SG to check, modify, and update the MSE. Currently, both the GitHub site and wiki are private and access permission has to be obtained from the ICCAT Secretariat.

The wiki is structured around the steps required to conduct a Management Strategy Evaluation; namely i) Identification of Management Objectives; ii) Selection of Hypotheses about Resource Dynamics; iii) Building the Operating Model; iv) Identification of alternative Management Strategies; v) Running the simulations; vi) Communicating the results of the simulations to stakeholders (including the use of a shiny-app) and vii) Agreeing the Management Strategies that best meet management objectives (after possibly iterating the previous steps).

When conditioning OMs on a full factorial design many assessment runs are conducted and it is difficult to evaluate each individual scenario without automation. Therefore a major task was to develop an automated procedure for validation, rejection and weighting of OM scenarios. When conditioning OMs on a full factorial design many assessment runs are conducted and it is difficult to evaluate each individual scenario without automation.

The work so far has identified a number of key issues that need to be addressed in the next stages of development of MSE, these relate to conditioning the OM and the choice of MPs:

- Validation of the OMs revealed poor fits to the data and potentially low prediction skill, if this is the case then use of the OMs developed so far might not provide a reliable measure of how robust the Management Procedures actually are. For example if retrospective analysis suggests a lack of stability, where an additional year (or years) of data changes the perception of resource dynamics, then it will not be possible to guarantee the robust performance of the MPs over a reasonable period. The initial MSE trials also uncovered problems with the stock assessment that may offer productive avenues for strengthening the assessment and make it more robust and stable in the future. This is especially true since the OM conditioning process and the stock assessment used for management advice are at this moment equivalent, hence validation of one serves as validation and insight into the other.

- The equilibrium dynamics of the OM, as represented by the production function, are very different from those of the biomass dynamic stock assessment model that was used in the past to provide advice. Previously a logistic production function was assumed for which B_{MSY} is 50% of virgin, however, in the current model B_{MSY} is less than 20% of virgin. This means that when the current MP is evaluated the target F has to be much less than F_{MSY} in order to stay above biomass limit reference point defined as ⅕ of virgin stock. This second issue is pertinent to the debate over definitions of limit reference points, as there is now a major difference whether these are based on virgin or MSY state. The MP based on a biomass dynamic assessment model overestimates the state of the stock relative to $B_{MSY.}$

- A consequence of the presumed OM production function is that the model implicitly considers the stock to be highly resilient to exploitation. So that even if the stock is reduced to just ⅕ of its unfished state, it is assumed to be at, or close to, MSY.

- Reliability of the model is visualised using a novel approach in the shiny app, in addition to an infographic contextualising the sources of uncertainty.

Overview

This report is a summary of the outcomes of the project, including all four deliverables; namely

I. Deliverable: Code
   A. GitHub (https://github.com/iccat/swonmse)
   B. (https://github.com/iccat/swonmse/wiki)

II. Deliverable: Design Document
   A. Operating Model (conditioned on SS, 9 main effects based on 4 sources of uncertainty)
   B. Observation Error Model
   C. Model Validation (developed and demonstrated a procedure that can be automated)

III. Deliverable: Management Procedures
   A. M: Biomass Dynamic Assessment with a hockey stick harvest control rule (HCR),
   B. P: Empirical MP where the total allowable catch (TAC) is set depending on the current value of an index relative to a reference period,
   C. D: Empirical MP where the TAC is set depending on a trend in the index.

IV. Deliverable: MSE Demonstrator
   A. (https://pl202.shinyapps.io/Swordfish_MSE_Vis/)
   B. Management Objectives
      1. Probability of being in the Kobe Green Quadrant (below target exploitation level and above target biomass level),
      2. Probability of being above limit reference point defined as the 20% of Virgin spawning stock biomass (SSB),
      3. Probability of being above 80% of target yield, here maximum sustainable yield (MSY),
      4. and stability of TAC, based on the coefficient of variation in simulations of future catches under different MPs.

The Code

A GitHub repository has been set up at a private gitHub repository https://github.com/ICCAT/swonmse. To request access please contact ICCAT secretariat at info@iccat.intThis repository is currently private but permissions could be set up to make it public, and in addition to all the code, it contains examples that can be helpful in reproducing the MSE results discussed in this report. The repository is set up to allow development of the MSE code and so it also includes prototypes and work in progress, it will be updated continuously. The code is implemented in R and FLR series of packages. The core of the code is well documented, includes online documentation, examples and has been extensively used in MSE evaluations worldwide.

The main MSE documents are written using reproducible reports, using R with markdown, knitr and pandoc. This means that the code, data and results are all integrated into the form of a report. This ensures that the work is reproducible in R.

There is also a page to help summarise the various steps required to conduct Management Strategy Evaluation, and the wiki is structured around the six main steps of this process:

- Identification of Management Objectives and mapping these to performance measures in order to quantify how well they are achieved.
- Selection of Hypotheses about Resource Dynamics for building Simulation Models, i.e. Operating Model (OMs).
- Building the Operating Model, i.e. conditioning models on data and knowledge, and rejecting and weighting different hypotheses and the Observation Error Model (OEM).
- Identification of alternative Management Procedures (MPs) or strategies, i.e. the combination of pre-defined data, stock assessment methods, reference points and Harvest Control Rules (HCRs).
- Running the simulations using the HCRs as a feedback control procedure; and
- Communicating and agreeing the Management Strategies that best meet management objectives.

Design Documents

The design documents collectively describe the OM and OEM that are part of the initial MSE trials. Following agreement at the Swordfish species Group (SG) meeting a validation procedure was developed and applied to the initial set of OMs.

The background information for the design document for the OM and the Observation Error Model had been presented in SCRS/2018/166, and the Management procedure section in this document is based on a more detailed but more generic document, SCRS/2018/167.

Conditioning/OM and OEM

The OM can be conditioned on a variety of data sets and hypotheses, however, in the first instance conditioning is based on the last N-SWO stock assessment conducted using Stock Syntheses (Methot and Wetzel, 2013). The OM is used to describe resource dynamics in simulation trials and the OEM (OEM) simulates pseudo data for stock assessments that are part of management procedures being tested.

Scenarios/OM

The working group had prioritised a number of uncertainties that the MSE should address. The Base Case Stock Synthesis assessment was used to evaluate scenarios that represent the main uncertainties. The sources of uncertainty are considered as factors with different levels, where the main effects are variations from the base case one factor at a time. For the initial set of runs 9 OM were selected **Table 1**).

There are four sources of uncertainty, modelled as factors:

1. Natural Mortality, i) Base Case (0.2), ii) low (0.1), iii) high (0.3) and iv) age varying based on the Lorenzen functional form set so that M (e.g. age at 50% mature) at a reference age is 0.2.
2. Steepness of the stock-recruitment relationship, i) Base Case (estimated at 0.88), ii) 2.5th percentile and iii) 97.5th percentile.
3. Selection pattern, i) Base Case, ii) all dome-shaped, iii) all flat-topped.
4. Effective sample size, the indices of abundance mainly affect the estimated stock trends, while the length compositions affect the absolute level of biomass; i) Base Case and ii) down-weighted length compositions.

These 9 Operating Models only account for a few of the important uncertainties and do not consider interactions between uncertainties. Various sources of uncertainty identified in previous studies and by the working group have been visualised in the MSE demonstrator app.

Figure 1 shows the probability density function of steepness as estimated by the Base Case assessment. While **Figure 2**, shows the estimate of natural mortality at-age derived from the Lorenzen vector, M at the age at 50% mature was set to have the same value as in the Base Case.

Model Validation

As part of the conditioning process, the OMs have to be validated and the plausibility of scenarios to be considered. Plausibility may be estimated formally based on statistical approaches, or based on expert judgement, and these measures of plausibility can be used to weight performance statistics when integrating over results for different scenarios (OMs).

Ideally a formal validation procedure should be agreed and adopted when conditioning OM Scenarios and not after simulations have been run. Otherwise there is a danger of an OM scenario being chosen based on the performance of the MP, rather than the reverse. It will always be tempting to argue that a particular OM is less plausible if it causes a MP which may seem preferable for some other reasons to perform poorly, therefore assigning measures of plausibility to OMs before seeing the MSE results is more prudent.

As described above the Base Case Stock Synthesis assessment was used to set up a factorial design, or grid, by the SG to evaluate scenarios that represent the main uncertainties. The tuna RFMO MSE WG, however, noted that using a limited number of OMs where only the main effects are run may be preferable to a complex grid design, as this provides enough contrast for testing the impact of most important uncertainties on management recommendations. Starting with a smaller grid design (i.e. a Reference Set of OMs) also makes more practical sense, as its difficult to check all diagnostics of all OMs run on a large grid. Less plausible or less influential uncertainties can be covered by robustness tests.

This allowed a variety of diagnostics tests to be considered and tried on the initial nine scenarios. Having such diagnostic tests is of particular importance when statistical models that integrate a variety of different data sets are used to condition OMs as these diagnostic tests are useful in identifying data conflicts and model misspecification that plague such statistical models, as argued by Maunder and Piner (2017).

Diagnostics include checks for model convergence, ascertaining that there is sufficient information in the data to actually estimate the parameters so that runs produce plausible values for time series and implied parameters such as r and K. Currently applied diagnostic techniques mainly include inspection of residuals and retrospective analysis, although other methods based on an evaluation of prediction skill are being developed by the tRFMO MSE working group.

Prediction skill is a evaluation of the of forecasts (Huschke, 1959). Throughout the remainder of this paper, “hindcast” (a term used to describe a forecast made retrospectively), “forecast” and “prediction” are used synonymously. Strictly speaking the results presented here apply only to retrospective prediction, since no results are presented for the evaluation of Management Procedures. In MSE the outputs from simulation trials are not actually predictions as they are not an attempt to describe future states but to evaluate the relative performance of alternative Management Procedures.

Various criteria are available for comparing the prediction skill of forecasts (see Hyndman and Koehler, 2006). Although root-mean-square error (RMSE) is widely used to compare model predictions it is an inappropriate and misinterpreted measure of average error (Willmott and Matsuura, 2005). RMSE is inappropriate because it is a function of 3 characteristics of a set of errors, rather than of one (the average error), RMSE varies with the variability within the distribution of error magnitudes and with the square root of the number of errors (*n*1/2), as well as with the magnitude of the average-error magnitude (MAE). MAE is a more natural measure of average error, and (unlike RMSE) is unambiguous. Scaling the average errors using the Mean Absolute Scaled Error (MASE) allows forecast accuracy to be compared across series on different scales. A scaled error is less than one if it arises from a better forecast than the average one-step naïve forecast, and it is greater than one if the forecast is worse than the average one-step naïve forecast. Where a “naïve” forecast (or random walk) is where the forecast is equal to the last observation.

Inspection of residuals may be subject to individual interpretation and hence lead to disagreements, further it is labour intensive if there are many parameters and models to consider. One alternative is to specify thresholds and automate the procedure, which is the approach taken here. In addition to questions over convergence, r and K plausibility, and residuals, a major uncertainty in stock assessment is whether dynamics are driven by a production function or variability in recruitment. Another advantage of a model validation procedure is that it can be used to explore the balance between assumptions over structure of the model versus assumptions about sources of variability based on how these appear to influence stock dynamics.

When conditioning OMs on a full factorial design many assessment runs are conducted and it is difficult to evaluate each individual scenario without automation. Unless there is a validation procedure the conclusions of MSE might be unreliable, for example, the choice of a MP may be based on scenarios that have low plausibility, e.g. carrying capacity is outside of priors limits inferred by the available habitat (Arrizabalaga et al., 2018), or where scenarios are not consistent with the observations making it difficult to generate data using the OEM. Whereas, unfairly excluding plausible OMs might lead to concluding that a MP is robust, hiding the risks, whereas a robust MSE process would have revealed the risks were a representative selection of OMs included in the analysis. A combination of accounting for various sources of uncertainty combined with a multi-test validation procedure are likely to reduce the potential for MSE to provide misleading evaluations (either declaring risky MPs safe or rejecting robust MPs).

The provision of fisheries management advice requires fitting a model to data to assess current stock status, however, the perennial question is to how do we know that a model that describes the past has anything pertinent to say about the future? Predicting the response of a stock to management requires checking that predictions are consistent with reality in some measurable way. The accuracy and precision of the prediction depend on the quality of the model, the soundness of knowledge, the information in the data and the prediction horizon (i.e. how far ahead we wish to predict). For example small changes to the input data or assumptions can result in substantial differences to advice (Collette, 2017). Therefore when conditioning OMs in order to develop robust MPs, a procedure is required for the validation of scenarios comprising alternative model structures and datasets. Where validation examines if a model family should be modified or extended, the validation method is complementary to model selection and hypothesis testing. Model selection searches for the most suitable model within a family, whilst hypothesis testing examines if the model structure can be reduced.

Therefore a variety of manual tests are reviewed for the main effects and an automated procedure is proposed that can be applied to the full set of interactions.

There are several questions that need to be addressed when employing a stock assessment paradigm to condition OMs:

1. Has model converged to a solution?
2. Are there data conflicts?
3. Is the model misspecified?
4. Does the model have prediction skill?

When using an integrated stock assessment model these questions are interrelated; exploring these questions is an active research area (Maunder and Piner 2014). The proposed four step model validation procedure described in this document addresses each of these questions in turn.

A potential validation procedure was developed by considering the above four questions in detail, and included exploring the following elements to select those that are proposed to be part of an automated procedure:

- Examining model convergence, e.g. jittering
- Checking that the runs produce plausible values for catch and biomass time series
- Performing retrospective analysis to check that an additional year of data does not greatly change the model dynamics
- Using residuals to identify problems due to data conflicts and possible model misspecification
- Checking that estimates of derived parameters such as r and K are within plausible ranges
- Checking that parameter distributions are not truncated in odd ways
- Likelihood profiles
- Performing cross-validation
- Using Age structured production model (ASPM)

Validation Procedure

Methods like SS3 are more computationally intensive than traditional methods because there are more data and parameters to consider. This means that development time is increased, especially as a model has usually to be run many times to check for implementation errors and to carry out sensitivity tests to evaluate the performance of an assessment and to determine how robust it is to its main assumptions. Therefore in this section, we propose a procedure for validating OMs and discuss how it can be potentially automated and translated into a qualitative measurement of OM reliability that can then be visualised alongside reliability of data and knowledge inputs (MSE demonstrator app, ‘Reliability’ tab).

The automated validation procedure consists of four steps: namely i) check for convergence, ii) residual runs tests, iii) checking plausibility of r and K, and iv) validation with model outputs and/or observations.

Jittering was performed for each of the Stock Synthesis runs to help ensure that a minimum had been found. There is a potential for jittering to be formally included in a validation procedure in the future. We therefore recommend that the SG should consider how the validation procedure can be amended.

The validation procedure was based on a subset of the procedures considered, several of which are active research areas under the tRFMO MSE WG and their utility is still being evaluated (Kell, et al., 2017; Carvalhao et al., 2017). Some of the potential validation measures under investigation are simulation-based and it would be difficult to run these for all potential scenarios.

In an integrated stock assessment, the indices of abundance mainly affect the estimated stock trends, while the size-composition data (i.e. the length or weight composition of the catches) affects the absolute level of biomass. The relative weighting of these data components can therefore have a large influence on the results of integrated stock assessment models especially in the presence of model misspecification, or conflicting information between data sets. Therefore the base case was reconfigured as an Age Structured Production Model (ASPM, Wetzel and Punt, 2015) to allow an evaluation the relative impact of the indices and the length composition data on the integrated assessment estimates, exploring, for example, whether dynamics are driven by a production function or the recruitment dynamics (Vera-Minte, 2017). This scenario is not intended to be used as an OM but to evaluate the impact of the data and model assumptions on the OM dynamics and allow alternative OM scenarios to be proposed.

Methods such as ASPM, hindcasting and MCMC are not intended to be used for all scenarios but are suited primarily for use with the base case to identify data conflicts and model misspecification. For example, the ASPM can be used to detect model misspecification in the growth function or whenever information on recruitment is lacking (as is the case for swordfish). ASPM could also be used to pose alternative hypotheses and scenarios for testing. Data conflicts may occur, for example, when the objective function components from different data sources achieve minima at different values for a given parameter. In which case likelihood profiling of individual data components can be used to evaluate the influence of data associated with model structure on estimated dynamics.

The fact that a large proportion of the indices of abundance failed the runs tests shows that the residuals are not IID, therefore running MCMC or a bootstrap for the scenarios is invalid. A preferred technique is therefore to use the hessian to derive the covariance matrix. Correlation between parameters may help identify problems with data or in the assumed model structure, while excessive CVs on estimated quantities or a large final gradient can be informative about misspecifications in model structure. (SCRS/2015/151 Collect. Vol. Sci. Pap. ICCAT, 72(5): 1186-1232 (2016)).

Therefore **Table 2**, evaluates problems with the assessment models. Model convergence can be based on whether or not the Hessian can be inverted; and correlation between parameters may result in an inability to invert the Hessian and so the correlation matrix was examined for highly correlated (> 0.95) and non-informative (< 0.01) parameters. Parameters estimated at a bound are also diagnostic for possible problems with data or in the assumed model structure, while excessive CVs on estimated quantities (> 50\%) or a large final gradient (> 1.00E-05) may also be indicative of uncertainty in parameter estimates or assumed model structure (Courtney, 2016).

Therefore runs tests and evaluation of the hessian/covariance matrix are more appropriate for all scenarios and lend themselves to automation, while methods like hindcasting, ASPM and MCMC and bootstrapping are best reserved for the base case or other scenarios for which the extra scrutiny can be justified. Therefore, validation tests using ASPM, MCMC and likelihood profiling were limited to the base case.

Test 1 Model convergence

Model convergence can be tested based on whether or not the Hessian can be inverted. Correlation between parameters may result in an inability to invert the Hessian, **Table 3**. This may not necessarily be a problem, but all highly correlated (> 0.95) and non-informative (< 0.01) parameters should be identified. The consequences depend on the model structure and the data used, and evaluation of model convergence in this way may suggest alternative scenarios that may help avoid issues such as high correlation or non-informative parameters.

Parameters at a bound is another diagnostic of possible problems with data or model structure. Further, a large maximum final gradient (> 1.00E-05) and excessive CVs on estimated quantities (≫ 50%) may also be indicative of inability to estimate parameter from given datasets or unresolved uncertainties in the assumed model structure (Courtney, 2016).

Model estimates should be calculated at a global maximum, if estimates change depending on the starting values it is likely that the algorithm is failing to find the global values. The stability of convergence can be assessed through jittering, a procedure when the model is rerun with new starting values.

Test 2 Residual Analysis

This step is important for validation of the model and for conditioning of the OEM.

The main approach here depends on the idea that residuals should be distributed in expected ways. We utilise a test that compares the probability of longest consecutive run (either positive or negative residuals) and the number of crossings with expected statistical behaviour.

Plotting data over time is a simple method to examine trends, patterns, and variation in data and to look at the effect of changing model specifications and data conflicts. A run chart is a simple line graph of a measure over time with the median shown as a horizontal line dividing the data points so that half of the points are above the median and half are below. Therefore if a series of catch per unit effort is an unbiased index of abundance then it should only show random variation, and the data points will be randomly distributed around the median. *Random* meaning that we cannot know if the next data point will fall above or below the median, but that the probability of each event is 50%, and that the data points are independent. *Independence* means that the position of one data point does not influence the position of the next data point, that is, data are not auto-correlated.

If the process shifts, these desired conditions for residuals are no longer true and patterns of non-random variation may be detected by statistical tests. Non-random variation may present itself in several ways. If the process centre is shifting due to improvement or degradation we may observe unusually long runs of consecutive data points on the same side of the median or notice that the graph crosses the median unusually few times. The length of the longest run and the number of crossings in a random process are predictable within limits and depend on the total number of data points in the run chart (Anhoej, 2014, 2015).

A shift signal is present if any run of consecutive data points on the same side of the median is longer than the prediction limit, round(log2(n)+3), where n is the number of useful data points, that is, data points that do not fall on the median. Data points that fall on the median do not count, they do neither break nor contribute to the run (Schilling, 2012). A crossings signal is present if the number of times the graph crosses the median is smaller than the prediction limit, qbinom(0.05, n-1, 0.5) (Chen, 2010). The shift and the crossings signals are based on a false positive signal rate around 5%. Therefore we have two statistical tests, based on residuals, that can be automated in order to detect problems in the estimation process for a number of select parameters in OMs.

Figure 4 shows run charts for the base case indices, for example although JPN_ERLY_5 has 24 points, the series only crosses the median line three times, and has a run of nine observations above the median. The shift and crossings signals are two sides of the same coin and will often signal together. However, either one is a diagnostic of non-random variation.

An index fails if the number of crossings its residuals make has less than 5% chance of occurring if it were IID. Since we are testing many indices a less stringent criteria for failure could be justified (for instance, less than 1%), this is to reduce the probability of failing the model that is good - in testing a large number of parameters it would be more likely that by chance a parameter will fail the 5% test even if all residuals were IID (it will be less likely that a parameter falsely fails 1% test by chance). What is the equivalent test to the 5% threshold for a given number of parameters should be discussed with by the SG. We can also look at the proportion of parameters that pass the residual tests under different OM, **Tables 4** and **5**, but these tests also enable us to query specific indices.

Three indices showed poor diagnostics, i.e. CAN_LATE_4, JPN_ERLY_5 and SPN_1, the conclusion is either that they are not tracking trends in abundance and so may just be adding noise to the assessment or there is some model misspecification. Under no scenarios did the CAN_LATE_4 or JPN_ERLY_5 show good diagnostic fits, while for SPN_1 for only 2 scenarios (M=0.3 and low length composition weighting) did the fits improve.

This raises two questions, i) should the assessment model be rerun for scenarios where the only indices included are those showing good diagnostic properties, and i) given some indices do not pass the test, how should we develop the Observation Error Model? In the context of the Observation Error Model, if an index has strong residual patterns then this implies it is not a good index of abundance and it could be argued that it should not be used in any Management Procedure.

There is uncertainty in the results of all stock assessments. This uncertainty arises because the observed data do not perfectly represent the population and the stock assessment model does not perfectly represent the dynamics of the population or of the fisheries.

Uncertainty in stock assessments can be expressed as parameter uncertainty, e.g. as confidence intervals and CVs through a probability distributions from a (as in this case) Monte Carlo Markov Chain simulations. These have been estimated under the assumption that the stock assessment model does perfectly represent the dynamics of the system. Since it is unlikely that this assumption is satisfied, however, these distributions may underestimate the amount of uncertainty in assessment results. That is why model structure uncertainty is investigated by running sensitivity analyses. However F**igure 5** shows the parameter uncertainty in current stock status is larger than that due to varying assumptions in the sensitivity analysis. This reflects that the data themselves are probably not that informative about productivity, this is particularly disconcerting when it comes to the issue of production function shape discussed in the beginning of the report. Therefore to develop robust advice will require testing the Management Procedures not just across a number of scenarios but to uncertainty in the parameter estimates. **Figure 5** clearly shows that the uncertainty in the derived quantities from a single scenario is greater than the difference between scenarios.

Test 3 Consider whether the implicit values of r and K parameters are plausible

Even integrated assessment models like SS have an implicit production function, i.e. where the dynamics are described by three parameters, the carrying capacity (K), the intrinsic population growth rate (r) and the shape (i.e. the ratio of B_{MSY} to K, see **Table 6)**.

A prior for the maximum intrinsic rate of increase (r) for North Atlantic swordfish was first formulated by using a stochastic life-table approach (McAllister et al. 2000) and had been applied in assessments since then. This is the base case prior for r that has been applied in ICCAT's BSP application to Atlantic swordfish since then. McAllister et al.'s (2000) methodology includes input distributions for the survival rate of pre-recruits and the natural mortality rate of fish that have recruited to the exploitable stock biomass. It presumes fixed assumptions about fecundity at age and the fraction mature at age. This methodology has been extended in Canadian BSP2 stock assessments (Stanley et al. 2009; Yamanaka et al. 2012) to a more easily parameterized protocol that applies the Euler-Lotka method to compute r which offers a near exact approximation of the Leslie matrix approach (McAllister *et al.* 2001). Stanley *et al.* (2009) replaces the prior for egg to age 1 survival rate with a prior for the steepness stock-recruit parameter. Steepness is a unitless parameter that reflects the fraction of average unfished recruitment achieved when spawning potential is reduced to 20% of unfished conditions. While it was possible to use literature based estimates of batch fecundity and frequency of spawning bouts to formulate a prior for egg to age zero survival rate (McAllister et al. 2000), the quantification of uncertainty in survival rate was entirely arbitrary. The formulation of a prior for the steepness parameter is more accessible due to its common usage in stock assessments and the numerous meta-analyses of stock-recruit data sets that have formulated priors for steepness whereby the central tendency and variance can have a rigorous empirical basis (Dorn 2002). Uncertainty was included for the stock-recruit steepness (h) parameter and the rate of natural mortality (M), means and variances for the female growth parameter estimates, the length to weight conversion factors, and parameters for the fraction maturity-at-age schedule (the prior covariances in parameter values can be assumed to be zero or can be empirically based).

A cross-test was performed to evaluate the bias in the stock assessment used in the MP. In the cross-test the OM was used to generate pseudo data and then the biomass dynamic stock assessment model (as used in the Management Procedure) was fitted to these (SCRS/2018/167). The results are summarised in **Figure 7.** In this example only five OMs were considered, i.e. the base case and two choices for M and steepness. For each OM 4 different sets of indices (total, exploited, mature and juvenile biomass) were used to fit the stock assessment. It can be seen **that there is a bias in** the estimates of stock status in the MP relative to the OM. The estimate of the stock status which is the basis for the decision in the MP is likely to be much higher than the 'reality', that is, this management procedure is likely to assume that the stock is doing well when in fact it can be overfished.

Test 4 Cross Validation

It is a desirable property that the OM is a reasonable representation of the system resource. Since a primary use of the OM is to project the resource forward for a variety of alternative MPs. Ideally the OM should represent the dynamics for a reasonable period of time. Therefore a model used as an OM scenario should be stable and have good prediction skill. Otherwise the MP cannot be tested over the likely future conditions if there is no way to model these reliably over a longer term, and the MP can only be implemented for a limited number of years before the MSE has to be rerun. Since the common assessment frequency of the main ICCAT stocks is about 3 years the OM should be stable for a least 5 years.

Therefore in step 4, we propose a series of tests that can be used to form a judgement whether a particular OM has prediction skill (Kell *et al.,* 2016).

To evaluate prediction skill, a cross validation was combined with a retrospective analysis - a method called 'hindcast'. The hindcast is a multi-step prediction where a model is fitted using a tail-cutting procedure, where data are deleted from year $t - n$ to t and then rerun using the data from year 1 to $t - n - 1$ to make predictions of what will happen in years $t - n$ to t. When conducting projections to provide managers with advice, such as a total allowable catch (TAC), the short term is of primary importance as usually the immediate consequences of management advice is a major concern of some stakeholders. However, managers are expected to take a longer-term view.

Retrospective analysis is a calculation whereby a year of data is removed from the data set before the model is rerun. The removed year of data can then be thought of as 'external data' and then predicted. More generally, to check that predictions are consistent with reality it is necessary to evaluate prediction skill (Walters and Punt, 1994; Patterson et al., 2001; Ralston et al., 2011); a statistical evaluation of the accuracy of a prediction relative to a reference model or dataset. Prediction skill can be used to compare alternative models or observations used for prediction to a reference set of estimates or data (Jin et al., 2008; Weigel et al., 2008; Balmaseda et al., 1995). If data are regarded as being representative of the dynamics of the stock then they can be used as a model-free validation measure (Hjorth, 1993), and the best performing scenarios (consisting of choices of models and data) can be identified by comparing predictions with external observations. Stock biomass cannot actually be observed so if estimates of population abundance were compared in the hindcast this would be model-based validation.

For the initial set of OM, a retrospective analysis was conducted to ensure that the assessments are not sensitive to only a few additional years of data. In general, the models performed poorly. Removing only a few years of data led most of the OM to falter on the first 3 of the model validation tests, as most showed problems with convergence, **Table 7**.

Further analysis options:

The steps described above represent an initial set of methods that can be automated and interpreted to convey reliability of individual OM and to enable these models to be prioritised into reference and robustness sets as well as rejected.

However, a wide range of other validation methods have been developed and applied in fisheries and relevant contexts. These are potential candidates for an expanded validation procedure, or as additional means of scrutiny for those OM that are considered most plausible.

Potential diagnostic tests to identify model misspecification include, i) residuals analysis and run tests; ii) retrospective analysis; iii) likelihood component profiles; and iv) Age structured production models. Carvalho et al. (2017) in a simulation exercise showed that residual analyses were easily the best detector of misspecification of the observation model while the ASPM test was the only good diagnostic for detecting misspecification of system dynamics model. While retrospective analysis and the likelihood component profile infrequently detected misspecified models. Applying multiple carefully selected diagnostics can increase the power to detect misspecification without substantially increasing the probability of falsely concluding there is misspecification when the model is correctly specified. Therefore, it is important for the validation procedure to have at least four tests, failing any one of which should raise a red flag for the OM.

Although many different datasets can be fitted to an integrated stock assessment model in some such stock assessments the index of abundance provides almost no information on population scale. Consequently, the estimates of the model outputs rely almost completely on the size- and age-composition data and model structure (Maunder and Piner, 2015). Configuring SS3 as an Age Structured Production Model (ASPM) could be used as a diagnostic tool, to in effect compare a complex model with a simplified one: ASPM can be used for comparison with a model estimating all of the model parameters and fitting to all the data as in the base case. If the ASPM cannot mimic the index, then either the stock is recruitment-driven, catch levels have not been high enough to have a detectable impact on the population, the model is incorrect, or the index of relative abundance is uncertain or not proportional to abundance (Minte-Vera, *et al*., 2017).

Fish stocks can fluctuate extensively over a large range of spatial and temporal scales independent of human exploitation (Hjort, 1926). Many studies have also shown that most exploited marine populations are not stationary, but display strong natural spatial and temporal variations (Cushing, 1976; Walters, 1987; Ottersen, 1995; Schwartzlose, 1999; Ravier, 2001; Dorner, 2008). While recent studies, such as Cury *et al.* 2014, Szuwalski, 2014, Punt, 2013 and Vert 2013, support the hypotheses that productivity may vary independently of SSB and discuss the importance of developing HCRs and reference points that take this feature of dynamics into account.

Stock status and reference points are likely to be strongly dependent on the assumptions made about the steepness parameter of the stock-recruitment relationship, the assumed levels of adult natural mortality, and the weighting assigned to the size-composition data. Therefore **Figures 8** and **9** summarise the recruitment dynamics as estimated by the Stock Synthesis runs. **Figure 8** shows the estimated stock recruitment pairs (S-R) and the stock recruitment relationships. While **Figure 9** shows the cross correlations in order to evaluate whether there is a monotonic relationship between recruitment and SSB (Szuwalski *et al*, 2104). The cross correlations were calculated based on Spearman's correlation. If there is compensation (for example recruitment follows a Beverton and Holt stock recruitment relationship) then there will be a significant correlation with a time lag equal to the age at recruitment. However, if stock biomass is driven by recruitment as a result of fluctuations in the environment then there will be negative lags at corresponding to the mature ages. Only if SSB has a larger and significant influence on recruitment than recruitment does on SSB then the existence of a S-R is supported.

ASPM

Indices of relative abundance provide information to estimate relative stock abundance trends and length composition to estimate recruitment variability and hence absolute abundance. Including composition data is necessary, and simply reweighting the size-composition data does not solve misspecification problems.

In some integrated stock assessments the indices of abundance provide almost no information on population scale. Consequently, the estimates of the model outputs rely almost completely on the size- and age-composition data and model structure. Maunder (2014) proposed a diagnostic tool that can be used to evaluate the information content of data about absolute abundance and assess whether the model is correctly specified. This diagnostic consists of comparing the results of an age-structured production model (ASPM) to those from a model estimating all of the model parameters and fitting to all the data (e.g., an integrated analysis). It is inferred that a production function is apparent in the data when the catch data explain indices with good contrast (e.g., declining and increasing trends), therefore providing evidence that the index is a reasonable proxy of stock trend. If the ASPM cannot mimic the index, then either the stock is recruitment-driven, catch levels have not been high enough to have a detectable impact on the population, the model is incorrect, or the index of relative abundance is uncertain or not proportional to abundance (Minte-Vera, 2017). This diagnostic has only begun to be implemented, and its utility remains unknown (Carvalho et al., 2017).

Figure 10 shows the time series of SSB from the two ASPM runs and the base case, the trends are similar. **Figure 11** shows the stock/recruitment relationships and **Figure 12** the corresponding production functions. While the fits to the indices are shown in **Figure 13.**

That there is little difference between the fits indicates that the changes in the abundance indices can be explained by the catches alone, i.e. the dynamics are not determined by shifts in recruitment, since a deterministic model would not be able to explain changes in abundance caused by periods of low or high recruitment.

Likelihood Profiles

Integrating multiple data sets into a single population dynamics model is contingent on being able to make assumptions that are internally consistent, explicit and that account for uncertainty. However, the integrated approach is sensitive to model misspecification and data conflicts.

For highly complex population models fitted to large amounts of often conflicting data, it is common for there to be difficulties in estimating absolute abundance. Therefore, a likelihood profile analysis was conducted for the marginal posterior likelihoods of R_0 (virgin recruitment) and steepness.

The R_0 likelihood component profile diagnostic estimates all model parameters while fixing R_0, i.e. the population scaling parameter, at different values and plotting the resultant likelihood (or commonly, the negative log-likelihood) value for each data component against this parameter (Francis, 2011; Lee et al., in press). The likelihood profiles of each data component usually follow a smooth parabolic curve, indicating the value with most support and the amount of uncertainty in that support. Different maxima (minima if the negative log-likelihood is used) among data components indicate possible conflict in the data sources about absolute abundance. The higher the gradient in the likelihood profile the more influential that data source on the model's estimate of scale (Maunder and Piner, 2015). The R_0 profile has been used to diagnose selectivity misspecification (Lee et al., 2019; Wang et al., 2109).

These show that there is a strong negative correlation between steepness and absolute abundance, e.g. the stock is either very productive or large. The profile of steepness can be interpreted to show that a high value (>0.9) is supported by most datasets, or that there is little information on year-class strength in the data, or that that the stock has never been low enough for compensatory dynamics to be evident, **Figure 14**.

For R0 the length data appear to support a low level of abundance, however, it is important to note that at values of log(R0) less than 6.6 the assessment failed to converge and that the prior appears to be driving the fit, **Figure 15** and **16**.

Automated Validation Procedure

Performing each of these model validation steps generates a lot of information, therefore a method for interpreting this information and communicating it to non-modellers is essential. We follow the lead of (https://www.wur.nl/en/show/JAKFISH-Judgment-and-knowledge-in-Fisheries-Management.htm) and (http://www.myfishproject.eu/) EU projects that pioneered the use of qualitative scales for data and knowledge model reliability to develop a four score qualitative scale for model validation onto which the results from each of the validation procedure step can be mapped, **Table 8**. The scale is the same for each of the four tests, and it ranges from ‘Failed’ to ‘Very Good.’

Jakfish developed a four point scale for evaluating key data and knowledge components for fisheries models, **Figure 17** and **Figure 18**.

Ratings have been slightly modified from the criteria used by Jakfish, see also ‘Reliability’ tab in the shiny app, and can be modified further with feedback from the working group in the future. Myfish has built on the communication platform for model reliability developed by Jakfish to include model validation and uncertainty tests (**Table 8**). We have developed this approach further to provide a visual summary of the quality of data, knowledge and model (https://pl202.shinyapps.io/Swordfish_MSE_Vis/, ‘Reliability’ tab).

We tested these validation procedures for the 9 OM and results are presented in **Table 9**, and averaged in the visualisation app, ‘Reliability’ tab (where half-filled circle indicate that the results were in between adjacent states, **Figure 19**).

Each step is described above in more detail, providing a manual for performing such validation assessment for the OMs in the future.

We propose that failing any of the first three tests (convergence, residual runs, r and K) disqualifies the OM from the list of plausible representations to which the management procedure need to be shown to be robust. In other words, achieving a ‘Failed’ score on any of the first three tests can translate into a decision to consider a specific conditioning of the OM as implausible. This does not mean however that an alteration to the model such as dropping some of the data or a slight modification of assumptions or structure would not be sufficient to pass the test and probably should be attempted before dismissing the model.

Similarly, we propose that performing “poorly” on any of the tests should prevent the OM from being included on the reference set. However, a model that performs in the ‘Poor’ to ‘Very Good’ range can still be considered for robustness trials.

The relative choice and importance of the scores across the elements of the validation procedure and the consequences is an area for discussion. For example, based on the scores, should a scenario be excluded (model selection), should some factors be dropped (hypothesis testing) or should alternative scenarios be developed?

Management Procedures

Alternative management strategies are specified in the form of an MP, which are the combination of pre-defined data, together with an algorithm to which such data are input to provide a value for a TAC or effort control measure. The intention is to demonstrate, through simulation trials, to show robust performance in the presence of uncertainties (Rademeyer, 2017).

It is generally necessary to simplify MPs in order to evaluate them quantitatively through MSE, for instance, the simulated stock assessment method that forms part of the MP is often much less complex than the real-world stock assessment upon which the Operating Model is commonly based.

There are two main types of MP, namely:

> ***Empirical:*** An MP where resource-monitoring data (such as survey estimates of abundance) are input directly into a formula that generates a control measure such as a TAC without an intermediate (typically population-model based) estimator; and
>
> ***Model-based:*** An MP where the process used to generate a control measure such as a TAC is based on a determination of stock status resulting from a population assessment model.

At this initial stage, we evaluated 3 MP, two empirical (one trend-based, and one conditioned on the historical catch data) and one model-based (biomass dynamic model). For each of the three management procedures, four variants were tested.

Implementation

The MPs are implemented in R and can be installed from the gitHub repository where examples are provided

Visualisation/MSE demonstrator

An interactive web application was developed for this project in collaboration with a graphic designer Jana Kleineberg who has joined the Sea++ team. The app is hosted on a shiny server and is publicly available at

The app is in an initial stage of development, as is the MSE for the North Atlantic Swordfish, and will be developed to allow stakeholder evaluate the impact of uncertainty on management outcomes. The app currently conveys information on uncertainties that are relevant to the MSE process, most of which are yet to be accounted for. Information on the reliability of modelling, that includes a qualitative assessment of the modelling inputs, model validation, and the degree to which uncertainties were included.

As noted by the tRFMO MSE WG major interactions with decision makers and stakeholders will best be conducted by illustrating trade-offs, so that they can choose between tangible options on the basis of actual projections rather than abstract concepts. The initial MP design and performance statistics, however, should be few, informative and based axes such as ‘stock status’, 'safety', 'stability' and 'yield'.

Therefore a only a limited number of MPs were evaluated based on procedures already in use by the tuna RFMOs.

Some initial results are presented in the form of the three alternative MPs namely M: the biomass dynamic MP with on a Hockey Stick HCR (as used for North Atlantic albacore), P: empirical HCR based on trend in CPUE, and D: Empirical HCR based on relative difference in CPUE. The latter two are adapted from the CCSBT MP.

Management Objectives

The performance of MP is judged based on four performance criteria:

1. Probability of being in Kobe Green, in all simulations for 2019 -2038
2. Probability of being above 20% of Virgin SSB,
3. Probability of being above 80% of MSY,
4. and stability of TAC

Of these four measures, meeting the 20% of Virgin SSB was the most difficult one to achieve in some of the scenarios but this is related to implausibly high estimates of Virgin SSB relative to the SSB at MSY. It is necessary to further examine the SS stock assessment to determine what causes such low estimates of SSB at MSY relative to the unfished population and whether this is reasonable. Given that in the past advice was based on a logistic production function that assumes that B_{MSY} is found at 50% of virgin biomass, it would be a reasonable first step in the MSE to evaluate the performance of such implicit management procedure.

For these four criteria we can define explicit acceptance thresholds, for example, we might want to be above LRP with at least 75% probability and we might want to be in the Kobe Green quadrant with at least 75% chance.

We can then use the MSE demonstrator to see which MP perform acceptably under all OM, or if such MP cannot be found, explore which scenarios are causing MPs to fail.

The most challenging feature of the current MSE is that the model conditioned on SS predicts that the MSY occurs when the stock is reduced by about 80% from a virgin level. All MSE results are dependent on this built-in dynamics that derive from the SS-based stock assessment. Furthermore, model validation procedures undermine our confidence in the validity of the OM conditioned on the SS stock assessment. Given uncertainties in data, knowledge on productivity parameters and model structure, exclusive reliance on an OM that represents a population that is unusually resilient to exploitation might be inconsistent with the precautionary approach and the spirit of the MSE.

Conclusions

The results from the ASPM suggest that dynamics are mainly determined by the production function, as the length compositions which inform the variability in recruitment and absolute abundance had little relative impact in model fits. The production function is determined by parameters for which there is little information in the data, i.e. M, steepness, R0 and growth. Therefore, the stock dynamics portrayed by the assessment is largely driven by prior beliefs and knowledge of the working group as expressed in the stock assessment model's structural assumptions. Therefore the next step should be validation of these assumptions and a consideration of scenarios that reflect alternative but plausible assumptions.

The production function in the OM is very skewed, while the production function used in the biomass dynamic model previously used for advice was based on the logistic form and so is symmetric (this form is the default shape in fisheries and has also been used in the past for this and other swordfish stocks). This means that although the assessment could detect trends, estimates of reference points obtained by the MP and those in the OM will be substantially different. The MP overestimates the state of the stock relative to BMSY, see **Figure 20** below (it includes an additional scenario for which we have preliminary results related to recruitment variability - this scenario, generally omitted from this report, will be more formally considered in the next stage of the MSE trials). The basis for the skewness of production function needs to be thoroughly investigated by the group.

Figure 20 shows that the F_{MSY} and B_{MSY} reference points in the OM and MP display different behaviour. In the OM, F_{MSY} and B_{MSY} are negatively correlated, and show large variability depending on the OM scenario. The corresponding values in the MP are higher than the OM but show less variability across scenarios, neither do they appear to be correlated with the OM. This implies that the performance of an MP could be very sensitive to the choice of OMs.

Two types of diagnostics need to be developed and agreed upon, namely how i) to weight or reject hypotheses and ii) to pose alternative hypotheses.

It appears that Parameter uncertainty is as great as the uncertainty between scenarios, this implies that the data used in the assessment are largely non-informative. The poor fits to the indices of abundance, as judged by the residual runs test, and the lack of impact of the length composition data support this. This reiterates the need to be inclusive with respect to different ideas about stock dynamics in the model specification and to communicate uncertainty in the stock assessment and MSE process, as well as their reliability, to decision makers.

A key insight from these initial runs are to do with the conditioning of the Operating Model. The high coefficients of variation (i.e. >50%) for some estimated parameters implies that the information in the data is lacking for some modelled processes (e.g. vulnerability and recruitment) and there also needs to be a full exploration of factors that affect the production function which make it seem that the stock is best exploited at levels much lower than 50% of unfished biomass. For these reasons alternative ways to develop Operating Models should be considered.

The current automated validation procedure is an initial proposal and will benefit from an engagement with the group and other members of the MSE community of scientists. Currently, only a few sources of uncertainty have been considered and without interactions. The next stage of the project should involve expanding the number and variety of scenarios subject to those scenarios passing the automated validation procedure.

The shiny app is a prototype, i.e.an initial proposal of how to represent the key aspects of the MSE and communicate results. The app would benefit from feedback and should be iteratively tested and improved and expanded in the next stage of the project.

One of the key developments is in using graphic design to aid communication of the important aspects of the MSE process, not just the results of modelling. Additional recommendation for future work could include a continued elaboration of the types of schemes pioneered by JAKFISH and MYFISH projects that sought to develop a way of capturing key aspects of models, so several can be easily compared and relative reliability of the models could be communicated.

In this project we added a visualisation of the validation results to a qualitative guide to data and knowledge inputs however as with any communication tool, feedback and iterative design are necessary steps. The scoring is based on a four point scale, with four always being the best option. The lowest score can denote either an absence of key data, a lack of knowledge of a key process, or a failure of a key validation test. The details of the validation procedure and the basis for assigning particular scores are works in progress, and the next steps should include seeking feedback, making appropriate changes and seeking broader consensus for this visualisation approach as well as facilitating its wider adoption in the MSE community.

References

Anhoej, J., Vingaard Olesen, V. (2014).. PLoS ONE 9(11): e113825.

Anhoej, J. (2015).. PLoS ONE 10(3): e0121349.

Arrizabalaga, H., Erauskin, M., and Coelho, R. 2018. Relative Habitat Size For Swordfish Stocks Based On AGlobal Habitat Model, SCRS/2018/043.

Balmaseda, M.A., Davey, M.K., Anderson, D.L., 1995. Decadal and seasonal dependence of ENSO prediction skill. J. Clim. 8 (11), 2705–2715.

Cass-Calay, S. L., Tetzlaff, J. C., Cummings, N. J., and J. J. Isely. 2014. Model diagnostics for stock synthesis 3: examples from the 2012 assessment of cobia in the us gulf of mexico. *Collect. Vol. Sci. Pap. ICCAT*, 70 (5): 2069–2081, 2014.

Carvalho, F., Punt, A. E., Chang, Y.-J., Maunder, M. N. and Piner, K. R. (2017) Can diagnostic tests help identify model misspecification in integrated stock assessments? *Fisheries Research*, 192: 28–40.

Chen, Z. (2010). A note on the runs test. Model Assisted Statistics and Applications 5, 73-77

Cury, P. M., Fromentin, J.-M., Figuet, S., and Bonhommeau, S. (2014). Resolving Hjort’s dilemma how is recruitment related to spawning stock biomass in marine fish? Oceanography, 27(4):42–47.

Courtney, D., 2016. Preliminary Stock Synthesis (SS3) model runs conducted for North Atlantic blue shark. Collect. Vol. Sci. Pap. ICCAT, 72(5), pp.1186-1232.

Cushing D. H. and Dickson. R. R. 1976.The biological response in the sea to climatic changes. Advances in Marine Biology, 14:1–122

Deroba, J.J., Butterworth, D.S., Methot Jr, R.D., De Oliveira, J.A.A., Fernandez, C., Nielsen, A., Cadrin, S.X., Dickey-Collas, M., Legault, C.M., Ianelli, J. and Valero, J.L., 2014. Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods. *ICES Journal of Marine Science*, *72*(1), pp.19-30.

Dorn, M.W. 2002. Advice on West Coast rockfish harvest rates from Bayesian meta-analysis of stock–recruit relationships. N. Am. J. Fish. Manage. 22(1): 280–300.

Dorner, B. Peterman, R., and S. Haeseker. S, 2008. Historical trends in productivity of 120 pacific pink, chum, and sockeye salmon stocks reconstructed by using a kalman filter. Canadian Journal of Fisheries and Aquatic Sciences, 65(9):1842–1866.

Francis, R.C., 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences, 68(6), pp.1124-1138.

Hjort, J., 1914. Fluctuations in the great fisheries of northern Europe viewed in the light of biological research. ICES.

Hjorth, J.U., 1993. Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap. CRC Press.

Huschke, R.E., 1959. Glossary of meteorology. American Meteorological Society

Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), pp.679-688.

Jin, E.K., Kinter III, J.L., Wang, B., Park, C.-K., Kang, I.-S., Kirtman, B., Kug, J.-S.,Kumar, A., Luo, J.-J., Schemm, J., et al., 2008. Current status of ENSO prediction skill in coupled ocean-atmosphere models. Clim. Dyn. 31 (6), 647–664.

Kell, L.T., Kimoto, A. and Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. *Fisheries research*, *183*, pp.119-127.

Kell, L., Levontin, P., 2018a. Design Document For The North Atlantic Swordfish Management Strategy Evaluation. Operating Model (Om) And Observation Error Model. Collect. Vol. Sci. Pap. ICCAT, 75(4): 624-676.

Kell, L., Levontin, P., 2018b. Potential Management Procedures For North Atlantic Swordfish. Collect. Vol. Sci. Pap. ICCAT, 75(4): 677-699.

Lee, H.H., Piner, K.R., Methot Jr, R.D. and Maunder, M.N., 2014. Use of likelihood profiling over a global scaling parameter to structure the population dynamics model: an example using blue marlin in the Pacific Ocean. Fisheries Research, 158, pp.138-146.

Lorenzen, K and Enberg, K. 2002. Density-dependent growth as a key mechanism in the regulation of fish populations: evidence from among-population comparisons. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269 (1486): 49–54.

McAllister, M.K., Babcock, E.A., Pikitch, E.K. and Prager, M.H., 2000, Application of a non-equilibrium generalized production model to South and North Atlantic swordfish: Combining Bayesian and demographic methods for parameter estimation). Collect. Vol. Sci. Pap. ICCAT, 51(5): 1523-1550.

McAllister M.K. 2014. A generalized Bayesian Surplus Production stock assessment software (BSP2). Collect. Vol. Sci. Pap. ICCAT, 70 (4):1725-1757.

Maunder, M. N. and Piner. K. R. 2015. Contemporary fisheries stock assessment: many issues still remain. *ICES Journal of Marine Science*, 72 (1): 7–18.

Maunder, M. N. and Piner. K. R. 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. *Fisheries Research*, 192: 16–27

Maunder, M.N and Punt, A.E. 2013. A review of integrated analysis in fisheries stock assessment. *Fisheries Research*, 142: 61–74

Methot Jr, R.D. and Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fisheries Research, 142, pp.86-99.

Minte-Vera, C., Maunder. M. N., Aires-da-Silva A. M., Satoh, K. and Uosaki, K. 2017. Get the biology right, or use size-composition data at your own risk. Fisheries Research 192 (2017) 114–125

Patterson, K., Cook, R., Darby, C., Gavaris, S., Kell, L., Lewy, P., Mesnil, B., Punt, A., Restrepo, V., Skagen, D.W., et al., 2001. Estimating uncertainty in fish stock assessment and forecasting. Fish Fish. 2 (2), 125–157.

Preece, A., Hillary, R. and Davies, C., 2011. Identification of candidate limit reference points for the key target species in the WCPFC.

Punt, A. E., A’mar, T., Bond, N. A., Butterworth, D. S., de Moor, C. L., De Oliveira, J. A., Haltuch, M. A., Hollowed, A. B., and Szuwalski, C. (2013). Fisheries management under climate and environmental uncertainty: control rules and performance simulation. ICES Journal of Marine cience: Journal du Conseil.

Rademeyer, R. A., Plagányi, É. E., & Butterworth, D. S. (2007). Tips and tricks in designing management procedures. *ICES Journal of Marine Science*, *64*(4), 618-625.

Ralston, S., Punt, A.E., Hamel, O.S., DeVore, J.D., Conser, R.J., 2011. A meta-analytic approach to quantifying scientific uncertainty in stock assessments. Fish. Bull. 109 (2), 217–232.

Ravier C. and Fromentin. J. , 2001.. Long-term fluctuations in the eastern atlantic and mediterranean bluefin tuna population. ICES Journal of Marine Science: Journal du Conseil, 58(6):1299–1317.

R. Schwartzlose, J. Alheit, A. Bakun, T. Baumgartner, R. Cloete, R. Crawford, W. Fletcher, Y. Green-Ruiz, E. Hagen, T. Kawasaki, et al. 1999. Worldwide large-scale fluctuations of sardine and anchovy populations. South African Journal of Marine Science, 21(1):289–347.

Stanley, R. D., McAllister, M., Starr, P. and Olsen, N. 2009. Stock assessment for bocaccio (*Sebastes paucispinis*) in British Columbia waters. DFO Can. Sci. Advis. Sec. Res. Doc. 2009/055: xiv + 200 p.1738

Schilling, M. F., 2012. The Surprising Predictability of Long Runs. Math. Mag. 85, 141-149

Szuwalski, C. S., Vert-Pre, K. A., Punt, A. E., Branch, T. A., and Hilborn, R. (2014). Examining common assumptions about recruitment: a meta-analysis of recruitment dynamics for worldwide marine fisheries. Fish and Fisheries.

Vert-Pre, K. A., Amoroso, R. O., Jensen, O. P., and Hilborn, R. (2013). Frequency and intensity of productivity regime shifts in marine fish stocks. Proceedings of the National Academy of Sciences, 110(5):1779–1784.

Walters, C., 1987. Nonstationarity of production relationships in exploited populations. Canadian Journal of Fisheries and Aquatic Sciences, 44(S2):156–165

Walters, C., Punt, A., 1994. Placing odds on sustainable catch using virtual population analysis and survey data. Can. J. Fish. Aquat. Sci. 51 (4), 946–958.

Wang, S.P., Maunder, M.N., Piner, K., Aires-da-Silva, A. and Lee, H.H., 2014. Evaluation of R0 profiling as a diagnostic for selectivity curve structure in integrated stock assessment models. *Fish. Res.*, *DOI*, *10*.

Weigel, A., Liniger, M., Appenzeller, C., 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q. J. R. Meteorol. Soc. 134 (630), 241–260.

Wetzel, C.R., Punt, A.E., 2015. Evaluating the performance of data-moderate and catch-only assessment methods for U.S. west coast groundfish. Fish. Res. 171, 170–187. doi:10.1016/j.fishres.2015.06.005

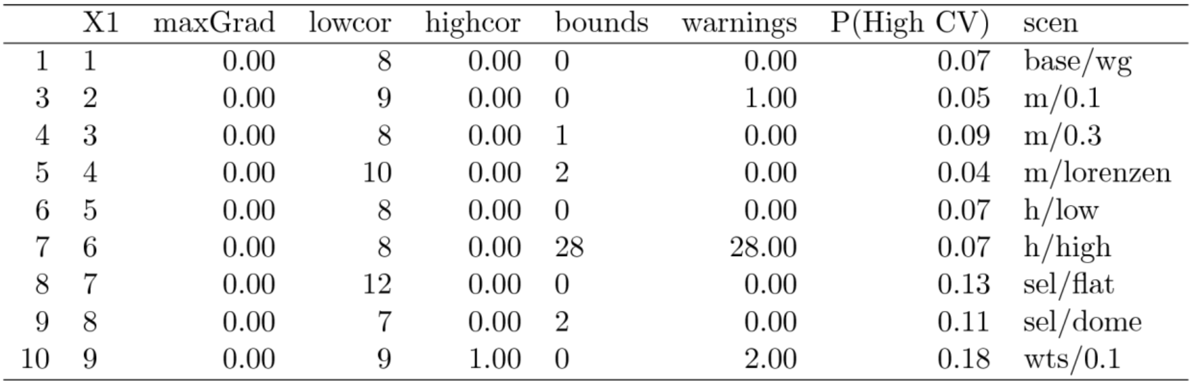
Willmott, C.J. and Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), pp.79-82.

Yamanaka, K. L., McAllister, M. K., Olesiuk, P. F. Etienne, M.-P., Obradovich, S. and Haigh, R. 2012. Stock Assessment for the inside population of yelloweye rockfish (Sebastes ruberrimus) in British Columbia, Canada for 2010. DFO. Can. Sci. Advis. Sec. Res. Doc. 2011/129. 131 p.

Table 1. Initial set of OM.

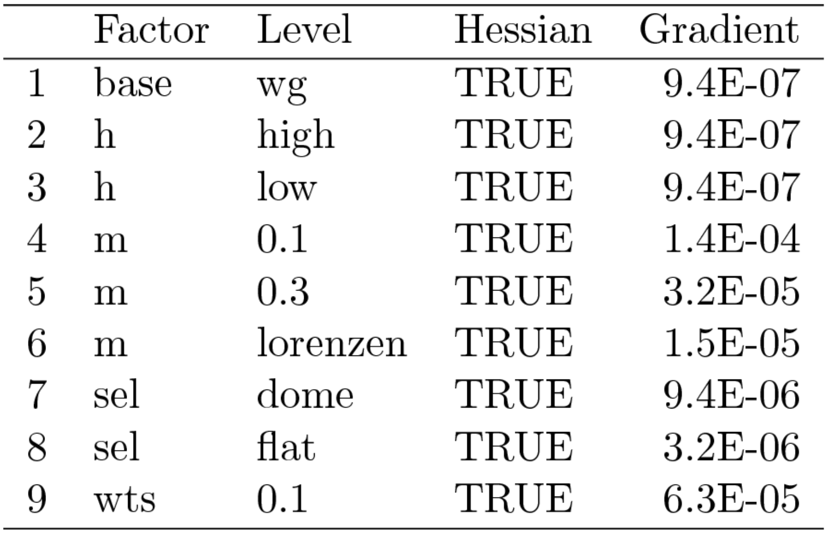
Operating Model	Scenario
OM 1	Base Case
OM 2	Steepness 97.5th percentile
OM 3	Steepness 2.5th percentile
OM 4	Low (0.1) natural mortality
OM 5	High (0.3) natural mortality
OM 6	Lorenzen M
OM 7	Selectivity dome-shaped
OM 8	Selectivity flat-topped
OM 9	Down weighted length compositions (0.1)

Table 2. Summary of parameter estimates.



	X1	maxGrad	lowcor	highcor	bounds	warnings	P(High CV)	scen
1	1	0.00	8	0.00	0	0.00	0.07	base/wg
3	2	0.00	9	0.00	0	1.00	0.05	m/0.1
4	3	0.00	8	0.00	1	0.00	0.09	m/0.3
5	4	0.00	10	0.00	2	0.00	0.04	m/lorenzen
6	5	0.00	8	0.00	0	0.00	0.07	h/low
7	6	0.00	8	0.00	28	28.00	0.07	h/high
8	7	0.00	12	0.00	0	0.00	0.13	sel/flat
9	8	0.00	7	0.00	2	0.00	0.11	sel/dome
10	9	0.00	9	1.00	0	2.00	0.18	wts/0.1

Table 3. Summary of hessian invertibility and gradients.



	Factor	Level	Hessian	Gradient
1	base	wg	TRUE	9.4E-07
2	h	high	TRUE	9.4E-07
3	h	low	TRUE	9.4E-07
4	m	0.1	TRUE	1.4E-04
5	m	0.3	TRUE	3.2E-05
6	m	lorenzen	TRUE	1.5E-05
7	sel	dome	TRUE	9.4E-06
8	sel	flat	TRUE	3.2E-06
9	wts	0.1	TRUE	6.3E-05

Table 4. Summary of number of Indices that past the runs crossing tests.

	Factor	Level	Proportion
1	base	wg	0.44
2	h	high	0.44
3	h	low	0.44
4	m	0.1	0.44
5	m	0.3	0.44
6	m	lorenzen	0.44
7	sel	dome	0.44
8	sel	flat	0.44
9	wts	0.1	0.44

Table 5. Summary of number of Indices that past the runs length test.

	Factor	Level	Proportion
1	base	wg	0.67
2	h	high	0.67
3	h	low	0.67
4	m	0.1	0.67
5	m	0.3	0.78
6	m	lorenzen	0.67
7	sel	dome	0.67
8	sel	flat	0.67
9	wts	0.1	0.78

Table 6. Summarises of values related to Virgin and BMSY states. Scenarios 6 and 8 imply that MSY occurs when the population is reduced by 88% percent from unfished or Virgin level, therefore the target for MPs in the MSE becomes 12% of Virgin stock.

	Factor	Level	BMSY	Virgin	MSY	BMSY.Virgin	MSY.BMSY
1	base	wg	21000	120000	13000	0.18	0.60
2	h	high	16000	110000	13000	0.14	0.84
3	h	low	29000	130000	12000	0.23	0.42
4	m	0.1	52000	290000	12000	0.18	0.23
5	m	0.3	17000	73000	13000	0.23	0.77
6	m	lorenzen	15000	130000	13000	0.12	0.85
7	sel	dome	19000	110000	13000	0.17	0.66
8	sel	flat	11000	92000	13000	0.12	1.13
9	wts	0.1	24000	120000	13000	0.19	0.54

Table 7. Values of Mean Absolute Scaled Error (MASE) from the hindcast for the OM length composition scenario; this was the only scenario for which it was possible to run the full 5 retrospectives.

	Factor	Level	Variable	MASE
1	wts	0.1	Biomass	2.59
2	wts	0.1	SSB	1.98
3	wts	0.1	Recruits	0.26
4	wts	0.1	Female	2.52
5	wts	0.1	Male	2.68

Table 8. Key to a qualitative assessment of OM reliability.

Order	Tests	Very Good	Good	Poor	Failed
1	Convergence	Maximum Gradients<1e-5 and Hessian invertible for all retrospective runs	Maximum Gradients<1e-5 and Hessian invertible for most recent run.	Maximum Gradients>1e-5 but Hessian invertible.	Hessian not invertible
2	Residual Runs Test	All Pass	<50% Fail	>50% Fail	All Fail
3	r and K	Both parameters within 1 sd range	The growth parameter r is within 1sd range but K is not	Both parameters are outside the 1 sd range	Both parameters are outside the 2 sd for this species
4	External Data/Cross validation	Mean Absolute Scaled Error test for 1, 2, 3, 4, 5 step ahead*	Mean Absolute Scaled Error test for 1 step ahead for 5 retrospectives	Can run 5 retrospectives	Can not run retrospectives

*Mean Absolute Scaled Error (MASE) has the desirable properties of scale invariance, predictable behaviour, symmetry, interpretability and asymptotic normality.

Table 9. Results of applying the validation procedure to the initial set of OM.

Scores	OM 1	OM 2	OM 3	OM 4	OM 5	OM 6	OM 7	OM 8	OM 9
Convergence	Good	Good	Good	Poor	Poor	Poor	Good	Good	Poor
Residual Runs Test	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor
r and K	V. Good	V .Good	Good	Good	Good	V. Good	V. Good	V. Good	Good
External Data/Cross validation	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Poor	Fail

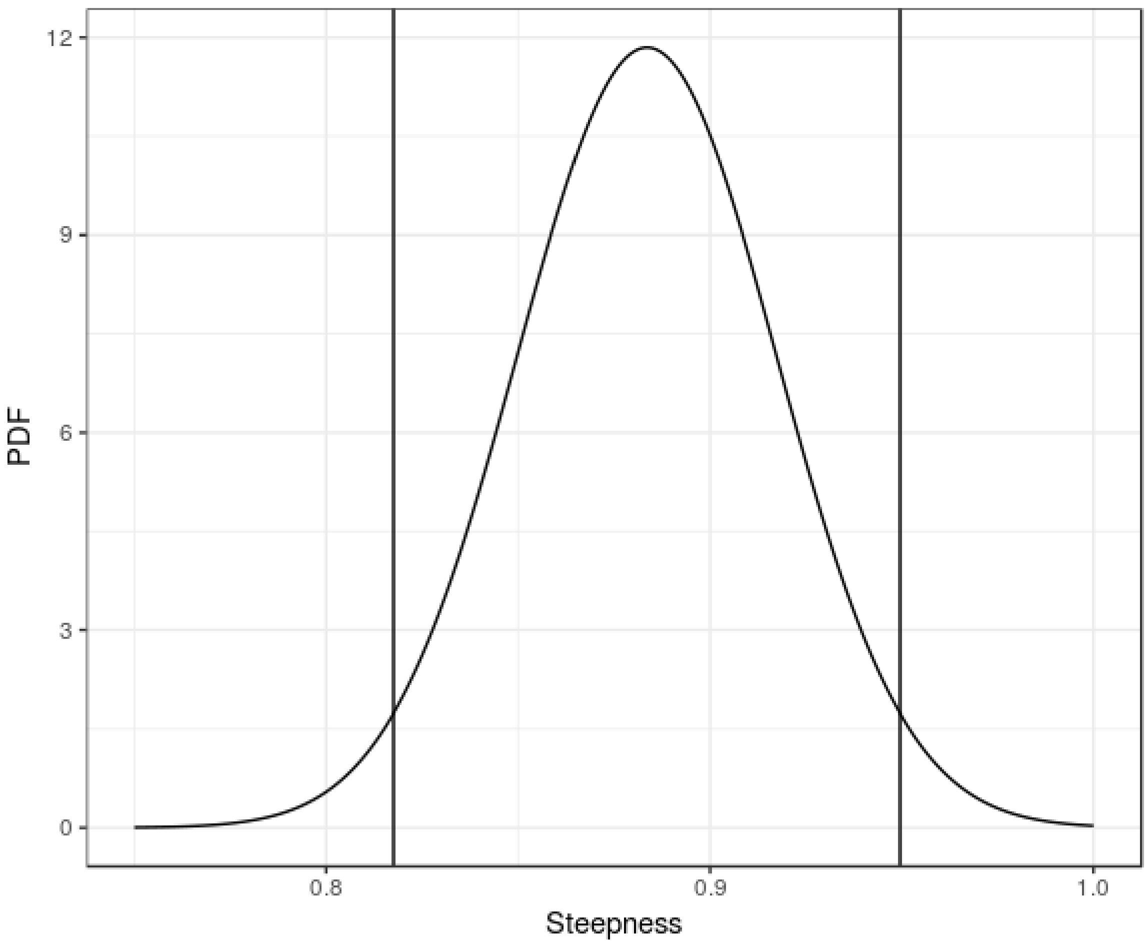


Figure 1. Probability distribution of steepness from the Base Case Assessment, showing the 2.5th and 97.5th percentiles. This distribution implies that there is zero chance that recruitment can decline by more than 25% percent of the unfished level if the biomass is decreased by 80% (the lower limit for the distribution is 0.75).

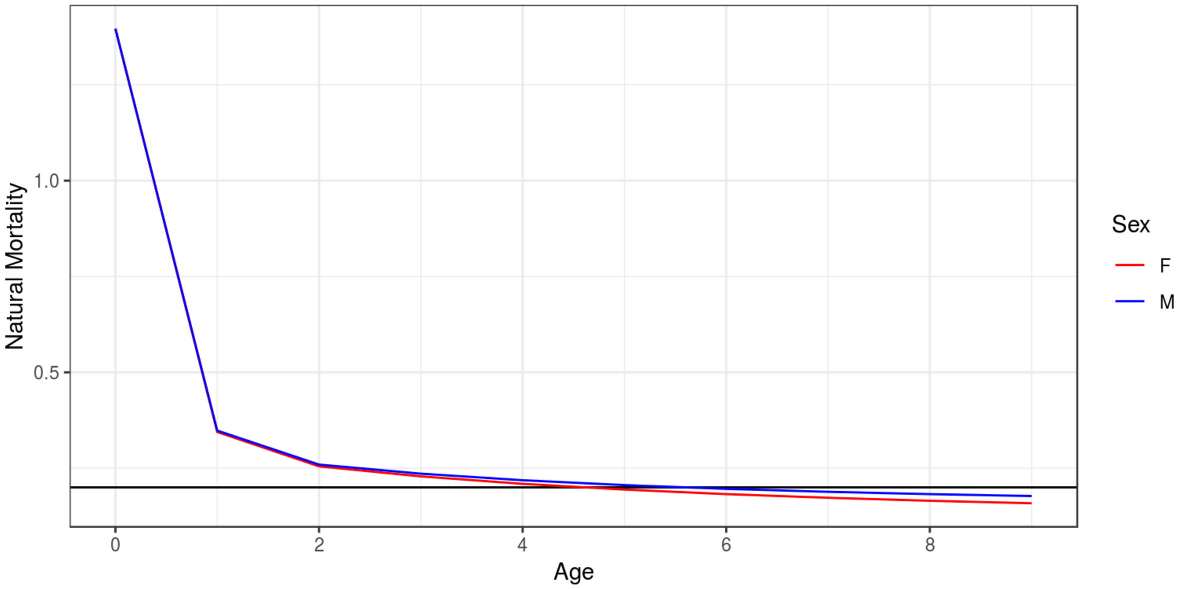


Figure 2. Natural mortality vectors-at-age for the Lorenzen relationship.

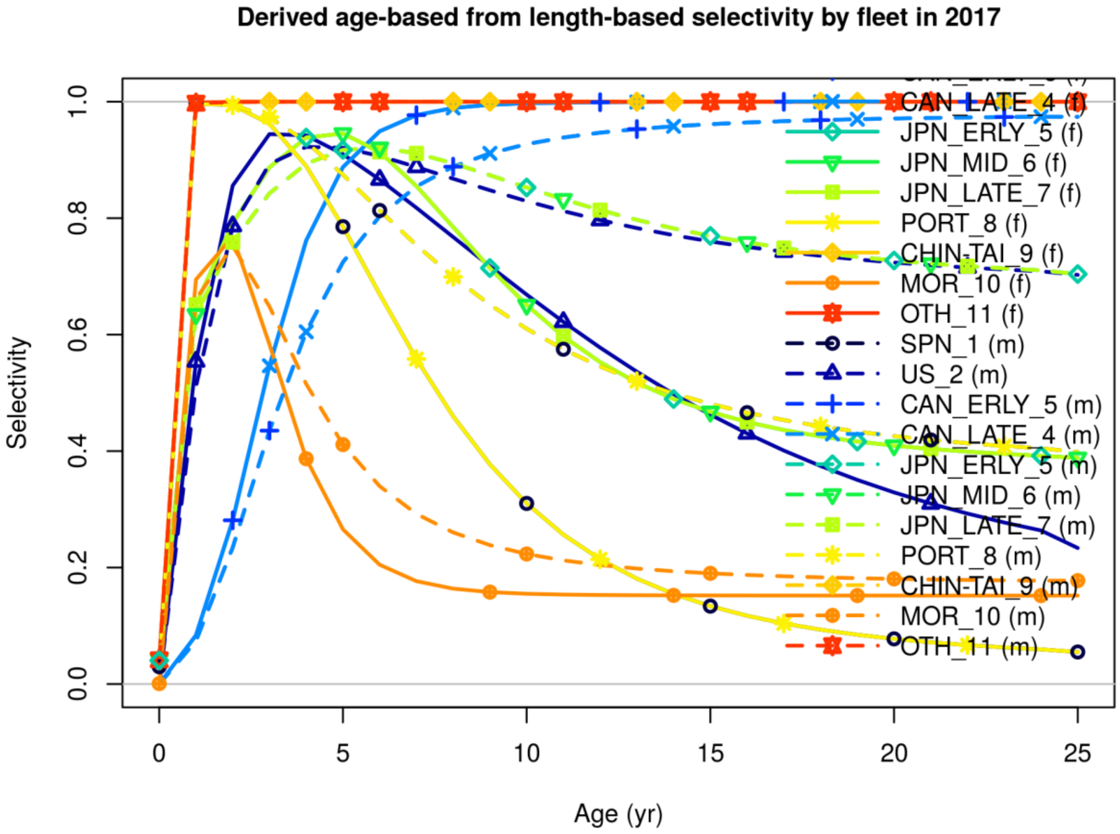


Figure 3a. The selection patterns by fleet as estimated by Stock Synthesis for the Base Case (OM1).

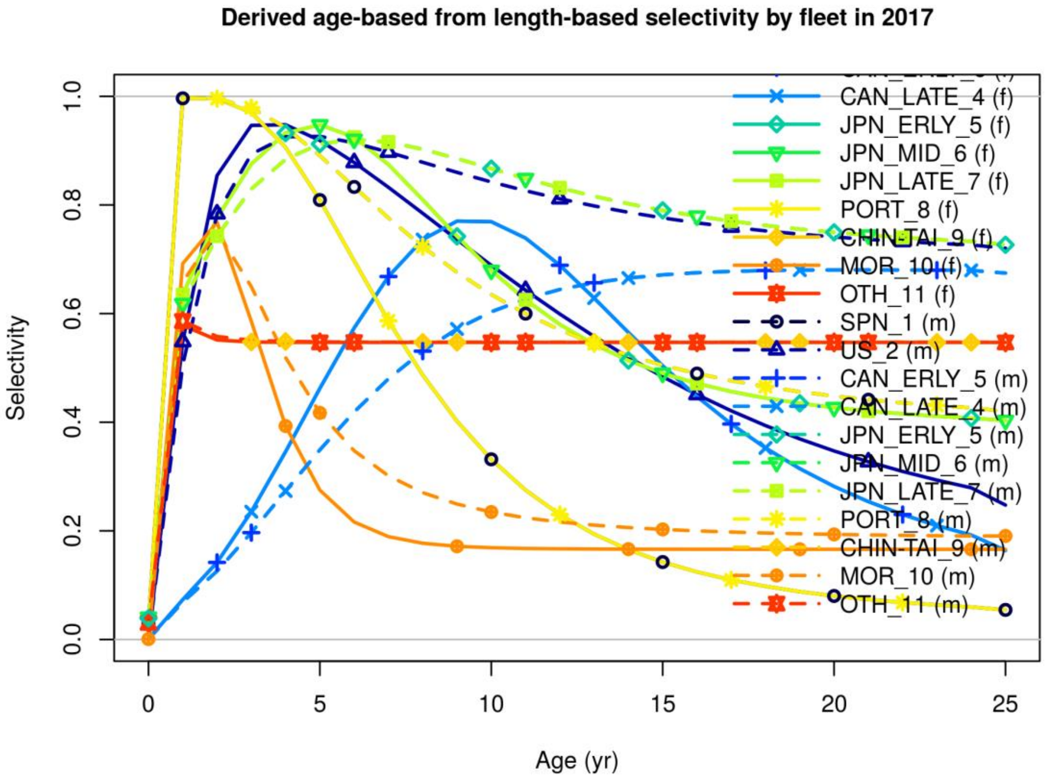


Figure 3b. The selection patterns by fleet as estimated by Stock Synthesis for the scenario with Dome Shaped. selection pattern (OM7).

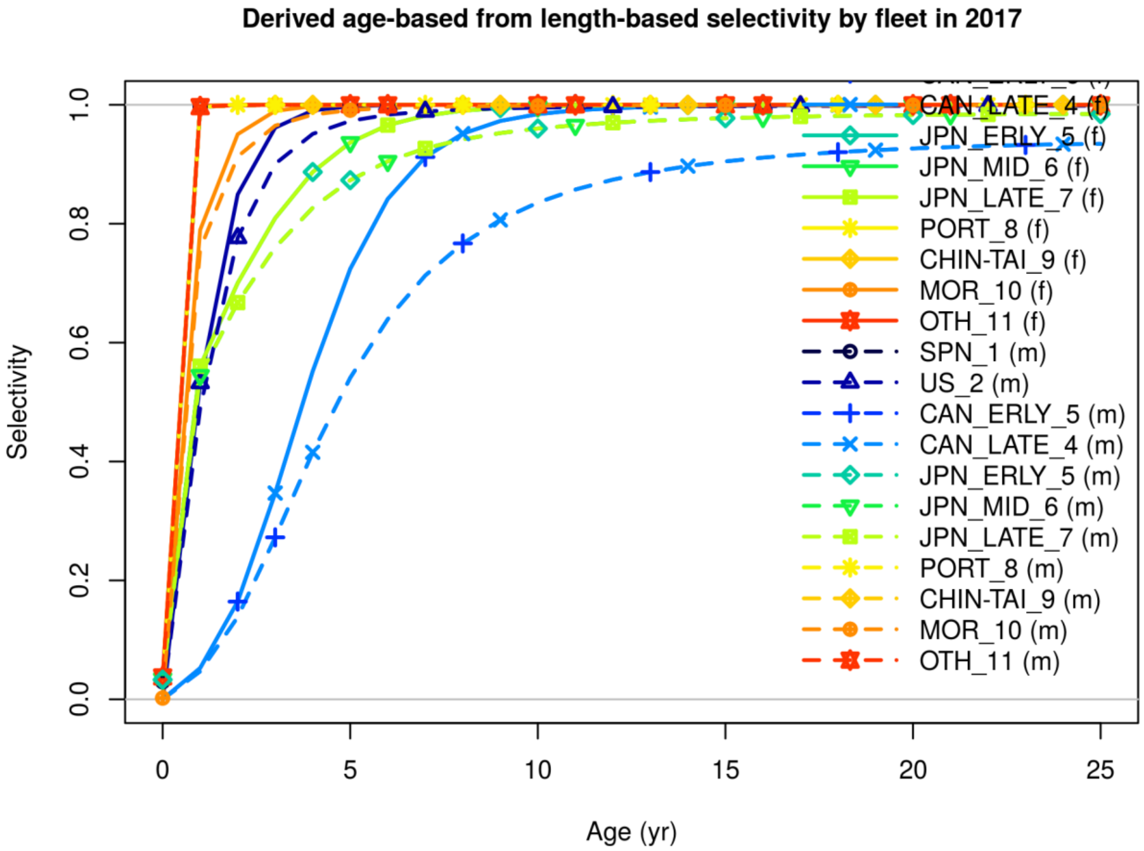


Figure 3c. The selection patterns by fleet as estimated by Stock Synthesis for the scenario with Flat topped selection pattern (OM8).

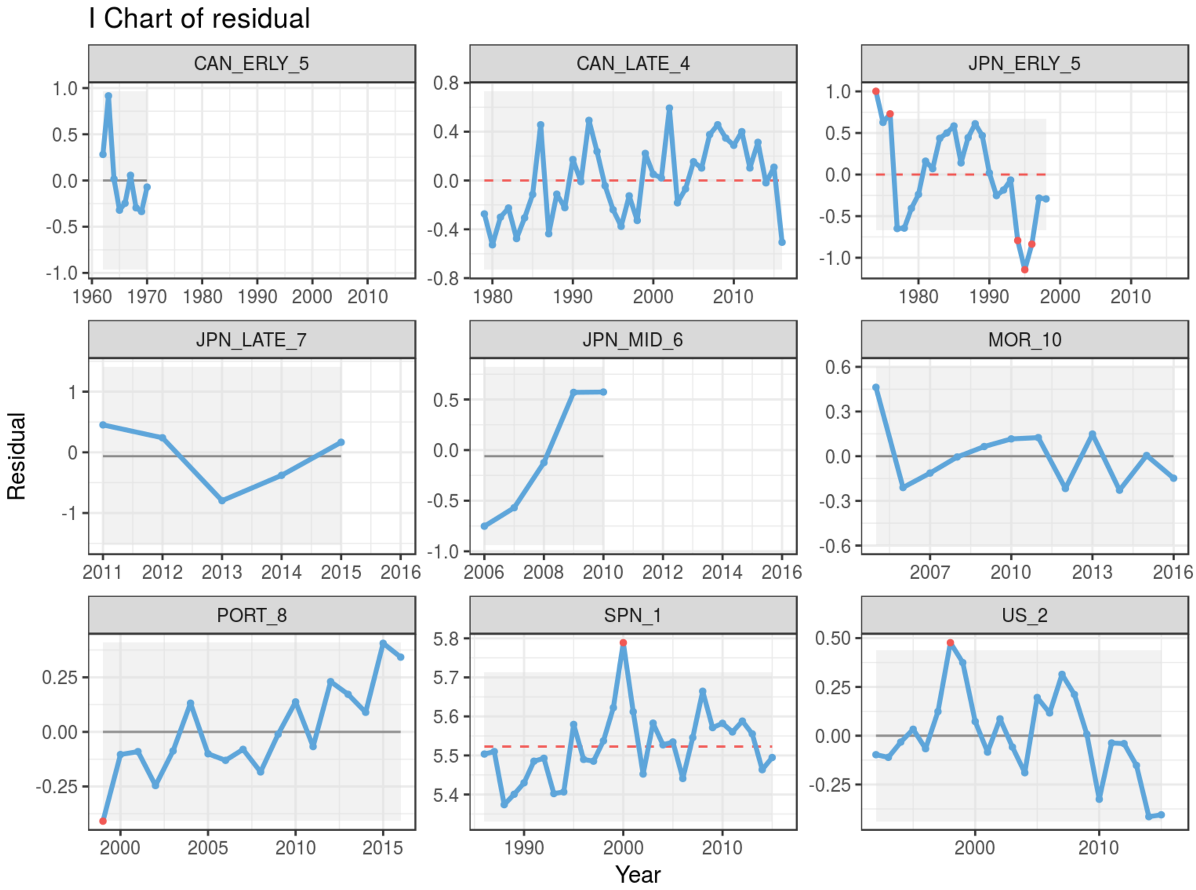


Figure 4. **S**hows run charts for the base case indices, for example although JPN_ERLY_5 has 24 points, the series only crosses the median line three times, and has a run of nine observations above the median. The shift and crossings signals are two sides of the same coin and will often signal together. However, either one is a diagnostic of non-random variation.

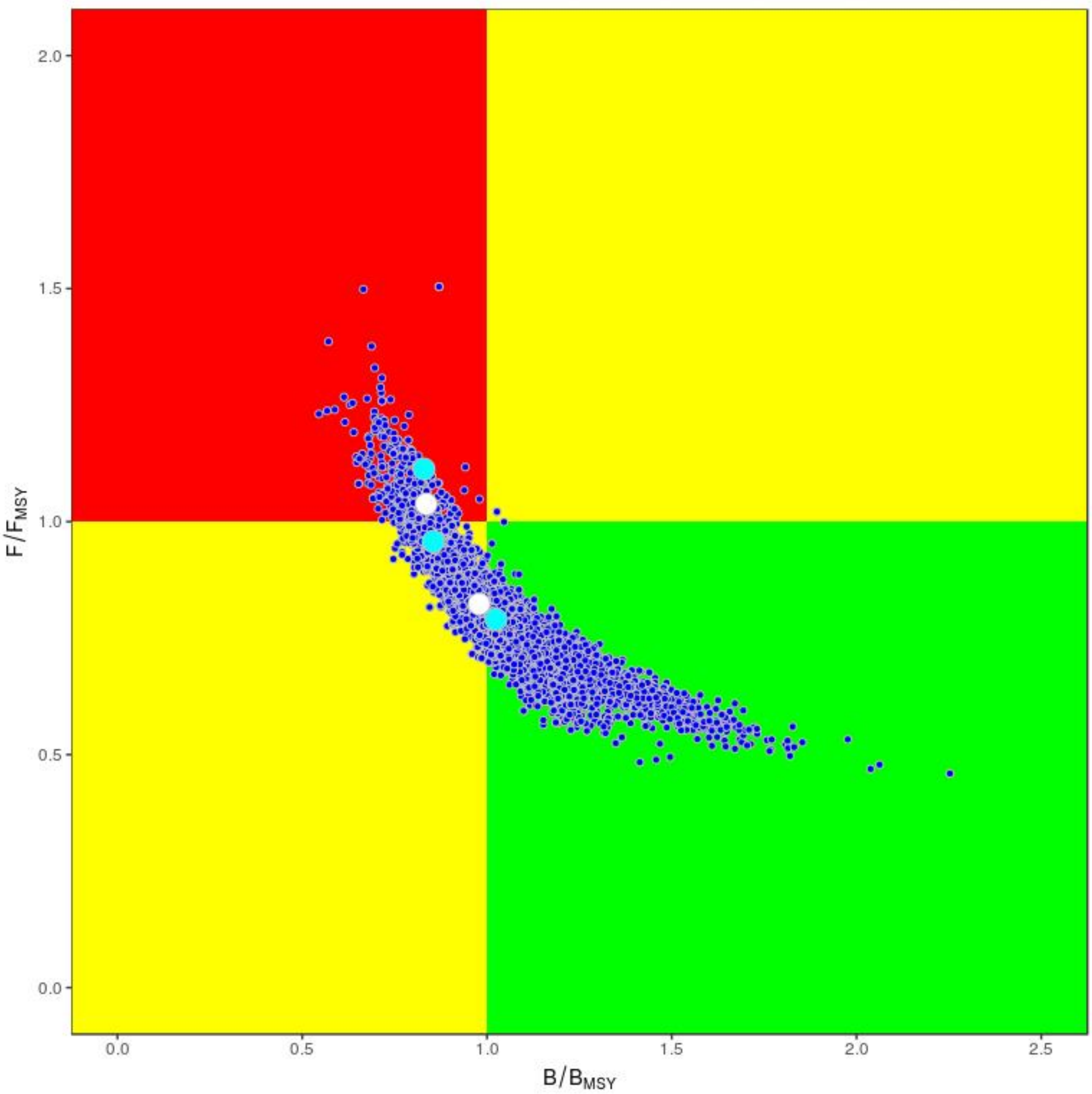


Figure 5. The Kobe phase plot derived from the base case, the small points show the MCMC estimates and the large points fits from different OMs as alternative runs.

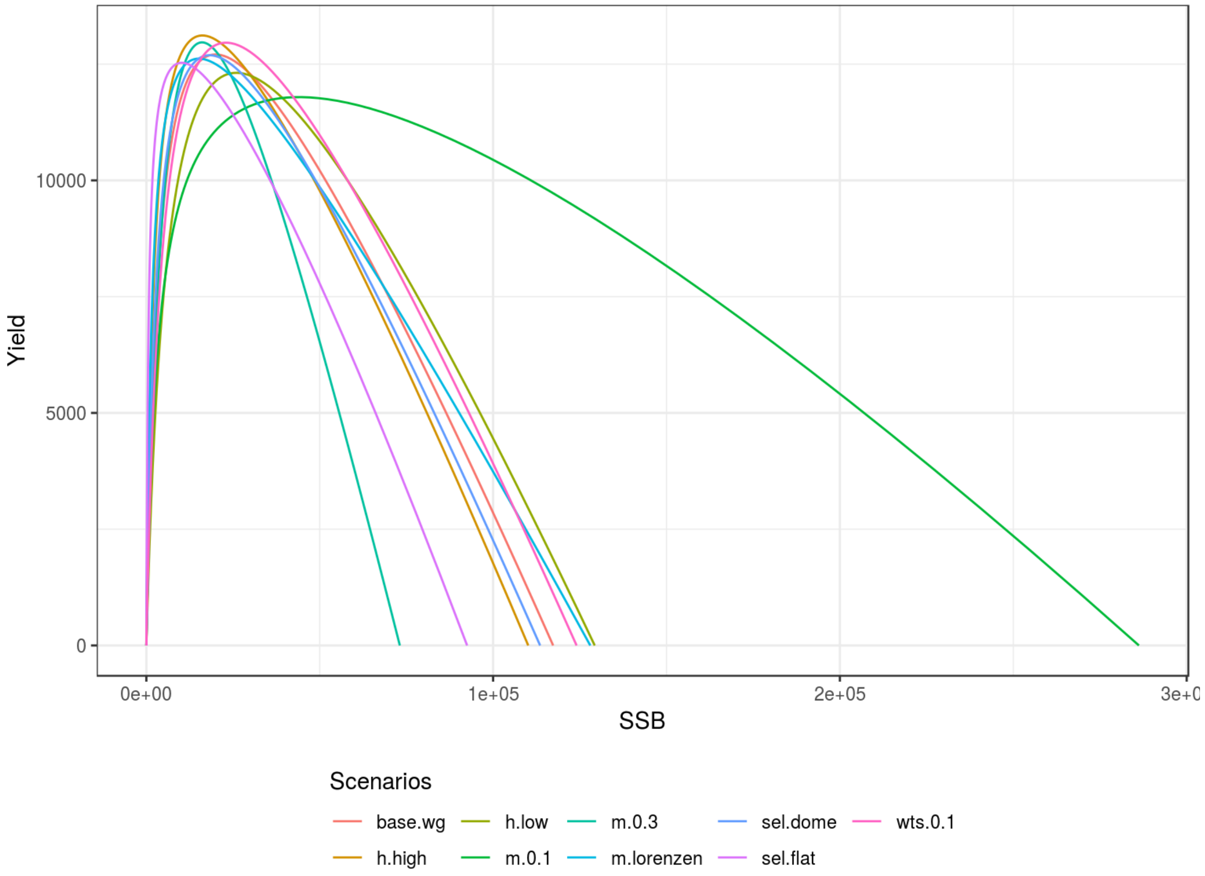


Figure 6. Production Functions.

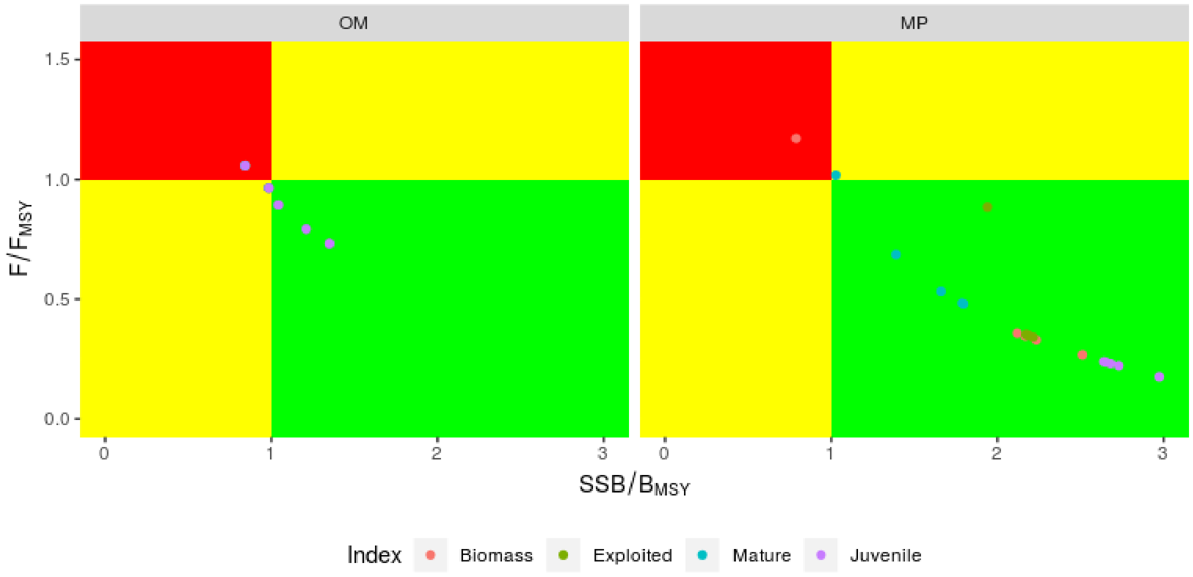


Figure 7. Comparison of current stock status from the OM and MP, taken from simulations run in Kell and Levontin, 2019. 5 OMs were used to simulate data for a cross-test of the biomass dynamic stock assessment used in the Management Procedure. In the cross-test indices were simulated for each OM representing, total, exploited, mature or juvenile biomass.

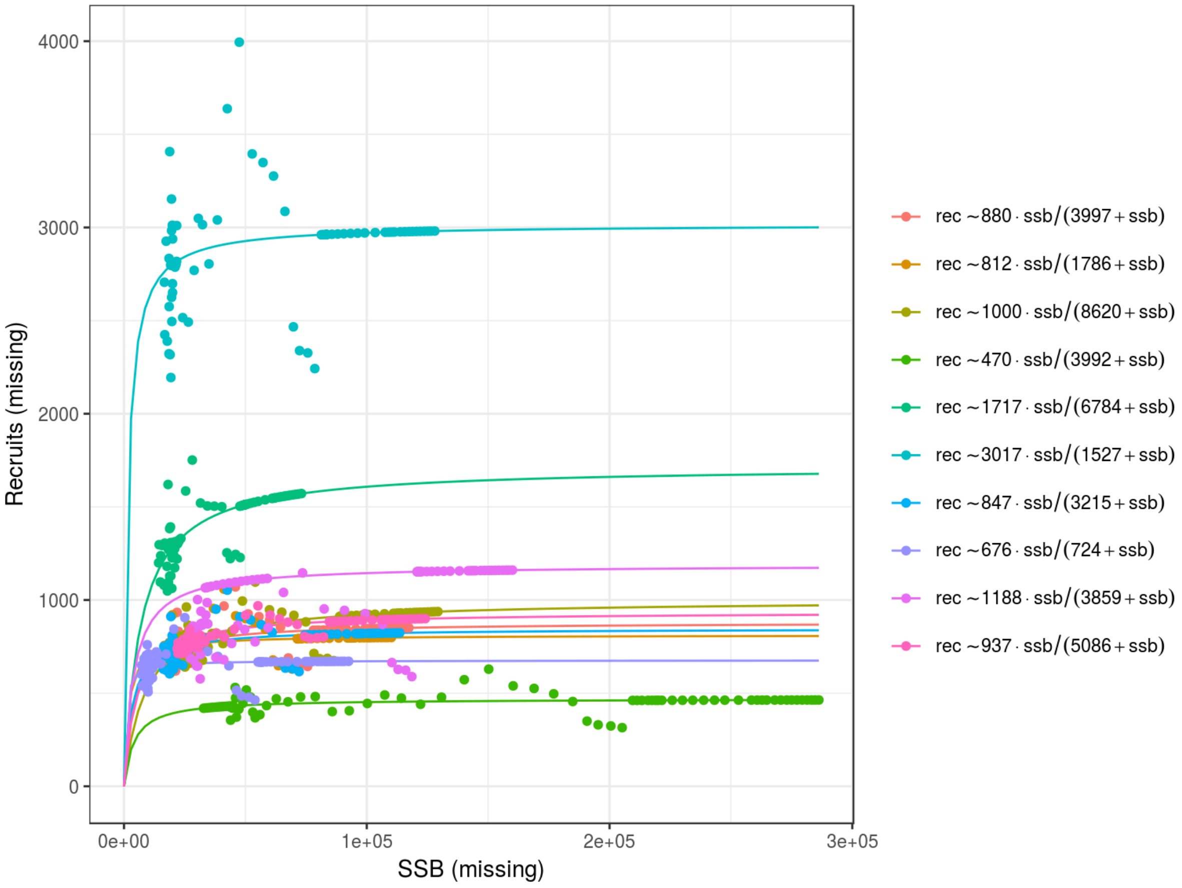


Figure 8. Stock recruitment relationships.

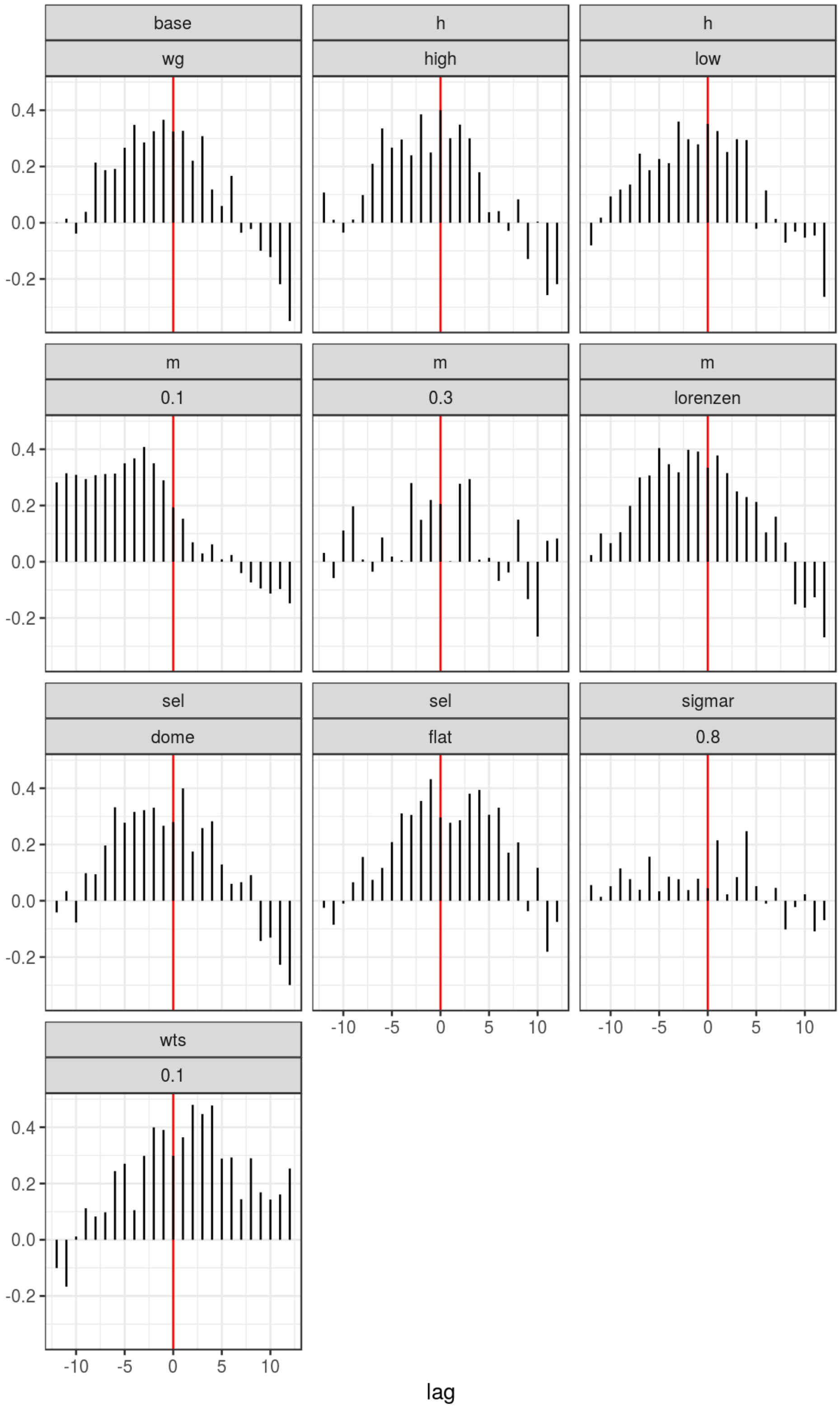


Figure 9. Cross-correlation showing the correlations between recruitment and SSB, the correlation at zero lag represents the influence of SSB on recruitment. Negative lags represent the influence of recruitment 1, 2, 3, years in the past on the current year's SSB.

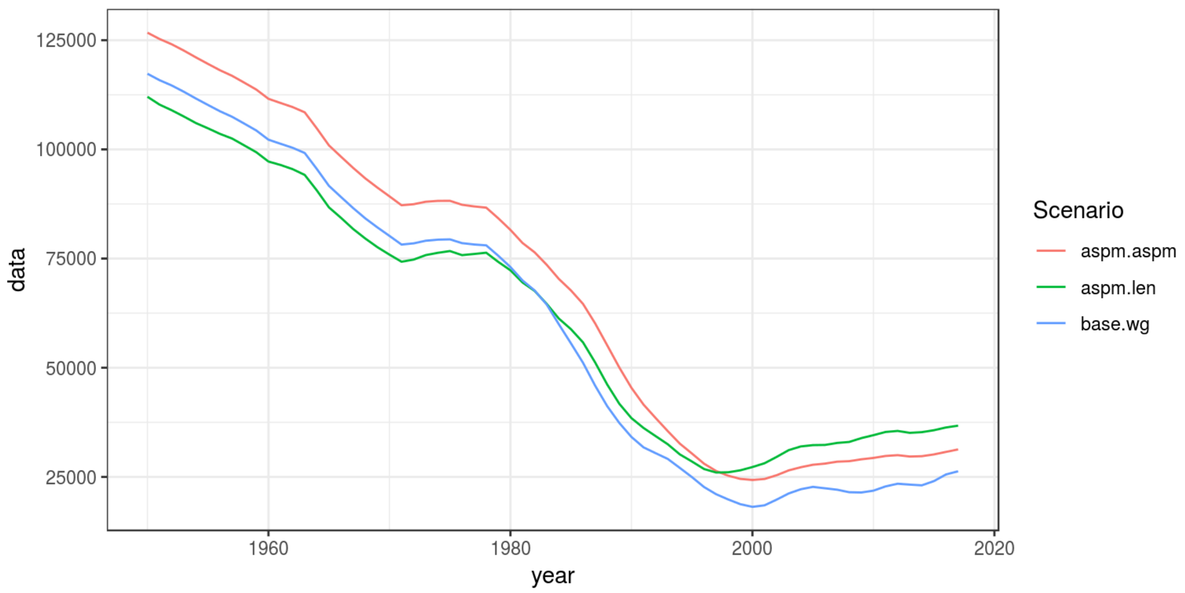


Figure 10. Time series of SSB from the two ASPM runs (aspm.aspm and aspm.len in the latter case length compositions were included) and the base case (base.wg as used by the working group).

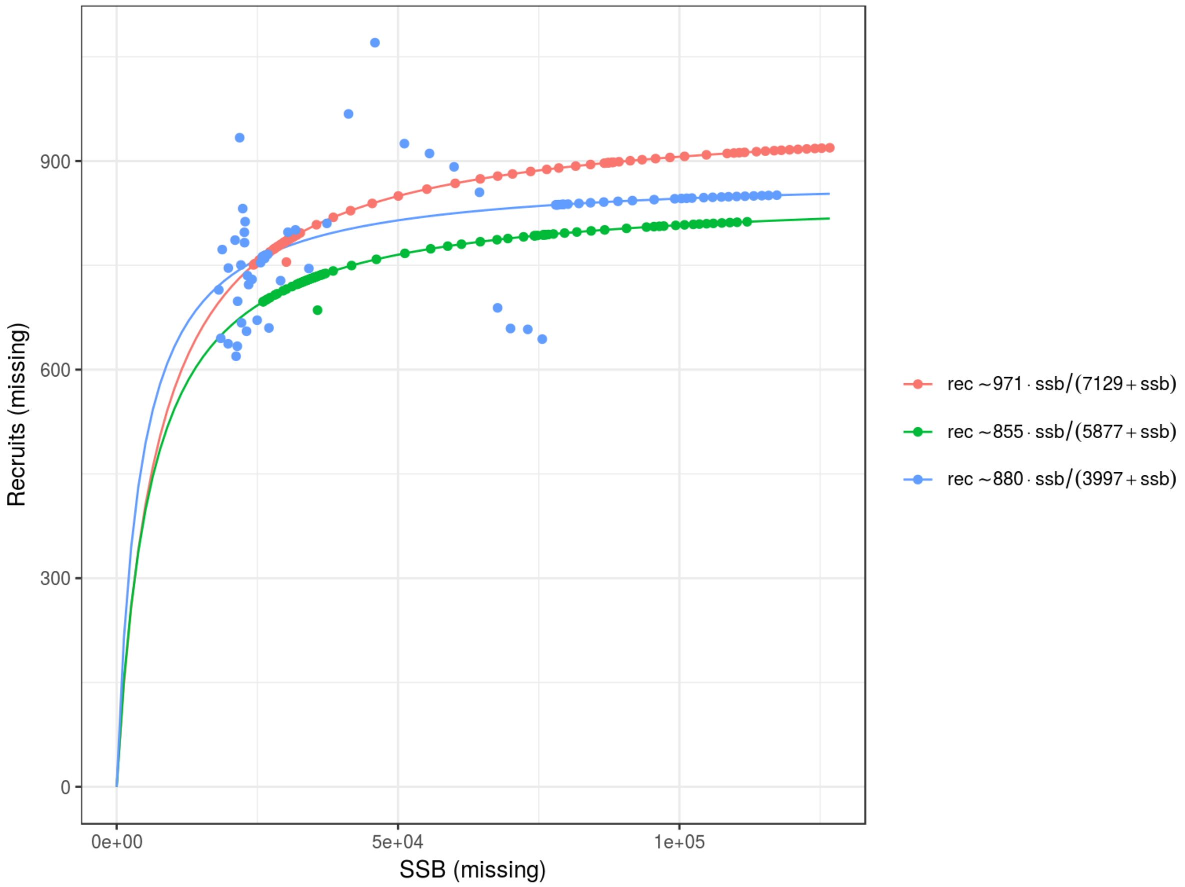


Figure 11. Stock recruitment relationships (red is the ASPM scenario, green the ASPM scenario where length composition data were used and blue is the base case)

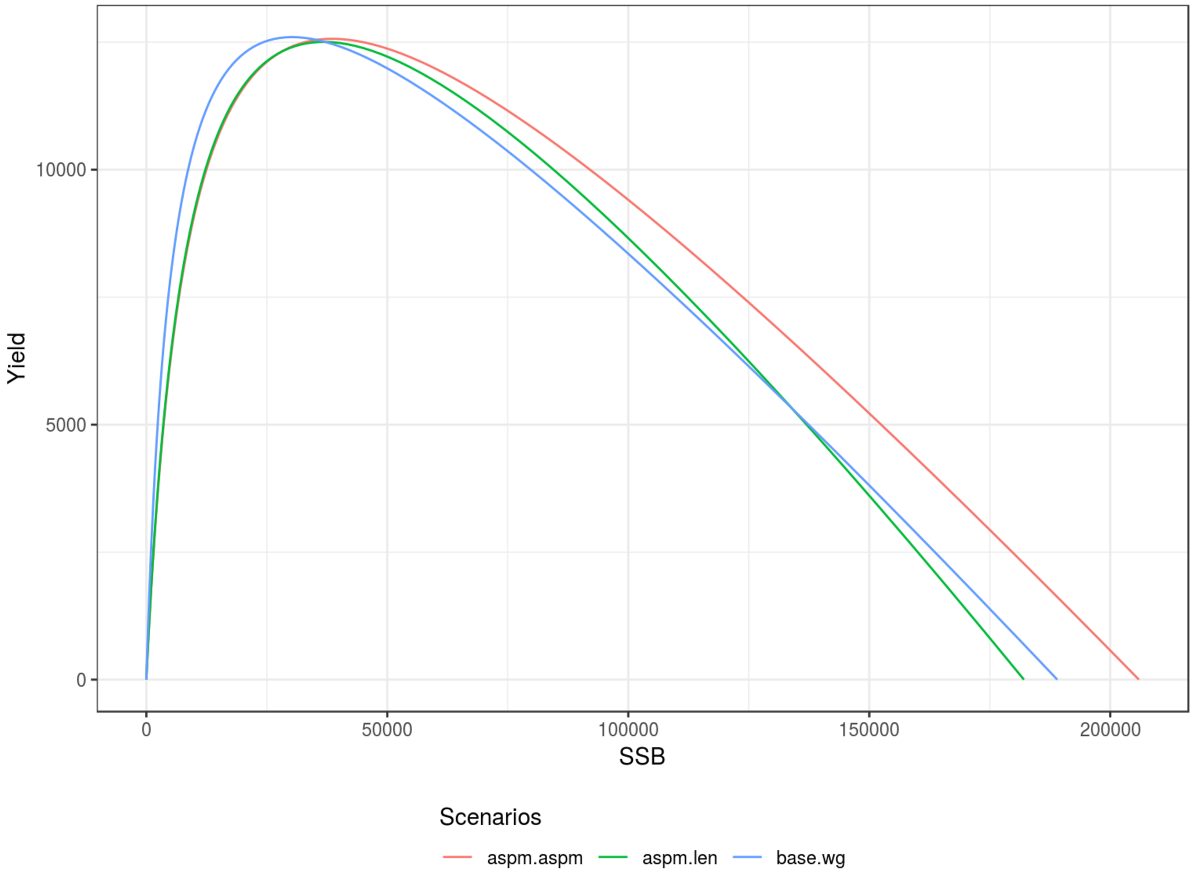


Figure 12. Production functions the two ASPM runs and the base case, (red is the ASPM scenario, green the ASPM scenario where length composition data were used and blue is the base case)

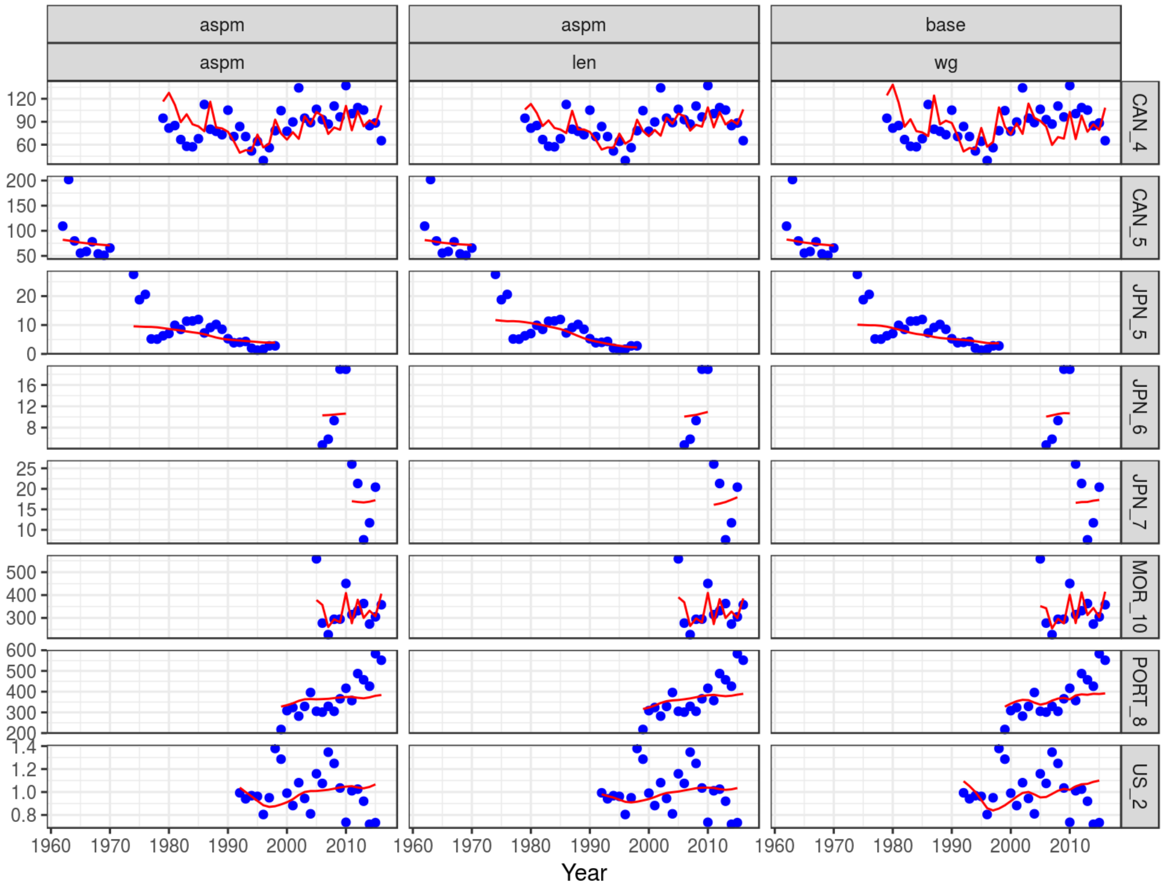


Figure 13. Observed indices of abundance (points) and predicted fits (lines) for two ASPM runs and the base case.

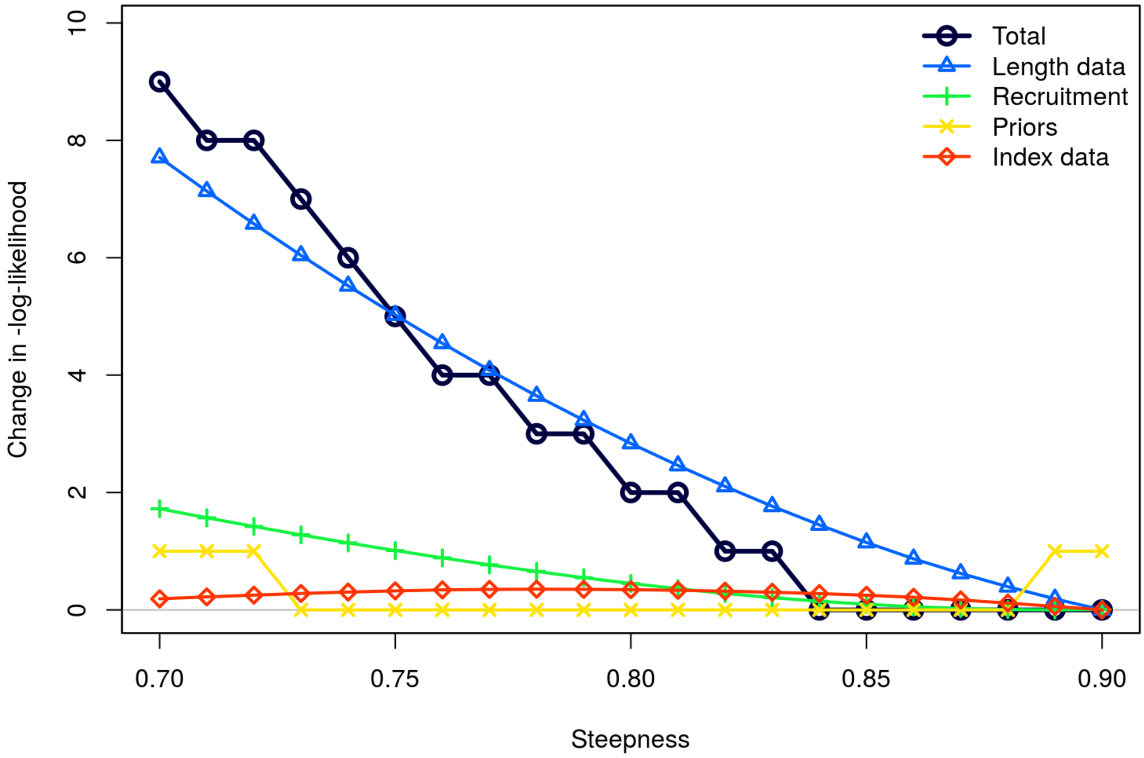


Figure 14. Likelihood profiles for steepness (from the Base Case); lines are the partial likelihoods for the different components at each fixed value of steepness.

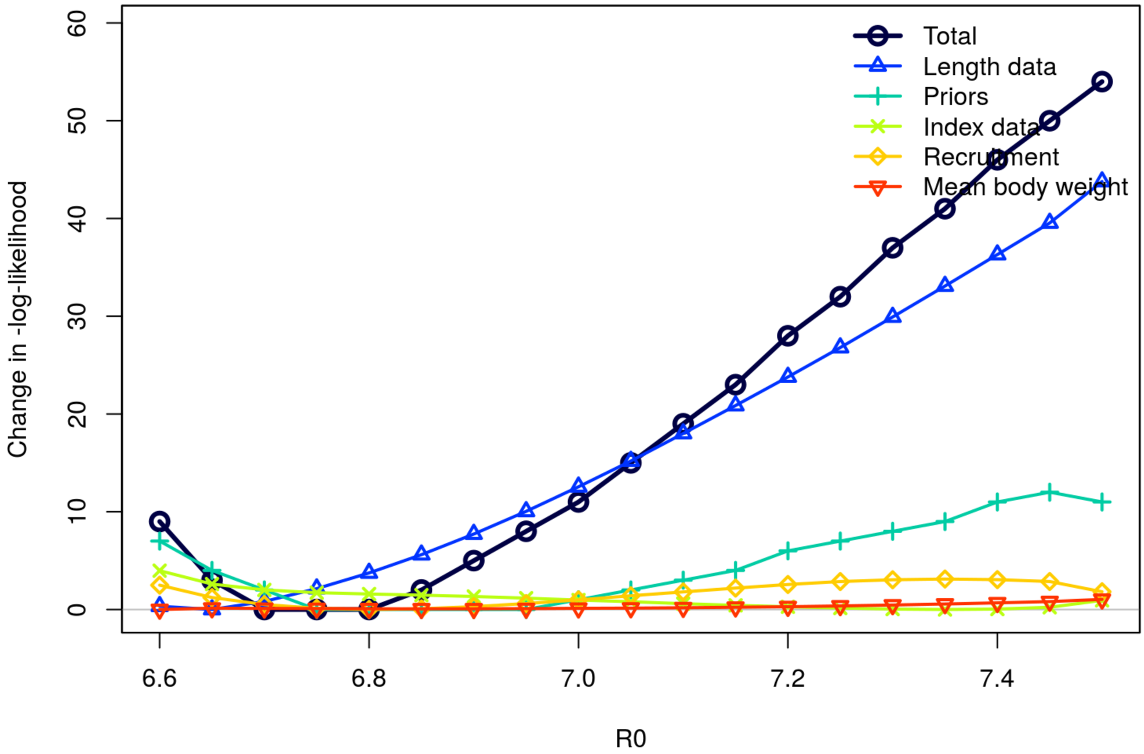


Figure 15. Likelihood profiles for R0 (from the Base Case); lines are the partial likelihoods for the different components at each fixed value of R_0.

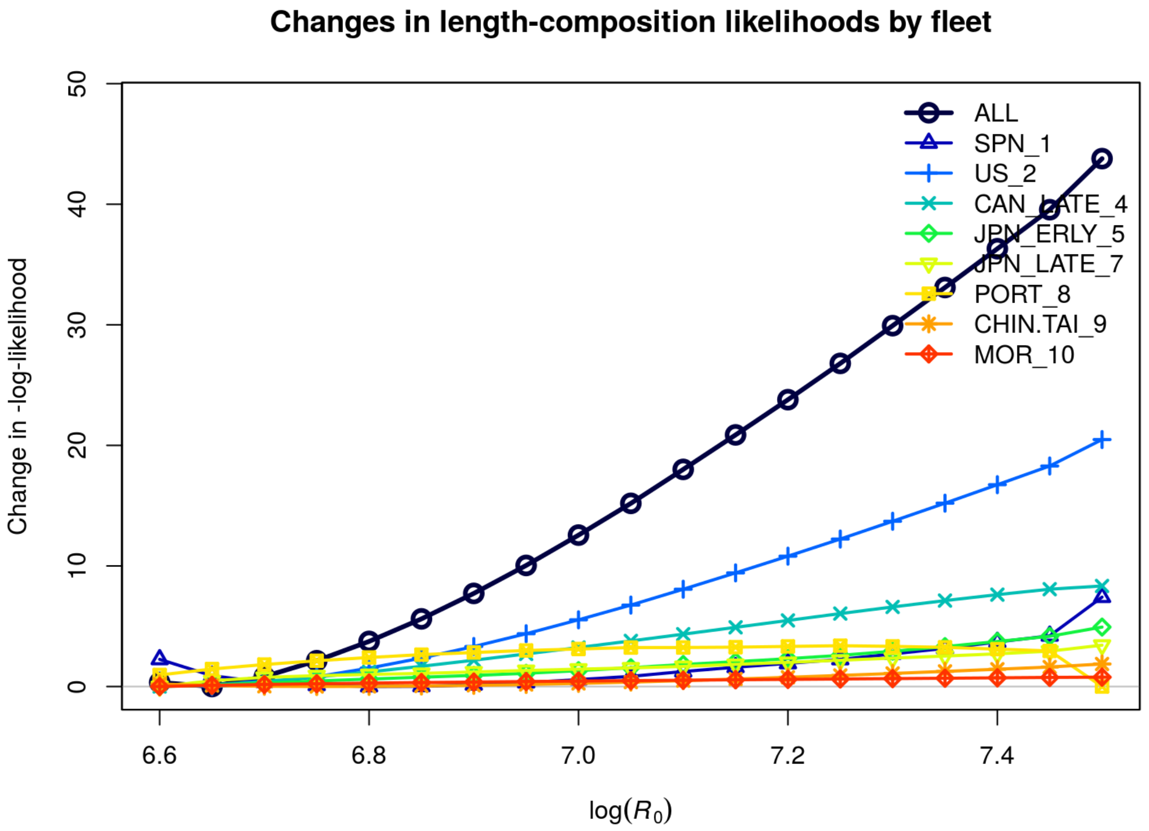
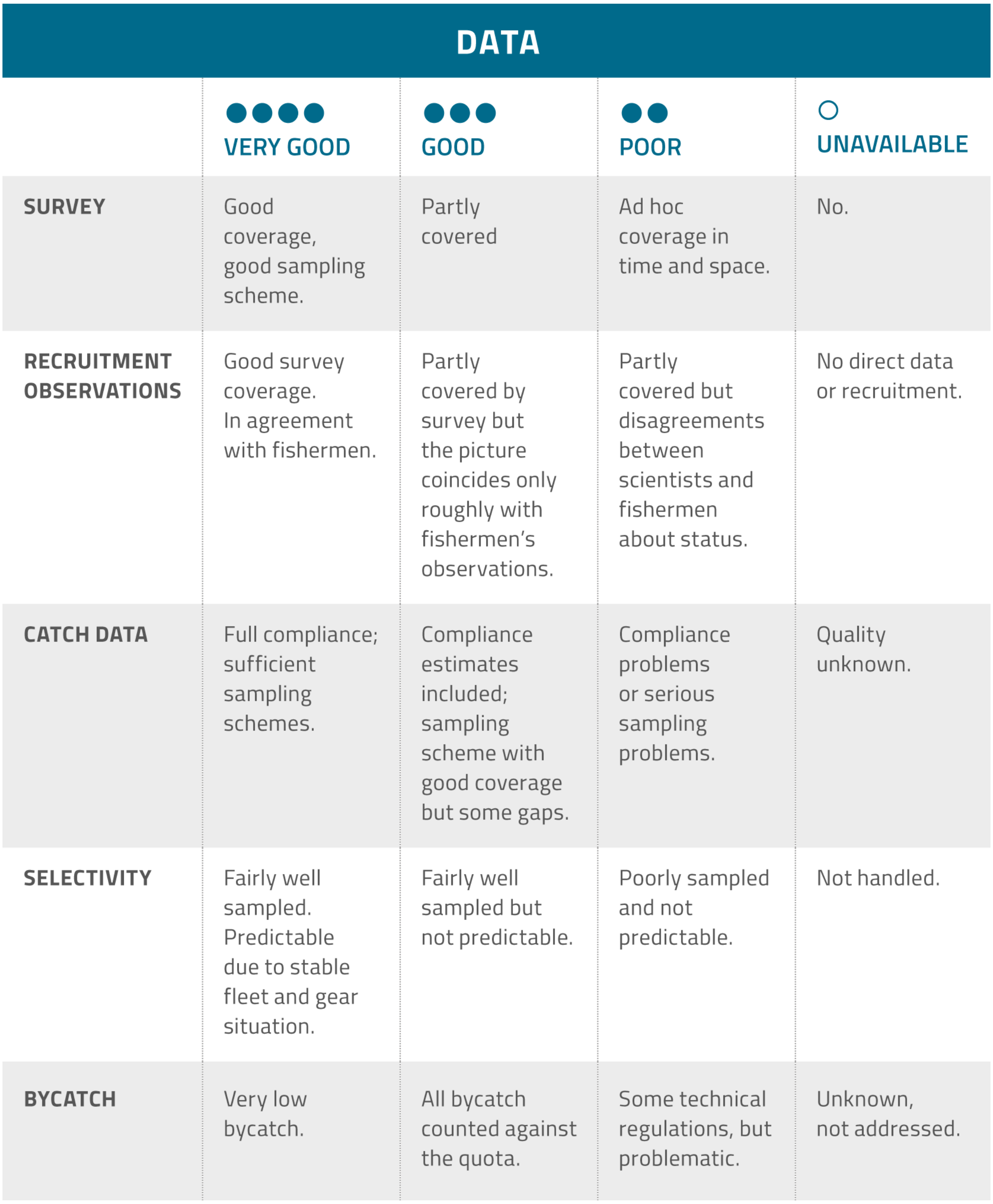
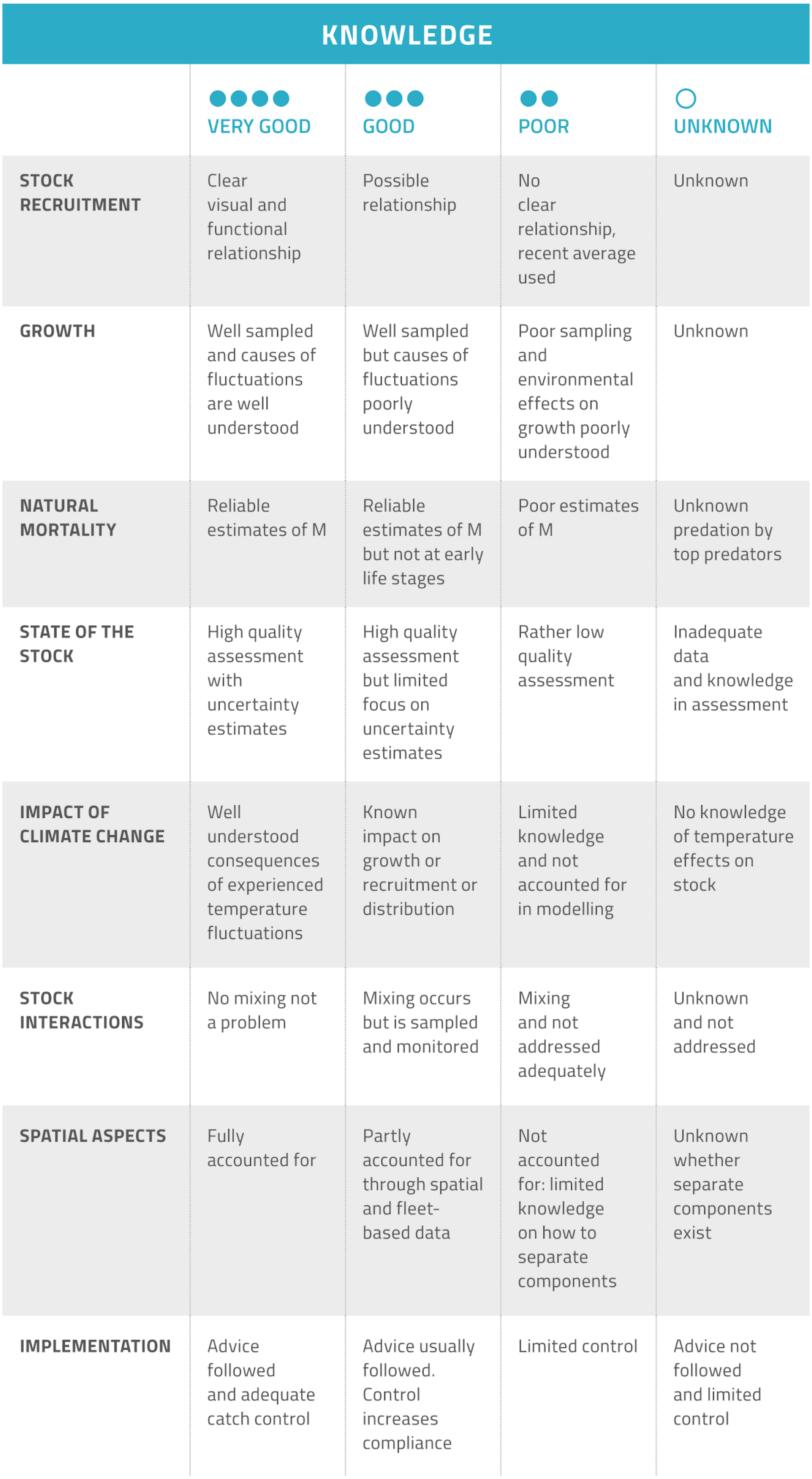


Figure 16. Piner plots showing likelihood component profiles over log(R_0) by fleet component.



DATA				
	●●●● VERY GOOD	●●● GOOD	●● POOR	○ UNAVAILABLE
SURVEY	Good coverage, good sampling scheme.	Partly covered	Ad hoc coverage in time and space.	No.
RECRUITMENT OBSERVATIONS	Good survey coverage. In agreement with fishermen.	Partly covered by survey but the picture coincides only roughly with fishermen's observations.	Partly covered but disagreements between scientists and fishermen about status.	No direct data or recruitment.
CATCH DATA	Full compliance; sufficient sampling schemes.	Compliance estimates included; sampling scheme with good coverage but some gaps.	Compliance problems or serious sampling problems.	Quality unknown.
SELECTIVITY	Fairly well sampled. Predictable due to stable fleet and gear situation.	Fairly well sampled but not predictable.	Poorly sampled and not predictable.	Not handled.
BYCATCH	Very low bycatch.	All bycatch counted against the quota.	Some technical regulations, but problematic.	Unknown, not addressed.

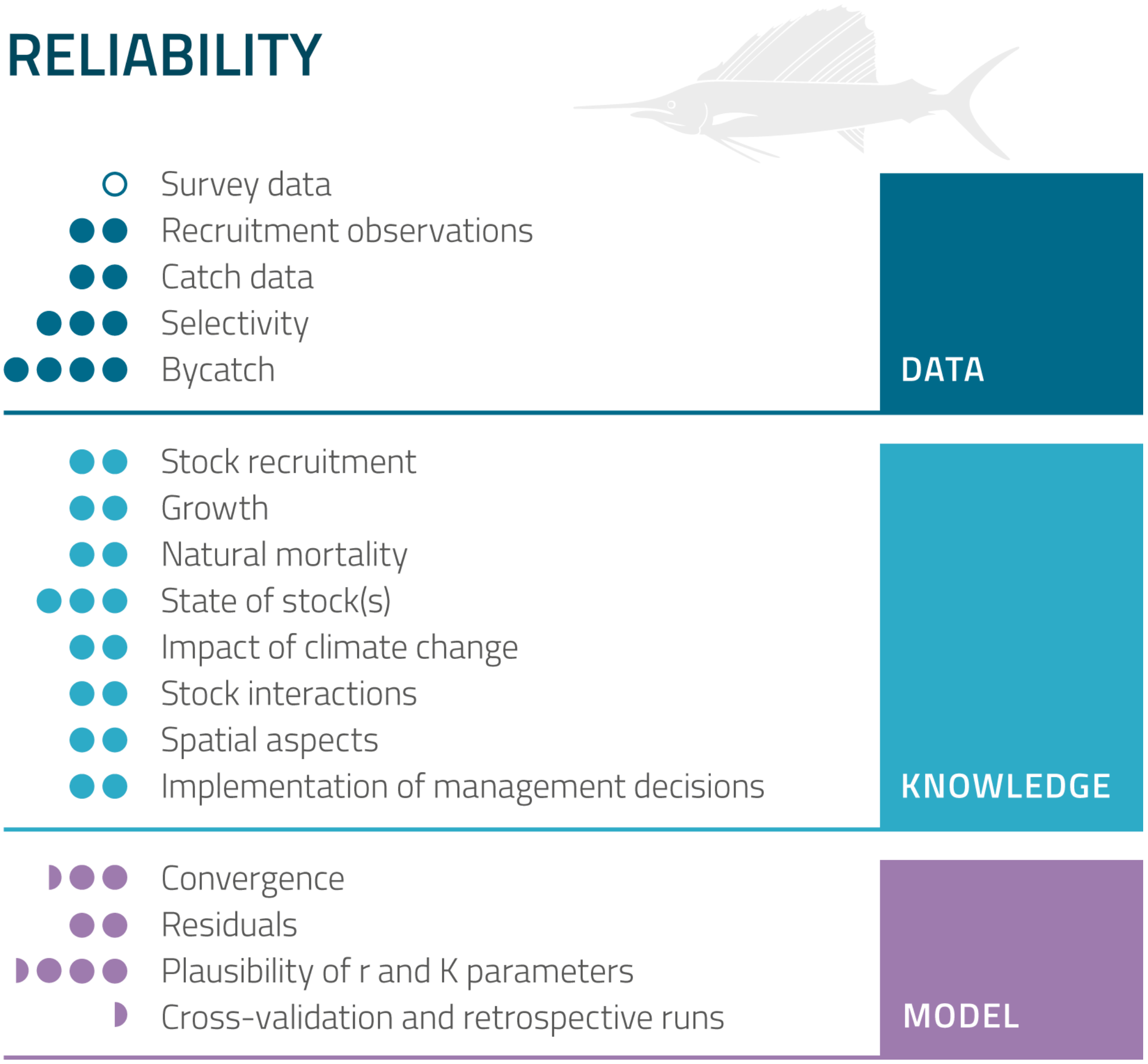
Figure 17. Key to evaluating data inputs, modified from Jakfish.



KNOWLEDGE

	●●●● VERY GOOD	●●● GOOD	●● POOR	○ UNKNOWN
STOCK RECRUITMENT	Clear visual and functional relationship	Possible relationship	No clear relationship, recent average used	Unknown
GROWTH	Well sampled and causes of fluctuations are well understood	Well sampled but causes of fluctuations poorly understood	Poor sampling and environmental effects on growth poorly understood	Unknown
NATURAL MORTALITY	Reliable estimates of M	Reliable estimates of M but not at early life stages	Poor estimates of M	Unknown predation by top predators
STATE OF THE STOCK	High quality assessment with uncertainty estimates	High quality assessment but limited focus on uncertainty estimates	Rather low quality assessment	Inadequate data and knowledge in assessment
IMPACT OF CLIMATE CHANGE	Well understood consequences of experienced temperature fluctuations	Known impact on growth or recruitment or distribution	Limited knowledge and not accounted for in modelling	No knowledge of temperature effects on stock
STOCK INTERACTIONS	No mixing not a problem	Mixing occurs but is sampled and monitored	Mixing and not addressed adequately	Unknown and not addressed
SPATIAL ASPECTS	Fully accounted for	Partly accounted for through spatial and fleet-based data	Not accounted for: limited knowledge on how to separate components	Unknown whether separate components exist
IMPLEMENTATION	Advice followed and adequate catch control	Advice usually followed. Control increases compliance	Limited control	Advice not followed and limited control

Figure 18. Key to evaluating knowledge basis for the Operating Model, modified from Jakfish.



RELIABILITY

○	Survey data	DATA
●●	Recruitment observations	
●●	Catch data	
●●●	Selectivity	
●●●●	Bycatch	
●●	Stock recruitment	KNOWLEDGE
●●	Growth	
●●	Natural mortality	
●●●	State of stock(s)	
●●	Impact of climate change	
●●	Stock interactions	
●●	Spatial aspects	
●●	Implementation of management decisions	
◗●●	Convergence	MODEL
●●	Residuals	
◗●●●	Plausibility of r and K parameters	
◗	Cross-validation and retrospective runs	

Figure 19. Representation of reliability of Operating Model for the MSE, accounting for data and knowledge inputs, as well as for how the models performed on average on validation tests.

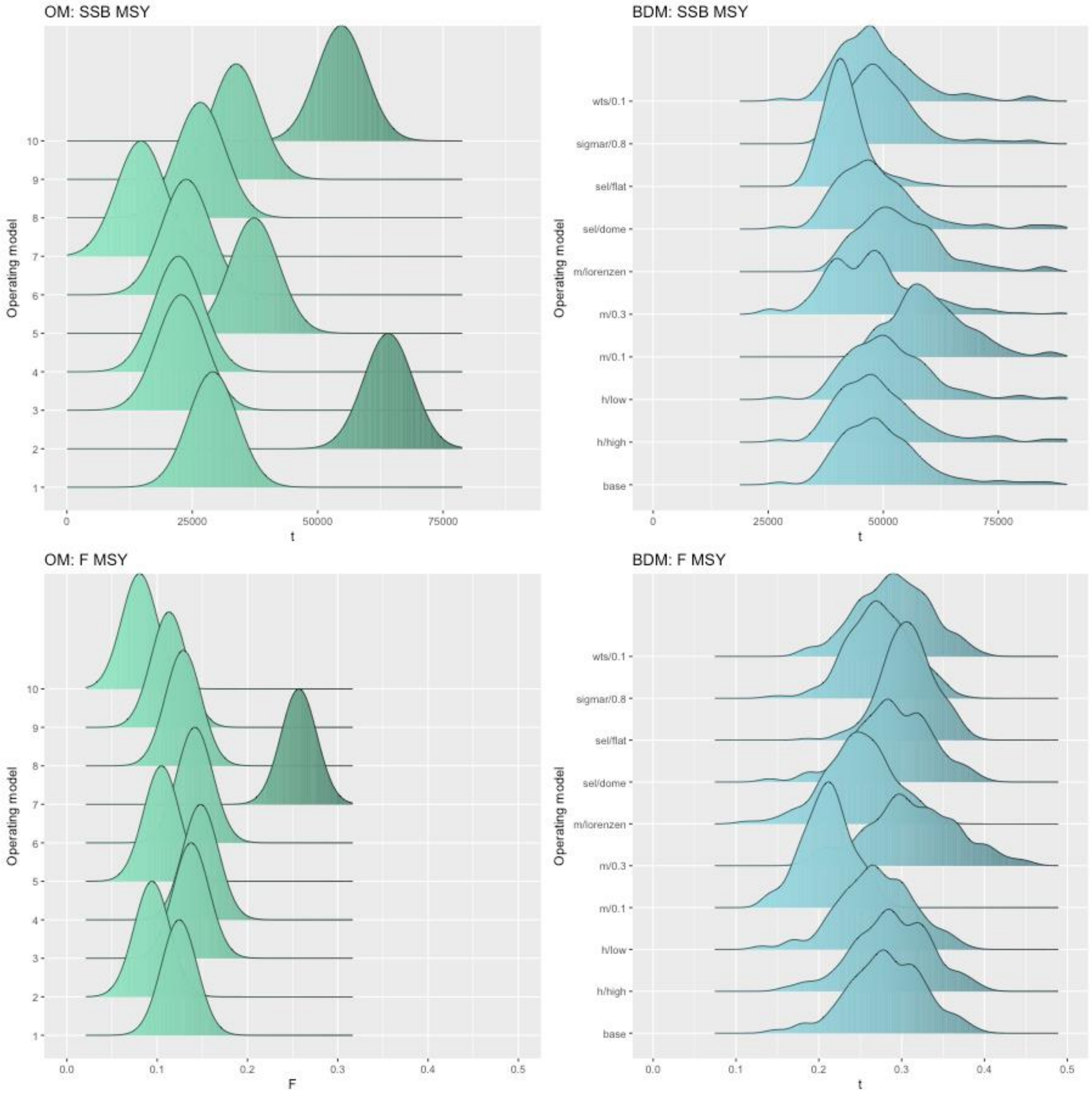


Figure 20. Reference points in Operating Models (left) versus 'stock assessment' Biomass Dynamic Model (BDM - right) that is part of a management procedure in the MSE.