OCEANS AND ATMOSPHERE
www.csiro.au
CSIRO

Update on IOTC Yellowfin Tuna Management Strategy Evaluation Operating Model development Oct 2019

Dale Kolody (dale.kolody@csiro.au)

Paavo Jumppanen

This paper was prepared for the Indian Ocean Tuna Commission Working Party on Methods and Working Party on Tropical Tunas, San Sebastien, Oct-Nov 2019.

COMMON OCEANS
FIAT PANIS
Food and Agriculture Organization of the United Nations
gef

Copyright

Important disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Acknowledgments

This work was jointly funded by the Global Environment Facility – Areas Beyond National Jurisdiction Project (administered through FAO) and CSIRO, Australia. Technical oversight and advice was provided by various IOTC Working Groups, and notably the participants of the IOTC MSE Task Force, including Iago Mosqueira, Toshihide Kitakado, Gerald Scott, Gorka Merino, Hilario Murua, Dan Fu and M. Shiham Adam. Operating model conditioning built upon the stock assessment work of Dan Fu, Adam Langley and others at the IOTC, with useful Stock Synthesis advice from Ian Taylor. The original R-based MSE code from phase 1 was adapted from the Atlantic Bluefin MSE work developed by Tom Carruthers (funded by the ICCAT GBYP project). Summary result graphics were adopted from the Fisheries Library in R (FLR) code provided by Iago Mosqueira.

Contents

1 Summary

This paper summarizes progress on the development of Operating Models (OMs) for IOTC yellowfin (YFT) tuna, highlighting priorities for technical feedback. A short stand-alone summary document describing the most recent reference set Operating Model (OM) is included at attachment 1. This paper focuses on OM developments since the IOTC MSE Task Force meeting in March 2019 (Kolody and Jumppanen 2019a,b). MP evaluation updates for yellowfin and bigeye tunas are described in Kolody and Jumppanen (2019c). The latest version of the MSE software and technical documentation is publicly available from github https://github.com/pjumppanen/niMSE-IO-BET-YFT/. Key developments include:

- The reference set YFT OM (OMref19.4.500) was updated from a reference case Stock Synthesis stock assessment presented to the 2018 WPTT (Fu et al. 2018), and expanded to represent uncertainty in 11 dimensions. The result is considerably more pessimistic than the reference set OM considered by the 2018 WPTT/WPM, however, the model is also more stable, and did not require the “bi-variate grid sampling” approach used to select a plausible model ensemble in 2018.

- The reference set OM proposed at this time was more comprehensively investigated than the OM which was used for the 2019 TCMP results and corrected a couple minor specification errors, but resulted in very similar MP evaluation performance.

- OM development requests from the various IOTC working groups have led to a potentially unwieldy number of yellowfin model specifications (4608, if all assumption interactions are evaluated). The computational burden is further exaggerated by numerical instability and sensitivity to initial parameter values. A few alternative approaches for grid specification were explored, which increase confidence that a relatively small number of models (e.g. 50-100) can probably provide similar MP evaluation advice to the full grid. A subset of 49 models (filtered from a main-effects fractional factorial design of 144) was able to produce very similar results to a much larger ensemble (2-way interaction fractional factorial design of 1152, filtered to 420). Similarly, the large grid was subject to a jitter analysis, which demonstrated substantial variability among many individual “converged” models. However, comparison of MP performance based on OM ensembles comprised of the best and worst fit of the converged models (i.e. selected from within the 3 replicates of each configuration) were very similar.

- The majority (~2/3) of the yellowfin models in the ensemble grids explored were rejected with numerical problems, identified by a relatively large SS3 catch penalty. We interpret this to mean that these models struggle to remove the observed catch and presumably explains much of the retrospective patterns that were noted in the 2018 WPTT. This is an issue in both the assessment and OMs, and suggests that a substantial rethinking of assumptions might be warranted. The yellowfin OM will need to take consider the findings

of the international yellowfin assessment review project. As part of this broader process, we explored a few new OM configurations and observed that:

  - The model did not provide compelling evidence for strong non-linearity in the LL CPUE – abundance relationship post-1972.

  - If the LL CPUE are aggregated into a single (regionally-weighted) index, the LL vulnerable biomass appears relatively stable for the past decade. When this is fit in a spatially-aggregated model (mimicked by forcing high movement rates), the result is somewhat more optimistic than the current assessment. However, it is unclear the extent to which this is primarily due to the simplification of regional processes or the removal of the tags (poor tag mixing renders them inappropriate for very large spatial regions).

  - If one assumes that i) northwest LL CPUE is reliable from 1972-2006 (pre- Somalian piracy period), ii) LL CPUE is reliable from 1972-2017 in the other regions, and iii) (standardized) PSFS catchability has increased at a continuous rate from 1986-2017, then a PSFS catchability increase of ~ 0.71% per year appears to largely reconcile the LL and PSFS CPUE series. When the PSFS CPUE series is included in this manner, the recent depletion level is estimated more optimistically than the assessment, but the productivity (MSY) is more pessimistic. At this time, we would be hesitant to include the PSFS series in this manner, because it requires an unexplained continuous catchability increase assumption, and the catch size composition of this fleet suggests that two modes of operation have been pooled (the rapid change between small and large fish observed in the quarterly catch statistics does not appear to be consistent with estimated recruitment patterns).

  - We briefly explored the Ricker stock recruit function, (requested as an OM robustness test). We would question whether it is worth pursuing further, since the estimates qualitatively resembled Beverton-Holt functions (possibly with steepness lower than was considered plausible for this population).

- We encourage further feedback from the WPTT and WPM on all elements of the Operating Model, especially the approach used to create the reference set ensemble, methods for dealing with numerical instability, process for evaluating plausibility (especially in an automated context), and specific requests for reference set and robustness tests. Some specific options are proposed for discussion, but it is recognized that the YFT review process may fundamentally shift the direction of the OM.

2 Introduction

The Indian Ocean Tuna Commission has committed to a path of using Management Strategy Evaluation (MSE) to meet its obligations for adopting the precautionary approach. IOTC Resolution 12/01 *"On the implementation of the precautionary approach"* identifies the need for fishery reference points and harvest strategies that will help to maintain the stock status at a level that is consistent with the reference points. Resolution 13/10 "*On interim target and limit reference points and a decision framework*" identified interim reference points and elaborated on the need to formulate management measures relative to the reference points, using MSE to evaluate harvest strategies in recognition of the various sources of uncertainty in the system. Resolution 15/10 supersedes 13/10 with a renewed mandate for the Scientific Committee to evaluate the performance of harvest control rules with respect to the species-specific interim target and limit reference points, no later than 10 years following the adoption of the reference points, for consideration of the Commission and their eventual adoption. A species-specific workplan was re-affirmed at the 2017 Commission Meeting, outlining the steps required to adopt simulation-tested Management Procedures for the highest priority species (IOTC 2017). Recognizing the iterative nature of the MSE process, 2021 is recognized as the earliest possible date for MP adoption. A draft yellowfin MP resolution (IOTC-2019-S23-PropP) was circulated to the 2019 TCMP and Commission for consideration and further development.

This paper:

1. attempts to quantify the implications of many operating model conditioning assumptions (in terms of the quality of fit to the data, stock status implications and assumption interactions)
2. explores the problem of OM complexity due to high dimensionality of assumption interactions and numerical instability of the Stock Synthesis (SS) conditioning models (Methot and Wetzel 2013).
3. explores some new structural assumptions (including robustness test requests) which should also be relevant to the broader YFT assessment review process
4. identifies the highest priority issues for feedback from the WPTT/WPM 2019.

This paper assumes familiarity with technical subject matter. More detailed explanations can be found in Kolody and Jumppanen (2016), Jumppanen and Kolody (2018) and various progress reports produced since the last YFT MSE update to the WPTT and WPM (Kolody and Jumppanen 2019b, c, d, e). MP evaluation updates for yellowfin and bigeye tunas are described in Kolody and Jumppanen (2019c).

2.1 Yellowfin MSE Requests from the 2018 WPTT/WPM

Requests for the yellowfin MSE development from the 2018 WPTT and WPM meetings were mostly addressed at the IOTC MSE Task Force meeting Mar 2019 (WPM 2019), as described in Kolody and

Jumppanen (2019b) and the following section. A few issues, notably related to robustness tests, were delayed until the current report (because it was agreed to not report robustness tests to the TCMP).

WPM (2018) requested the following yellowfin robustness tests:

> *1) Annual aggregated CPUE CV = 0.3 (auto-correlation = 0.5) (projections only) [High priority]*
>
> *2) 10% reported over-catch (projections only) [High priority]*
>
> *3) 10% unreported over-catch (projections only) [High priority]*
>
> *4) 2%, 3% LL catchability trend (projections only) [High priority]*
>
> *5) dome-shaped longline selectivity (noting potential for interaction with M and growth) (conditioning and projections) [Low priority]*
>
> *6) Recruitment shock (projections only) [High priority]*
>
> *7) Ricker recruitment (conditioning and projections) [Low priority]*

Items 1-4 and 6 are addressed in Kolody and Jumppanen (2019c) with the OMs defined in Table 1. Item 5 (dome-shaped selectivity) was elevated to the reference set OM as an additional dimension.

There is some question of the value of item 4 (2-3% per year CPUE catchability trend scenarios). Unless there is some specific insight to the fishery, presumably it would be more realistic to include the catchability trend in the historical conditioning as well. But these are very pessimistic scenarios that would require a very conservative MP to avoid adverse outcomes. If the IOTC community genuinely considers this to be a plausible scenario, then the IOTC probably needs to urgently rethink its reliance on longline CPUE data in the stock assessment and MP.

Ricker stock-recruit function (Item 7) is explored in section 6.

WPTT (2018) requested the following robustness tests:

> *241. The WPTT **NOTED** the need to modify the assumed time required to achieve mixing of tagged YFT with the untagged population to 4 quarters (from 3 quarters) based on decisions taken for the 2018 YFT stock assessment. Further, the WPTT **ENCOURAGED** that the MSE work consider the importance of also assuming the time needed for mixing of the tagged and untagged populations of 8 quarters for use in examining robustness of MPs to this assumption.*
>
> *242. The WPTT **ENCOURAGED** that the MSE work consider the importance of alternative growth for yellowfin tuna based on the growth model estimated by Dortel et al. (2014) for use in examining robustness of yellowfin MPs to alternative growth models.*
>
> *243. The WPTT further **ENCOURAGED** that the MSE work also consider the importance of adding the Purse Seine Free School CPUE as documented in IOTC–2018–WPTT20–36_Rev1, assuming a 1% per year cumulative increase in catchability (q) for the time period, for use in examining robustness of yellowfin MPs.*
>
> *244. The WPTT also **NOTED** that the decisions taken for the 2018 YFT assessment regarding short-term and chronic tag loss differed from the YFT Operating Model grid and **REQUESTED** that the 2018 YFT assessment assumptions be mimicked in the Operating Model grid.*

Items identified in paras 241 and 244 were addressed by adopting the new reference case assessment as the base for the reference set OM. We note that the request to increase the tag mixing period from 3 to 4 quarters was based on interim assessment results at the WPTT, in which the 3 quarter mixing model had converged to a local minimum, and the problem was subsequently resolved without the extended mixing period.

The alternative growth curve (para 242) was adopted into the reference set OM with equal probability to the reference case growth assumption (because there was no reason identified as to why the analysis was likely to be less valid than the original growth assumption).

2.2 Yellowfin MSE Requests from the 2018 WPTT/WPM and 2019 IOTC MSE Task Force

Requests for the yellowfin MSE development from the 2018 WPTT and WPM meetings were mostly addressed in Kolody and Jumppanen (2019b) and reviewed at the IOTC Task Force meeting March 2019 (meeting report expected to be tabled as an IOTC MWG information paper). The main exception was some of the robustness tests which are discussed in section 5. The 2019 MSE Task Force made the following recommendations for the next iteration:

- *The proposal for the reference case OM for the 2019 TCMP is aiming for between 72 and 144 models (depending on the fractional design adopted)*
    - *3 X steepness: h = 0.7, 0.8, 0.9*
    - *3 X M*
    - *2 X tag weight λ = 0.001, 1.0*
    - *2 X growth options (original and Dortel et al 2015)*
    - *2 X LL CPUE catchability trend 0, 1% per year*
    - *2 X tropical CPUE standardization method: HBF, cluster analysis*
    - *2 X LL CPUE CV: 0.1, 0.3*
    - *2 X regional scaling factors*
    - *2 X CL assumed sample sizes: ESS=10, (1 iteration of post-fit reweighting)^0.75*
    - *2 X LL selectivity function: logistic, double normal*
    - *2 X tag mixing period: 4, 8 quarters*
- *It was noted that the double normal selectivity function was requested as a robustness test for YFT, but was proposed for the reference set here to be consistent with ALB and SWO (if results are plausible).*
- *It was suggested that catchability trend estimations were done after the piracy period. This estimated trend will be used for the projections.*

These items were addressed in the OM presented to the 2019 TCMP (Kolody and Jumppanen 2019e) with the following exceptions:

- The original CL assumed sample sizes ESS=5 was retained instead of ESS=10 (this was a mistake in the MSE Task Force report)
- The post-piracy catchability trend estimation was not included. This was originally an oversight. However, the request was subsequently deferred, because the intent of the request is not clear, and there is a substantial international collaboration currently reviewing many elements of the yellowfin assessment, including CPUE series interpretation (which is also examined in this report).

2.3 Yellowfin MSE Requests from the 2019 TCMP and Commission meetings

Several requests were made at the 2019 TCMP and Commission meetings, but are all MP-related and are addressed in the companion paper Kolody and Jumppanen (2019c).

3 MSE software developments

The latest version of the MSE software is publicly available from github, along with all of the project reports and a technical description and user manual (https://github.com/pjumppanen/niMSE-IO-BET-YFT/). We recommend checking with the authors to ensure that the latest items have been uploaded. The software has been reasonably stable for the past several months, with notable minor changes:

- bug fix related to the choice of CPUE series to be used within an MP.
- Recalculation of MSY-based reference points in relation to user-defined catch allocation among fisheries. To date, all reporting has been based on recent F ratios among fisheries. This extension will enable MSY-based reference points to be compared under different catch allocation scenarios.
- Additional functionality for changing existing OM parameters, renaming MPs, etc.

The production model-based MPs have undergone considerable development to improve convergence reliability and improve MP behaviour, as discussed in Kolody and Jumppanen (2019c).

4 Yellowfin Reference Set OM Conditioning

4.1 Relationship between the stock assessment and Operating Models

The intention has always been to maintain a close relationship between the stock assessment modelling and the conditioning of OMs. The two processes are analogous in several respects, i.e. similar population dynamics models are fit to the same data, subject to the same concerns about model formulation and assumption violations, etc. It would be difficult to justify the two initiatives evolving in different directions from the same scientific process. Accordingly, the reference case yellowfin assessment provided by Dan Fu (IOTC Secretariat, see Fu et al. 2018) provides the core of the OM conditioning process as described in attachment 1. Originally there had been no intention to update the OM conditioning in relation to this assessment, but it was recognized that some substantial data changes had occurred, and the OM would also provide a useful platform for potentially contributing to the YFT stock assessment review process.

One structural difference between the yellowfin assessment and the OM models relates to seasonality in movement. The assessment linked movement rates to environmental indices. This would add an additional complication for the OM, because it would require projections of environmental indices (or the net effect of environmental indices). It also remains unclear whether these indices were helpful for the assessment in disentangling seasonal movement from catchability. The assessment inferences were not substantively changed when this extra complexity was removed. If this approach is explored in the future, we would recommend testing whether inter-annual variability associated with real environmental indices has any explanatory power over and above fixed seasonal effects.

Relative to the traditional stock assessment, OM conditioning has an increased emphasis on uncertainty quantification and projections required to develop robust feedback-based MPs through the MSE process. The reference set OM is an ensemble of assessment models that includes several alternative plausible assumptions. The approach to uncertainty quantification adopted here is similar to that used in the CCSBT, in which the emphasis is on model structural uncertainty (including parameters about which the data are expected to be uninformative), and stochastic recruitment uncertainty (and observation error) in the projections. The Maximum Posterior Density Estimates (best point estimates for parameters and population states) for the individual models are collated, with the expectation that the uncertainty among point estimates will generally be greater than the parameter estimation uncertainty conditional on any individual model. Once an adequate OM ensemble has been defined, it should not need to be updated with the frequency expected for the traditional stock assessment process. Unless new evidence emerges to indicate that the uncertainty encompassed by the OM no longer captures reality, we would hope that an MP would remain valid for something on the order of 5-10 years (i.e. until the next thorough MP review scheduled as part of the adoption process).

Robustness OMs are generally considered less likely than the reference set. They are defined to represent plausible, troublesome situations, that may help identify pathological MP behaviour in particular circumstances, and assist in choosing among MPs that are otherwise equivalent. An MP cannot be expected to be robust to every imaginable outcome (attempting to do so would likely result in an extremely conservative MP and considerable lost economic opportunity). Because some extreme events are unpredictable, a normal part of the MP approach involves regular oversight (e.g. simple analyses to determine if "exceptional circumstances" have arisen which render the MP inappropriate, at least temporarily), and a scheduled review period, at which point a detailed evaluation should determine if the MP testing remains valid, and whether there have been other changes in circumstance, e.g. changing Commission objectives, better assessment tools, etc.

For the purposes of this paper, we refer to a number of OM ensembles, and option abbreviations as defined in Table 1 and Table 2.

Table 1. Yellowfin reference and robustness set OM ensemble definitions reported to the WPTT and WPM 2019. OMrefY19.4.500 was the reference set used for the most recent MP evaluations in Kolody and Jumppanen (2019c)

Model Name	Definition (assumption abbreviations are defined in Table 2)
OMrefY19.3.500	The reference set OM used for the TCMP 2019 results, based on 48 SS model configurations, a subset of a 144 model fractional factorial grid, filtered for convergence, catch likelihood < 1E-5, and annualized aggregate CPUE CV < 0.3. 500 realizations were randomly sampled with replacement. Subsequent to the TCMP, some minor specification errors were identified and corrected in the OMs below.
OMrefY19.3b.500	As OMrefY19.3.500, with minor errors corrected (recruitment CV, MP data lag changed from 3 to 2 years). One model was added to the list so that the “preferred” CPUE series was used for MP testing. The intent here was to show that the full 2 way interaction grid used in OMrefY19.4.500, should not be required in future iterations.
OMrefY19.4.500	**The reference set OM used for the WPTT and WPM 2019 results, based on 420 SS model configurations, a subset of a 1152 model fractional factorial grid, filtered for convergence, using the best fit from repeated convergence, catch likelihood < 1E-5 and annualized aggregate CPUE CV < 0.3. Multinomial sample of 500 realizations were taken from the uniformly weighted set of 420. The best fit models were adopted from the repeated convergence tests.**
OMrefY19.4.420	Identical to OMrefY19.4.500, except that all 420 SS configurations were included exactly once. The multinomial sampling in OMrefY19.4.500 means that about 30% of SS specifications are not sampled at all. This OM was intended to demonstrate that random

	sampling of realizations should not substantively affect MP evaluation results (except perhaps in the extended tails of the distributions)
OMrefY19.4BF.368	Identical to OMrefY19.4.420, except that the worst fit of the repeated minimization results were adopted (and additional models were rejected due to failing the catch likelihood criterion, resulting in 368 SS specifications). This OM was intended to demonstrate that the sensitivity to initial values in the minimization should not substantively affect MP evaluation results based on a large sample (even though individual specifications might be disturbingly sensitive).

Robustness tests

OMrobY19.4.ICV30	Robustness scenario OM with projected CPUE observation error CV = 0.3 (annualized aggregate). (conditioning unchanged from OMrefY19.4.500)
OMrobY19.4.10overRep	Robustness scenario OM with consistent 10% overcatch for all fleets (catch is accurately reported) (conditioning unchanged from OMrefY19.4.500)
OMrobY19.4.10overIUU	Robustness scenario OM with consistent 10% unreported overcatch for all fleets (conditioning unchanged from OMrefY19.4.500)
OMrobY19.4.recShock	Robustness scenario OM with 8 consecutive quarters of poor recruitment (55% of expected values, similar to estimates for YFT in the early 2000s). (conditioning unchanged from OMrefY18.1)
OMrobY19.4.qTrend2	Robustness scenario OM with longline CPUE catchability trend of 3% per year in projections (conditioning unchanged from OMrefY18.1)

Table 2. Model assumption option abbreviations (as used in the text and figures). Bold indicates the assessment reference case assumption. Some abbreviations may relate to explorations that have not yet been examined, or are not reported in the current document.

Abbreviation	Definition
	Stock-recruit function (*h* = steepness)
h70	Beverton-Holt, *h* = 0.7
h80	**Beverton-Holt, *h* = 0.8**
h90	Beverton-Holt, *h* = 0.9
Rh70	Ricker, *h* = 0.7
Rh80	Ricker, *h* = 0.8
Rh90	Ricker, *h* = 0.9
	Recruitment deviation penalty
sr4	σ_R = 0.4
sr6	**σ_R = 0.6**
sr8	σ_R = 0.8
	Natural mortality multiplier relative to SA-base
M10	**1.0**
M08	0.8
M06	0.6
	Tag recapture data weighting (tag composition and negative binomial)
t00	λ = 0
t0001	λ = 0.001
t001	λ = 0.01
t01	λ = 0.1
t10	**λ = 1.0**
t15	λ = 1.5
	Assumed longline CPUE catchability trend (compounded)
q0	**0% per annum**
q1	1% per annum
q3	3% per annum
q5	5% per annum
	Tropical CPUE standardization method (error assumption for all series)
iH	**Hooks Between Floats (σ_{CPUE} = 0.3)**
i10H	Hooks Between Floats (σ_{CPUE} = 0.1)
iC	Cluster analysis (σ_{CPUE} = 0.3)
i10C	Cluster analysis (σ_{CPUE} = 0.1)
	Tag mixing period
x3	3 quarters
x4	**4 quarters**
x8	8 quarters
	Longline selectivity

SL	**Stationary, logistic, shared among areas**
SD	Double-normal (potentially dome-shaped), shared among areas
S4	LL selectivity independent among areas
NS	Temporal variability estimated in 10 year blocks
ST	Logistic selectivity trend estimated over time
Sdev	15 years of selectivity deviations estimated (most recent years)
Sspl	Cubic spline function (to admit possibility of dome-shape)
	Size composition input Effective Sample Sizes (ESS)
ESS2	ESS = 2, all fisheries
ESS5	**ESS = 5, all fisheries**
CLRW	ESS = One iteration of re-weighting; the output ESS from the reference case assessment analogue, capped at 100.
CL75	ESS = One iteration of re-weighting; the output ESS from the reference case assessment analogue raised to the power of 0.75 and capped at 100.

4.2 The reference set YFT Operating Model OMrefY19.4

MP evaluation results for the 2019 TCMP (Kolody and Jumppanen 2019e) were based on OMrefY19.3, a fractional factorial design that was intended to estimate main effects only (i.e. if it was applied in the context of classic experimental design and ANOVAs). It started from a balanced grid of 144 models, of which approximately two thirds were rejected either because of repeated convergence failures or relatively large catch likelihoods (indicating implausibly small biomass in some demographic stratum). Additionally, a small number of outliers with very high MSY were rejected (using a sharp threshold of a mean CPUE RMSE (annualized) > 0.3, there was a perfect correspondence between outlier MSY and high CPUE RMSE). There were some minor specification errors in OMrefY19.3 that affected the MP results presented to the 2019 TCMP (related to recruitment CV and the observed CPUE data available for the MP). The data lag time to MP application was also reduced from 3 to 2 years as per WPTT/WPM request. However we note that the draft YFT MP resolution proposes a 3 year lag (IOTC-2019-S23-PropP), and this discrepancy should be resolved for future testing.

As part of a broader test of how to deal with high dimensionality in the OM ensemble, and the issue of numerical instability (both discussed subsequently), the preferred reference set OMref19.4.500 was developed:

- A fractional factorial grid of 1152 models (main effects and all 2 way interactions estimable in an ANOVA context). The full factorial design would have had 4608 models.
- Each model specification was refit from jittered initial conditions until either 3 successful convergences were achieved, or 10 attempts were completed. This analysis took several weeks of computing time on a cluster (and would have taken more than 5 years on a single CPU). It was not intended as something that would be done routinely– rather the hope was that it would demonstrate that the shortcuts undertaken for OMrefY19.3 were valid and will suffice for future iterations.
- 1050 models reached the 0.01 maximum gradient convergence criterion. The catch likelihood distribution is clearly bimodal (Figure 1), and we opted for the 1E-5 rejection point, which results in a subset of 487 models. There is some outlier behaviour in MSY, that is associated with several interacting assumptions, e.g. always M10, t0001, I3, CL75 and SD (Figure 2). These outlier models were all associated with relatively poor fit to the CPUE (Figure 3), and accordingly we removed all models in which the mean annualized CPUE RMSE > 0.3. This is admittedly arbitrary and discontinuous, but removes the outliers and is consistent with the general principle that a failure to fit the CPUE means that the model is not consistent with the most informative data in the assessment. This resulted in the ensemble of 420 models that form the basis of reference set OMrefY19.4.
- Standard (r4ss) diagnostic plots at the 4 “corners” (defined as the highest and lowest MSY and depletion) of the OM ensemble were qualitatively inspected without obvious evidence of model failure (not shown). The assumption is that if the extremes are not outliers, then intermediate values are likely to be reasonable, or not likely to badly skew MP evaluation results if they turn out to be flawed.

Figure 4 - Figure 15 illustrate various features of the model fits to data and stock status characteristics in relation to model assumptions, from which we observe:

- The annualized fit to the CPUE is usually very good, i.e. median CV ~0.20 (region 3 CPUE is somewhat worse than the aggregate with median CV ~ 0.24). Not surprisingly, models with assumed input CV of 0.1 are consistently better fit than CV 0.3 assumptions. However, the CV 0.1 models are also rejected at a higher rate than the CV 0.3 models (attachment 1 Table 2), and this is presumably part of the conflict that leads to the retrospective assessment pessimism. But the issue is probably not that simple, or we would have expected the q1 (1% catchability trend) to also be disproportionately rejected.
- The quality of fit to the size composition data is more variable among fleets than among models. There is not much outlier behaviour that would be obviously useful for model retention/rejection.
- Many of the model assumptions affect stock status in a manner that is qualitatively predictable (i.e. low h, high M and 1% catchability trend are associated with more pessimistic outcomes). The least influential assumptions appear to be the CPUE standardization approach (HBF or cluster), the CPUE regional weighting factors and the tag mixing period. The tag mixing period is misleading in the sense that 50% of the observations are irrelevant (the tags are not included in 50% of the models).
- These diagnostics can be misleading for examining the implication of tag fits. i.e. if the tags are highly down-weighted, we are essentially saying that we should ignore their influence. Furthermore, if tags are subject to different mixing periods, the aggregate diagnostics are not comparable. However, it is clear that including the tags tends to be associated with more pessimistic stock status.

OMrefY19.4.500 projection characteristics are shown in Figure 13 with a catch moratorium and constant current catch. Current catches are estimated to be unsustainable in 90-100% of the projections.

4.3 Implications of fractional factorial design and numerical stability on the reference set YFT Operating Model

Calculating the 1152 SS model grid in the previous section, allowed us to revisit some important assumptions related to the formulation of a large OM ensemble:

- The current yellowfin stock assessment minimization tends to be unstable as indicated by jitter analyses. Is this instability likely to affect aggregate MP evaluation results?
- Are we confident that a main-effects fractional factorial design will yield MP evaluation results that are representative of a grid with estimable higher-order interactions?

Figure 14 and Figure 15 compare stock status inferences (marginalized by assumption) for the best and worst fit of the (converged) models (obtained from the 3 replicate jitter analysis) from OMrefY19.4.420. Figure 16 and Figure 17 show the deviation in stock status between the best and worst fit models from, from which we note:

- The MSY CV attributable to within model minimization instability is 4.6%, while the CV among the best fit models is 16%.
- The B(T)/BMSY CV attributable to within model minimization instability is 11%, while the CV among the best fit models is 28%.

This calculation was repeated for the subset of 368 models in which the worst of the converged models were also filtered with respect to the catch likelihood (in the preceding calculation this filtering was only applied to the best fit case). In this case, the CVs attributable to minimization error were only moderately reduced to 3.7% (MSY) and 8.3% (B/BMSY).

The minimization stability appears to be much more of a concern with yellowfin compared to bigeye (for bigeye, the minimization error CV was ~10% of the among model CV for both stock status quantities). It remains generally true that the variability among models is considerably greater than the variability introduced by minimization error, but it would not be surprising if minimization error is large enough to have a non-negligible influence in the tails of the MP evaluation distributions.

Figure 18 compares the MP performance from two contrasting MPs (at two tuning levels), and the alternative OM reference sets described in Table 1. The MP performance is very similar, from which we tentatively conclude:

- 500 multinomial samples from the 420 model grid yields similar results to uniformly testing all 420 models (even though a third of models are omitted in the multinomial sample)
- The minimization sensitivity does not appear to make much difference, i.e. The best and worst fit model ensembles yield very similar MP evaluation results.
- The fractional factorial design with main effects only (49 models, 144 before plausibility filtering) yields very similar results to the much larger grid with all 2 way interactions (420 models, 1152 after plausibility filtering).

To date, we have assumed that the results above would hold true from general statistical principles, but this gives us confidence that future OM conditioning can be streamlined. In this case, the fractional factorial design is probably a useful exercise for stratification but not particularly useful in the manner originally intended. i.e. The hope was that it would enable the importance of different assumption levels to be formally quantified, but since 2/3 of the models are doubtful for numerical reasons, the orthogonal design principles are presumably compromised. However, from general statistical principles, we would expect that randomly sampling from assumption combinations (and successfully fitting 50-100 models) would yield an OM with similar MP evaluation characteristics, even if we cannot be confident in reliably quantifying the relative importance of specific assumptions and interactions.

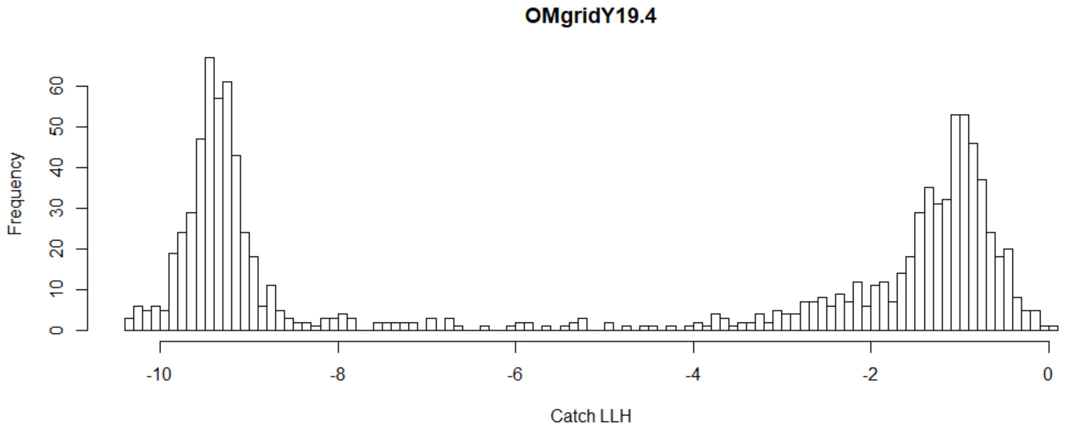


Figure 1. OMgridY19.4 catch penalty distribution.

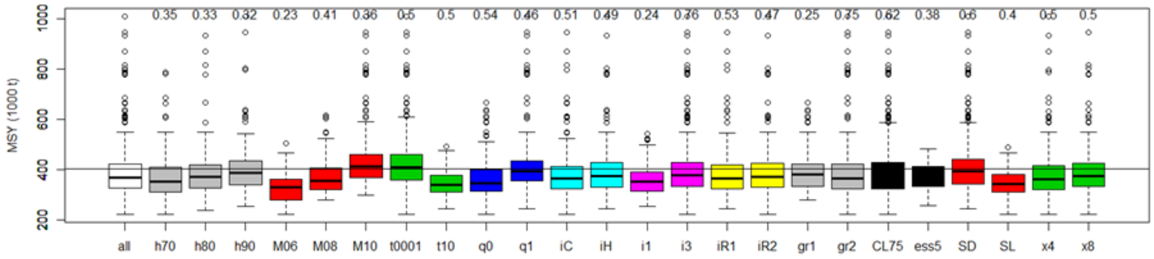


Figure 2. OMgridB19.4 (filtered by catch penalty) MSY distribution marginalized by model assumption.

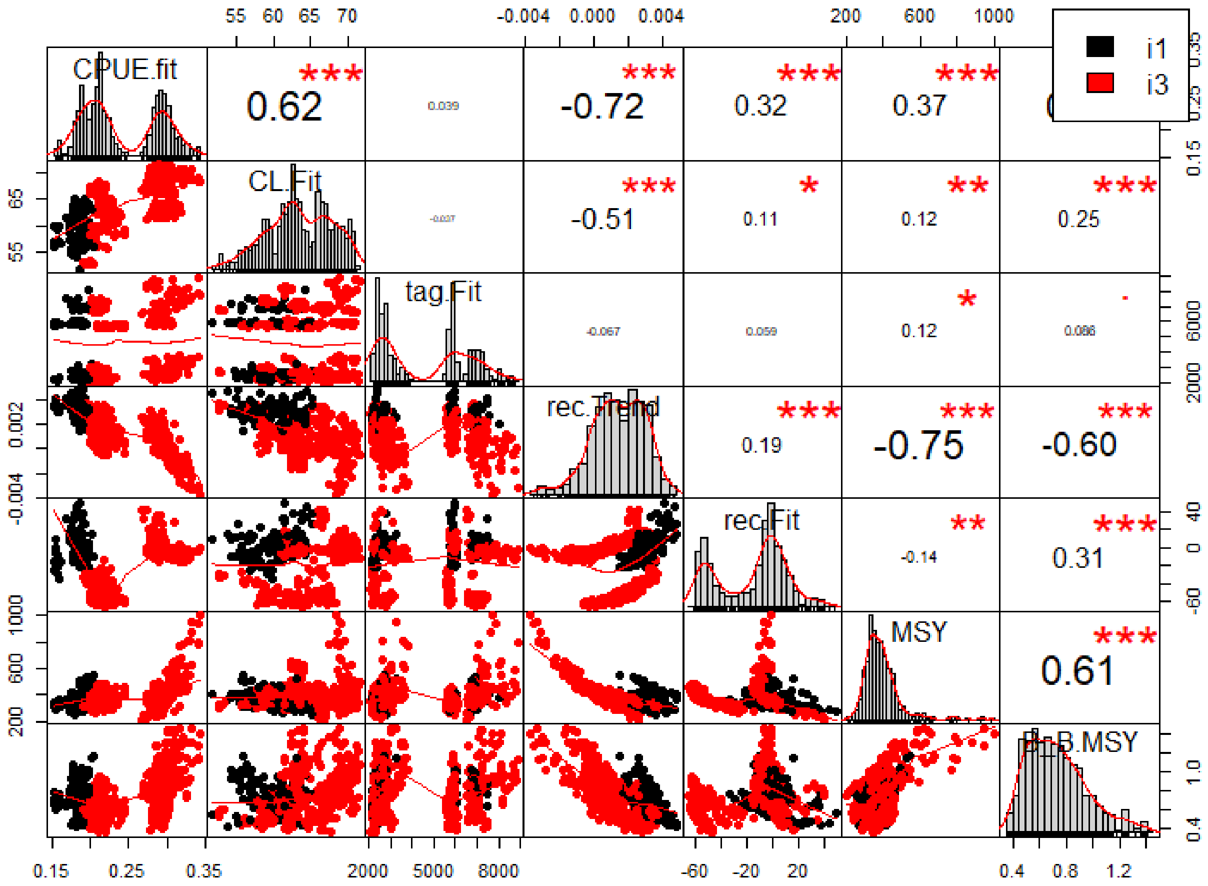


Figure 3. OMgridB19.4 (filtered by catch penalty) relationship among several stock status and quality of fit indicators.

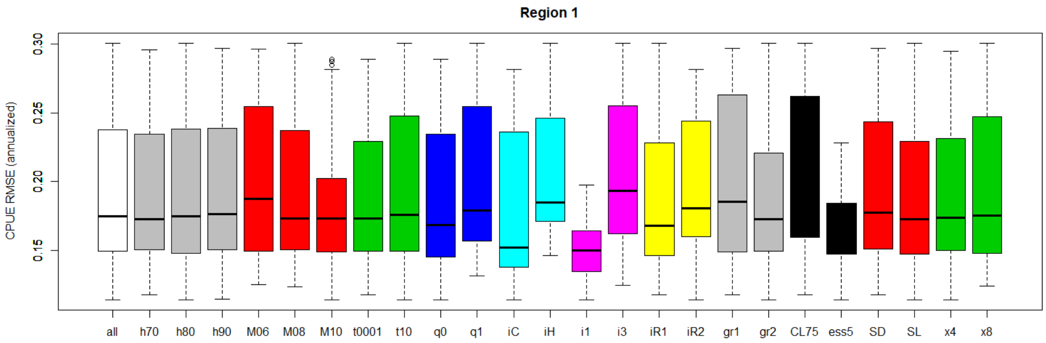


Figure 4. OMrefY19.4 quality of fit (RMSE) for the CPUE series in region 1 (annualized).

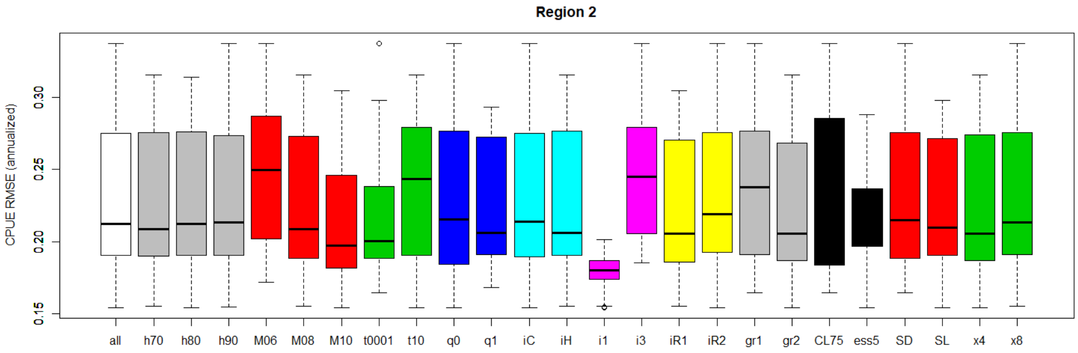


Figure 5. OMrefY19.4 quality of fit (RMSE) for the CPUE series in region 2 (annualized).

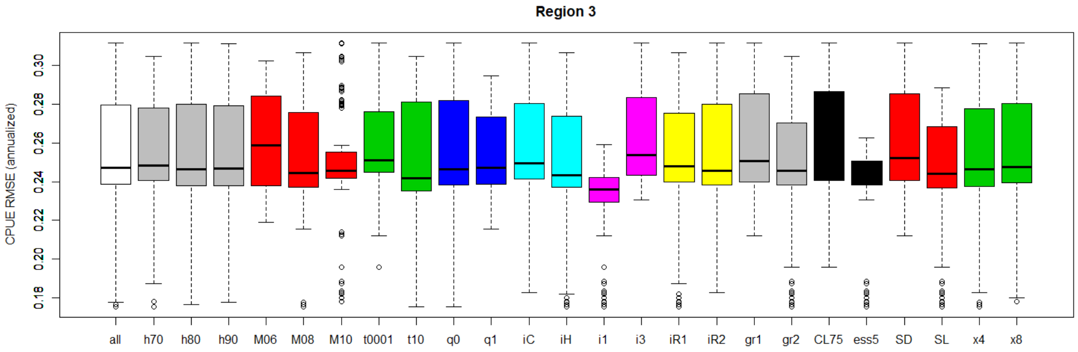


Figure 6. OMrefY19.4 quality of fit (RMSE) for the CPUE series in region 3 (annualized).

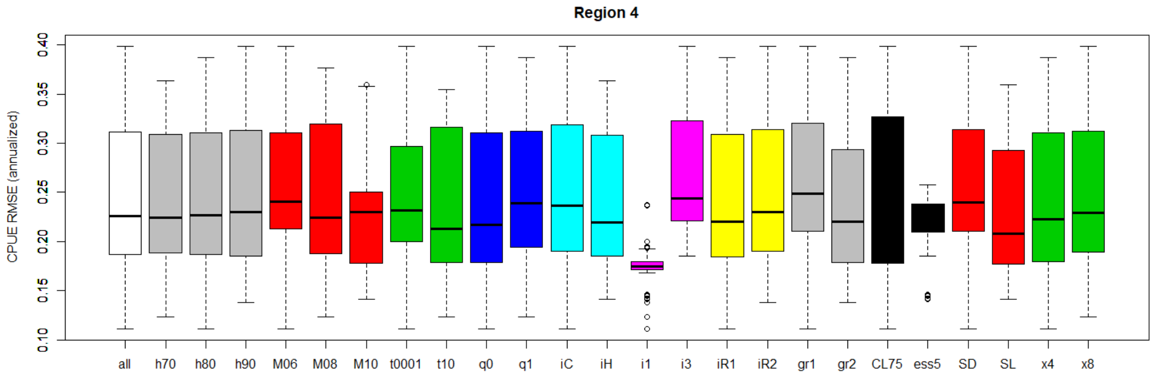


Figure 7. OMrefY19.4 quality of fit (RMSE) for the CPUE series in region 4 (annualized).

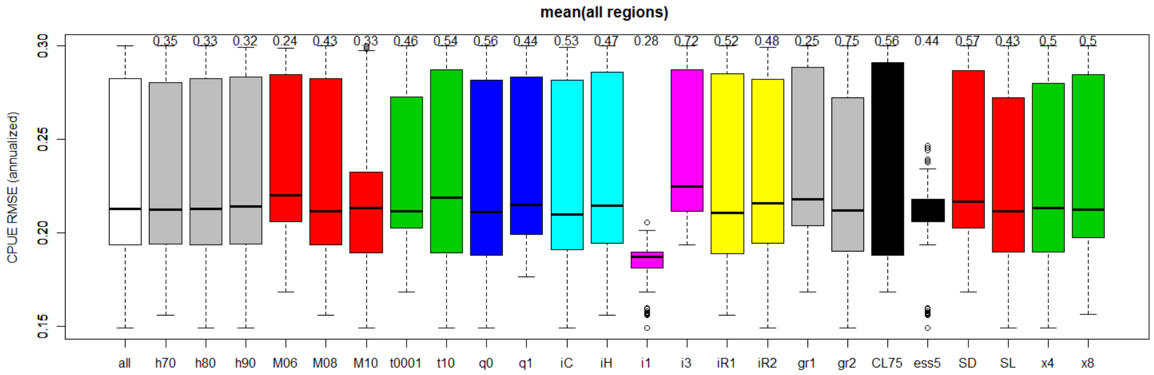
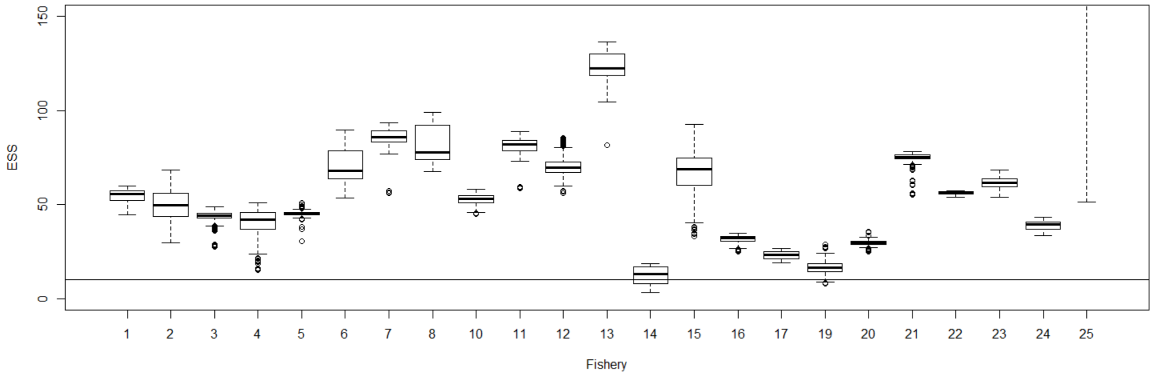


Figure 8. OMrefY19.4 Quality of fit (RMSE) for the mean of all CPUE series (annualized).



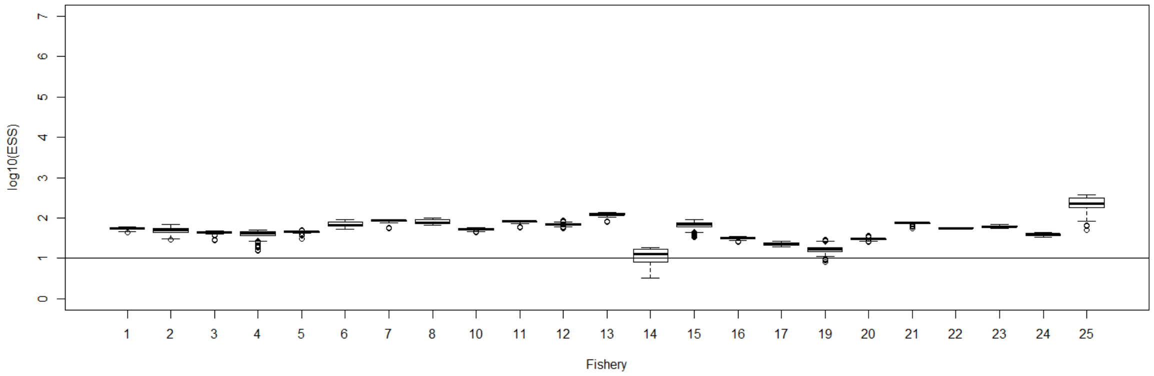


Figure 9. OMrefY19.4 quality of fit (post-fit Effective Sample Size) for the size composition data by fishery (all models combined). The key point is that the different model assumptions do not have much effect on the fit to the size data for the most part.

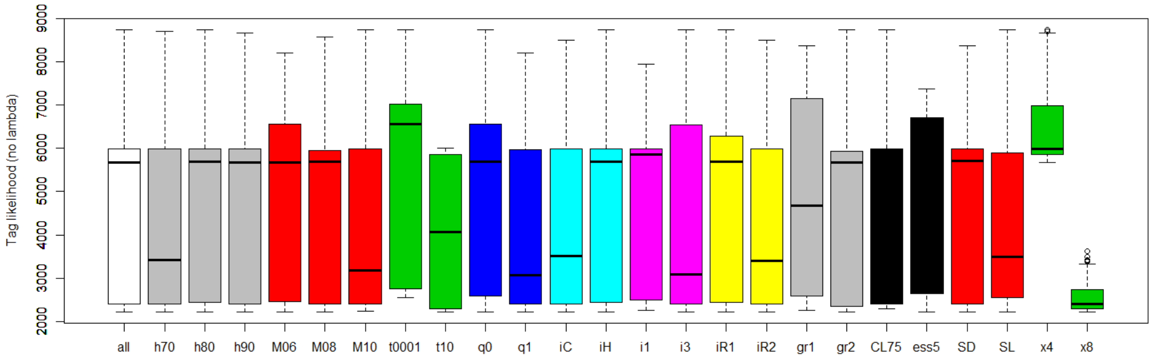


Figure 10. OMrefY19.4 Tag likelihood summaries marginalized over assumption levels.

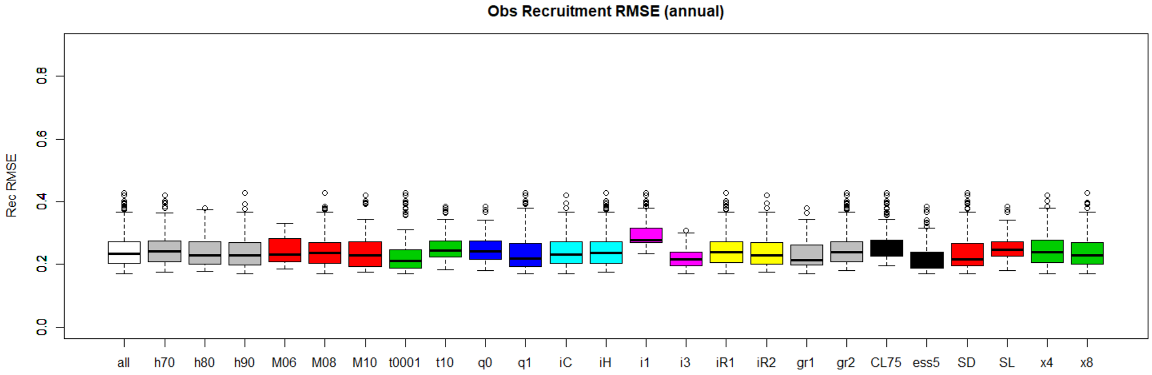


Figure 11. OMrefY19.4 recruitment RMSE (annualized - deviations aggregated across regions and seasons).

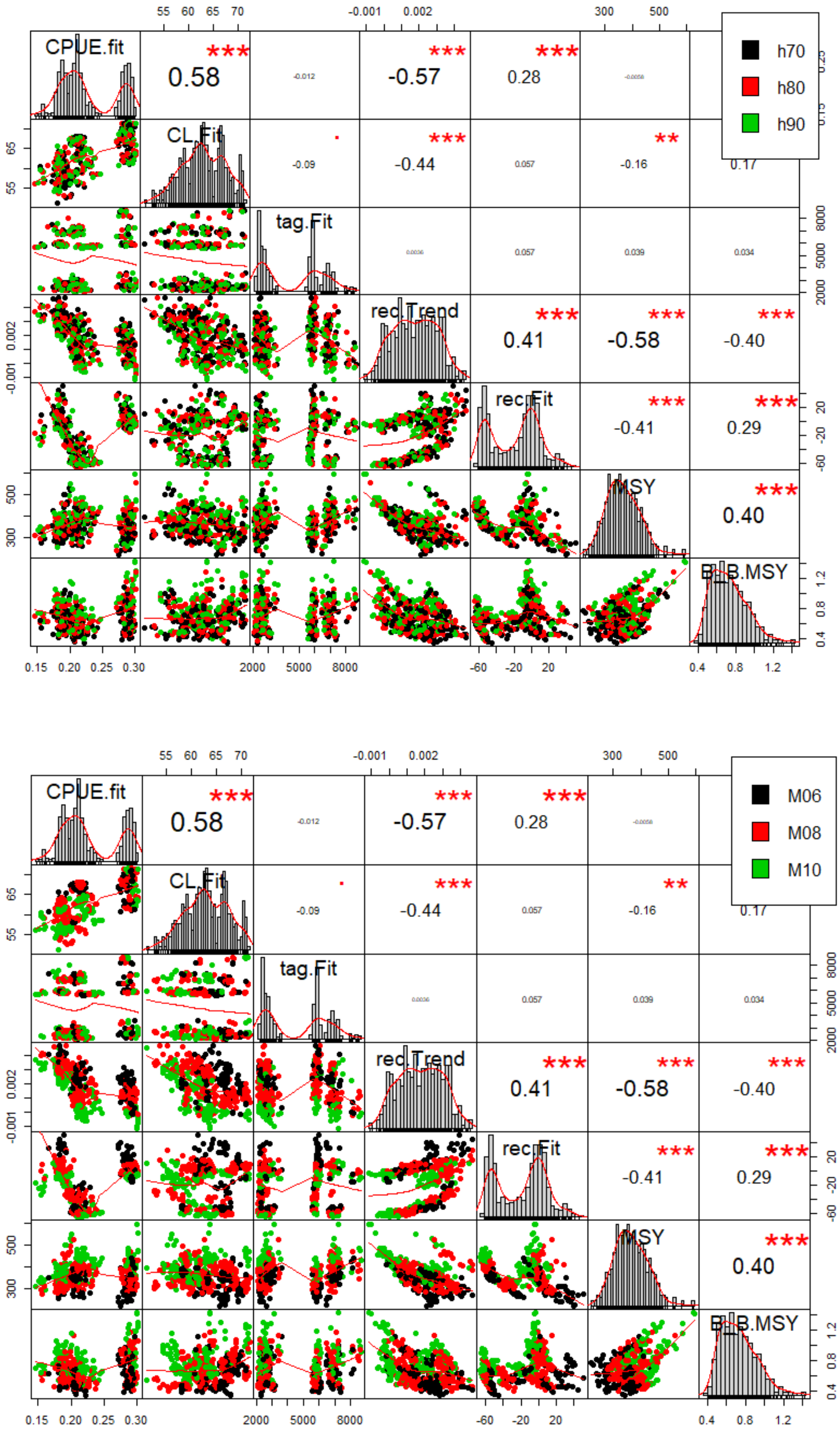
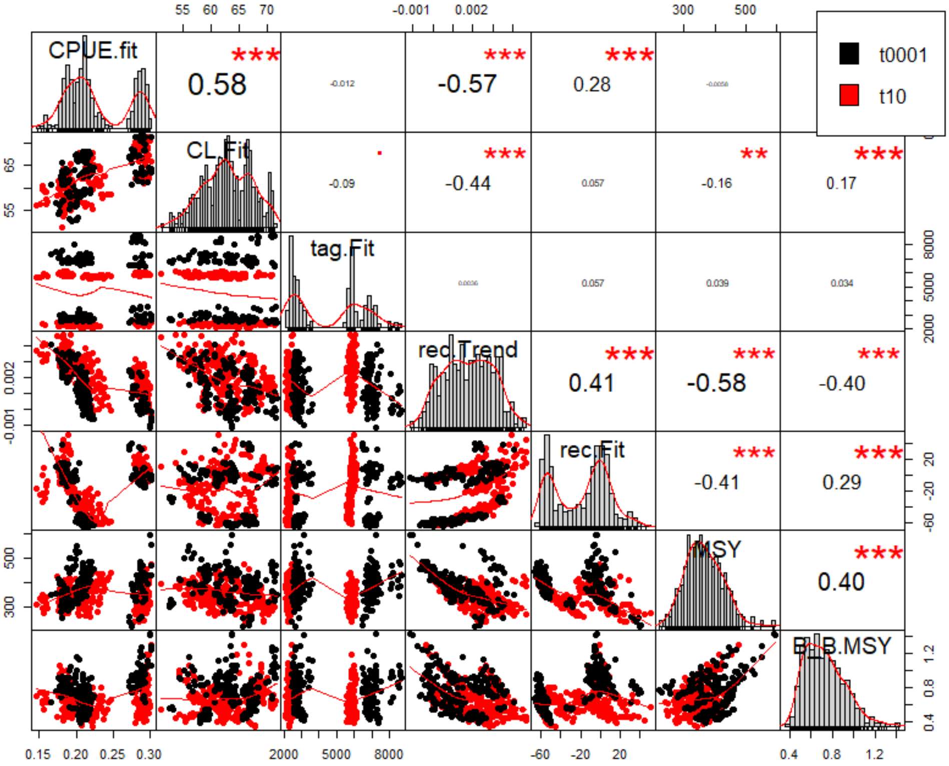


Figure 12. Operating Model OMrefY19.4 relationships among various quality of fit and stock status summary indices, partitioned by assumptions indicated in legend.



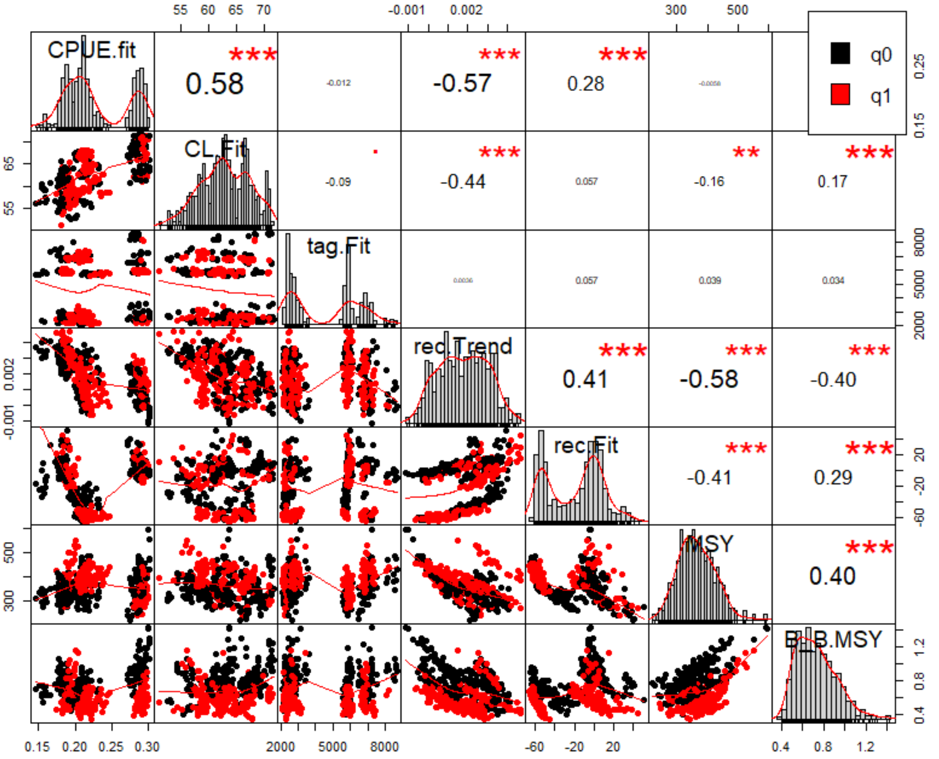


Figure 12 (cont.)

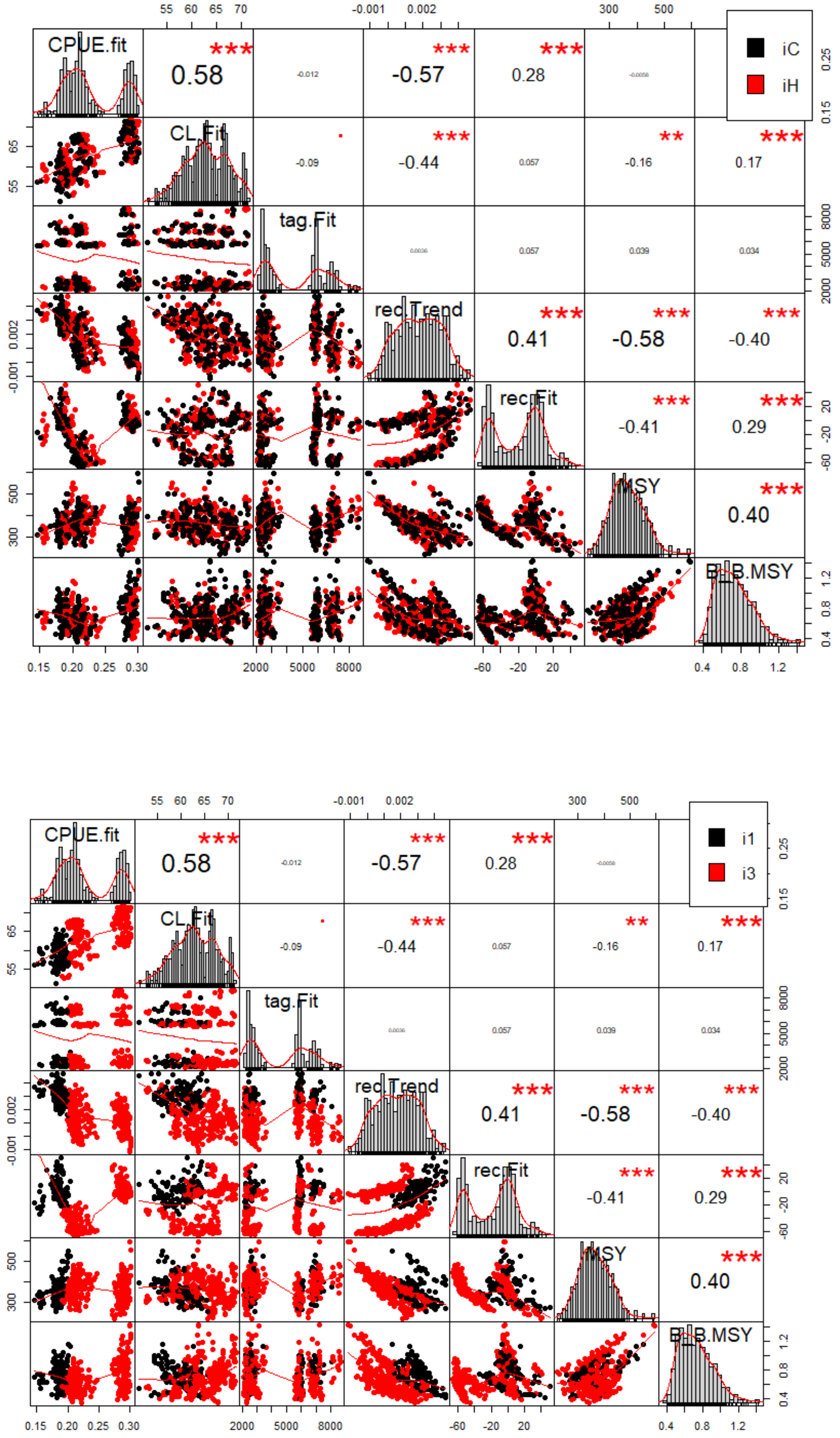
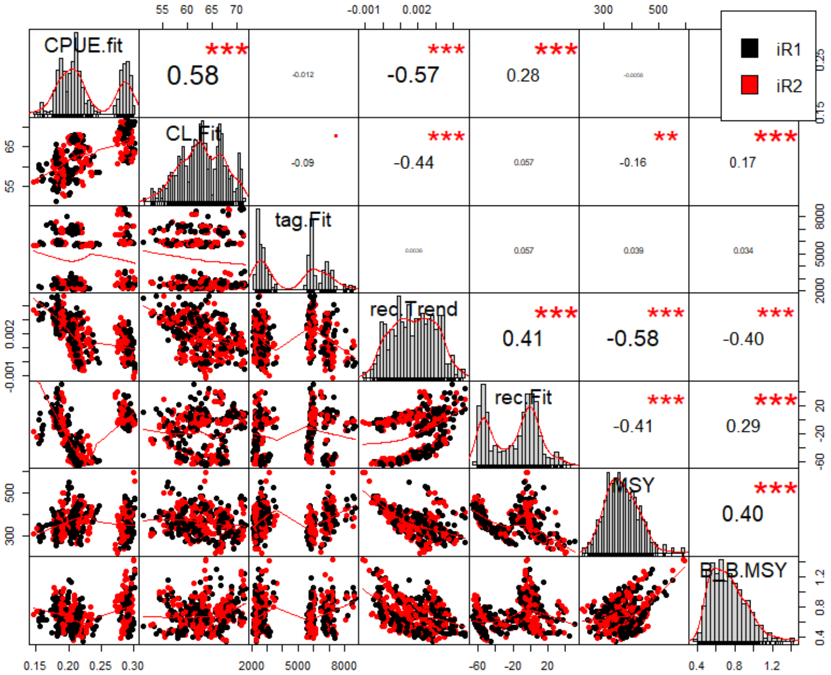


Figure 12 (cont.)



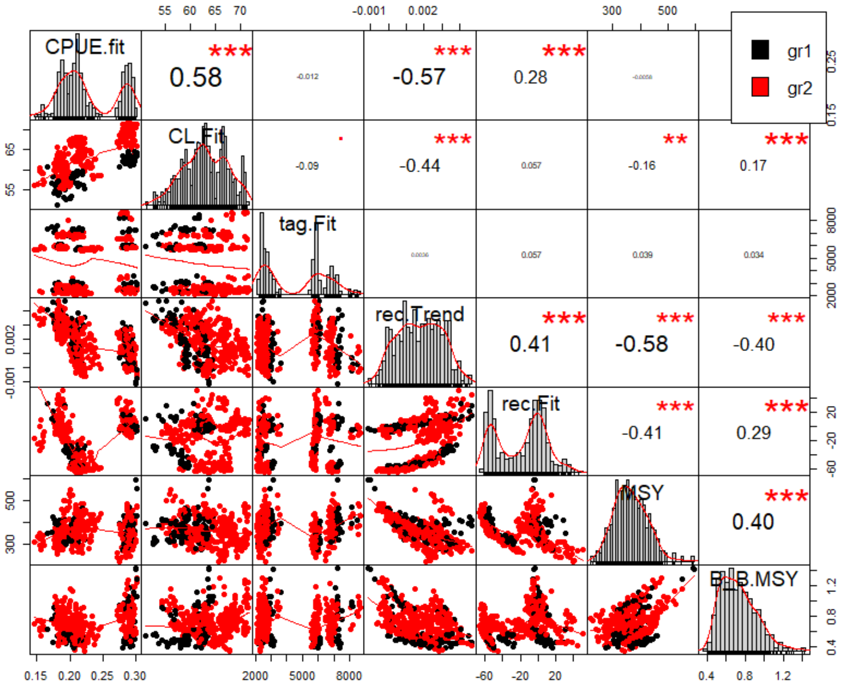
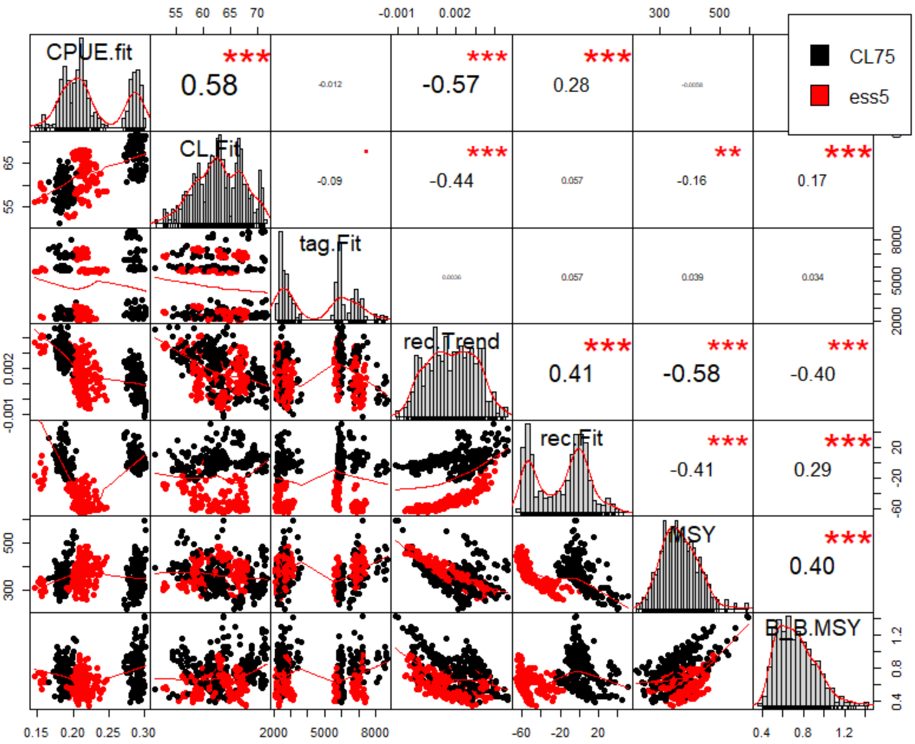


Figure 12 (cont.)



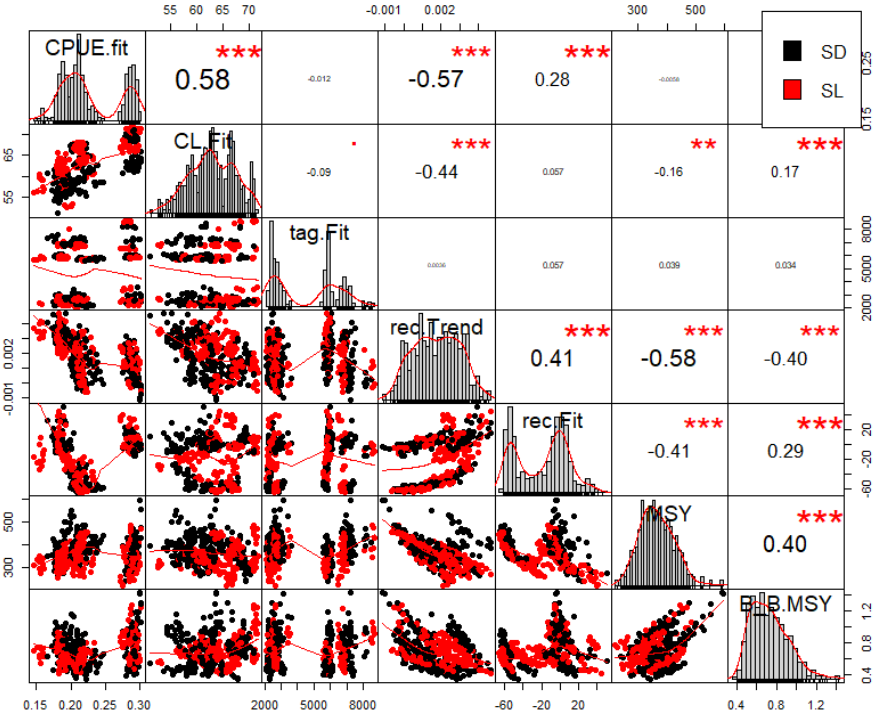


Figure 12 (cont.)

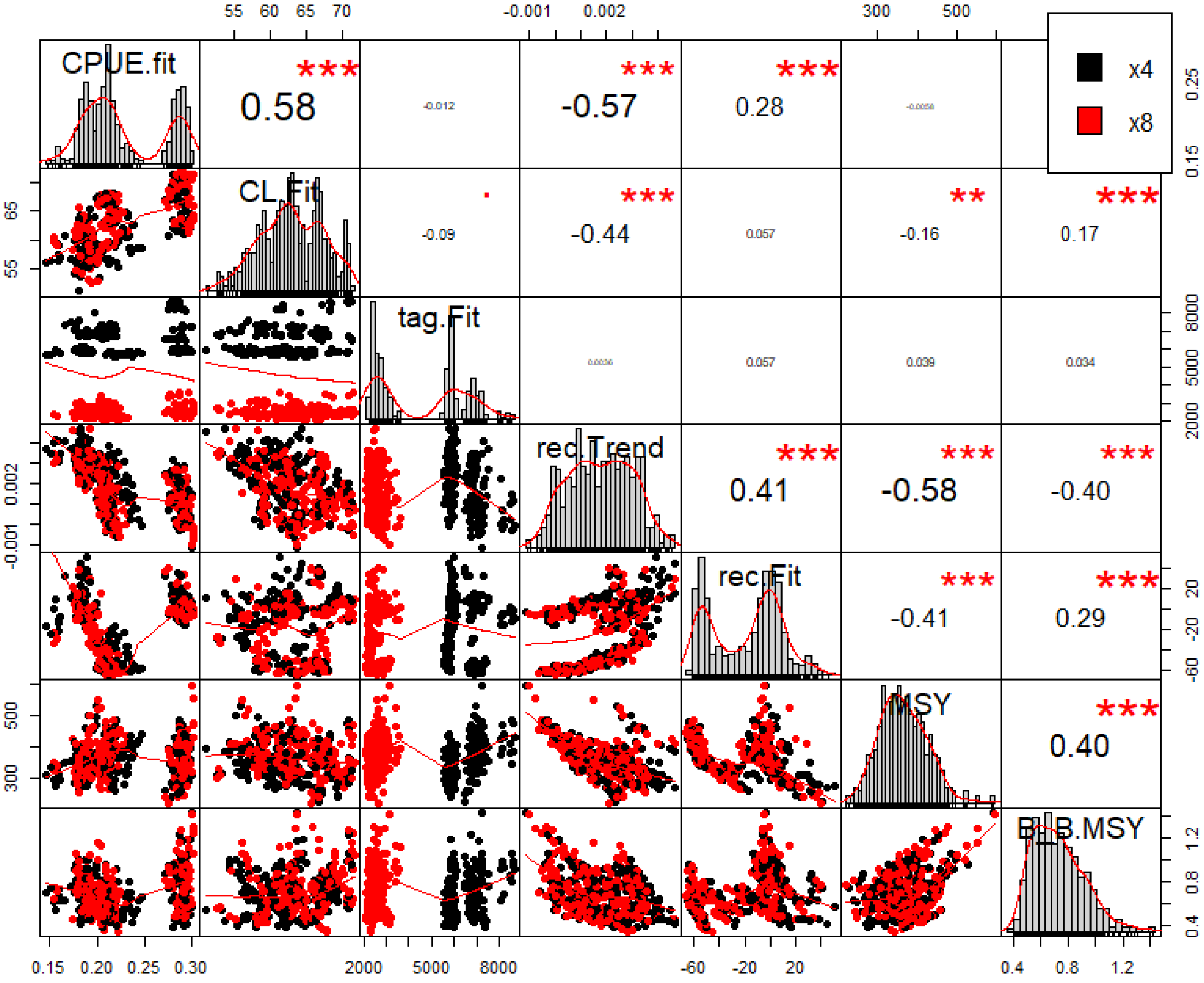


Figure 12 (cont.)

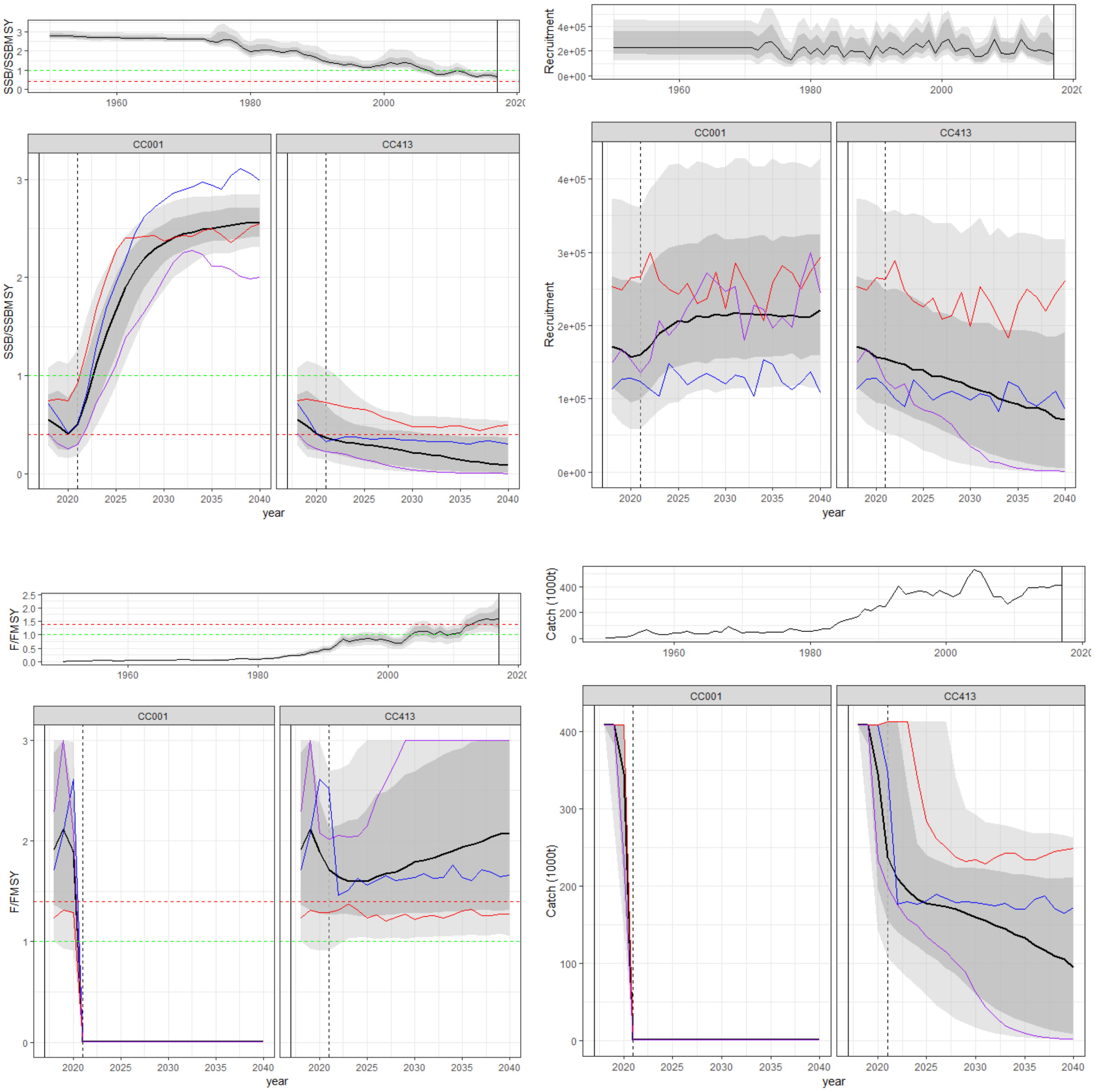


Figure 13. OMrefY19.4.500 constant catch projections (fishing moratorium and recent catch 413 Kt).

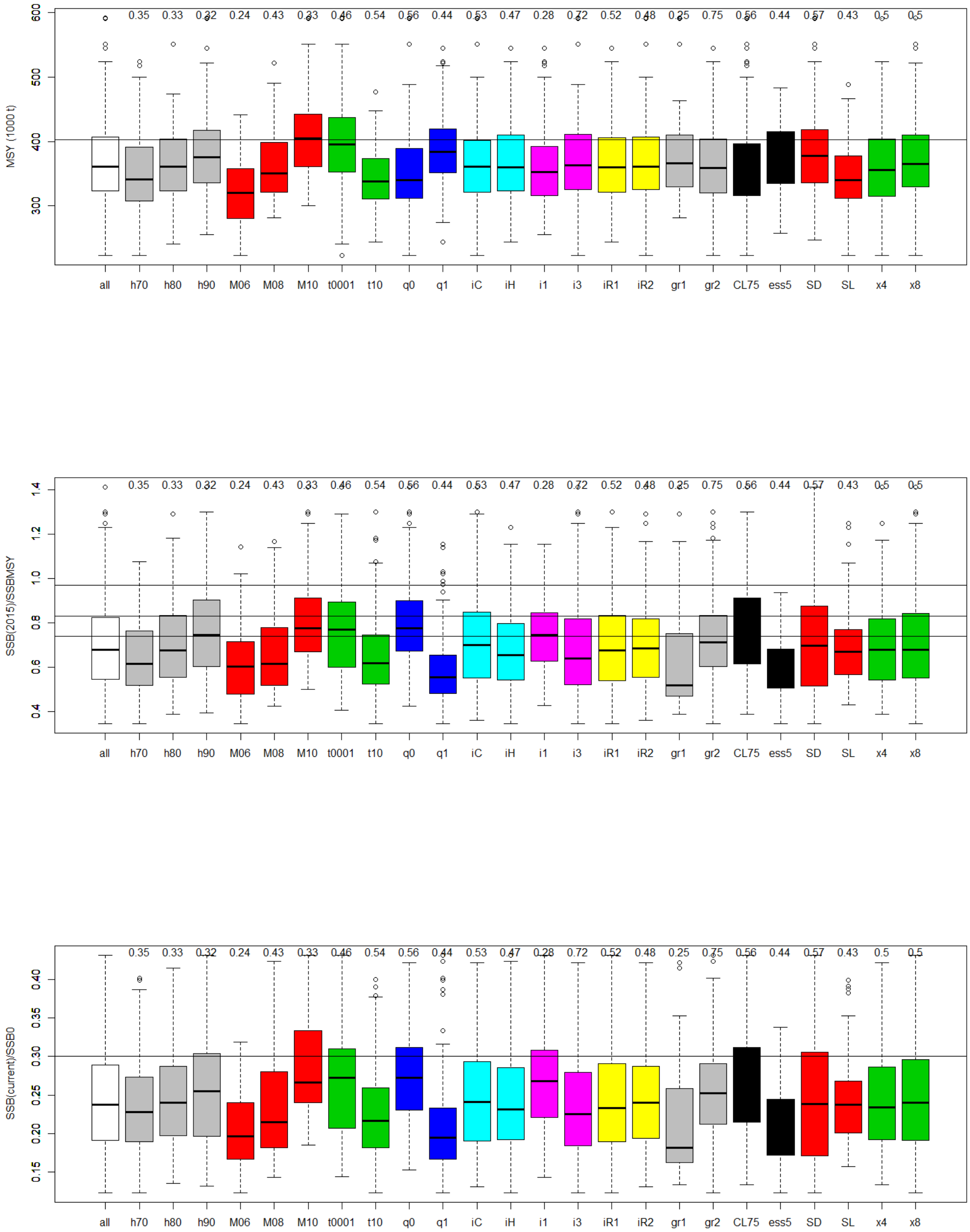


Figure 14. OMrefY19.4 key stock status inferences marginalized over model assumptions.

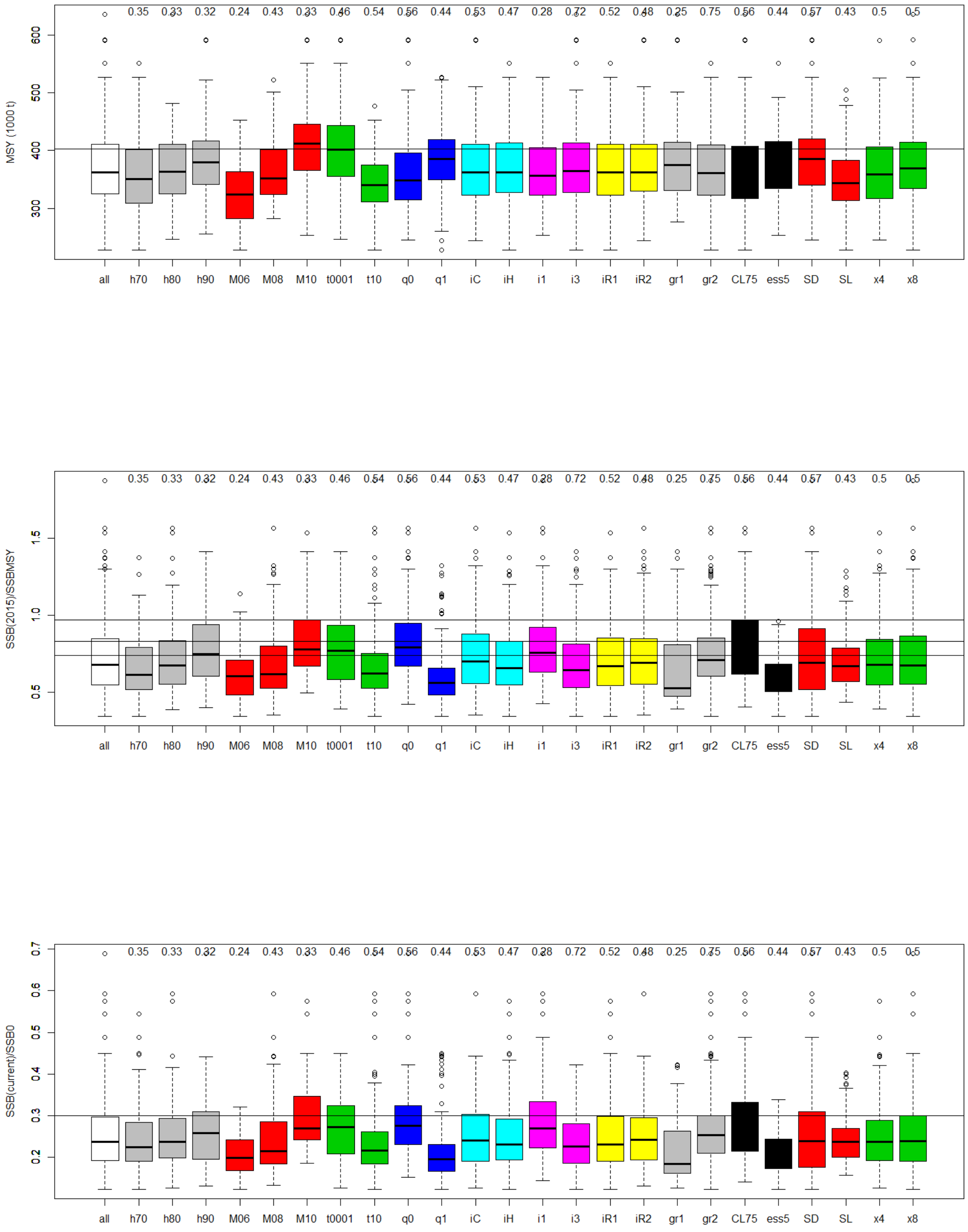


Figure 15. OMrefY19.4BF - key stock status inferences marginalized over model assumptions for the grid of models that is based on the worst-fit of the converged models that were attained from the jittered starting points.

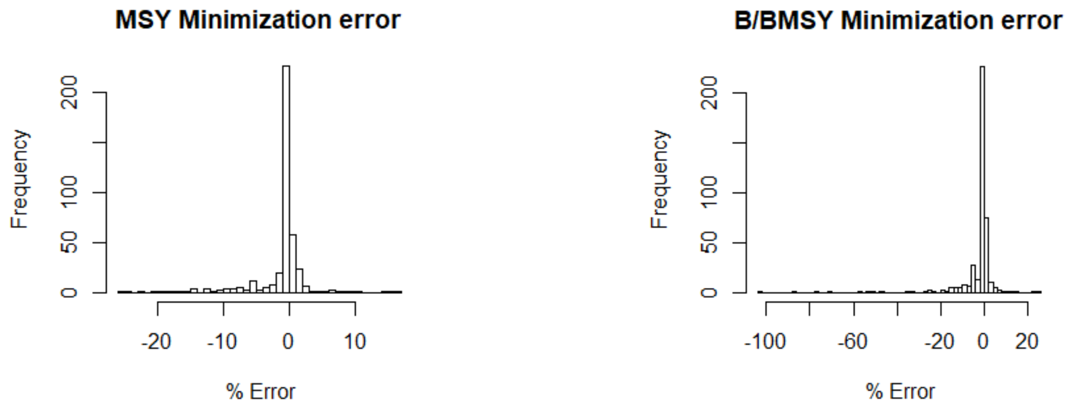


Figure 16. Relative error in stock status associated with the best and worst of the 420 converged minimizations resulting from jittered initial conditions (OMrefY19.4.420).

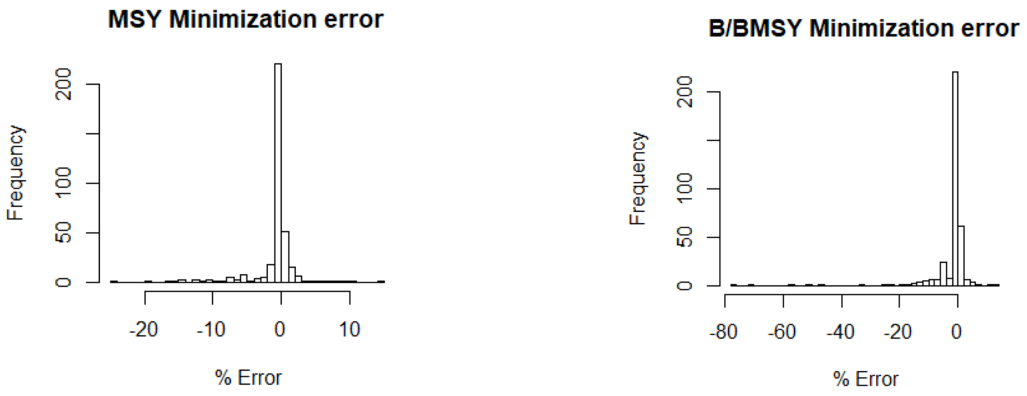


Figure 17. Relative error in stock status associated with the best and worst of the 368 converged minimizations resulting from jittered initial conditions (OMrefY19.4.420 vs OMrefY19.4.420BF), removing all models in which the catch likelihood component exceeded 1E-5 in either OMrefY19.4.420 or OMrefY19.4.420BF.

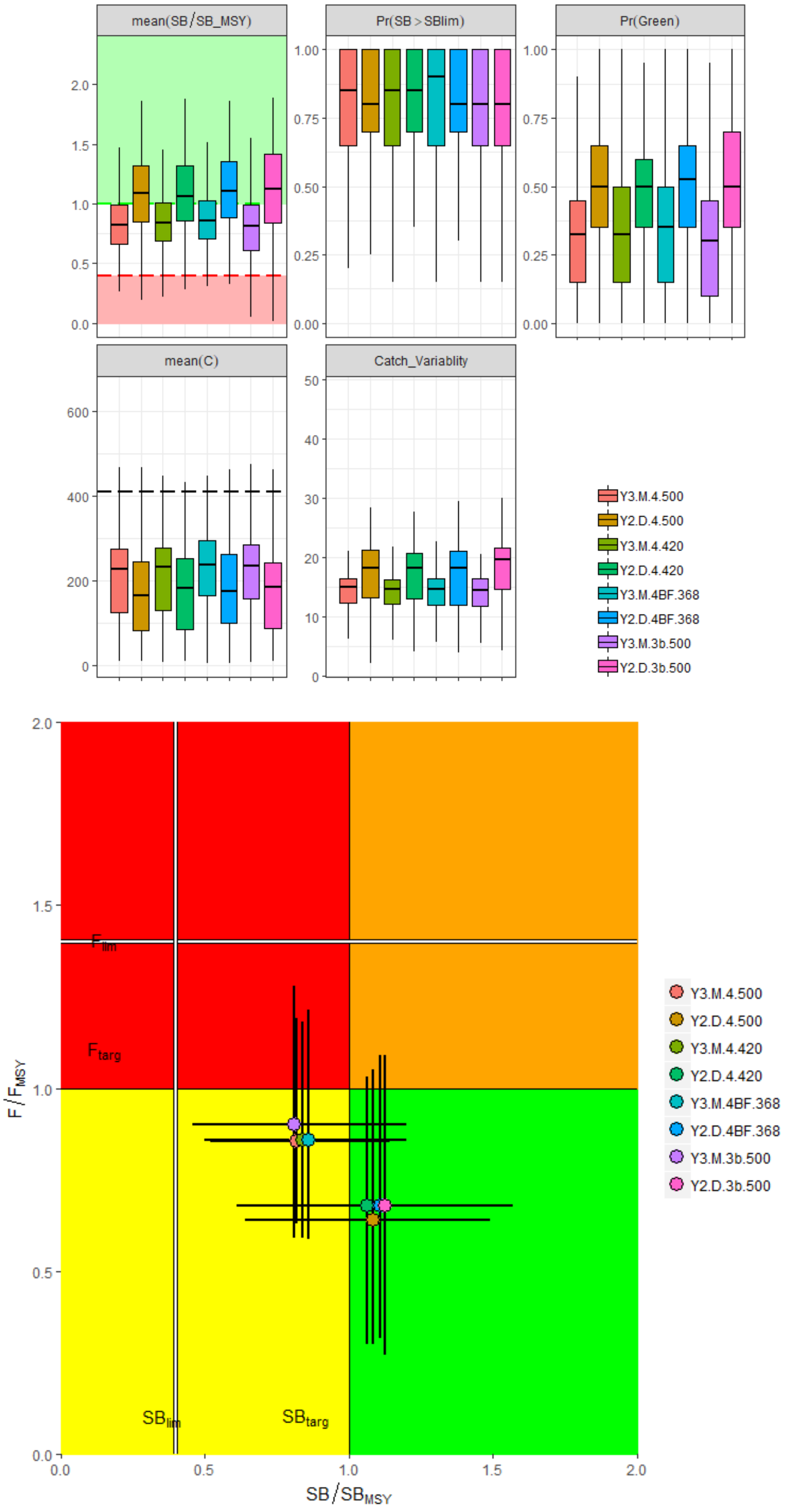


Figure 18. Comparison of MP evaluation performance for 2 contrasting MPs, derived from 4 different approaches to model sampling for the OM ensemble (defined in Table 1). The key point is that the MP performance was not sensitive to the OM sampling approach – jitter analyses and a large grid with all two-way (or higher) interactions is probably not essential.

5 Yellowfin Robustness OMs

Robustness test requests (e.g. section 2.1) were addressed in one of 4 ways:

- Uncertainty dimensions were elevated to the reference set for conditioning and MP evaluation (e.g. alternative growth curve, dome-shaped longline selectivity).
- Robustness tests were included as independent MP evaluations reported in Kolody and Jumppanen (2019c). In this case, the MP is tuned for the reference set, and the tuned MP is applied to the robustness test (e.g. recruitment shock test - the magnitude of the recruitment shock scenario is shown in Figure 19).
- Some basic exploration of a new uncertainty dimension was undertaken with respect to the reference case OM, but not taken further, pending feedback from the appropriate working groups (e.g. PS CPUE, Ricker function in the following section).
- Some ambitious robustness test proposals have not been addressed, because further clarification would be required before undertaking substantive code changes or data manipulation (e.g. non-stationary dynamics, higher resolution spatial structure).

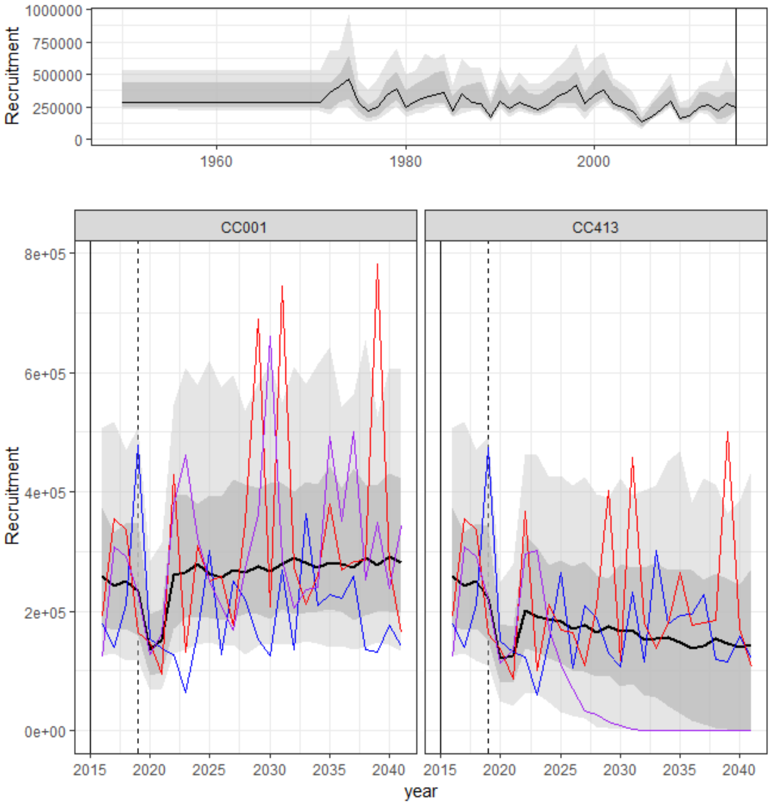


Figure 19. Yellowfin recruitment time series for the robustness scenario OMrobY18.1.recShock and two constant catch projections (figure is retained from an earlier document – the magnitude of the recruitment shock is unchanged).

6 Exploration of alternative Yellowfin Assessment/Operating Model Options

Recent IOTC yellowfin assessments and Operating Models have troubling features. Most notably, the stock appears to be of a size that can just barely sustain the recent observed catches. In many cases, this is evident from the Stock Synthesis catch penalty being "substantially" > 0. Retrospective analyses (Matsumoto et al. 2018) demonstrate that this situation extends back several years. i.e. Each successive model (with new data added) demonstrates that the previous model must have been unduly pessimistic (because subsequent catches can be extracted without the problems expected by the preceding model with fewer data). In the following we explore some elements of the apparent inconsistency between the IOTC yellowfin catch and standardized longline CPUE series.

Recognizing that i) catch records (from the artisanal fisheries in particular) may be substantially biased, and ii) there are species composition concerns, there have been requests from the working parties to consider alternative catch series. However, developing alternative plausible catch series is a non-trivial undertaking for the secretariat, and no concrete proposals on how to do this have been forthcoming to date. Because of this, we focus on alternative CPUE interpretations at this time.

The alternative CPUE interpretations explored here have not been integrated into the most recent OM grid, but are presented to help deliberations focused on improving the yellowfin assessment, which will in turn have follow on implications for the OM. A single model was fit for each, i.e. with the minimum required modifications to the reference case assessment (i.e. without consideration of minimization sensitivity that we know can be substantial in the YFT OM). These models are compared with respect to the Fu et al (2018) reference case assessment (labelled "ref2018"), and were conducted using the same software (SS3.24z). An important element that is not addressed here is the effect of the tags. In model formulations resembling the reference case assessment, inclusion of the tags tends to lead to a more pessimistic assessment (e.g. Figure 14). Ignoring the interaction between CPUE and tags tags in the discussion that follows might overstate the importance of the CPUE series.

The ref2018 CPUE series (with regional-scaling factors applied) are shown in Figure 20. Table 3 compares some key features of the models explored, which are explained in the subsequent text.

Two additional models were added here to examine the effect of assuming the Ricker stock-recruit function as per the 2018 WPTT/WPM robustness scenario request.

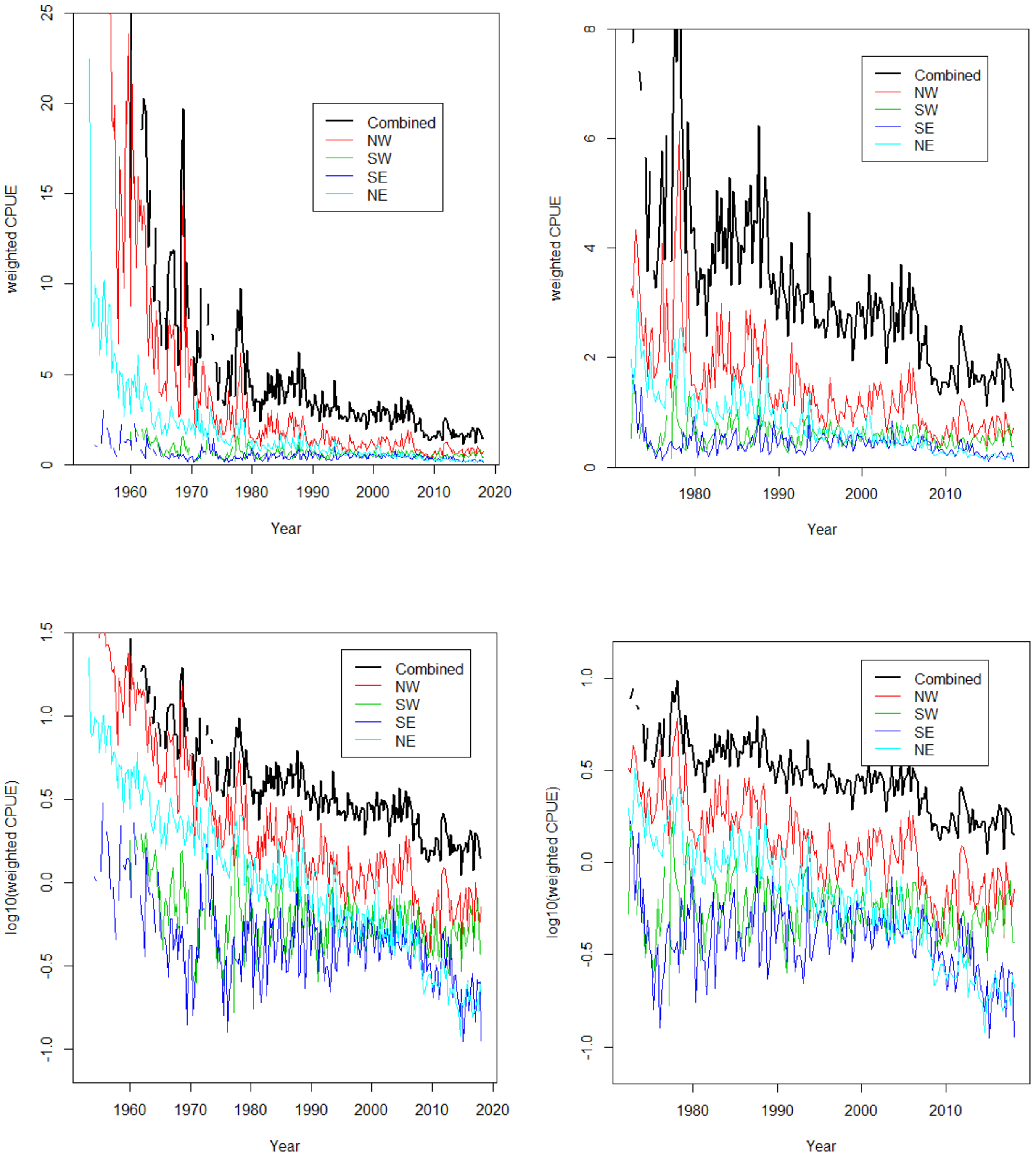


Figure 20. Standardized longline yellowfin CPUE, as adopted in the Fu et al (2018) reference case assessment, by region, and combined (individual series weighted by the regional scaling factors and summed). Top panels are linear-scale, bottom panels are log-scale. Left panels start in 1954, right panels start in 1972 (as used in the reference case assessment).

Table 3. Key outputs from the one-off exploratory model runs.

Model	MSY	B(2017)/B(MSY)	*Likelihood (group)	**Numerical Issues
ref2018 Reference case assessment	352	0.73	9360.7 (A)	
H0.1972	356	0.73	9388.5 (A)	Catch penalty = 0.03
Hest.1972	353	0.73	9355.5 (A)	Okay
H0.1.1972	366	0.78	9373.0 (A)	Okay
H0.2.1972	400	0.91	9440.4 (A)	Convergence fail
Hest.1954	350	0.68		Convergence fail
Hest.1987	363	0.78		Catch penalty = 0.04
R1	447	1.03	6031.9 (B)	Okay
R1b	434	0.98	5818.2 (B)	Okay
PSq0	330	0.83	9325.3 (C)	Okay
PSqEst	327	0.99	9236.2 (C)	Okay
PSHEst				Convergence fail
Ricker (h=0.8)	244	0.45	9413.5 (A)	Okay
Ricker (h =1.69)	372	0.67	9358.2 (A)	Convergence marginal

*reported likelihood values (maximum posterior density at the minimum identified) are only comparable within the same LLH group (i.e. those fitting to the same data). Missing values means that there is no useful comparison in the table.

** *Okay* indicates "reasonable" numerical convergence (max. gradient < 0.015) and catch penalty < 0.001.

CPUE Hyperdepletion/Hyperstability

In the yellowfin tuna (and most other) stock assessments, it is usually assumed that standardized longline CPUE is proportional to selected abundance, and this provides the primary information on changing stock size. In the case of IOTC YFT (and many other tuna populations), there is strong evidence of hyperdepletion, i.e. the rate of decline of CPUE in the early stages of fishery development is faster than the rate of decline of the population. In the absence of hyperdepletion, the population would probably not be able to sustain the much larger catches observed in subsequent years. In the case of IOTC YFT, 1972 is accepted as the year that hyperdepletion stops. As far as we understand, this is a subjective historical decision based on visual inspection of the observed time series (though there may have been other considerations, e.g. in southern bluefin, a starting date was adopted based on the year that the Japanese fleet started to achieve broad and relatively consistent time/area coverage). Recognizing that the proportionality assumption may not be appropriate, we used the in-built Stock Synthesis feature to examine non-linearity in the relationship, e.g. $I = qN^H$, where the observed CPUE (*I*) is proportional to the selected abundance (*N*), raised to the power of *H*. *H* > 1 indicates hyperdepletion and *H* <1 indicates hyperstability (i.e. CPUE does not decline as quickly as abundance). The following scenarios were examined, with key summary statistics in Table 3:

- refYFT2018H0.1972 – H fixed at 0 to confirm the new Q parameterization is equivalent to the assessment – the difference was negligible as expected.
- refYFT2018Hest.1972 – H estimated with the CPUE time series 1972-2017 – the model estimated trivial hyperdepletion, H = 0.0170. The total biomass trend was very similar to the reference case, though there was a moderate redistribution of biomass among areas.
- refYFT2018 Hest.1954 – H estimated with the CPUE time series 1954-2017 – the model estimated moderate hyperdepletion (H = 0.119, though did not converge adequately). The results attained indicate that this approach does not resolve the early hyperdepletion problem (Figure 22). Whatever non-linear relationship exists does not appear to be consistent before and after ~1972. The model estimates very little contrast in biomass prior to the 1980s, which is consistent with the large catches starting in the 1980s (Figure 23). We do not consider the alternative interpretation to be very plausible, i.e. yellowfin biomass could have been extremely high at the beginning of the fishery due to high recruitment that has not been observed since.
- refYFT2018Hest.1987 – H estimated with the LL CPUE time series 1987-2017. The model estimated trivial hyperstability, H= 0.014, similar to the period 1972-2017. The consistency of the results starting in 1972 and 1987 might provide some reassurance that missing operational data from the early part of the LL CPUE data history may not be very important for current stock status estimates.
- refYFT2018H0.1.1972 – H fixed at 0.1 (refYFT2018H0.2.1972 tested H = 0.2, but failed to converge). Since CPUE is highly informative to an assessment, it may not be reasonable to expect that the other data are sufficiently informative to provide information about *H* (otherwise estimation would probably be common practice). To the extent that we believe the objective function values, it appears that H = 0.1 is a better fit than H = 0. However, the assessment inferences were not substantially different between H = 0, H = 0.1 and the estimated value H = 0.017.

Overall, these results do not suggest that a non-linear relationship between abundance and observed longline CPUE will be very useful or influential to the yellowfin assessment. It is also worth noting that estimating H appears to greatly increases the minimization time (many parameters may be highly correlated with H).

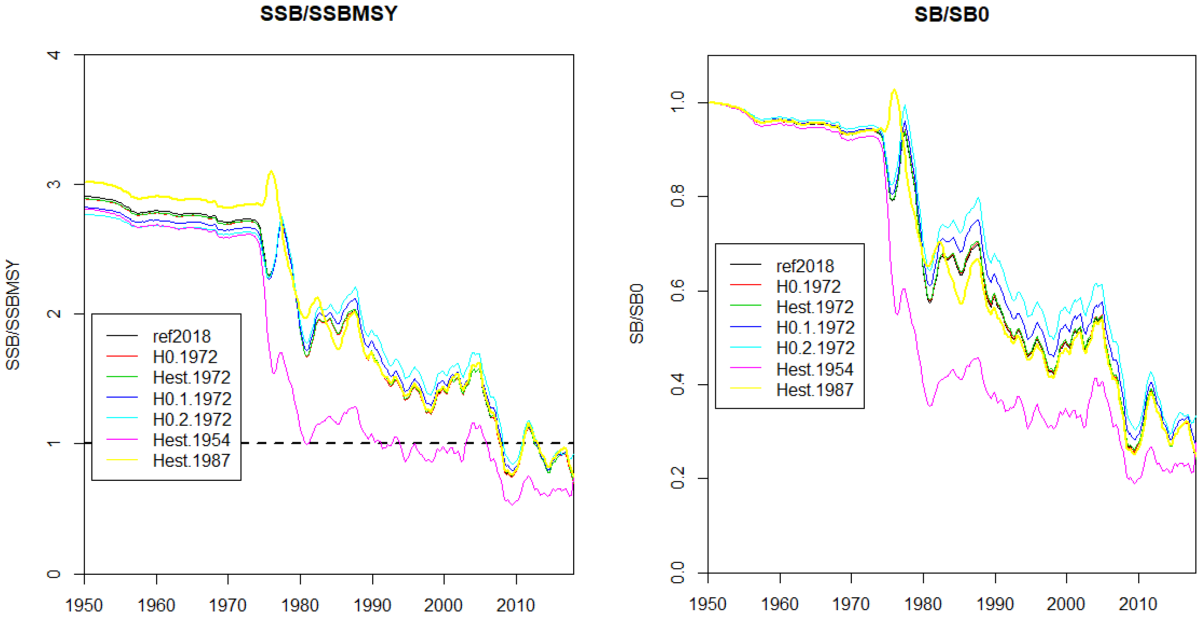


Figure 21. Biomass time series for models with alternative CPUE interpretations related to longline hyperstability.

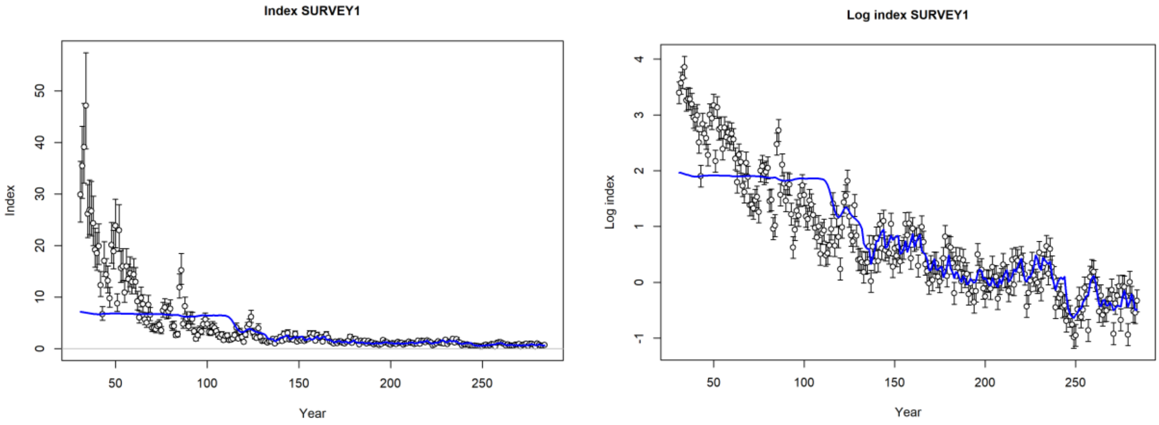


Figure 22. Failure to fit the early (1954-1972) LL CPUE series using estimated hyperdepletion parameter.

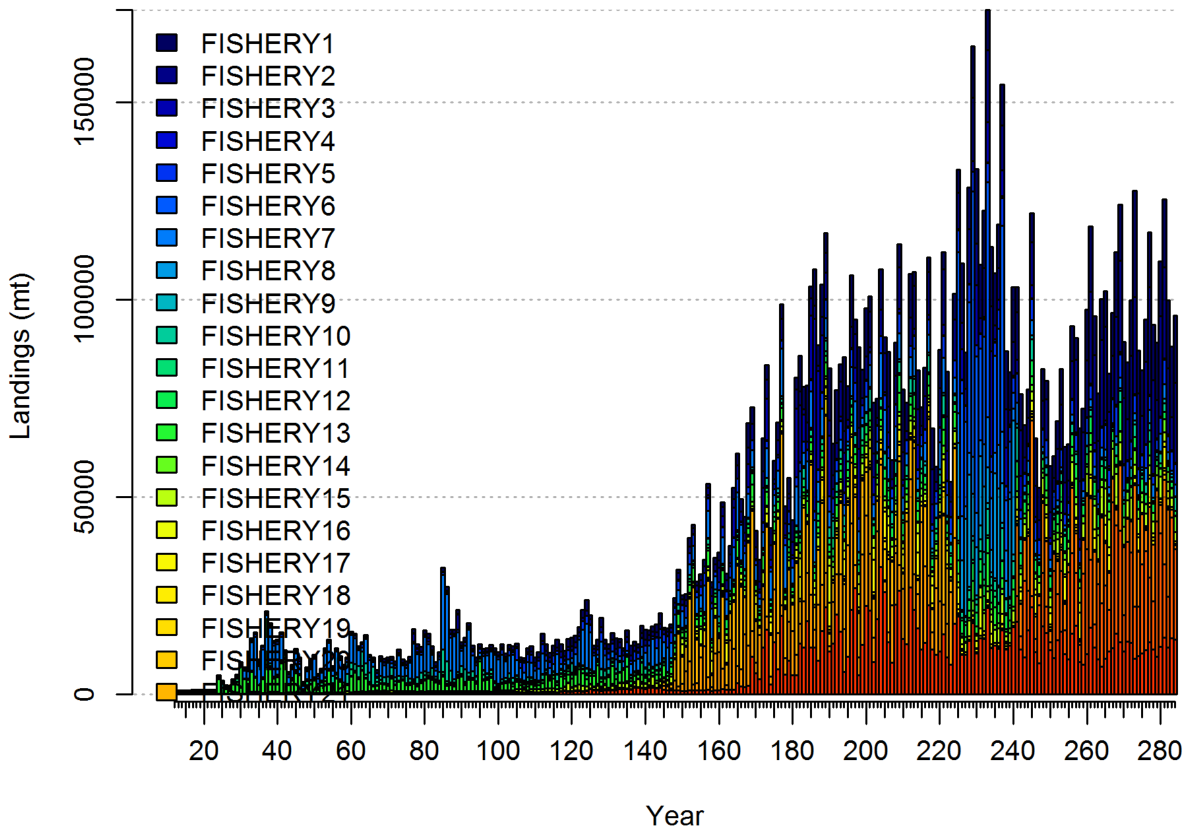


Figure 23. Indian Ocean Yellowfin catch time series. The large increase in catches starting around YrQtr 150 is ~1982.

Simplified spatial structure

Some or all of the OM and assessment model configurations struggle to remove the recently observed catches. This could be because of unrealistic movement assumptions that impact only a small subset of quarter/region strata (probably combined with poor selectivity, age/length, and/or M assumptions that do not leave a sufficient number of large fish in the population to be caught). Effort spent trying to fit the disaggregated data may be losing track of the big picture. e.g. the CPUE likelihood terms are structured such that all areas are given equal weight in the objective function, even though some regions represent more of the population, and the quality of the standardization presumably differs with the extent and consistency of fishery operations. It is perhaps notable that the area-weighted aggregate longline CPUE does not suggest much trend over the past decade (Figure 20). The following runs focused on fitting the aggregate CPUE trend:

- refYFT2018R1 - Reduce/Remove the spatial structure
    - movement fixed at high rates between adjacent regions, i.e. such that differential relative depletion among regions must be negligible (environmentally-linked movement is removed).

  - CPUE is only fit as an aggregate in region 1 (it should not matter which region is used if mixing rates are very high). Shared catchability constraints among regions are relaxed (and should be uninformative).
  - The region-specific CPUE series were retained with an insignificant weighting, for the purposes of visualizing the model fit.
  - Shared longline selectivity among regions is relaxed.
  - Tags are removed because the tag mixing assumptions are definitely inappropriate at the basin scale.
- refYFT2018R1b – as above, except recruitment is uniformly distributed in all areas and movement rates are uniform among all adjacent regions. i.e. Of the two, this is the preferred analogue of a spatially-aggregated model.

The two models above have very similar aggregate biomass dynamics estimates and fit to the CPUE (Figure 24 and Figure 25, key summary statistics in Table 3). The fit to the CPUE is very good except for the first couple years, which may again relate to the unresolved issues in the early time series. The first model has differing absolute biomass by region, but relative CPUE trends are almost the same. The second model essentially has identical biomass in all regions.

These are the most optimistic models of those explored here, however, it remains unclear the extent to which this is simply a consequence of removing the spatial structure, or removing the tags (i.e. it has been recognized that the tags generally have a pessimistic influence on the YFT assessment, and the tag assumptions are questionable).

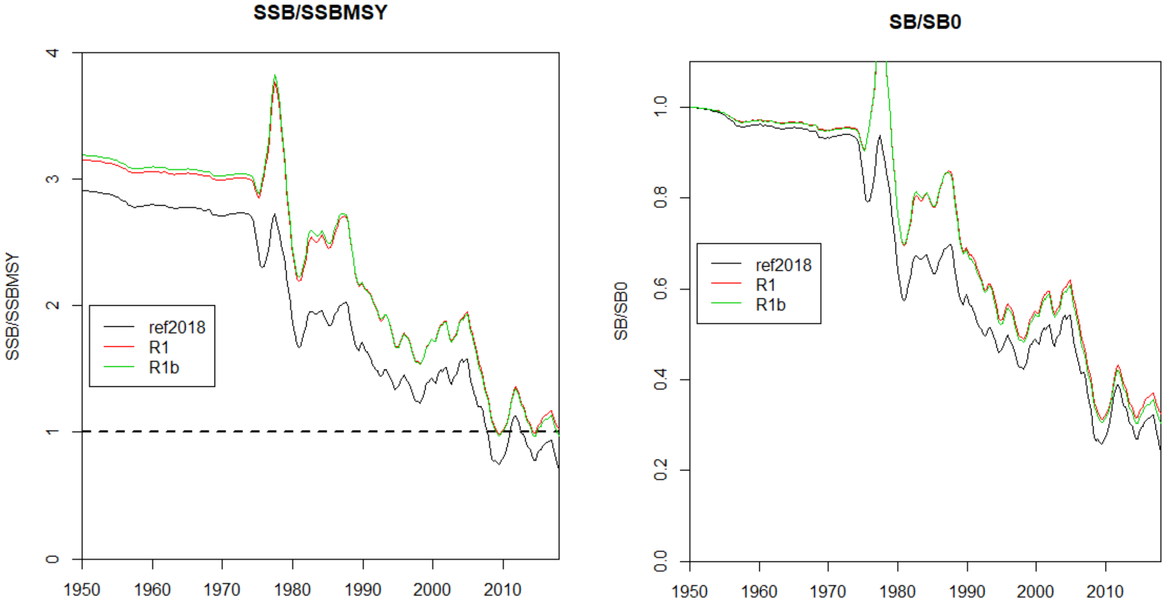


Figure 24. Biomass trajectories for exploratory spatially-aggregated models.

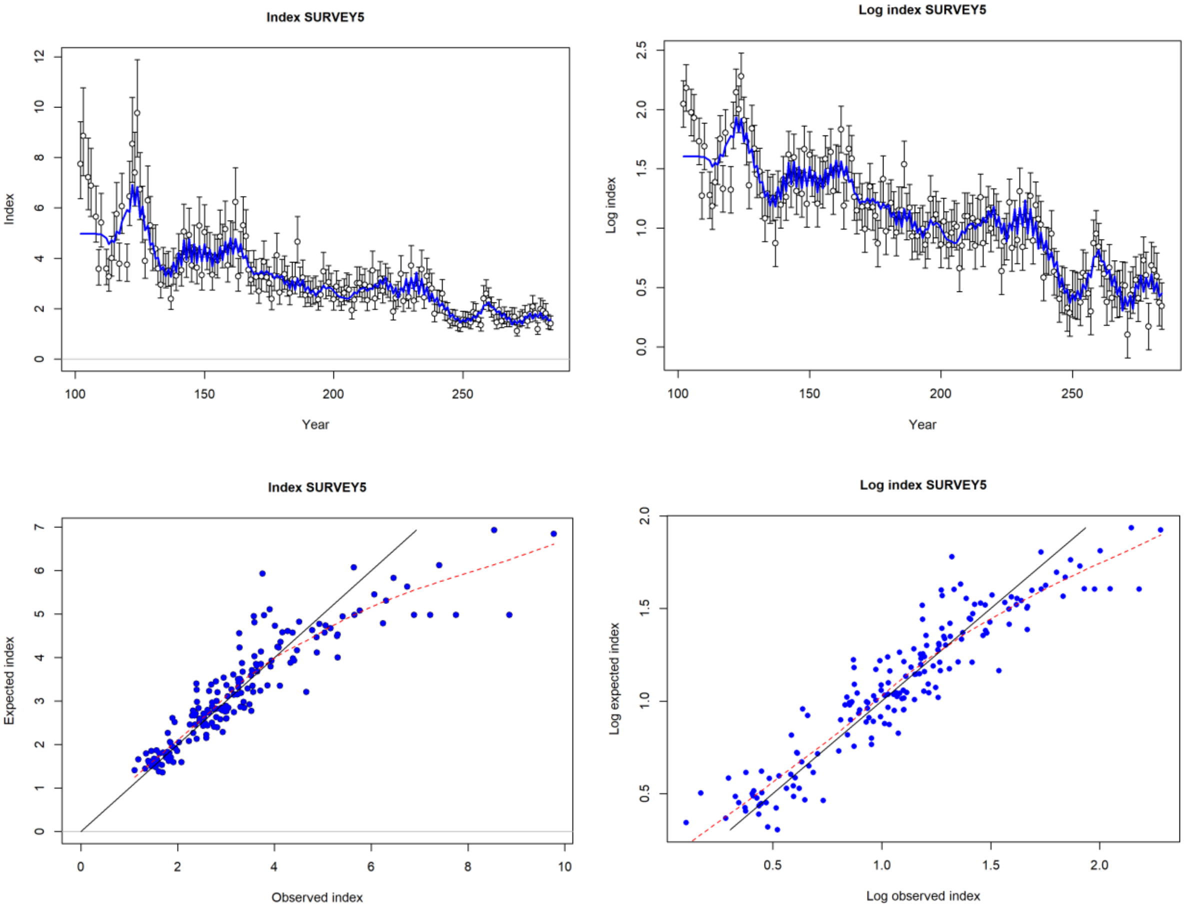


Figure 25. Model refYFT2018R1b fit to the aggregate longline CPUE series for the model with aggregate spatial dynamics (implemented as extremely high and uniform mixing).

Purse Seine CPUE

WPTT (2018) requested the testing of free school PS CPUE (Figure 26) in the yellowfin operating models as a robustness scenario. This standardized CPUE series was still considered to be a work in progress. There is a generally increasing PSFS CPUE trend over the past three decades, despite the large increase in total yellowfin catches during that period, and the general perception that populations would usually decline in these sorts of situations. The WPTT suggested adopting a 1% per year catchability trend for this time series. However, there was no scientific justification for the 1% assumption, and whatever assumption is imposed will presumably be the strongest driver of the assessment (depending how it is weighted relative to the LL CPUE). The effect of catchability trends from 0 - 4 % (per year, compounded) are shown in Figure 26, spanning a broad range from dubiously optimistic to pessimistic. Rather than arbitrarily choosing a catchability trend option, we tried to find a more quantitative approach. We attempted to include the PS FS CPUE in a manner that potentially addresses what we consider to be the highest concern identified for the LL CPUE - the effect of Somalian piracy on catchability. This involved the following assumptions:

- Piracy affects the LL fishery operations in region (1) only (NW), starting in 2007 and persisting to the end of the time series. The NW LL CPUE is assumed to be informative up to 2006 and given (essentially) zero weight 2007-2017. The other LL CPUE series are assumed to be unaffected by piracy.
- Standardized PS FS CPUE is assumed to have a continuously increasing catchability trend of unknown magnitude from 1986-2017.
- Standardized PS FS CPUE is included in the model with a high weighting (CV = 0.1). We believe this to be unrealistic precision, and is only intended to magnify the effect of the PSFS CPUE for exploratory purposes.
- The PSFS catchability trend is estimated using an environmental-link to catchability in stock synthesis, Q(t) = Q *exp(env_link * env_data(t)), and a dummy environmental time series (representing a 1% per year trend compounded quarterly). i.e. the effective PSFS CPUE series should resemble something in the range shown in Figure 26 depending on the estimated link parameter.
- PSFS selectivity is linked to fishery 21 (note that the NW PSFS fishery does have changes in selectivity estimated in the early 2000s, which are ignored here).

refYFT2018PSItrend0 – This model included the standardized PSFS series without the catchability trend estimated (accidentally). All of the CPUE series fit better than had been expected (Figure 27). The NW LL CPUE 2006-2017 model predictions are systematically higher than the observations, but not as much as might have been expected. The PSFS CPUE predictions are lower than observations for ~3 years in the early 2000s. Overall, the model is quite pessimistic, despite the influential and optimistic PSFS series.

refYFT2018PSItrendEst – as above, except the PSFS CPUE catchability trend was included as a free parameter, and estimated at 0.71 % per year. The PSFS CPUE fit is clearly better with the catchability trend estimated, as would be expected. However, B(2017)/B(MSY) is higher with the increasing catchability trend, which is not what we would have predicted. This could indicate unstable minimization, or subtle influences throughout the time series.

refYFT2018PSHest – as above, except instead of estimating a time trend in PSFS catchability, a hyperdepletion parameter was estimated (PSFS CPUE only). A few attempts to fit this model resulted in either significant hyperstability (almost flat predicted CPUE) or numerical failures.

Key summary statistics are compared in Table 3. These results suggest that the PSFS CPUE is not as incompatible with the LL CPUE as had been anticipated. However, it is also not clear whether, or how, to include this information in the assessment. Is it reasonable to expect that catchability has increased uniformly for the past 3 decades in a manner that the standardization cannot account for? It is also not clear that including the PSFS CPUE data in this way resolves the fundamental assessment issue of concern. i.e. The stock is estimated to not be as depleted as the reference case assessment, and the catch penalty was not invoked, however, the stock productivity (MSY) is the lowest observed in Table 3 (aside from the Ricker stock-recruit test).

The WPTT has also periodically considered whether the PSFS fishery should be disaggregated somehow to represent the two very different selectivities that seem to be operating (e.g. aggregate size composition has two modes, but some observations have only one mode or the other, as shown in Figure 29). It would seem natural to expect that pooling two very different types of sets would have implications for catchability.

Summary observations from the alternative CPUE Interpretation explorations:

- These results are all one-off tests relative to the reference case 2018 YFT assessment. We do not know the extent to which the results are affected by minimization sensitivity, or how they would interact with other plausible assumptions in the current OM, or proposals arising from the YFT assessment review process. Only a superficial visual inspection of model fits to data was undertaken to identify obvious model failures.
- The longline CPUE-abundance hyperdepletion relationship that is known to exist in the early part of the fishery does not appear to extend across the full history of the stock, at least not in the simple manner that SS3 is capable of representing. If estimated as a free parameter, the relationship is estimated to be almost proportional from 1972-2017. It remains unclear that 1972 is the best choice for the start year for assuming that the relationship is proportional, but the assessment inferences (and hyperdepletion estimates) are similar regardless if the LL CPUE series starts in 1972 or 1987.
- The spatially-aggregated model is somewhat more optimistic than the disaggregated reference case model, in terms of current status (depletion) and productivity (MSY). At the aggregate level, there may not be an irreconcilable conflict between catch and CPUE. The disaggregated model is more satisfying from the perspective of attempting to explain abundance trends that differ by region, and including the tag dynamics. However, it is worth considering whether the extra assumptions in the disaggregated model (movement, recruitment distribution, tag mixing, etc) are appropriate and lead to improved inferences about the big picture (e.g. there is a general principle that the optimal model complexity

should strike the right balance between model bias (overly simple models are biased because they fail to describe important system features) and variance (complicated models might be able to represent more important system features, but lack the data to estimate the additional parameters with adequate precision). Unfortunately it is hard to know where the optimum is located.

- The standardized LL and PSFS CPUE trends conflict in the NW region, though not to the extent that we were expecting. If one assumes that i) LL CPUE is reliable from 1972-2006 in the NW, ii) LL CPUE is reliable from 1972-2017 in the other regions, and iii) PSFS catchability has increased at a continuous rate from 1986-2017, then a PSFS catchability increase of ~ 0.71% per year appears to largely reconcile the LL and PSFS CPUE series. When the PSFS CPUE series is included in this manner, the recent depletion level is estimated more optimistically, but the productivity (MSY) is more pessimistic.

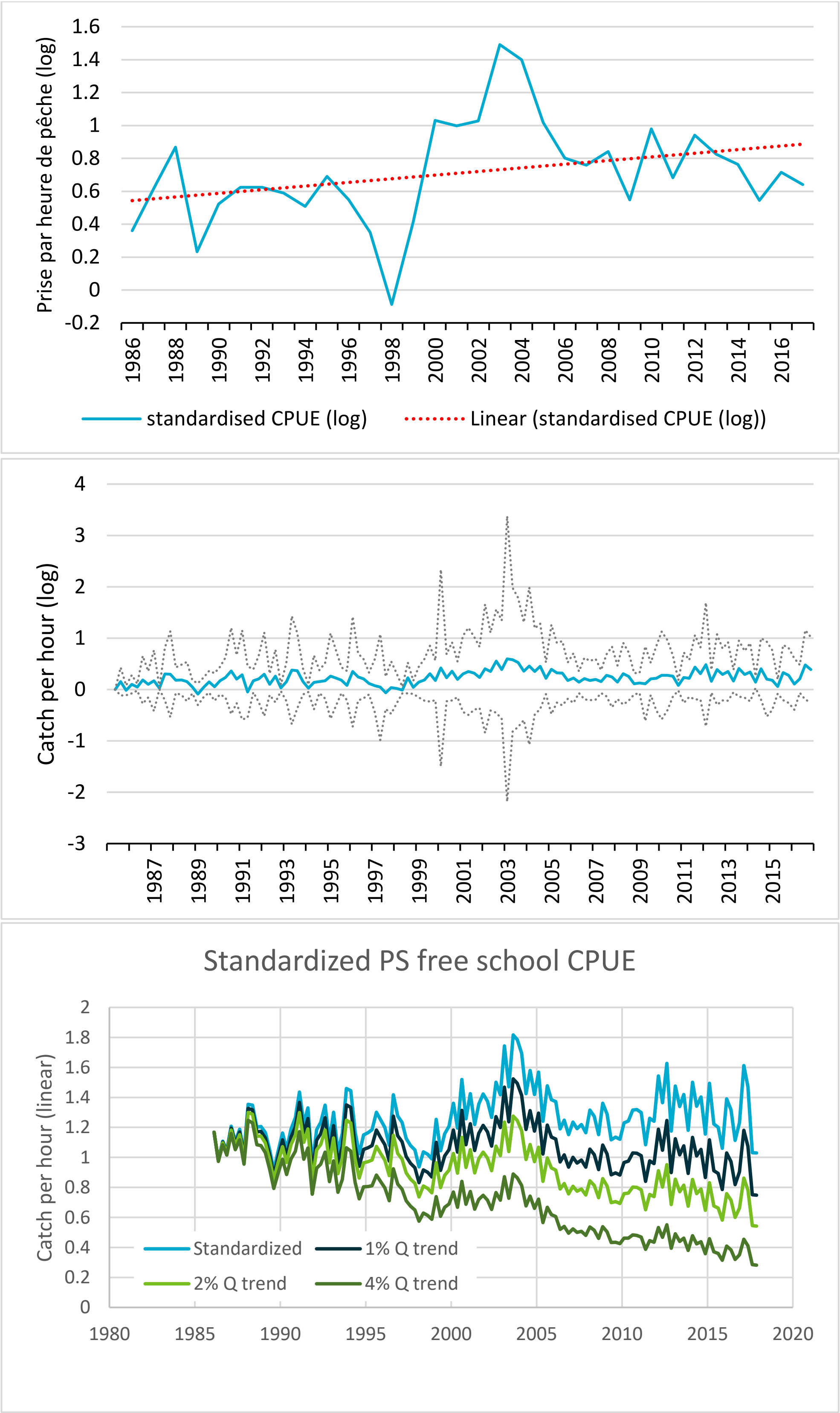


Figure 26. Standardized Purse seine free-school CPUE time series (provided by Francis Marsac, IFREMER, 12Nov2018), with alternative catchability trend assumptions compared in the bottom panel.

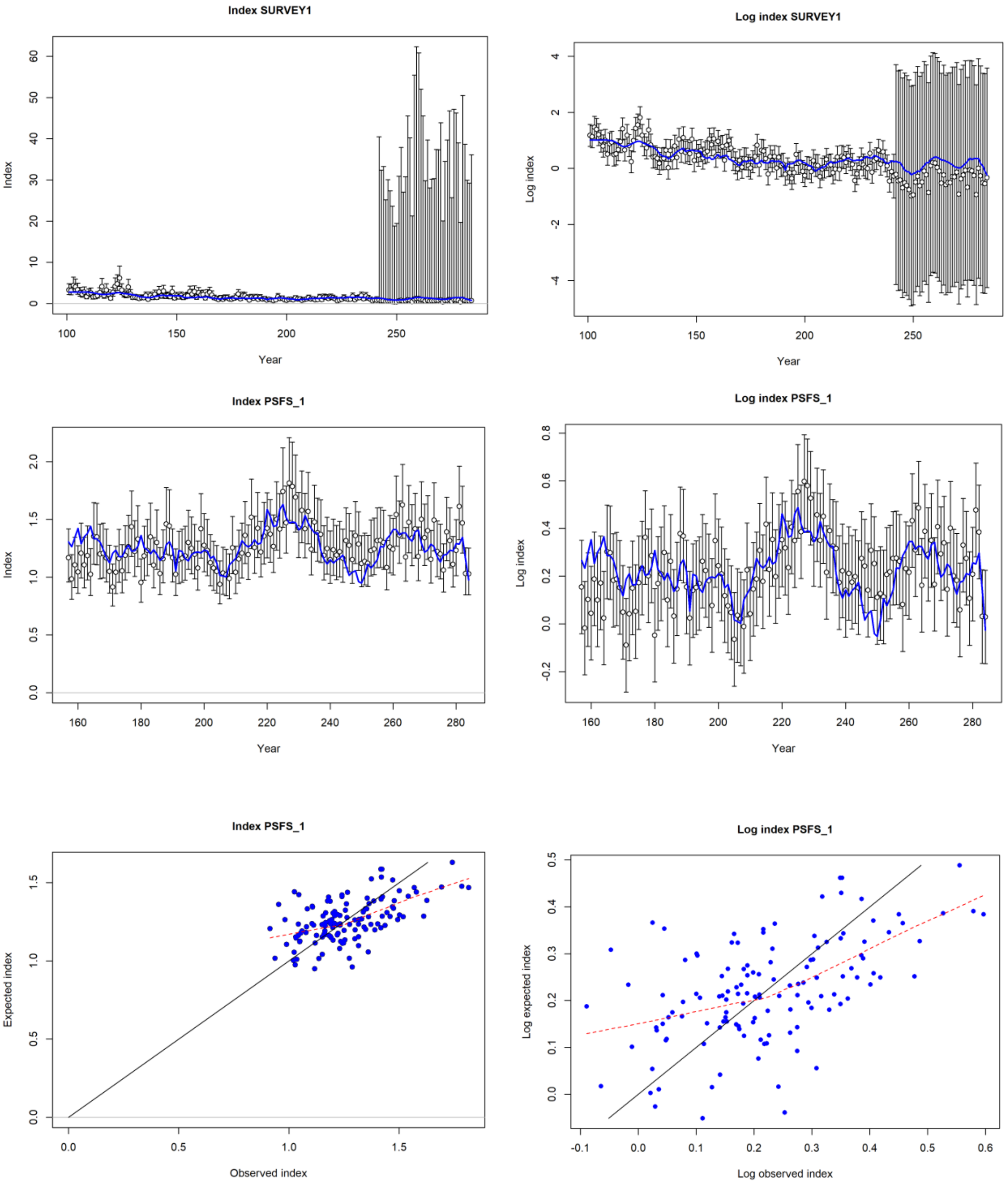
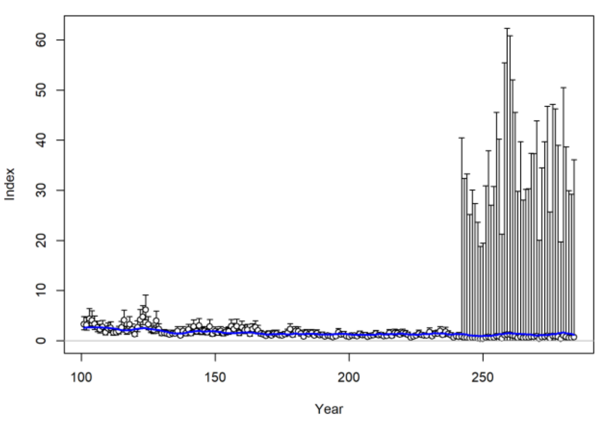
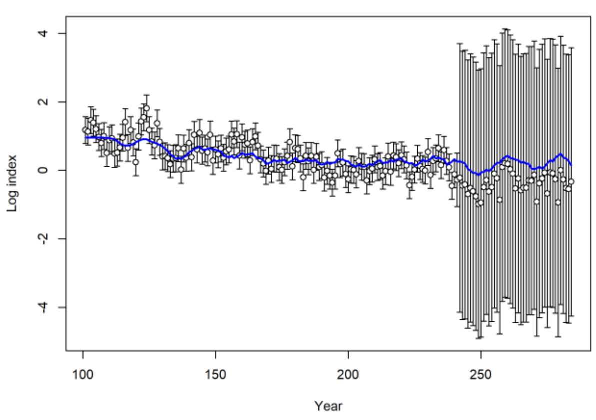
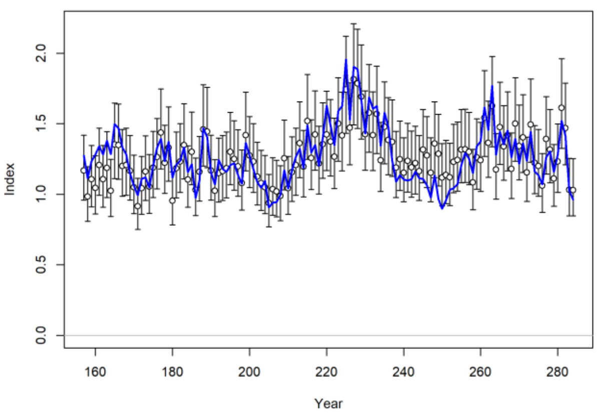
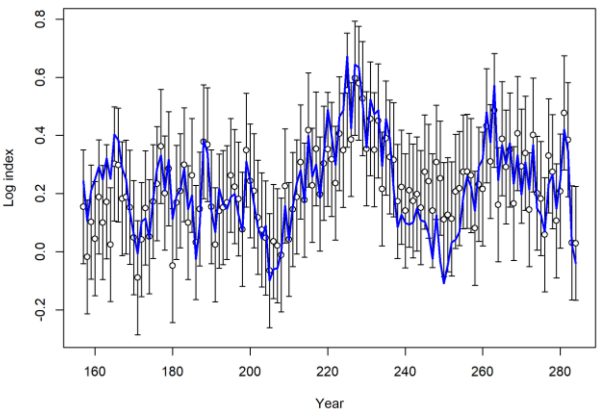


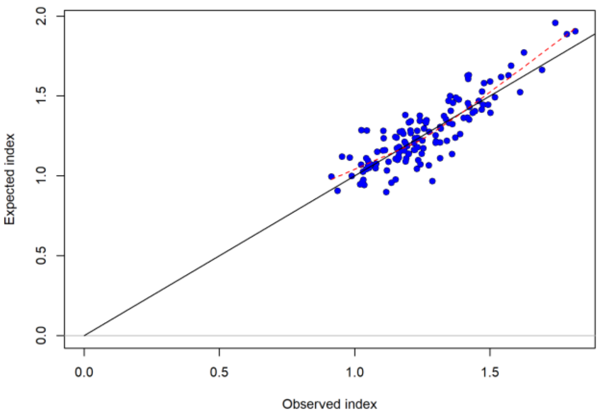
Figure 27. Model refYFT2018PSItrend0 fit to the Area 1 (NW) LL CPUE (top) and PS Free school (middle-bottom) (linear-scale left and log-scale right). NW (region 1) LL CPUE weight is negligible 2007-2017, no catchability trend is estimated for any index.


Index SURVEY1


Log index SURVEY1


Index PSFS_1


Log index PSFS_1


Index PSFS_1

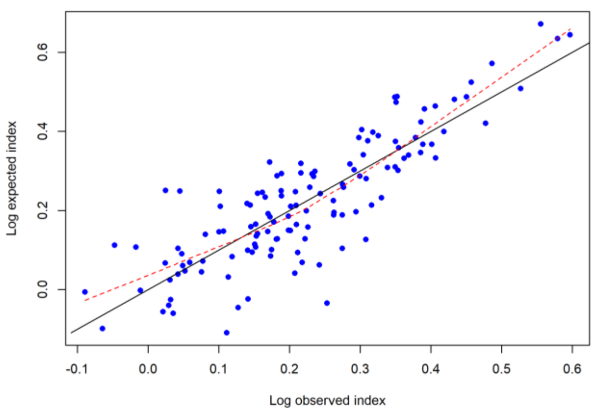

Log index PSFS_1

Figure 28. Model refYFT2018PSItrendEst fit to the Area 1 (NW) PS Free school (top) and LL CPUE (bottom) (linear-scale left and log-scale right). NW (region 1) LL CPUE weight is negligible 2007-2017. PSFS CPUE catchability trend is estimated at 0.71% per year.

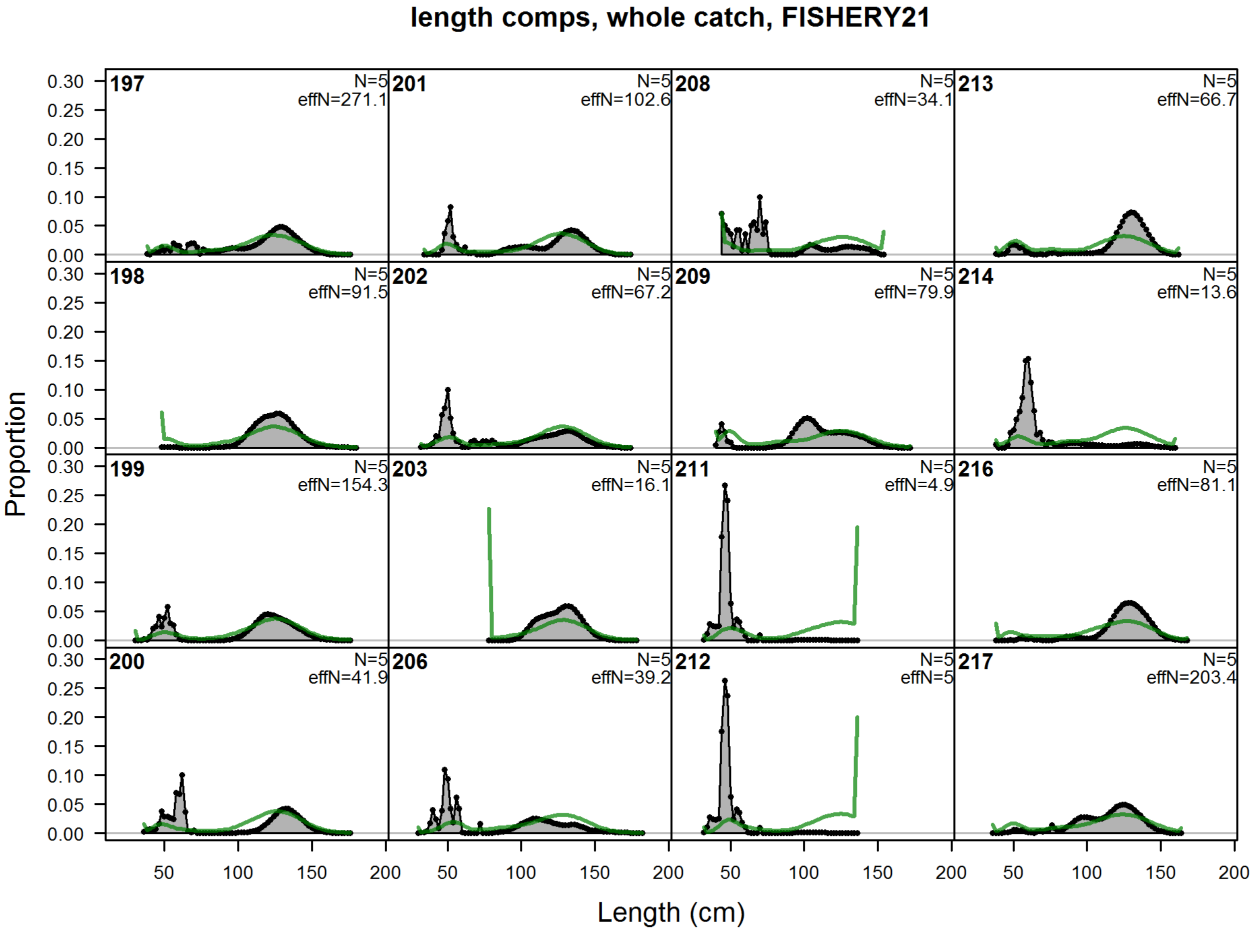


Figure 29. Example of size composition fit in the PSFS fishery (reference case assessment), noting that the fishery often seems to operate in two very different ways, that cannot be easily reconciled with a single selectivity function, e.g. In year-quarter 203, only large fish are landed, while in year-quarter 211, the catch is almost entirely small fish.

Ricker Stock Recruit function

Models *refYFT2018Ricker* and *refYFT2018RickerhEst* refit the reference case stock assessment with a Ricker stock-recruit function (h=0.8 fixed, and h=1.69 estimated, respectively). The estimated stock recruit functions and recruit deviates are shown in Figure 30.

The model with h=0.8 represents a considerably worse fit than the reference case (in terms of the likelihood) and was considerably more pessimistic in terms of stock status (lowest MSY and B(2017)/BMSY in Table 3 by a substantial margin). The estimated recruitment function does not show the Ricker bend (i.e. dome shape with decreasing recruitment as spawning biomass increases to the right of the peak). It appears that adopting the Ricker curve with h=0.8 would be largely analogous to entertaining a Beverton-Holt curves with steepness lower than 0.7 (the current lower value in the reference set).

The Ricker model with h estimated represents a very similar (slightly better) fit than the reference case, with similar MSY (slightly higher) and B/BMSY (slightly lower). We are not familiar with the interpretation of the steepness parameter (h = 1.69) in the context of a Ricker function (e.g. h > 1.0 is not defined in a Beverton-Holt function). The estimated SR function demonstrates a mild Ricker bend, but qualitatively does not appear to deviate substantially from what the Beverton-Holt function would predict.

There is some systematic lack of fit in the recruitment deviates for both models (slight increasing trend in the first 25 years of estimated deviates). But qualitatively, it does not appear that a Ricker curve would represent a fundamentally different challenge for an MP than a Beverton-Holt function, over the range of spawning biomass and recruitment that are likely to be relevant in the medium term for IOTC yellowfin. Including the Ricker curve in formal robustness tests would require further consideration of the appropriate steepness parameters, and we doubt that it would offer new insight to the MP selection process.

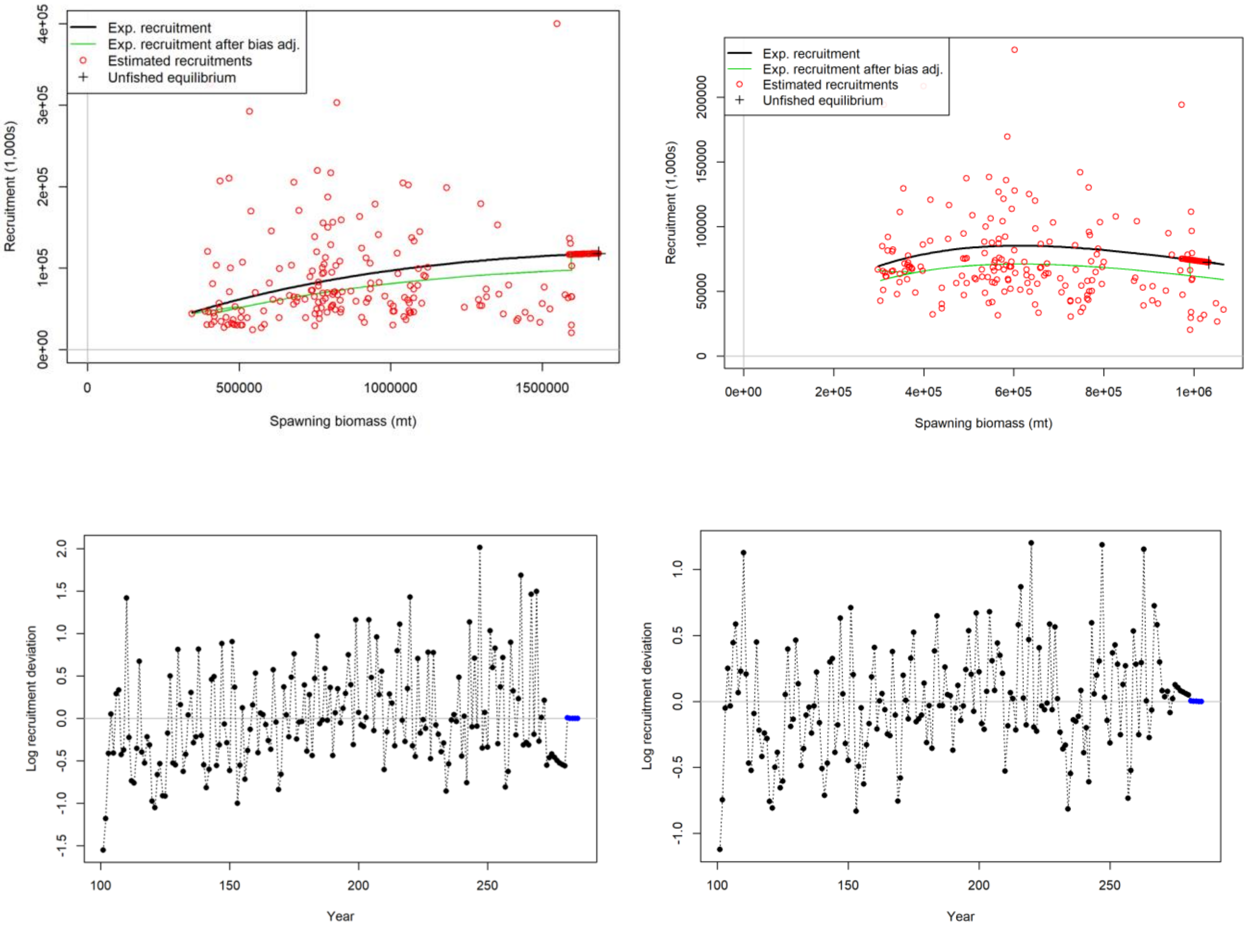


Figure 30. refYFT2018Ricker (left) and refYFT2018RickerhEst (right) stock and recruitment functions (top) and time series of recruitment (bottom).

7 Discussion

We continue to welcome feedback on any aspect of the OM formulation, software or MSE workplan. It should be recognized that a number of subjective decisions need to be made in an MSE process that might have important implications for MP evaluation and selection (and fishery management performance if the MP is adopted). Ideally, MSE in an RFMO context should be undertaken with the active engagement of many stakeholders, including at the technical level, to represent the broad scientific experience within the working parties. We continue to encourage other member scientists to download the source code, and scrutinize OM assumptions, performance characteristics and MP formulations, and present alternative views where appropriate (please contact the authors ahead of time, to ensure that the latest version of the code is available from github).

We highlight the following priority points for feedback/endorsement for the YFT MSE to move forward in the next iteration:

YFT reference set OM

- Given the tests done for both yellowfin and bigeye, we are reasonably comfortable with using a highly fractional (main effects) OM ensemble of 50-150 models to approximate the MP performance that would be expected from a much larger grid of models that explicitly quantifies higher level interactions.
- The numerical instability in the current YFT assessment configuration is substantial and does not give us much confidence in the results obtained from any individual model. Fortunately, there did not appear to be any substantial bias among model ensembles that would undermine MP performance evaluation and selection if sub-optimal models were used (of course we can never be certain that the best models that we identified represented global solutions, but they were markedly better and different from the worst solutions identified). However, if the numerical stability cannot be improved and target grid size can be kept reasonably small, it is probably worth retaining some jitter analysis as standard practice in OM specification.
- We will be looking to compare these results with the 2019 YFT assessment review process and adopting relevant insights, particularly in relation to:
    - Spatial Structure and fixed parameter assumptions
    - Tagging data (inclusion/exclusion or weighting)
    - CPUE data (particularly observation error assumptions)
    - Improvements to numerical stability
    - Potentially influential parameter bounds and priors
    - Model diagnostics, particularly those that can be applied in an automated context, and how they might be used to improve model weighting (including model retention/rejection).

YFT robustness tests:

We note that the term robustness test is often used in two ways i) "likely" uncertainty options that are worth testing to see if they affect MP performance, in which case they should be added to the reference set, and ii) "less likely" but plausible and troubling scenarios which are used to test MPs independent of the reference set OM. Robustness tests of the first sort might best be covered under the dot points in the reference set above. Given that we are already facing the situation of too much uncertainty in the reference set grid (with the majority of models being discarded with implausible results), it is not clear what will be gained from simply expanding the yellowfin uncertainty grid further.

Additional robustness scenarios that require modifications to the conditioning and projection code, should be considered and specified carefully. i.e. Do they represent genuine concerns arising from the stock assessment deliberations? Can they be meaningfully quantified? Do they need to be tested as a full dimension within the reference case grid, or can they be defined by a representative subset of dimensions?

In the interest of clear communication, its worth considering which robustness tests should be presented to the TCMP. Unless the tests offer additional information that will be useful in helping the TCMP/Commission to select among MPs, they may create additional confusion.

8 References

Fu, D, Langley, A, Merino, G, Urtizberea, U, 2018. Preliminary Indian Ocean yellowfin tuna stock assessment 1950-2017 (stock synthesis). IOTC–2018–WPTT20–33.

IOTC. 2019. Report for the 23rd session of the Indian Ocean Tuna Commission. IOTC-2019-S23-RE

Jumppanen, P, Kolody, D. 2018. User manual for IOTC Yellowfin and Bigeye Tuna MSE Software. https://github.com/pjumppanen/niMSE-IO-BET-YFT/.

Kolody, D, Jumppanen, P. 2016. IOTC Yellowfin and Bigeye Tuna Management Strategy Evaluation: Phase 1 Technical Support Project Final Report. Indian Ocean Tuna Commission Working Paper IOTC-2016-WPTT18-32.

Kolody, D, Jumppanen, P. 2019b. Indian Ocean Yellowfin Tuna MSE Update March 2019. IOTC–2019–WPM10-INF02.

Kolody, D, Jumppanen, P. 2019c. IOTC Bigeye and Yellowfin Management Procedure Evaluation update Oct2019. IOTC-2019-WPM10-11.

Kolody, D, Jumppanen, P. 2019d. IOTC bigeye Management Procedure evaluation update June 2019. IOTC-2019-TCMP03-10.

Kolody, D, Jumppanen, P. 2019e. IOTC yellowfin Management Procedure evaluation update June 2019. IOTC-2019-TCMP03-11.

Methot, R.D., Wetzel, C.R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research 142 (2013)* 86–99.

TCMP. 2019. Report of the 3rd IOTC Technical Committee on Management Procedures. IOTC-2019-TCMP03-R[E].

WPM. 2018. Report of the 9th Session of the IOTC Working Party on Methods. Beau Vallon, Seychelles, 25-27 October 2018. IOTC–2017–WPM09–R[E].

WPM 2019. Report of the 8th Workshop on Management Strategy Evaluation in Working Party on Methods of Indian Ocean Tuna Commission. IOTC–2019–WPM10-INF01.

WPTT. 2018. Report of the 20th Session of the IOTC Working Party on Tropical Tunas. Seychelles, 29 Oct -= 3 Nov 2018. IOTC–2017–WPTT20–R[E].

CONTACT US

t 1300 363 400
+61 3 9545 2176
e csiroenquiries@csiro.au
w www.csiro.au

AT CSIRO, WE DO THE
EXTRAORDINARY EVERY DAY

We innovate for tomorrow and help improve today – for our customers, all Australians and the world.

Our innovations contribute billions of dollars to the Australian economy every year. As the largest patent holder in the nation, our vast wealth of intellectual property has led to more than 150 spin-off companies.

With more than 5,000 experts and a burning desire to get things done, we are Australia's catalyst for innovation.

CSIRO. WE IMAGINE. WE COLLABORATE.
WE INNOVATE.

FOR FURTHER INFORMATION

Oceans & Atmosphere
Dale Kolody
t +61 6 6232 5121
e Dale.Kolody@csiro.au
w www.csiro.au

9 Attachment 1. Current (Sep 2019) State of the IOTC Bigeye reference set Operating Model for MP evaluations

State of the IOTC Yellowfin Reference Set Operating Model for Management Procedure evaluation Sep 2019

Dale Kolody (dale.kolody@csiro.au)

Paavo Jumppanen

CSIRO, Australia

Introduction

This document provides a brief description of the most recent state of the yellowfin tuna reference set Operating Model (OM) used for Management Procedure (MP) evaluation, *OmRefY19.4.500*. The documentation for the latest version of the MSE software, technical documentation, and series of project reports is publicly available from github https://github.com/pjumppanen/niMSE-IO-BET-YFT/. The iterative and sometimes circuitous decision process undertaken by the IOTC technical working groups and analysts to reach the current state of the OM are not described here. These may be found in various IOTC working papers, information papers and meeting reports, along with various model results and diagnostics that were used to guide the OM development process.

OmRefY19.4.500 was proposed by the MSE task force in March 2019, and represents a modest improvement to the OM used to provide MP evaluations for the TCMP in May 2019 (more extensive uncertainty grid structure and some minor bug/specification fixes). The definition and role of the robustness tests is continuously evolving and not described here.

Conditioning Software

This version of the OM is an ensemble of models conditioned using the *Stock Synthesis* assessment software version SS3.24z.exe (e.g. Methot and Wetzel 2013).

Projection Software

The projection software is available from https://github.com/pjumppanen/niMSE-IO-BET-YFT/. The population dynamics equations conform to fairly standard assumptions, and are fully documented in the technical reference (also on github).

Reference Case OM

The various models in the OM ensemble are derived from the reference case stock assessment (supplied by Dan Fu, IOTC secretariat, and defined in Fu et al (2018)). Key assumptions include:

- 4 regions (Figure 1) with age-dependent movement
- Quarterly dynamics (implemented with calendar quarters as SS model-years)
- 25 fisheries (Table 1) - 21 with some temporal variation represetned as independent fisheries
- Parameter estimation objective function includes
    - Total catch penalty (if some component of catch cannot be removed)
    - Standardized longline CPUE (one series per region)
    - Size composition data
    - Tags (down-weighted to be essentially excluded in some OM scenarios)

  - Recruitment penalties on deviations from stock recruit relationship and mean spatial distribution
  - Diffuse priors on all estimated parameters
- Estimated parameters:
  - Fishery selectivity (various functional forms, parameters shared among some fleets)
  - Longline catchability (in aggregate - regional scaling factors are used to scale relative density to relative abundance among regions)
  - Virgin recruitment
  - Recruitment deviations from the Beverton-Holt stock-recruit relationship, recruitment spatial partitioning among tropical regions (1 and 4) and deviations from the mean spatial distribution.
  - Juvenile and adult movement rates
  - Initial fishing mortality
- Modifications to the reference case for the base of the OM included removing the movement-environment link and relaxing some bounds (i.e. so that an arbitrary hard bound does not constrain parameter estimates).

OM Reference Set Grid

- Model structural and parameter uncertainty is introduced to the OM through the alternative assumptions listed in Table 2. Only the point estimates (maximum posterior density) of parameters and states from each model specification are retained for the OM.
- A fractional-factorial experimental design was used to calculate a subset of 1152 models, which would allow the estimation of all main effects and 2 way interactions in the context of a GLM (the full factorial grid with all interactions would require 4608 models).
- In recognition that the IOTC yellowfin assessment model parameter estimates can be sensitive to initial starting conditions, minimization was repeated from randomly jittered starting conditions until either (i) successful minimization was achieved 3 times (maximum gradient of the objective function with respect to the estimated parameters <0.01) or (ii) 10 attempts were made without reaching 3 successful minimizations. Approximately 10% of models failed to converge.

OM Reference Set *OMrefY19.4.500*

- Within an individual model configuration, the version with the lowest objective function value (from the jittered minimizations) were retained (initially)
- The best fit models were subsequently rejected from the reference grid if:
  - Minimization unsuccessful (max. grad. >0.01)
  - SS3 Catch Penalty >1E-5 (i.e. model struggles to remove the observed catch, which is assumed to be related to the pessimistic retrospective patterns)
  - The aggregate annualized CPUE CV exceeded 0.3. This is an arbitrary value that affected a small number of models, most of which were associated with high MSY outliers.
- All retained models were subject to a qualitative comparison of simple diagnostics to identify outlier behaviour or polymodal stock status inferences (no major problems were

noted). The four most extreme models (highest and lowest depletion and productivity) were visually examined in more detail, without obvious evidence for blatant model failure.

- The OM reference set grid is fully balanced with respect to each factor, while the retained grid has 420 of the original models, with the factor level distribution shown in Table 4. Rejected models were disproportionately associated with options M06, i1 and gr1.
- Each SS model was assigned a plausibility weighting (to date, models have only been assigned a weighting of 0 or 1, such that all retained models are uniformly weighted). *OmRefY19.4.500* consists of 500 models randomly sampled (with replacement) from the grid of retained models.
- Key projection assumptions are summarized in Table 4 (values not presented conform to the SS model assumptions or estimates).

References

Fu, D, Langley, A, Merino, G, Urtizberea, U, 2018. Preliminary Indian Ocean yellowfin tuna stock assessment 1950-2017 (stock synthesis). IOTC–2018–WPTT20–33.

Methot, R.D., Wetzel, C.R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research 142 (2013)* 86–99.

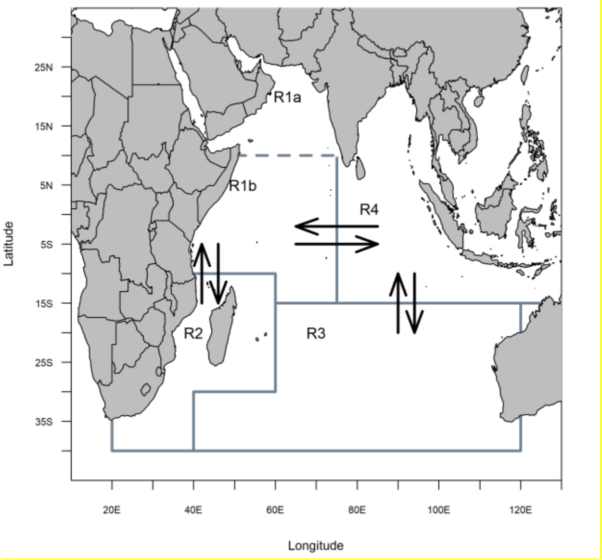


Figure 1. Spatial structure for the yellowfin tuna OM (figure from Fu et al. 2018).

Table 1. IOTC Yellowfin assessment fishery definitions.

Fishery	Definition	Region
1	Gillnet (GI)	1
2	Handline (HD)	1
3	Longline (LL)	1
4	Other (OT)	1
5	Baitboat (BB)	1
6	Purse-seine - free schools (FS) 2003-2006	1
7	Longline (LL)	1
8	Purse-seine - log schools (LS) 2003-2006	1
9	Troll (TR)	1
10	Longline (LL)	2
11	Longline (LL)	3
12	Gillnet (GI)	4
13	Longline (LL)	4
14	Other (OT)	4
15	Troll (TR)	4
16	Purse-seine - free schools (FS)	2
17	Purse-seine - log schools (LS)	2
18	Troll (TR)	2
19	Purse-seine - free schools (FS)	4
20	Purse-seine - log schools (LS)	4
21	Purse-seine - free schools (FS) pre 2003	1
22	Purse-seine - log schools (LS) pre 2003	1
23	Purse-seine - free schools (FS) post 2006	1
24	Purse-seine - log schools (LS) post 2006	1
25	Longline - fresh tuna (LL)	4

Table 2. Assumptions in OMrefY19.4.500 Stock Synthesis conditioning. Bold indicates the reference case assumption from the Fu et al (2018) assessment.

Abbreviation	Definition
	Stock-recruit function (*h* = steepness)
h70	Beverton-Holt, *h* = 0.7
h80	**Beverton-Holt, *h* = 0.8**
h90	Beverton-Holt, *h* = 0.9
	Natural mortality multiplier relative to reference case M vector
M10	**1.0**
M08	0.8
M06	0.6
	Tag recapture data weighting (tag composition and negative binomial)
t0001	λ = 0.001
t10	**λ = 1.0**
	Growth curve
gr1	Fonteneau (2008)
gr2	Dortel et al. (2015)
	Assumed longline CPUE catchability trend (compounded)
q0	**0% per annum**
q1	1% per annum
	Tropical longline CPUE standardization method
iH	**Hooks Between Floats**
iC	Cluster analysis
	Longline CPUE error assumption (quarterly observations)
i3	**σ_{CPUE} = 0.3**
i1	σ_{CPUE} = 0.1
	Tag mixing period
x4	**4 quarters**
x8	8 quarters
	Longline fishery selectivity
SL	**Stationary, logistic, shared among areas**
SD	Stationary, double-normal (potentially dome-shaped), shared among regions
	Size composition input Effective Sample Sizes (ESS)
ESS5	**ESS = 5, all fisheries**
CL75	ESS = One iteration of re-weighting from reference case model (fishery-specific), raised to the power of 0.75, capped at 100.

Table 3. Frequencies of reference set grid assumptions retained after convergence and plausibility criteria are applied. If all models were retained, each assumption would be equally represented (i.e. either 0.33 or 0.5, depending on the number of assumption levels). Assumption abbreviations are defined in Table 3.

Model Assumption (proportion represented in OMrefY19.4)		
h70	h80	h90
0.35	0.35	0.32
M06	M08	M10
0.24	0.43	0.33
t0001	t10	
0.46	0.54	
i1	i3	
0.28	0.72	
iC	iH	
0.53	0.47	
iR1	iR2	
0.52	0.48	
q0	q1	
0.56	0.44	
gr1	gr2	
0.25	0.75	
CL75	ESS5	
0.56	0.44	
x4	x8	
0.50	0.50	
SD	SL	
0.57	0.43	

Table 4. OM Projection assumptions in the yellowfin reference set and robustness sets. Reference set values not listed are identical to the model-specific conditioning assumptions/estimates. Robustness case values not are identical to the reference set except as noted.

OM	Projection assumption	Value
OMrefY19.4	Reference set OM	
	Initial population error CV (a = age in quarters)	0.6exp(-0.1a)
	Recruitment deviation penalty Recruitment deviation lag(1) auto-correlation (annual equivalents)	σ_R = 0.42 ρ_R = 0.21
	CPUE observation error CPUE observation error lag(1) auto-correlation (annual)	σ_R = 0.2 ρ_R = 0.5
	Multinomial Catch-at-length sample size (all fisheries)	100
	Selectivity stationary for all fisheries	
	Quota Implementation error	CV = 0
	First MP quota year	2021
	Assumed catches 2018-2020	409 Kt
	MP data lag (i.e. data from 2018 informs 2021 quota)	2 years*
	Quota allocation (average observed over)	2016-2017

*Results presented to TCMP 2019 erroneously had a 3 year data lag.