OCEANS AND ATMOSPHERE
www.csiro.au

Indian Ocean Bigeye Tuna Management Procedure Evaluation Update March 2020

Prepared for the IOTC MSE Task Force originally scheduled for March 2020

Dale Kolody, Paavo Jumppanen, Jemery Day
CSIRO Oceans and Atmosphere, Castray Esplanade, Hobart TAS 7000, Australia

Citation

Kolody, D, Jumppanen, P, Day J. 2020. Indian Ocean Bigeye Tuna Management Procedure Evaluation Update March 2020. Report prepared for the Indian Ocean Tuna Commission Informal Management Strategy Evaluation workshop 2020.

This document will be assigned a number in the IOTC archive at a future date.

Contents

Acknowledgments

This work was jointly funded by the Australian Department of Foreign Affairs and Trade and Australia's Commonwealth Scientific and Industrial Research Organization (CSIRO - Oceans & Atmosphere). Technical oversight and advice was provided by various IOTC Working Groups, and notably the participants of the IOTC MSE Task Force, including Toshihide Kitakado, 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. Thanks to Simon Hoyle and Dan Fu for various files and clarifications related to the 2019 bigeye tuna assessment and CPUE standardization.

Summary

This working paper describes developments on the Indian Ocean Tuna Commission (IOTC) bigeye (BET) reference set and robustness test operating models (OMs), with key Management Procedure (MP) evaluation results, since the 2019 Working Party on Tropical Tunas (WPTT) and Working Party on Methods (WPM). In the following (for historical reasons), we mostly use the term MP and Management Strategy (MS) interchangeably, though we subscribe to the specific definition of MP as a subset of MS (as defined in the CCSBT and IWC, in which the MP aims for full specification and simulation testing of data collection and analytical methods). Management Strategy Evaluation (MSE) is the simulation testing process, using complex operating models, for evaluating performance of alternative MSs (or MPs). The intent was to obtain feedback on presentation requirements for the 2020 Technical Committee on Management Procedures (TCMP) meeting, and recommendations on further analyses and revisions for the OMs in preparation for the WPM and WPTT 2020 (but priorities changed due to the Covid-19 pandemic and remain uncertain).

Key points include:

- The BET OM was updated with respect to the WPTT/WPM 2019 requests, using the new 2019 assessment for the core data and structural assumptions, subject to the following modifications:
    - Uncertainty in the CPUE standardization method was not included in this iteration, because the CPUE group did not produce the CPUE series that were used in the previous iteration. Furthermore, after consulting with the leader of the CPUE group, it was agreed that the representation of CPUE uncertainty in the MSE requires further focused consideration. This is probably the most important input to the OM (and assessment), and should be considered carefully. Accordingly, the decision was taken to only use the CPUE series from the assessment at this time, and it was recommended that the Terms of Reference for the CPUE group should be expanded in 2020 to include the provision of explicit recommendations for the MSE work. CPUE uncertainty in the OM is retained in terms of other dimensions (alternative regional scaling factors, catchability trends) and the alternative weightings for different data sources.
    - The Stock Synthesis (SS) maximum fishing mortality setting used in the assessment (Fmax = 2.9), corresponds to an exploitation rate of 95% (for the most highly selected age class), and results in a “non-trivial” catch likelihood term for the majority of OM specifications (and the assessment cases examined). i.e. There is a discrepancy between predicted and observed catch, because there are not enough fish for the observed catch to be taken in at least one time/fishery/age strata, unless F exceeds this arbitrary value. In the previous iteration of the OM, this was interpreted to be an indicator of an implausible model. Raising the max F constraint to 6.0 allows the majority of BET models to avoid this problem and the difference in stock status characteristics between the grid of models with Fmax=2.9 and Fmax=6.0 was very small. However, Fmax=6 corresponds to an even more

dubious exploitation rate of 99.5% (for the most highly selected age). We consider this to be a warning that there are probably structural or data problems in the stock assessment and OMs, but have not identified an obvious solution. In this iteration, we have retained the Fmax = 6.0 value , and not rejected any models on the basis of the catch likelihood criterion (following the approach implicitly endorsed in the stock assessment).

- The 4 seasonal CPUE series in the southern region (R3), were merged into a single series (each season first renormalized over the years with non-missing values for any season). A substantial grid of models was run comparing this approach with the 4 season CPUE approach, and the stock status differences were found to be negligible.
- The recruitment deviates were highly constrained for the most recent 12 quarters in the OM – this avoids the problem of some models estimating very large recent recruitments that are not supported by much data (the problem was more serious for yellowfin). The OM introduces variability to the estimated initial numbers-at-age in the projections to compensate for this constraint.
- At WPTT/WPM 2019, an OM problem was reported related to the discontinuity between historical and projected CPUE arising because the CPUE series available for the MP may differ from the CPUE series used for conditioning. We have addressed this problem with an improved method for linking historical and projected CPUE. The OM now uses the model-specific (spatially-aggregated, annualized) MP CPUE RMSE for the projection CV, including the lag(1) autocorrelation. The first simulated observation is linked to the last real observation error deviation. Thus systematic historical lack of fit is interpreted as correlated observation error, and consistently carried forward in the projections. If the calculated MP CPUE RMSE is <20%, the projected MP CPUE CV is set to 20%.

- The 2018 WPTT discussed the problematic retrospective patterns observed in the yellowfin assessment – removing x years of data consistently resulted in a more depleted stock status estimate for year T-x relative to that observed in T-x with all data. There is a similar pattern for bigeye. When future catches are taken, the population did not decline as much as would have been expected, so the stock status appears to have been more optimistic than previously estimated. Since this pattern is consistently repeated, it seems reasonable to expect that it might continue into the future, meaning that the most recent assessment can probably be expected to be deemed too pessimistic when examined at some future date. Given that CPUE are the most informative data in the model with respect to relative abundance, one mechanism for introducing this retrospective pattern might be a non-linearity between CPUE and abundance (i.e. hyperdepletion, which is commonly observed in at least the early development of many tuna fisheries). We imposed several different values for the SS non-linear abundance-CPUE relationship parameter H equally for all longline fleets (where Index = QN^{1+H}). If the retrospective pattern is a simple result of this non-linearity, we would have expected to see the retrospective pattern become more exaggerated with negative values of H, and diminish with increasing H (possibly reversing

direction at some point). The different H values (-0.5 to 0.5) had an impact on the stock status inferences and retrospectives, but is not a simple solution to the problem.

- There was some experimentation with iterative reweighting (i.e. adjusting variance-related parameters to achieve internal consistency between model predictions and observations) using the formalized approaches that are commonly used by the Pacific Fishery Management Council and the Australian Commonwealth Southern and Eastern Scalefish and Shark Fishery stock assessments. These approaches had an effect on point estimates of the models examined, but the differences were very small relative to the uncertainty encompassed by the overall OM grid (and the approach used is not clearly preferable to the approaches already used in the IOTC assessments, though it may be more consistent and reproducible if tags were to be included in the algorithm).
- The mechanics of generating the OM grid were similar to the previous iteration:
    - In recognition of the numerical instability of these models, the minimization was automatically repeated from jittered initial parameter values, until convergence (maximum absolute gradient < 0.01) was achieved in 3 independent runs (or at least 10 minimization failures occurred). All reference case OM configurations were able to meet this criterion for BET (though this was not the case for YFT). Only the lowest objective function iteration was retained for each OM specification. Bounds were relaxed when (important) parameter bounds were hit. In a few cases, other bounds were hit in the second iteration, but there was not time to repeat the process.
    - Fractional factorial design was used to create a grid of 72 models with orthogonal contrast in factors (uncertainty dimensions).
    - The final reference set OM is identical to the reference set grid, i.e.
        - No models were removed due to convergence failures (max. gradient > 0.01 following repeated attempts to minimize from jittered initial conditions)
        - No models were removed due to the catch likelihood (i.e. though the fishing mortality required to achieve this is questionable in the majority of cases)
        - High level model diagnostics, including fit to data, and trends in recruitment deviations, did not reveal any obvious outlier behaviour.
- The stock status inferences from the reference set OM (OMrefB20.1) appear to be somewhat more pessimistic than the stock assessment. The largest factor contributing to this is the 1% per year CPUE catchability trend option.
- The MP tuning objectives requested by the 2019 TCMP appear to provide reasonable MP behaviour. Both tuning objectives appear to be attainable, with the expectation of a modest increase in realized catches relative to 2018, over the medium term.
- Five robustness tests were conducted, which degrade the performance of the MPs in a qualitatively predictable manner.

We welcome feedback or endorsement on all elements of the MSE work, with some key points for consideration highlighted in the discussion. We continue to encourage other members of the IOTC

scientific community to engage with the MSE process. This would have the added benefits of i) improving the reliability of the code, ii) increasing fail-safe redundancy of the MSE process, and iii) possibly developing new MPs that have better performance with respect to specific, subtle objectives.

1 Introduction

This paper represents a progress update on key technical elements of the IOTC bigeye tuna (BET) Management Procedure (MP) evaluation project to obtain feedback in preparation for the 2020 IOTC TCMP, WPM and WPTT. This document is primarily an update on Kolody and Jumppanen (2019), and attempts to address the specific requests from WPM (2019) and WPTT (2019), along with other general issues in the BET MP and stock assessment processes. The target audience is already familiar with the scope of the work and technical jargon. Other interested parties should consult the more accessible project reports found in https://github.com/pjumppanen/niMSE-IO-BET-YFT/. The general approach and many of the ongoing issues are similar to those discussed in the yellowfin companion paper (Kolody et al. 2020), which may provide a different level of detail and emphasis in some cases.

Table 1 lists the OM grid model options described in the text, and Table 2 lists the OM configurations discussed in this paper and the rationale for each. The development requests from WPM (2019) and WPTT (2020) for bigeye (or bigeye and yellowfin implicitly) are attached in Appendix A.

The current phase of YFT/BET MSE support has funding until June 2021, however, staff allocations are reduced over the Jun-Oct 2020 interval, as this was anticipated to be the slow period for MSE development. Along with general scientific and technical feedback, the authors are seeking guidance from the IOTC MSE Task Force about how to revise project priorities given the uncertainty about the IOTC meeting schedule as a consequence of global Covid-19 disruptions.

Table 1. Model specification abbreviations. Bold indicates the reference case assessment assumption. Some abbreviations may relate to additional explorations that were either not completed, reported in earlier iterations, or pertain to YFT.

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
	CPUE regional-scaling factors
iR1	preferred estimate from Hoyle (2018) – 7994_m8
iR2	alternate from Hoyle (2018) – 8000_m8
	mean Age-length relationship (growth curve)
gr1	**original from assessment**
	Recruitment deviation penalty
sr4	σ_R = 0.4
sr6	**σ_R = 0.6**
sr8	σ_R = 0.8
	Future recruit failure
r55	3 years of poor recruitment (2021-2024); mean dev = -0.55, consistent with 2015 YFT assessment
	Natural mortality scaling factor relative to SA baseline level

M10	1.0
M08	0.8
M06	**0.6**
	Tag recapture data weighting (tag composition and negative binomial)
t00	$\lambda = 0$
t0001	$\lambda = 0.0001$
t001	$\lambda = 0.01$
t01	$\lambda = 0.1$
t10	**$\lambda = 1.0$**
t15	$\lambda = 1.5$
	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
iH	**Hooks Between Floats**
iC	Cluster analysis
	CPUE observation error
i1	annual $\sigma_{CPUE} = 0.1$
i2	**annual $\sigma_{CPUE} = 0.2$**
i3	annual $\sigma_{CPUE} = 0.3$
	Tag mixing period
x3	3 quarters
x4	**4 quarters**

x8	8 quarters
	Longline selectivity (in conditioning)
SL	**Stationary, logistic, shared among areas**
SD	Stationary, logistic for region 1N, double normal for other regions
S4	Eestimated independently, LL selectivity independent among areas
NS	Temporal variability estimated in 10 year blocks
ST	Logistic selectivity trend estimated over time
Sdev	15 years of recent selectivity deviations estimated
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
ESS10	**ESS = maximum of 10 for BET fisheries 9, 11 and 15 (as defined in the SS files), maximum of 1.0 for all fisheries.**
CLRW	ESS = One iteration of reweighting; the output ESS from a reference case assessment specification (capped at 100)
CL75	ESS = One iteration of reweighting; the output ESS from SAref raised to the power of 0.75 (capped at 100); includes the initial ESS10 assumption for some missing early years of ESS outputs

Table 2. Operating Model definitions. The OMs are listed in the order discussed in the text, reflecting the sequence of development.

OM Ensemble	(factor abbreviations are defined in Table 1)
OMgridB20.1	72 models (Fmax constraint = 6.0) with 7 factors in a "main effects" fractional factorial design Assumption levels: h70, h80, h90 (SR steepness) M10, M08, M06 (M) t0001, t01, t10 (tag-weight) q0, q1 (catchability trend) iR1, iR2 (regional scaling factors applied correctly) ess10, CLRW (CL assumed sample sizes) SL, SD (longline selectivity function) (i2 - LL CPUE CV 0.2 only) (iH – HBF CPUE standardization method only) (gr1 – original growth curve only)
OMrefB20.1.500	500 realization OM, stochastically sampled from equally-weighted grid OMgridB20.1.
OMgridB20.1Fmax2.9	As gridB20.1 except using the maximum F constraint from the assessment (2.9)

OMgridB20.2	As gridB20.1, except the original CPUE structure from the assessment was used (i.e. 4 independent seasonal CPUE series for the temperate region)
OMrobB20.1.ICV3	A robustness scenario with longline CPUE CV (spatially-aggregated annual = 0.30, auto-correlation = 0.5)
OMrobB20.1.10overRep	A robustness scenario in which every fishery has a 10% over-catch implementation error, with accurate catch reporting
OMrobB20.1.10overIUU	A robustness scenario in which every fishery has a 10% over-catch implementation error, that is not reported.
OMrobB20.1.qTrend3	A robustness scenario in which there is a 3% per year LL CPUE catchability trend starting in the projections (conditioning unchanged from the reference case)
OMrobB20.1.recShock	A robustness scenario with 8 consecutive quarters of poor recruitment (55% of expected values, similar to estimates for YFT in the early 2000s). (conditioning and sampling is unchanged from OMrefB20.1.500)

2 Core Operating Model Assumption deviations from the Stock Assessment

The 2020 BET OM is derived from the Fu et al. (2019) stock assessment, which was based on a grid of Stock Synthesis (SS) models (SS3.24Z, Methot and Wetzel 2013). Changes to the 2019 assessment relative to 2016 include:

- The data were updated to include 3 additional years, and were re-aggregated to properly partition all elements of the Northwest region into northern and southern sub-regions.
    - The shift in purse seine operations in recent years from PSFS to PSLS (Free-School to FAD set) particularly for the Spanish fleet) may have an important impact on perceived stock status and MP performance. Further investigation into the reliability of the reported Spanish catch distribution was proposed at the 2019 WPTT.
    - The region 1 CPUE is down-weighted immediately following the peak piracy years (i.e. because there is a spike in CPUE that is localized, short-lived and probably associated with a small number of vessels operating in an atypical fashion). This may be caused by the same poorly-understood mechanisms that caused hyperdepletion in the early development of many tuna fisheries, and the large BET CPUE spikes in the late 1970s.
- The length-mass relationship was updated.
- The double normal longline selectivity option in the new assessment was implemented:
    - Fishery 1 (FL2) - logistic
    - Fishery 13 (LL-1N) – logistic
    - Fishery 2 (LL-1S) - double normal, independent
    - Fishery 3 (LL2) - double normal, independent
    - Fishery 4 (LL3) - double normal, independent
    - Fishery 12 (Line2) – shared with Fishery 2

The main structural difference in the OM conditioning relative to the assessment arises from the reduction of 4 temperate CPUE series (partitioned by season) into 1 combined series as discussed below. Other minor differences include: i) the number of iterations used to solve the catch equations is raised from 4 to 7 in the OM and the maximum F is raised from 2.9 to 6.0 (this improves performance in high fishery mortality situations and is discussed further below) and ii) some parameter bounds and priors are relaxed in the OM (i.e. so they are not unintentionally informative and do not constrain model fits in arbitrary and perhaps misleading ways).

As in the previous iteration of the OM, temperate region CPUE assumptions have been simplified relative to the assessment. Instead of a different CPUE series for each quarter (with independent

catchability), the OM uses a single CPUE series, in which each seasonal series is independently normalized (based on all years 1980-2018 for which there were no missing values in any season). A comparison of the temperate series with and without the renormalization is shown in Figure 5. The renormalization removes some of the seasonality in the time series, but there is still a large degree of within-year variability. This might reflect the fact that fishing activity does not consistently align with seasonal bigeye movements. It is recognized that the ability to partition movement and catchability in these models is not very reliable at present (e.g. Fu et al. 2019). Two models are compared below:

- SAref – a reference case assessment *cSci_sL_TagLambda1_h80*, provided from the Fu et al. (2019) stock assessment, includes 4 independent seasonal CPUE series in the temperate region (catchability shared with other regions only for season 4).
- OMref – mostly identical to SAref, except for the temperate CPUE series are merged, with a CV = 0.2 for quarter 4 and CV = 0.4 for quarters 1-3 (catchability shared with other regions).

These models are not considered to have special weight in a stock status context, but are hopefully representative for the purposes of discussing general model behaviour.

As shown in Figure 1, there is some sensitivity to the treatment of the temperate CPUE, notably in terms of the relative biomass distribution among areas. Key stock status inferences are very similar (Table 3), and differences are small relative to the uncertainty introduced from other model assumptions. Neither approach fits the temperate CPUE particularly well (Figure 2, Figure 3). Setting the OMref temperate CPUE CV to 0.2 or 0.4 for all seasons resulted in further divergence from SAref that bracketed the stock status quantities of SAref in Table 3 (not shown). The MSE code was modified to accept the seasonal CPUE approach, but it was not adopted in the OM at this time because i) the discrepancy appears to be small relative to other sources of uncertainty, and ii) it is not obvious that the assessment approach is necessarily better than the OM approach (e.g. It might be argued that the application of the regional-scaling factors is not consistent within the seasonal CPUE interpretation - the regional scaling should presumably be applied over the same time intervals upon which the catchability is shared among regions, but it currently is not).

The OM increases the number of iterations used in the SS hybrid F calculation (from 4 to 7) – it has been reported that this may improve the precision of the catch extraction, particularly in high F situations (and this seems likely given the results in Table 4, discussed in 4.1).

For future reference, we note some minor differences between the OM and the assessment documentation, (typos that do not affect the assessment results):

- CPUE input files:
    - Joint_regB2_R2_dellog_vessid_79nd_yq.csv
    - Joint_regB2_R3_dellog_vessid_79nd_yq.csv
- 7994_m08 Regional scaling factor for region 3 = 0.86 (not 0.63)

Table 3. Comparison of stock assessment characteristics for SAref and OMref (best likelihood among 3 replicates from jittered initial values).

	SAref (Fmax=2.9)	SAref (Fmax=6)	OMref
SB(2018)	486 Kt	480 Kt	481 Kt
SB(2018)/SB(MSY)	0.98	0.96	0.98
MSY	79 Kt	79 Kt	76 Kt

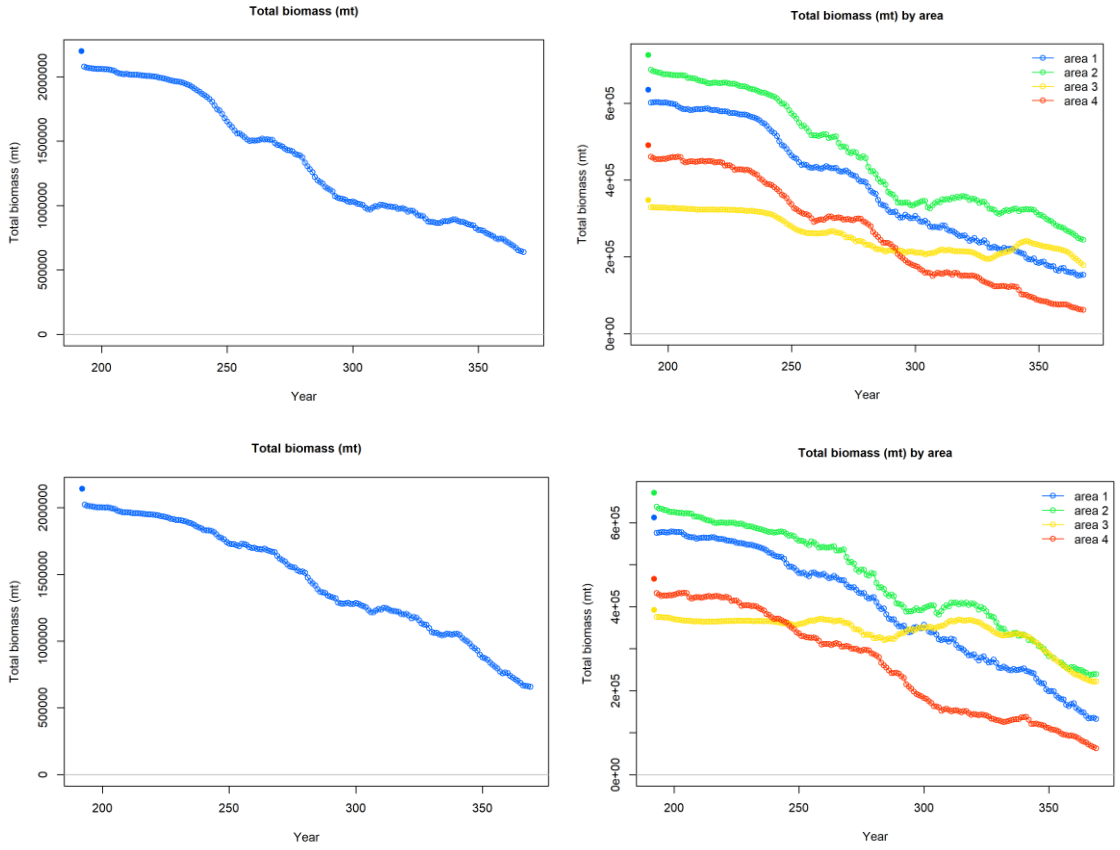


Figure 1. Comparison of SAref (top) and OMref (bottom) total biomass aggregated (left) and by region (right).

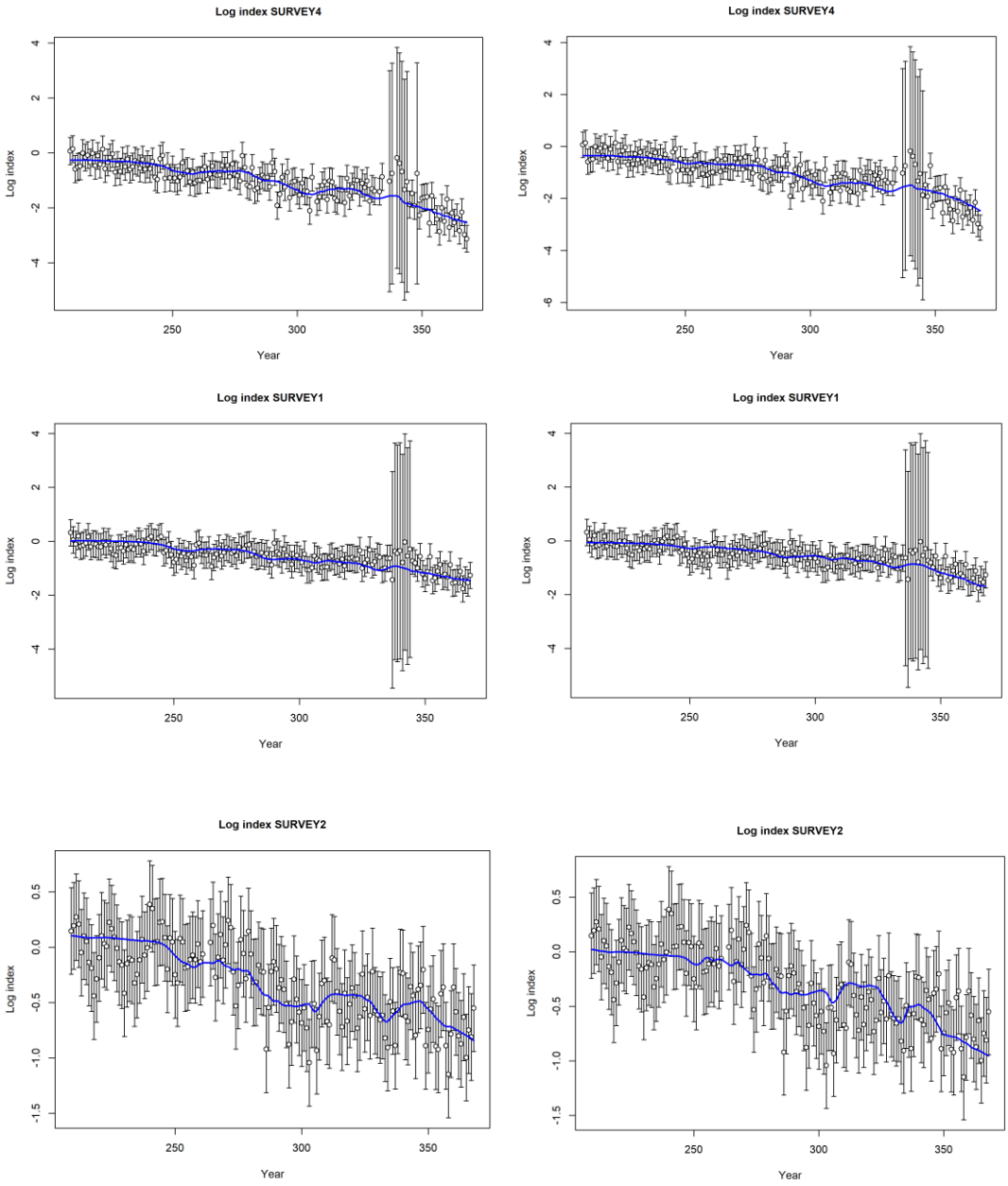


Figure 2. Comparison of SAref (left) and OMref (right) CPUE fits for the tropical regions.

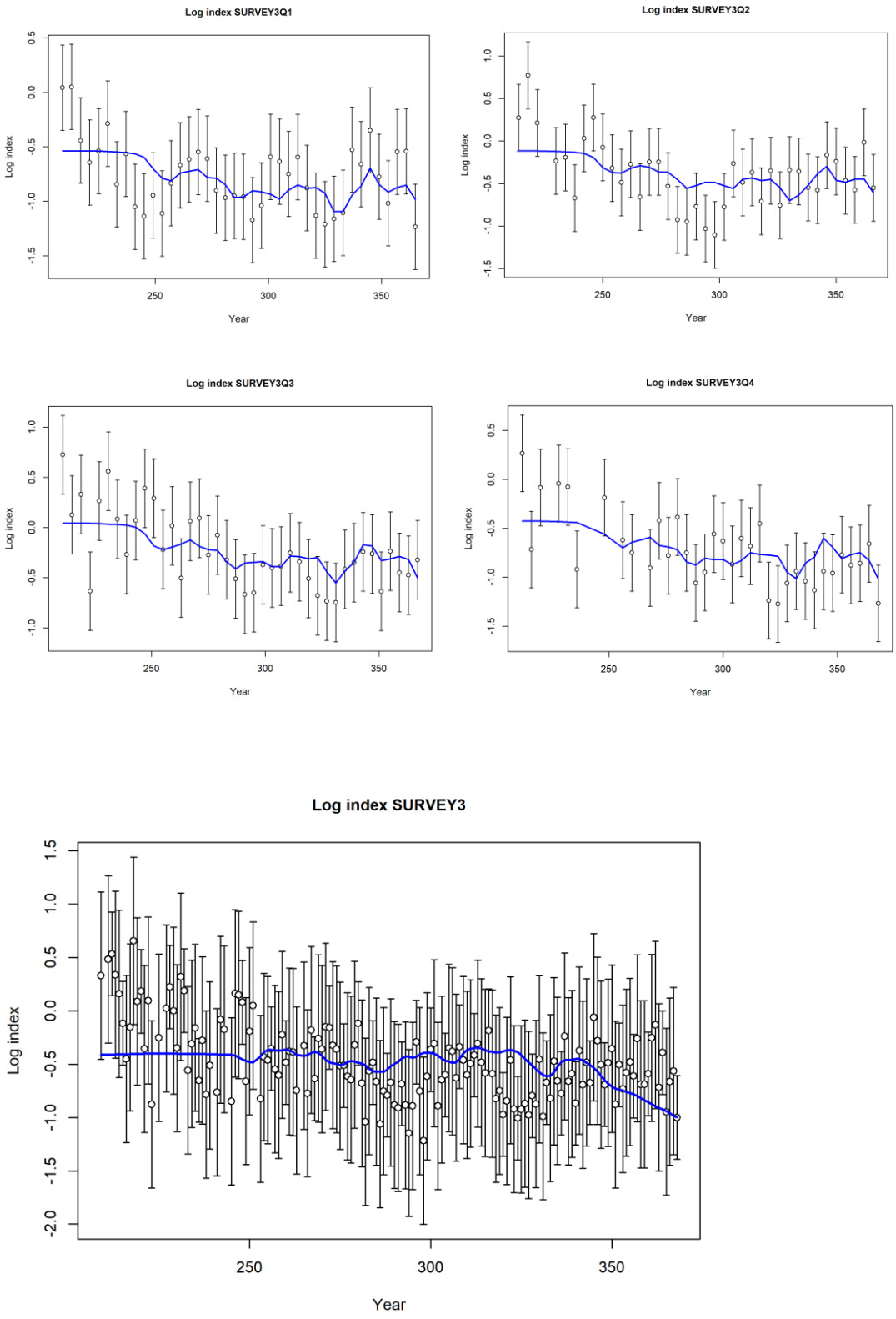


Figure 3. Comparison of temperate CPUE series fits - SAref independent CPUE by season (top 4 panels) and OMref merged CPUE series (each season independently normalized).

3 Conditioning Assumptions in the March 2020 Bigeye Operating Model

The new requests for the BET OM arising from the WPM/WPTT in 2019 are minor and summarized in Appendix A (features not discussed conform to the assessment, as the actual assessment input files were adopted as the template configuration). Some additional modifications were made as noted below:

- Re-introduction of the intermediate tag-weighting assumption ($\lambda = 0.1$)
- The recruitment CV for the most recent 12 quarters was constrained to essentially zero. Four quarters were used in the OM previously (and in the assessment), however, it was found (for yellowfin in particular) that the most recent recruitment events sometimes take on large values that are not strongly supported by data. In the MSE projections, stochastic error is introduced back into the initial population age structure.
- We eliminated the alternative CPUE standardization series. This decision was taken following recognition that the CPUE group did not produce the alternate CPUE series that were used in the previous MSE iteration. Furthermore, following discussion with the consultant for the CPUE group (Simon Hoyle, pers. comm.), it was decided that the CPUE uncertainty in the MSE is too important to simply use an ad hoc decision as was done in 2018. Accordingly, the decision was taken to only use the (implicitly "best") CPUE series from the assessment at this time. It has been recommended to the Secretariat and CPUE working group that the Terms of Reference for the CPUE group should be expanded in 2020 to include the provision of explicit MSE recommendations.

CPUE uncertainty in the OM is retained in other dimensions (alternative regional scaling factors, catchability trends) and the alternative weighting of different data sources. The CPUE uncertainty introduced by the catchability trend is shown in Figure 4 and regional-scaling factors in Figure 5. The catchability trend is clearly very influential. It is not clear that the regional-scaling uncertainty is important, however, it is clear that option 8000_m8 is more different from the assessment assumption (7994_m8) than 7494_m8. Hence 8000_m8 was retained for the grid. The previous iteration used 7494_m08 for consistency with yellowfin.

The CPUE projection assumptions are discussed in the following sections.

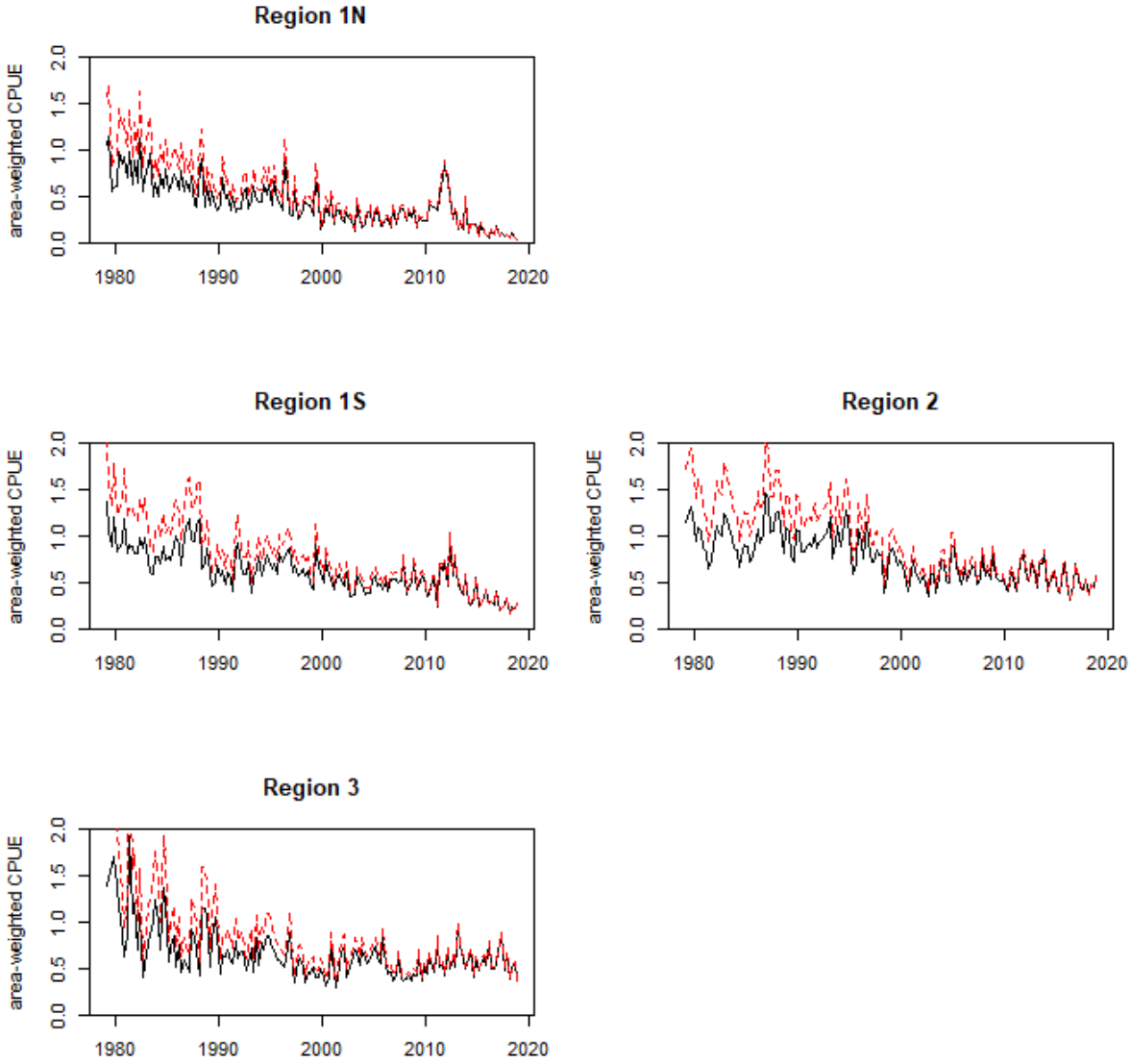


Figure 4. BET CPUE series comparing the catchability trend 0 (black) and 1% per year compounded annually (red).

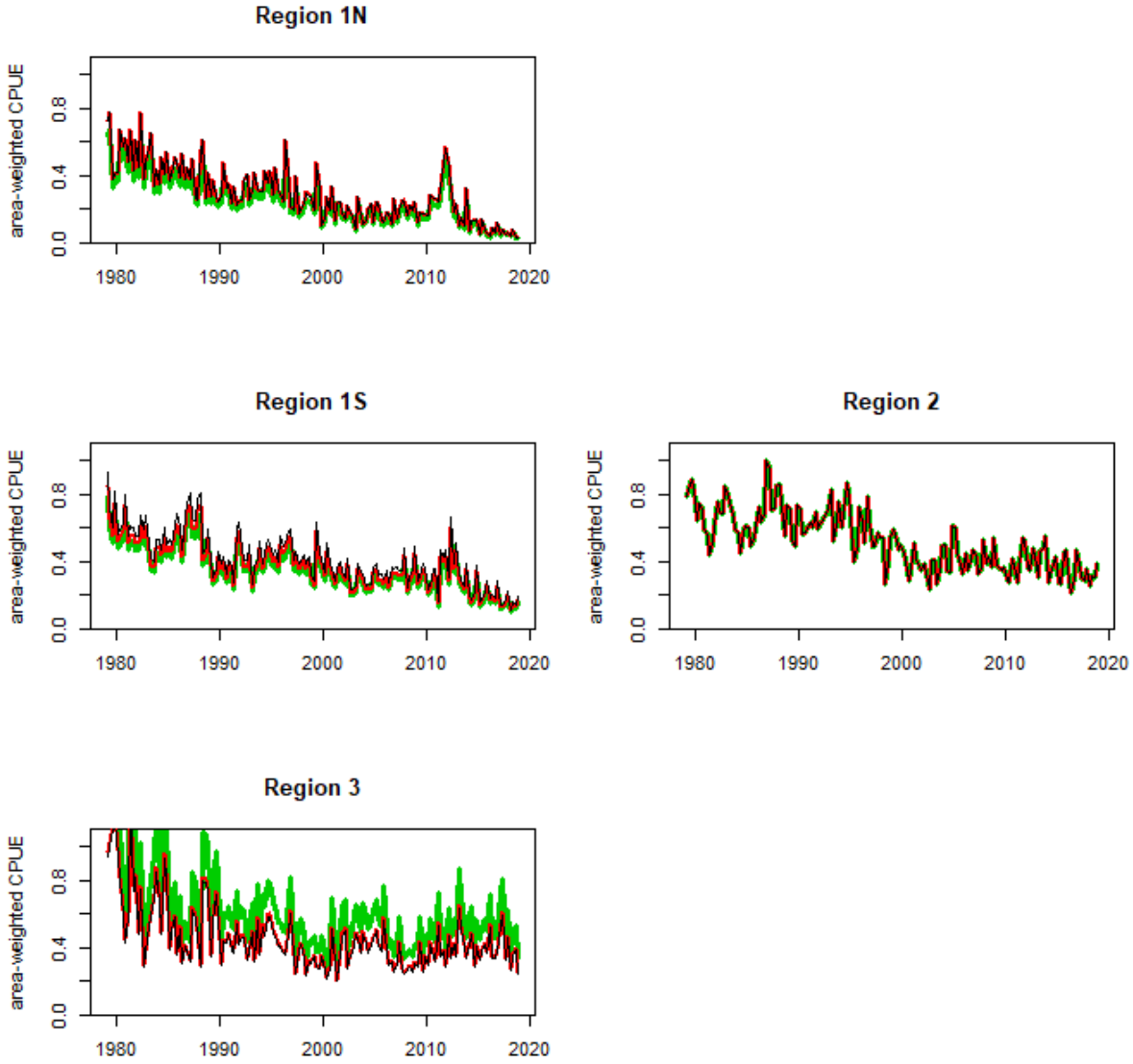


Figure 5. BET CPUE series comparing the different regional-scaling factors (black = assessment case 7994_m8, red= 7594_m8 and green= 8000_m8)

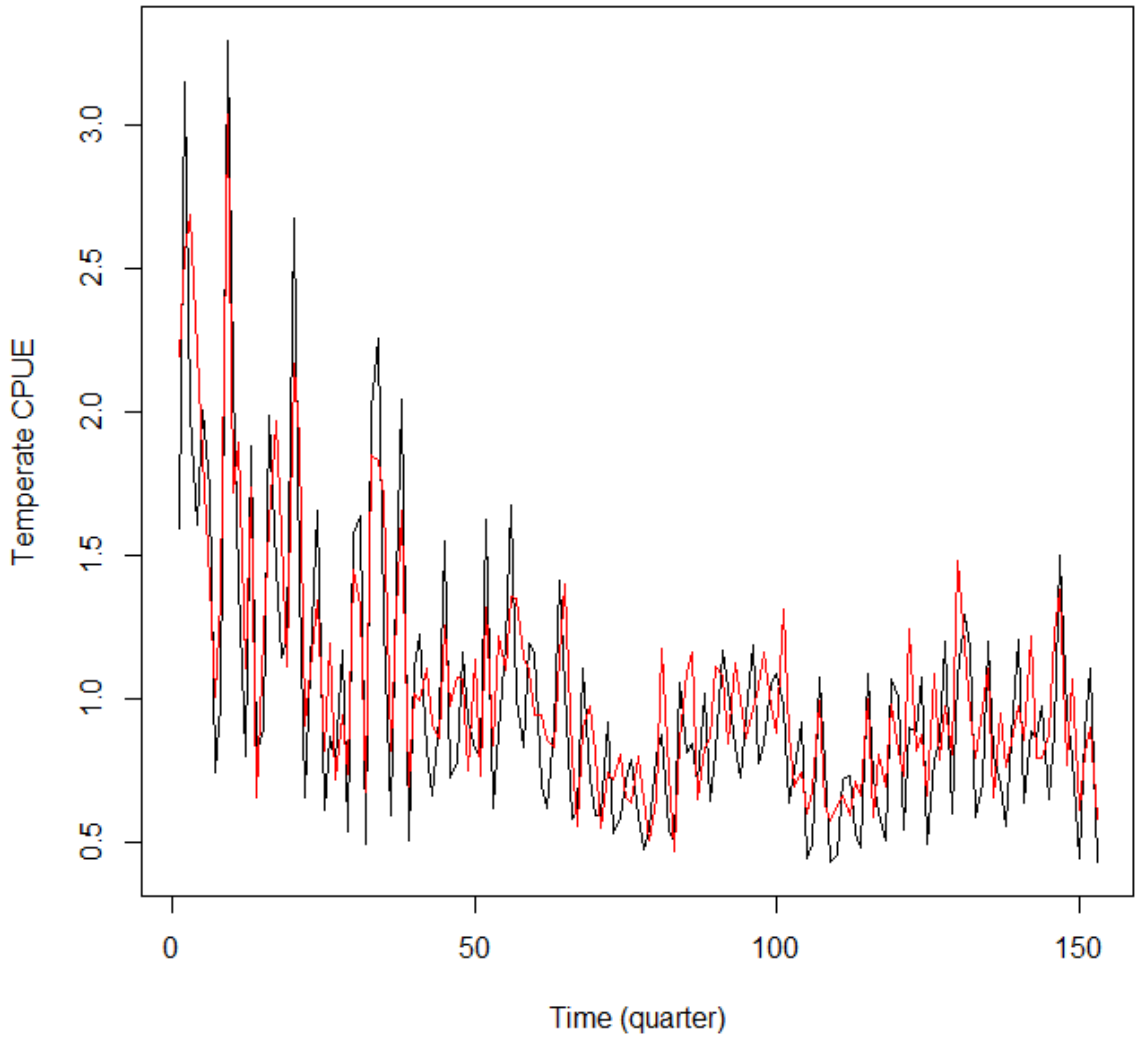


Figure 6. BET temperate CPUE as reported in the standardization (black) and with each season independently re-normalized over the common time period (red).

3.1 New approach for projecting CPUE series in MP evaluation

At WPTT/WPM 2019, we reported an OM problem related to the discontinuity between historical and projected CPUE, arising in part because the CPUE series available for the MP often differs from the CPUE series used for conditioning. We compare three approaches for dealing with this issue below. For convenience of discussion, we refer to a set of CPUE assumptions as a single CPUE series. i.e. Each set of CPUE series actually consists of multiple regional series, but the calculations of interest here are aggregated over quarters and regions (with regional-scaling factors), to

produce an annual aggregate time series (as used in all MPs explored to date). The catchability trend (0 or 1 % per year) is omitted below. The OM simulates new CPUE observations based on the relationship:

$$I_y^{proj} = qB_y^v \exp(\tau_y)$$

Where:

y = year (>year of last observed CPUE)

I_y^{proj} = the simulated CPUE in year y,

B_y^v = vulnerable (longline-selected) numbers-at-age in year y,

$q = exp\left(\frac{1}{t2-t1}\sum_{Y=t1}^{t2} log\left(\frac{I_Y^{obs}}{(B_Y^v}\right)\right)$ = catchability co-efficient over the period t1:t2,

I_Y^{obs} = the standardized historical CPUE, Y ≤ 2018,

σ = CPUE observation error standard deviation,

τ_y = a random normal deviate with lag(1 year) auto-correlation ρ, i.e.

$\tau_y = \rho\tau_{y-1} + \omega_y\sqrt{1-\rho^2}$,

$\omega_y \sim Normal(\mu = 0, \sigma)$.

The three approaches were:

1) The approach used for the MP evaluation results presented to the 2019 TCMP. CPUE catchability, q, is calculated over the whole time period of valid (i.e. non-missing and used in the assessment) CPUE observations. The CPUE observation error (CV and auto-correlation) in the projections was the same among all models, and the historical correlation in error was not carried forward from the historical period. i.e.
    - t1:t2 = 1980:2017
    - σ = 0.2
    - ρ = 0.5
    - $\tau_{y=2017}$ = model-specific observation error estimate

2) The approach proposed by the WPM 2019 involved simply changing the catchability calculation period to a more recent period:
    - t1:t2 = 2015:2017
    - σ = 0.2
    - ρ = 0.5
    - $\tau_{y=2017}$ = model-specific observation error estimate

3) The third option estimates model-specific CPUE error characteristics based on the model-specific degree of discrepancy between model predictions and the MP CPUE series (i.e. irrespective of which CPUE series was used in model conditioning):
    - t1:t2 = 1980:2017
    - σ = max(model-specific estimate, 0.2)
    - ρ = model-specific estimate
    - $\tau_{y=2017}$ = model-specific observation error estimate

Figure 7 contrasts the 3 approaches for 6 random examples from the OM. The discontinuity of concern is most evident in the approach used previously (left column, in which CPUE can be seen to drop by ~50% in the first projection year). The WPM (2019) recommended approach (middle column), greatly reduces the initial projection discontinuity, but tends to create a large systematic lack of fit to the historical data, that becomes amplified further back in time, and the historical CPUE RMSE is greatly increased (Figure 7, Figure 8). The third approach is shown in the right column. It reduces the discontinuity of the first projected CPUE, and avoids the systematic lack of fit to the historical data.

The contrast among columns in Figure 7 suggest that the choice of CPUE approach can be very important in some cases (e.g. case C shows the sustained systematic lack of fit that the projected CPUE can have in the projections if it is consistent with the historical lack of fit). However, we did not compare the overall performance difference that would be realized in an MP. Further investigation revealed that the very large problems identified in 2019 were partly due to a small number of models that had large recent recruitment events, usually 5-8 quarters before the end of the assessment period (particularly for yellowfin). These (poorly-informed) recent recruitment spikes have been removed by constraining recent recruitment for 12 quarters in the conditioning (both species).

We consider the third CPUE option to be preferable. If we are as uncertain about the appropriate CPUE assumptions as the OM requests from the WPM/WPTT indicate, then we should be careful about not overstating the information content in the MSE testing. We opted to set the MP CPUE CV to the SS model-specific RMSE, or 0.2, whichever is larger.

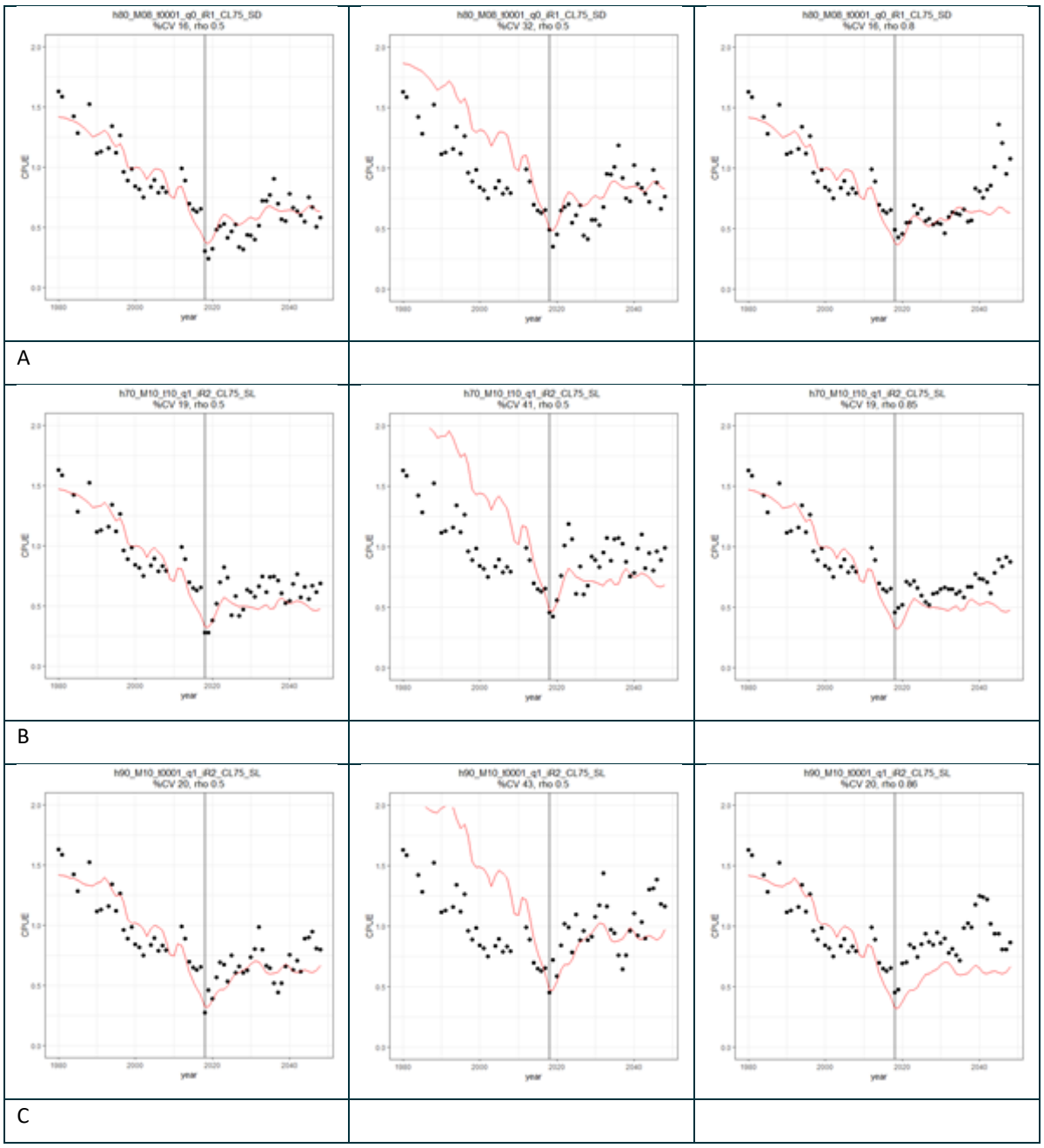
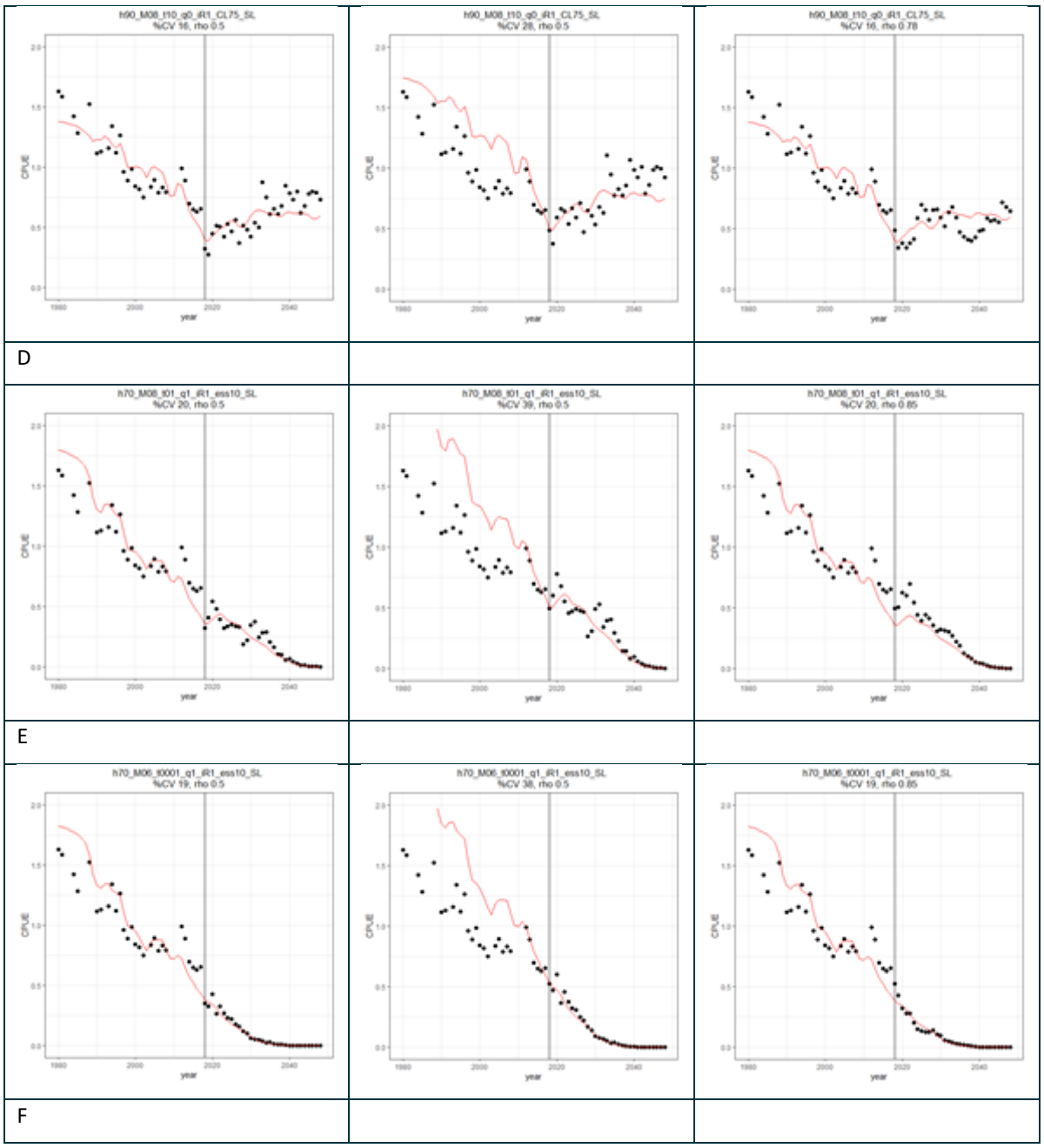


Figure 7. Some examples from the BET OM ensemble contrasting the predicted (lines) and observed (circles) CPUE series for the MPs, derived from the three CPUE projection options outlined in the text. Vertical line indicates the first projection year. Left column – q calculated over whole period, projection CV = 20% and auto-correlation = 0.5 (for all models as used in previous iterations); middle column – q calculated over final 3 years only, projection CV = 20% and auto-correlation = 0.5; right column – q calculated over whole period, CV and auto-correlation correspond to the individual model. Vertical line is the last year of real observations. Note that CV values are not consistent with Figure 8 due to minimization sensitivity. **(Figure 7 continued on next page)**



(Figure 7 cont.)

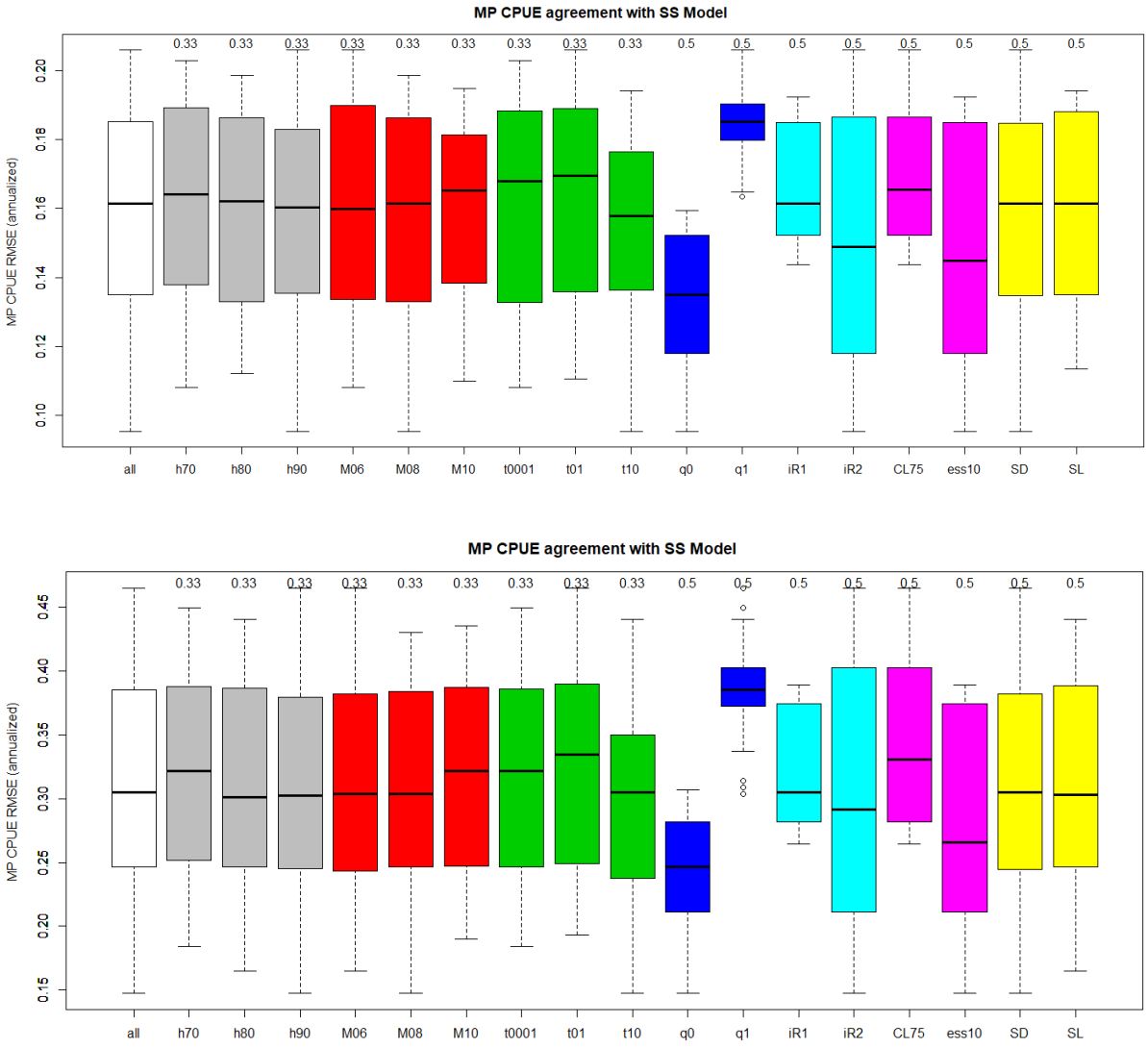


Figure 8. MP CPUE relationship with OM predicted CPUE – catchability calculated over all years (top) and final 3 years only (bottom).

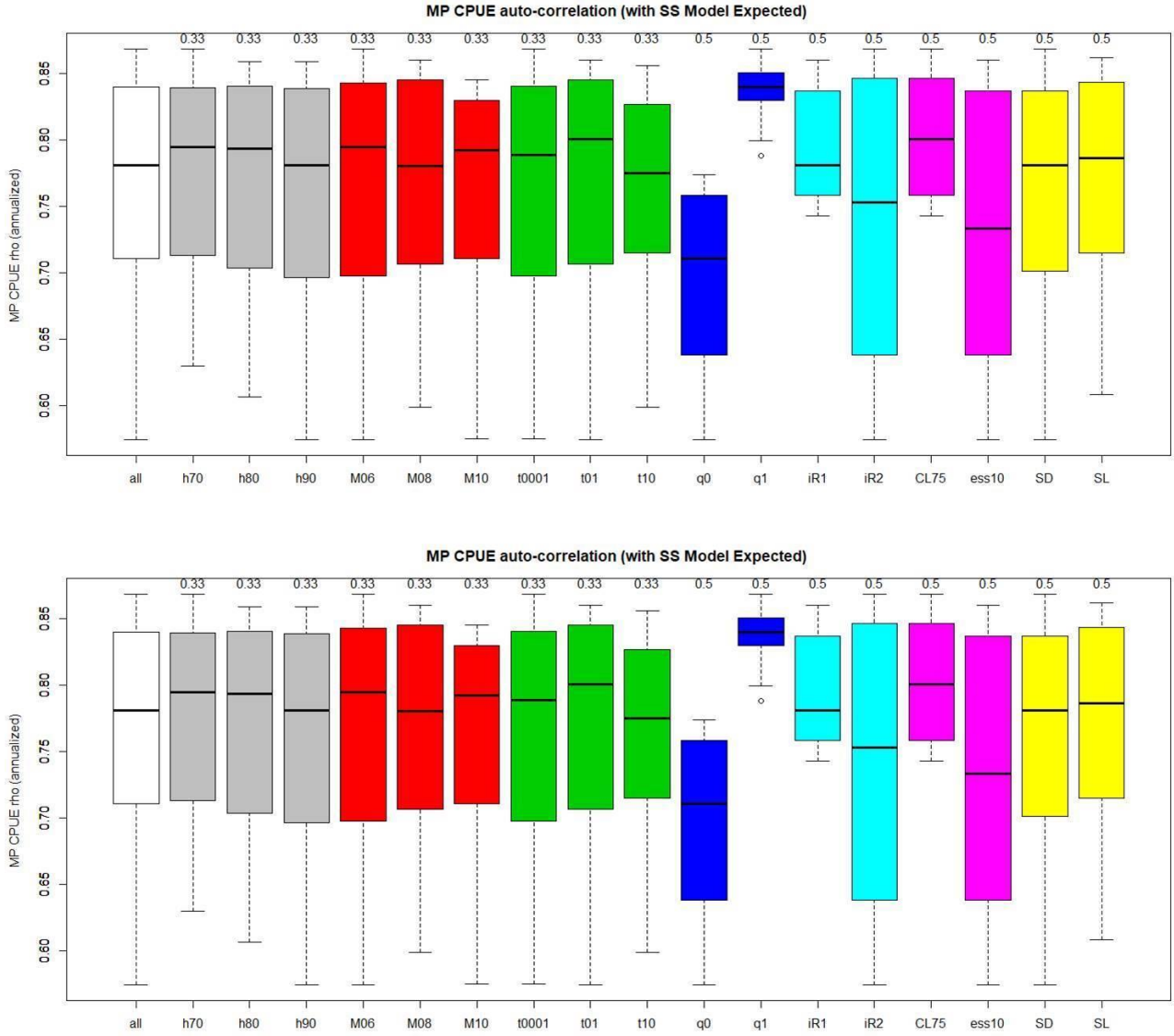


Figure 9. MP cpue relationship with OM predicted CPUE – catchability calculated over all years (top) and final 3 years only (bottom). i.e. Auto-correlation is about the same.

4 Revisiting the catch likelihood as a diagnostic of model plausibility

Both SAref and OMref have non-trivial catch likelihood terms. In previous iterations of the OM development for both bigeye and yellowfin, the catch likelihood terms have shown clearly bimodal distributions (e.g. with ~95% of models either having values <1E-7 or >1E-2, similar to that shown in the YFT companion paper Kolody et al. 2020). We have been interpreting the catch likelihood as an indicator that some models struggle to remove the observed catch (for at least one time/area/age strata) and hence are probably unduly pessimistic. The large gap in the bimodal distribution appears to offer an almost binary basis for retention or rejection. However, the initial 72 model ensemble (OMgridB20.1Fmax2.9), has a unimodal catch likelihood distribution (Figure 10), with almost all of the models failing the rejection criterion adopted in previous iterations. This clearly makes the likelihood rejection/retention criteria critical to how we proceed.

The SS catch likelihood does not appear to manifest itself like other likelihood components in a hybrid-F configuration. From what we have seen, a "high" catch likelihood seems to be associated with a failure to remove the observed catch, for one of two reasons: i) the number of pre-specified iterations in the catch equation solver may be too low to reach a suitable solution, and ii) the maximum F constraint (2.9 in SAref) may be too low. In other models, a catch likelihood with a moderate CV tends to result in failure to fit the catch series in a way that trades off with other data in a more direct way (e.g. the model might improve a bad fit to the tag recoveries through a compensating error in the predicted catch, which could be an over-catch or under-catch relative to the observed).

To further illustrate this issue, the interaction among a few SS assumptions related to the catch equations are presented in Table 4 (in all cases, the results pertain to the best fit model among 3 jittered minimizations). It appears that Fmax is the biggest determinant of the catch likelihood. Increasing the limit from 2.9 to 6 effectively shifts the catch likelihood between the two modes that were observed in previous iterations, such that the majority of models achieve a negligible catch likelihood (Figure 10). Table 4 and Figure 11 - Figure 12 suggest that there is not a big difference in the stock status estimates associated with Fmax = 2.9 or 6.0 (though the lower F constraints presumably should result in at least slightly more optimistic outcomes, and this seems to be the case, particularly evident with Fmax=1.4, added for contrast). But this does not resolve the issue of whether or how to use the catch likelihood as a plausibility constraint.

Harvest rates associated with different OMref fisheries are shown in Figure 13. In some cases, we might expect that an implausibly pessimistic model could be identified if the harvest rate changes at a rate that is not compatible with the effort (particularly a rapid increase in F as a fishery might be collapsing). If we could show a disconnect between F for the LL fisheries that deviated from the catch rate standardization, this might be evidence for implausibility, but this is not obvious. It is the PSLS fisheries that exhibit the highest Fs (20 years ago, though recent harvest rates are almost as high), and which exceed the Fmax 2.9 threshold (but not 6.0). It seems questionable whether the PSLS fishery could achieve F>2.9, as this corresponds to a harvest rate ~95% (for the most highly selected age class in a particular quarter/region stratum). Figure 10 suggests a maximum

F>2.9 for ~90% of models and F>6 for ~20% of models in the ensemble. This probably provides information that could be applied as a plausibility index, but there is not an obvious threshold value or weighting function to adopt.

Part of the problem could relate to the model not being able to represent seasonal movement adequately. E.g. If, in reality, a large proportion of the fish move back and forth, followed by the fishing fleets, a model without seasonal movement might not have enough fish in the right region at the right time (but plenty of fish in total). However, we note in the YFT companion paper (Kolody et al. 2020) that a comparison of YFT models with and without seasonal migration (implemented via an environmental link) had little effect on stock status estimates (but also tended to have limited seasonal movement as tested).

Table 4. Investigation of F-related options in relation to the catch penalty and stock status indicators. It appears that the default maximum F assumption is frequently breached, leading to the non-trivial catch likelihoods (red would fail our retention criterion and green would pass). The stock status results are not very sensitive to this constraint, and increasing the F ceiling removes the problem, but this leaves open the question of whether F >= 2.9 (or some other arbitrarily high value) should constitute a plausibility threshold?

SS model	Catch s.e. (log)	Hybrid F iterations	Max F	Catch LLH	SSB(2018) 1000 t	SSB(T) / SSB(MSY)	MSY 1000 t
1	0.1	4	2.9	1.4	472	0.98	77
2	0.1	7	2.9	1.5	479	0.99	77
3	0.01	9	2.9	0.17	487	0.98	76
4	0.1	7	1.4	16	546	0.99	84
5	0.1	7	6	1E-11	481	0.98	76
6	0.01	9	6	1E-9	465	0.95	76

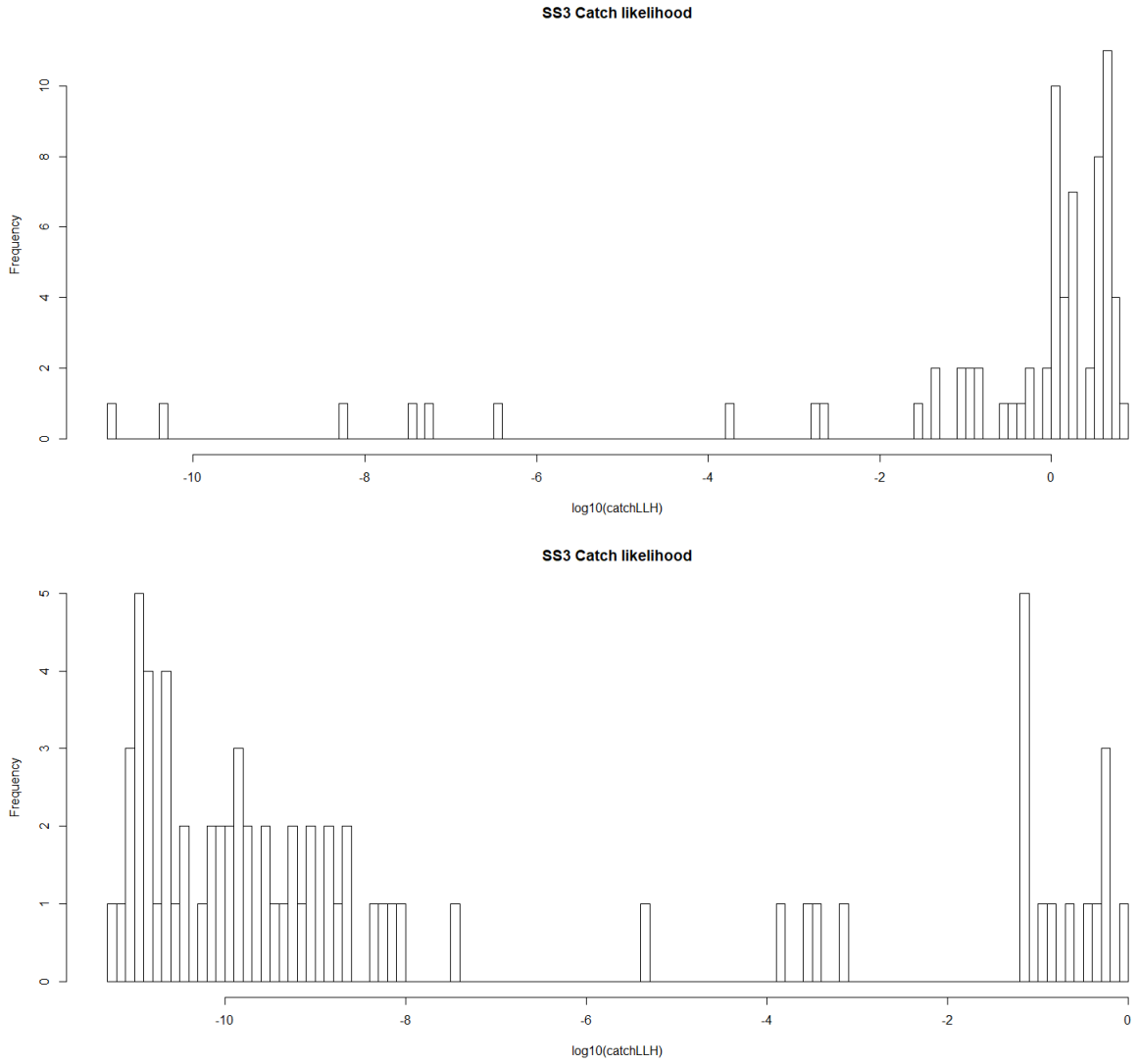


Figure 10. Catch likelihoods associated with OMgridB20.1Fmax2.9 (Fmax = 2.9, top) and 6.0 (Fmax = 6.0 bottom).

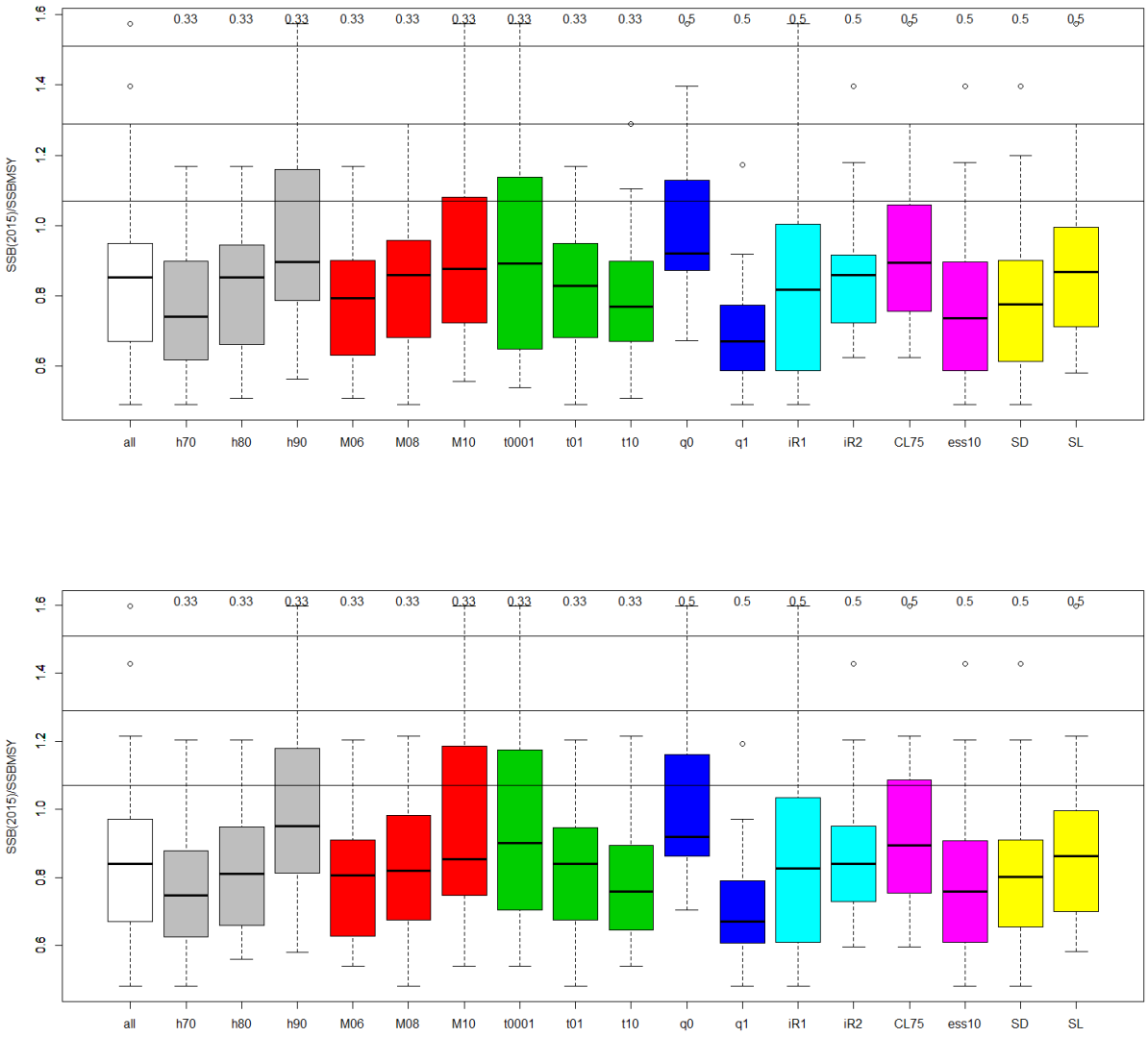


Figure 11. Comparison of depletion from OMgridB20.1Fmax2.9 (Fmax = 2.9, top) and OMgridB20.1 (Fmax = 6.0 bottom).

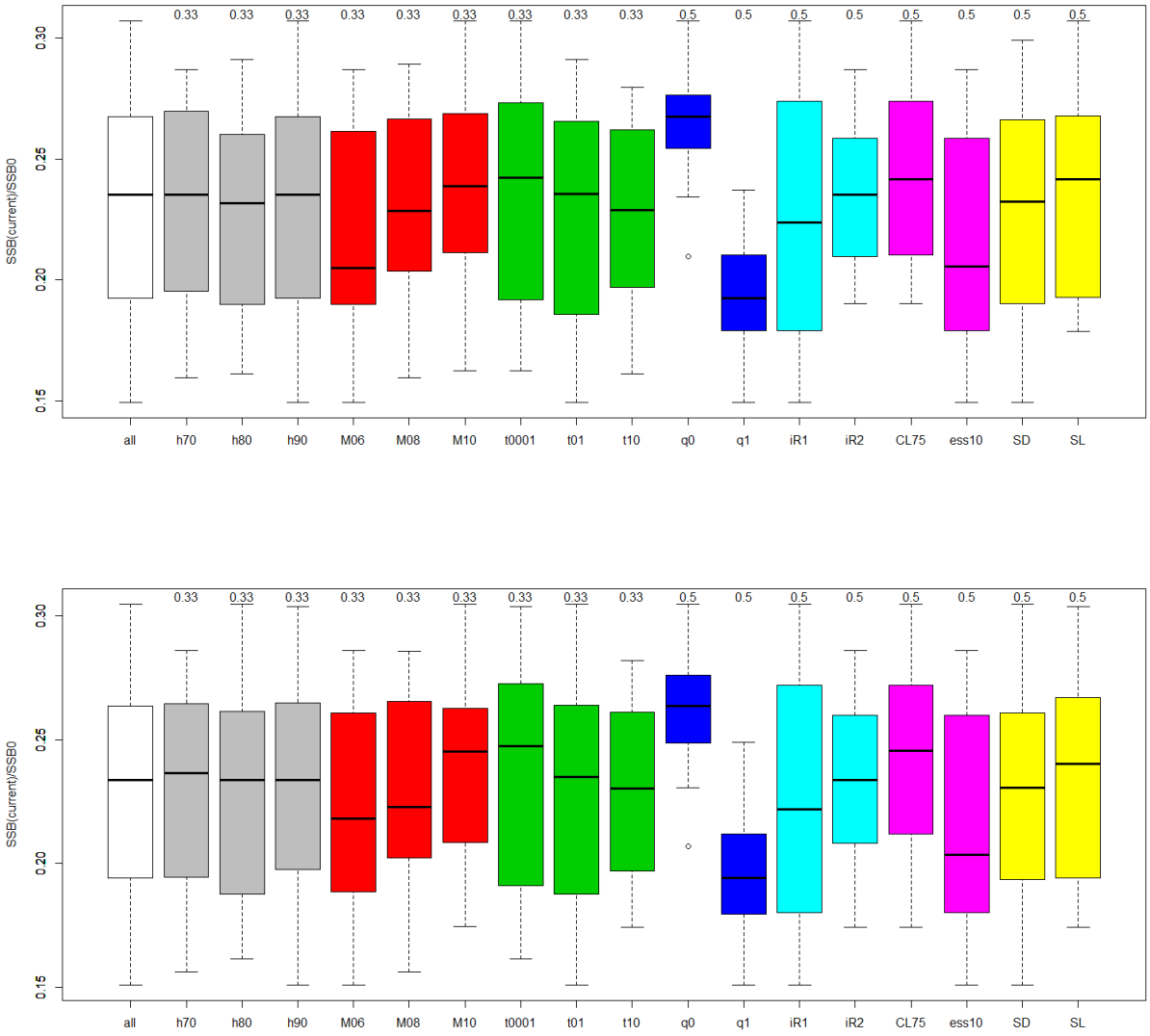


Figure 12. Comparison of depletion from OMgridB20.1Fmax2.9 (Fmax = 2.9, top) and OMgridB20.1 (Fmax = 6.0 bottom).

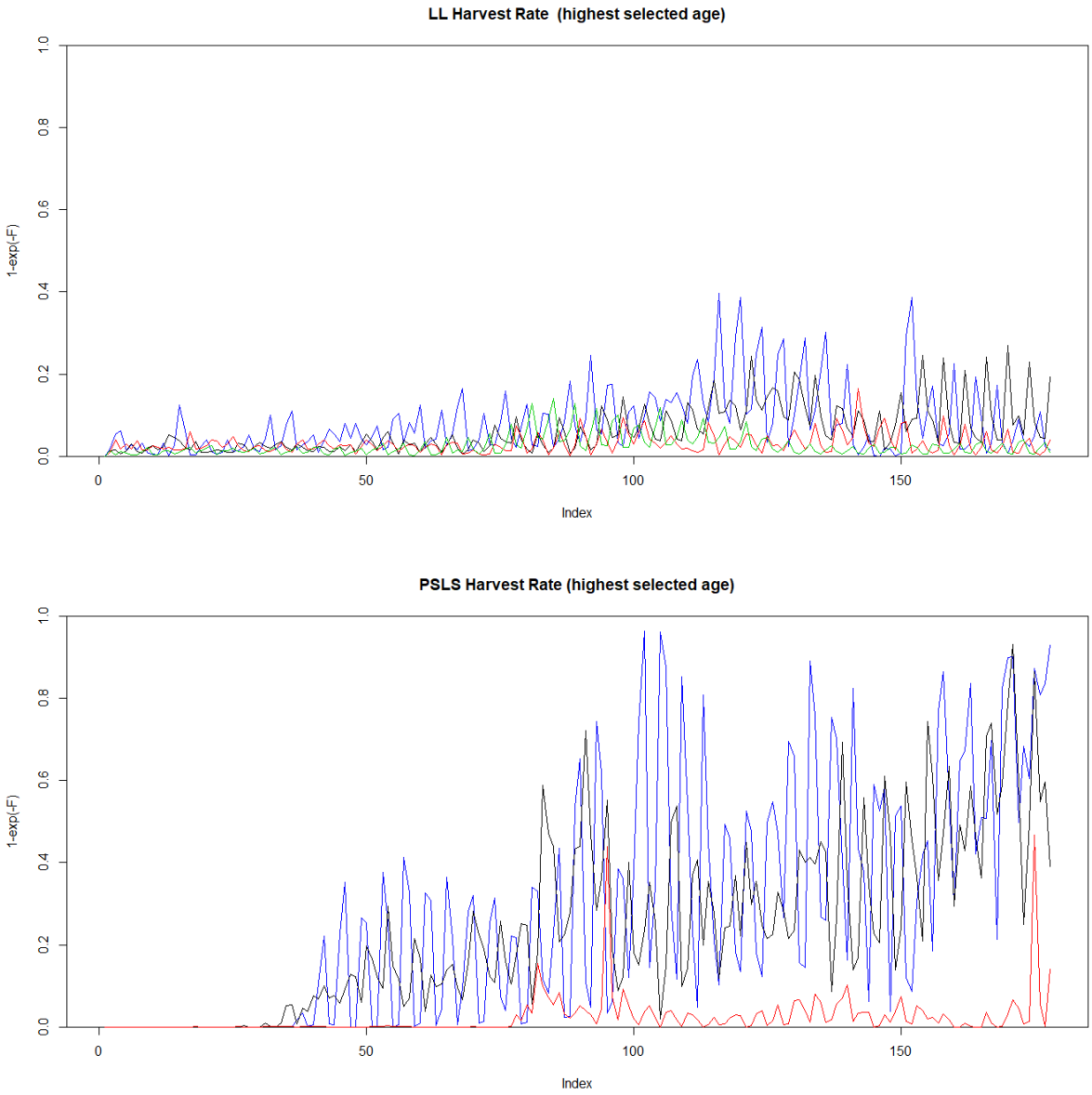


Figure 13. Quarterly time series of OMref Harvest Rate (most highly selected age for individual fishery). Blue is northern NW region, black is southern NW region, red is NE and green is the southern region.

4.1 Retrospective patterns in the BET assessment

The 2019 bigeye assessment has an undesirable retrospective pattern, similar to that identified for the yellowfin stock assessments illustrated by Matsumoto et al (2018). Removing x years of data results in a more depleted stock status estimate for year T-x relative to that observed in T-x when all data are included. When subsequent catches were taken, the population did not decline as much as would have been expected had the T-x assessment been "correct", so the stock status seemingly must have been more optimistic than previously estimated. Since this pattern is consistently repeated, it seems reasonable to expect that it will probably continue into the future, and the most recent assessment will probably be deemed too pessimistic when examined in the future.

Given that CPUE are the most informative data with respect to relative abundance, one mechanism for introducing the retrospective pattern might be a non-linear relationship between CPUE and abundance (i.e. hyperdepletion is the situation in which CPUE exaggerates the level of depletion, and is commonly recognized in the early development of many tuna fisheries). We imposed several different values for the SS non-linear abundance-CPUE relationship parameter H (equally for all longline fleets), where $I = QN^{1+H}$ (see Figure 14). If the retrospective pattern is a simple result of this non-linearity, we would predict that the retrospective pattern would be more exaggerated with negative values of H (hyperstability), and diminish with increasing H (eventually reversing direction, such that historical assessments would be shown to be too optimistic).

Different H values (-0.5 to 0.5) had an impact on the stock status inferences and retrospectives, but not in a consistent monotonic pattern in either case. It might be argued that H=0.1 has the best retrospective pattern for spawning output, and H=0.5 has the best pattern for depletion. However, the negative log-likelihood favours H = 0, which argues against adopting H > 0. This may be worth further investigation, but we would argue against adopting H <> 0 for the OM at this time. If the CPUE-abundance non-linearity is operating, it could be far more complicated, e.g. differing by region, probably varying as a function of time and confounded with other CPUE factors such as technological change and the actual amount and distribution of effort within regions.

We also recognize that temporal trends in reported catch bias might contribute to a problematic retrospective pattern, but this has not been examined. We also did not examine the extent to which the retrospective results could be an artefact of unstable minimization.

Table 5. Bigeye SS objective function values for a range of fixed values for the hyperdepletion parameter H.

H (non-linearity parameter)	Objective function value	Relative objective function
-0.5	5752.07	1135.96
-0.2	4715.88	99.77
-0.1	4645.61	29.5
0 (assessment value)	4616.11	0
0.1	4648.82	32.71
0.2	4701.94	85.83
0.5	4892.52	276.41

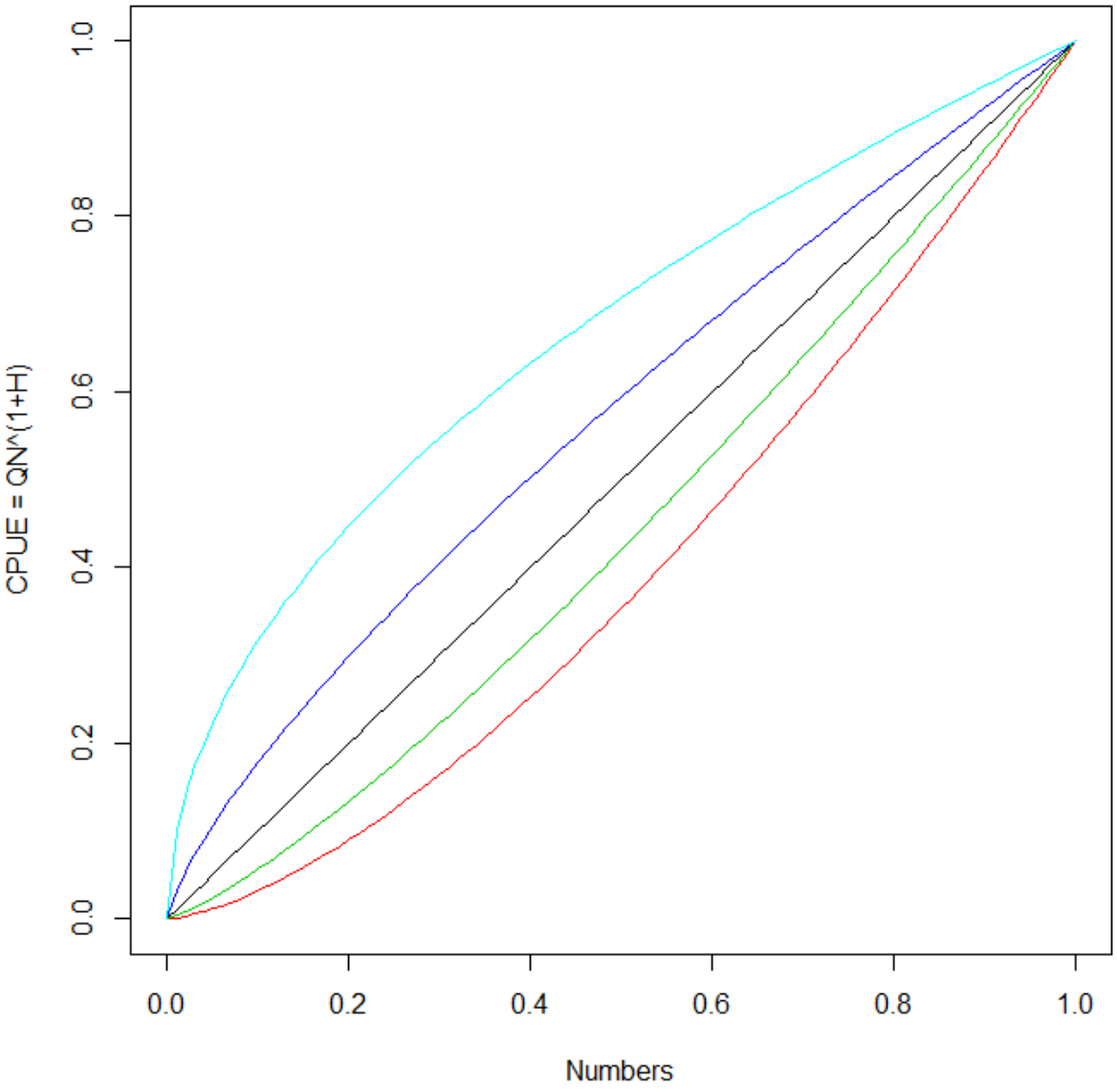


Figure 14. SS non-linear CPUE – abundance relationship for values of H from +0.5 (red, hyperdepletion) to -0.5 (pale blue, hyperstability).

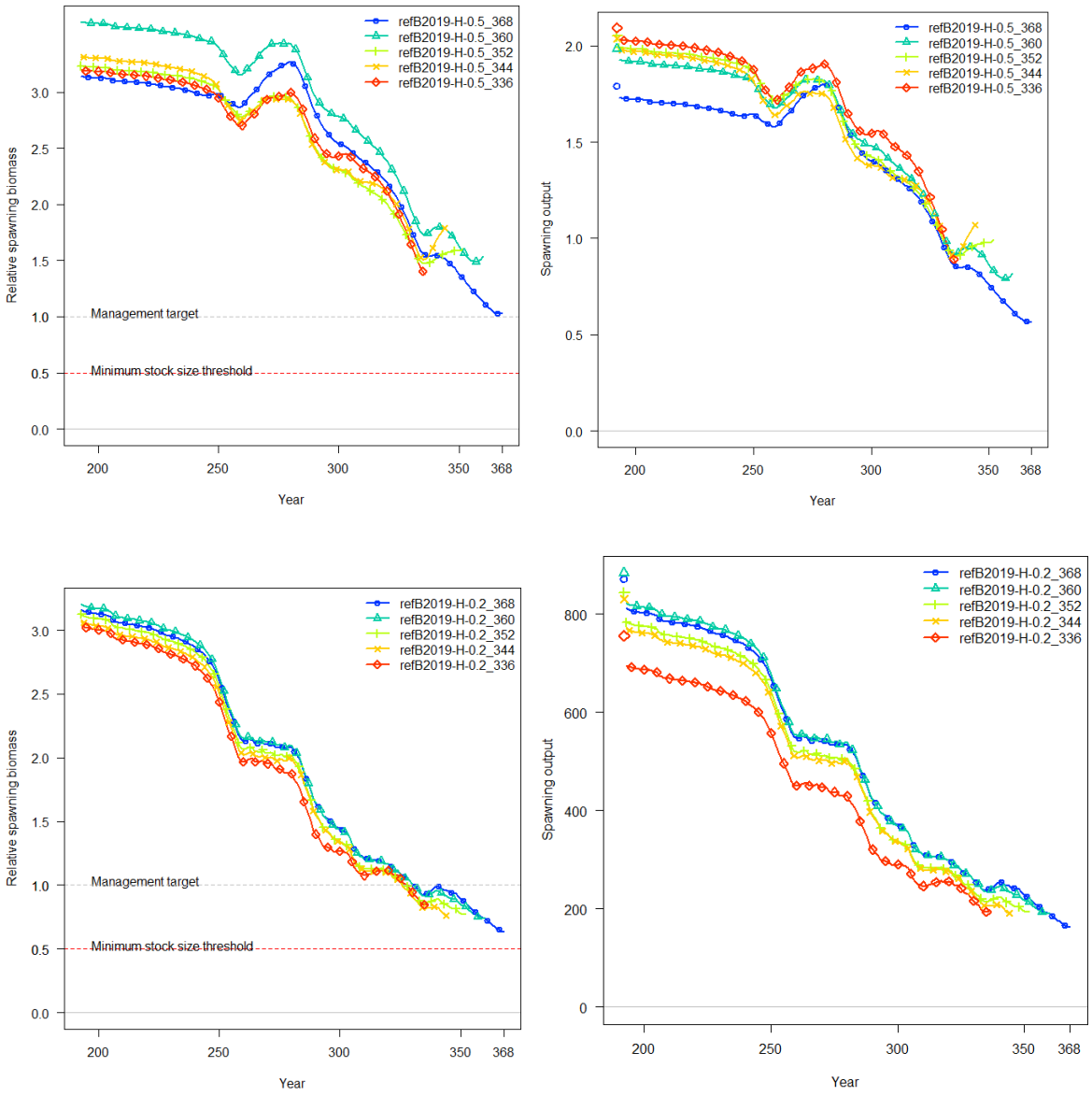


Figure 15. r4ss Retrospective biomass patterns for a reference case model with non-linear CPUE relationship to abundance from strong hyperstability (H = -0.5) to strong hyper-depletion (H = 0.5). Plot continues below. Note that we are not sure why the r4ss "Spawning output" Y axis units vary by a factor of 800. (Figure 15 continued on following pages)

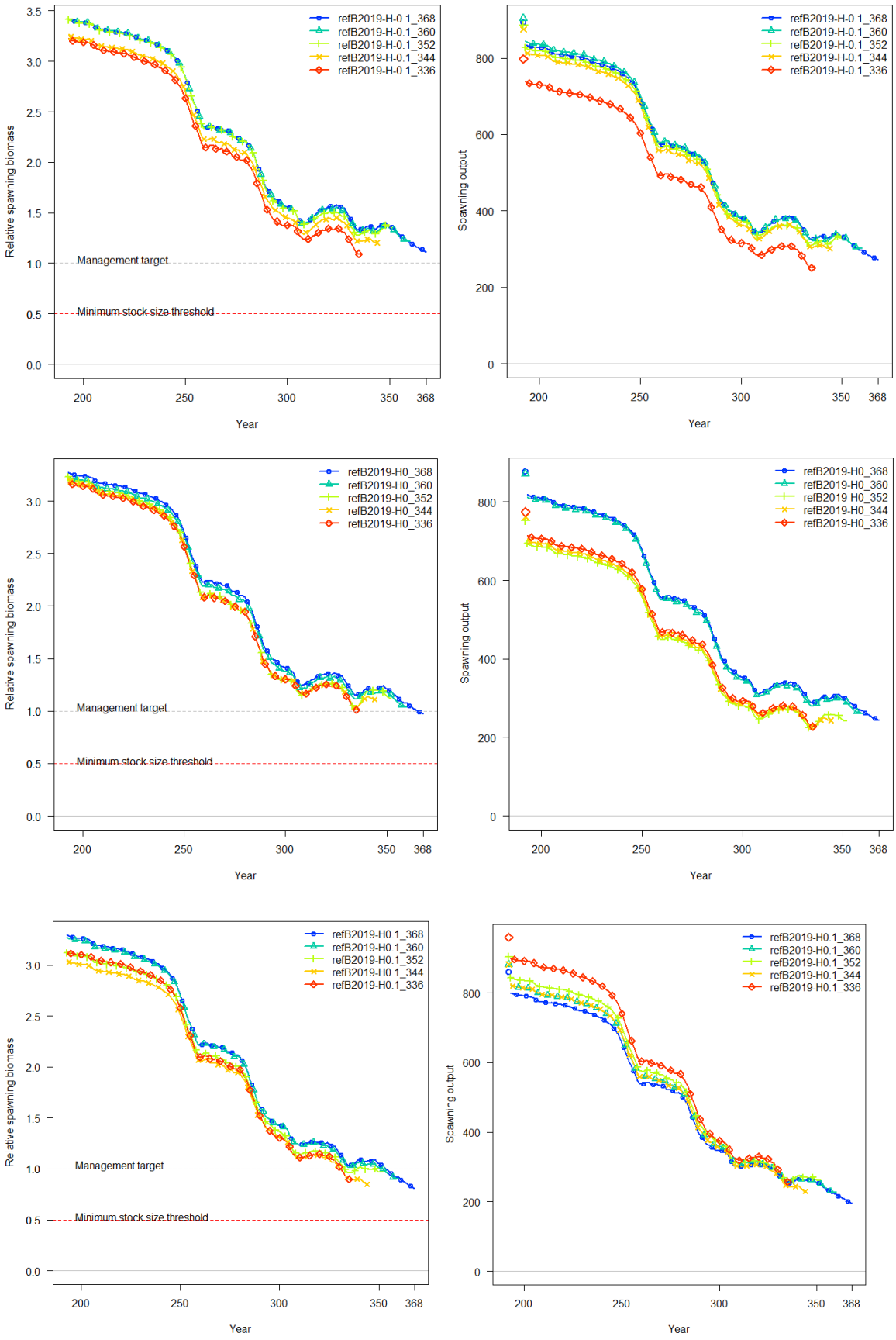


Figure 15 (cont.)

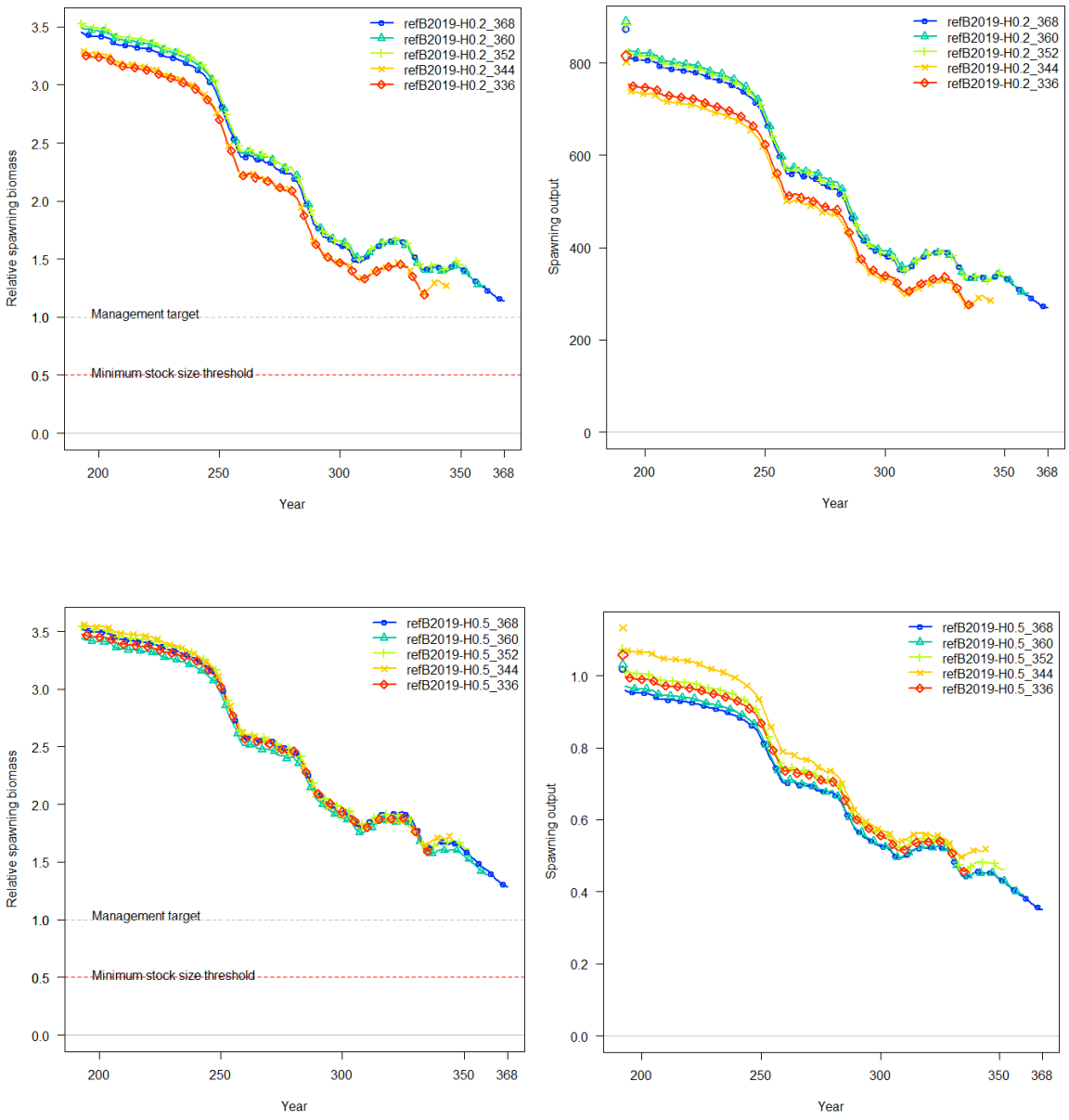


Figure 15 (cont.)

5 Iterative reweighting

Given the large amounts, varying quality, and disparate types of data in these assessments, combined with the structural simplifications that are required to produce tractable estimators, it is inevitable that different data sources will influence model inferences in different ways (i.e. resulting in conflicts). Iterative reweighting (sometimes called rescaling, or tuning – not to be confused with MP tuning referred to in section 8) is a method of improving the consistency between variance-related assumptions and the model quality of fit to the data. Most of the indices (CPUE, surveys, tags and composition data) used in fisheries underestimate their true variance by only reporting measurement or estimation error and not including process error. The variance-related assumptions include the recruitment deviation penalty, CPUE observation error, size composition sample sizes and tag-related assumptions, etc. If the internal consistency of the model is improved, it is expected that the statistical properties of the model will be improved. However, we note that the approach i) does not directly address issues related to systematic lack of (i.e. due to model specification errors and/or data biases), and ii) tag data were omitted from the process.

The specific approach tested is recommended as current best practice in the Pacific Fishery Management Council (2018), and described in more detail in the Kolody et al. (2020) YFT companion paper. We do not assert that the approach is necessarily better than the ad hoc approaches that have been applied to bigeye to date – it is more a test of sensitivity, to see if this issue requires more urgent attention.

For bigeye, the initial iterative reweighting steps involved adjusting only the bias ramp and the catch-at-length (CL) sample sizes without estimating any additional standard error on the CPUE series (for yellowfin this was done in a single step). The approach explored here strictly should have involved reducing some CL sample sizes lower than 1.0 for some fisheries. But in a multinomial context, N=1 is already minimally informative). We mostly consider this to be a reminder that the model and data are simply not very compatible for some fleets. There is probably not much that can be done to improve the sampling biases in the historical data for some fleets, and improved sampling should continue to be encouraged in the future. But there may also be model structural issues, e.g. the current growth curve may not be appropriate for the whole Indian Ocean, and stationary fishery selectivity is probably a tenuous assumption in some cases.

The impact of four steps of iterative reweighting for the bigeye tuna reference case is shown in Figure 16. Differences in the estimates of recruitment events are shown in Figure 17 with differences in the fits to CPUE (in both normal and log scale) shown in Figure 18, for fleet 16. It is notable that more iterations of reweighting tended to converge toward the original model (e.g. refBET2019Tune4).

Allowing estimation of additional standard error to the CPUE was subsequently added to refBET2019Tune4, resulting in refBET2019Tune5cpue. We would not necessarily advocate this step, i.e. if it results in a poor fit to the CPUE, this conflicts with the general principle that assessments of this sort are probably not very useful if the relative abundance index is not well fit (or alternatively, if the relative abundance indices are biased, it is not likely that the other data are

going to be able to compensate for the bias). However, it is reassuring that adding this final step resulted in stock assessment inferences that appear to be even closer to the original assessment.

It is difficult to conclude much from this analysis, except that the implicit or explicit iterative reweighting conducted in the recent BET assessment appears to yield very similar point estimate results to that achieved using the approach advocated by the Pacific Fishery Management Council (2018). It is also reassuring that the CPUE appears to be slightly over-weighted and the size composition somewhat under-weighted, which helps ensure that the information in the relative abundance indices is dominating. The argument might be made that the analysis is worth repeating on more models in the OM ensemble, but this adds a big computational overhead, and we do not have evidence to suggest that this is urgent (and a systematic approach for adding the tags would be warranted).

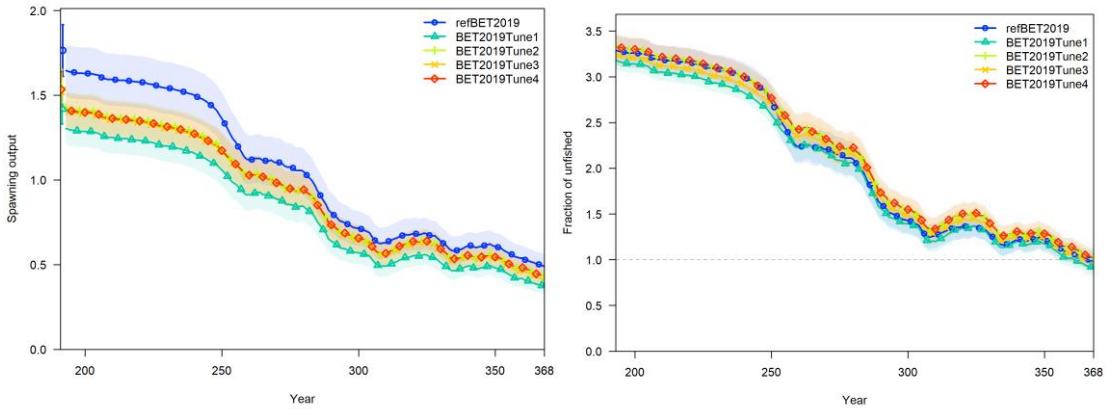


Figure 16. Comparison of the absolute (left) spawning output and relative to SBMSY (spawning biomass) (right) from the reference case (refBET2019, blue) and from four steps in iteratively reweighting the model, with asymptotic confidence intervals.

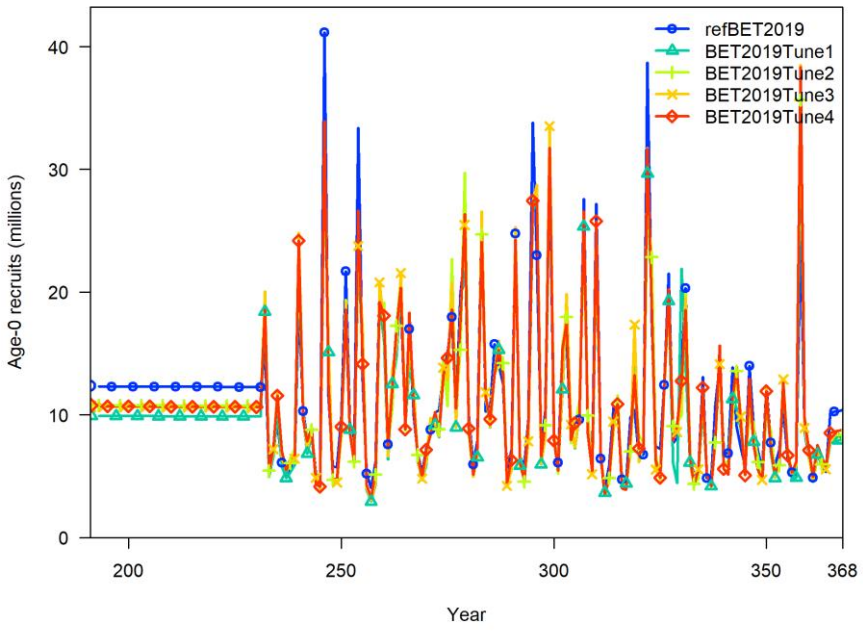


Figure 17. Comparison of the recruitment deviations from the reference case (refBET2019, blue) and from four steps in iteratively reweighting the model.

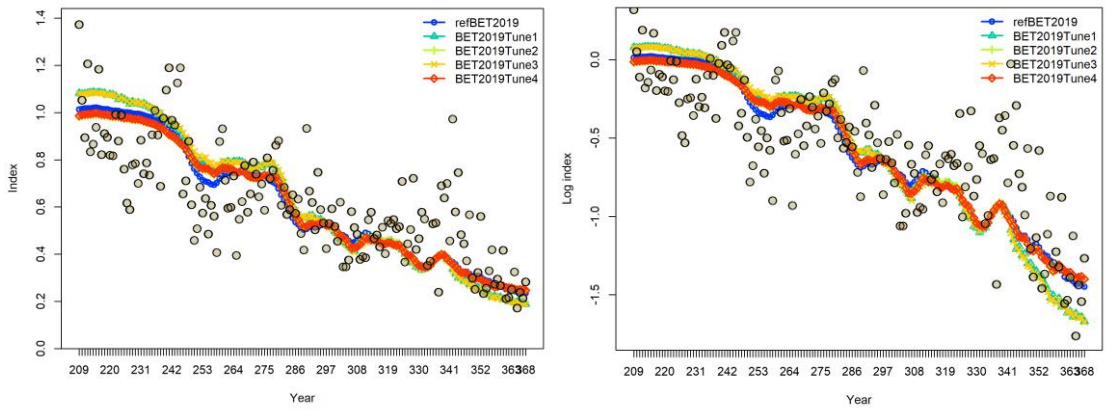


Figure 18. Fits to the CPUE for fleet 16 the reference case (refBET2019, blue), and from four steps in iteratively reweighting the model.

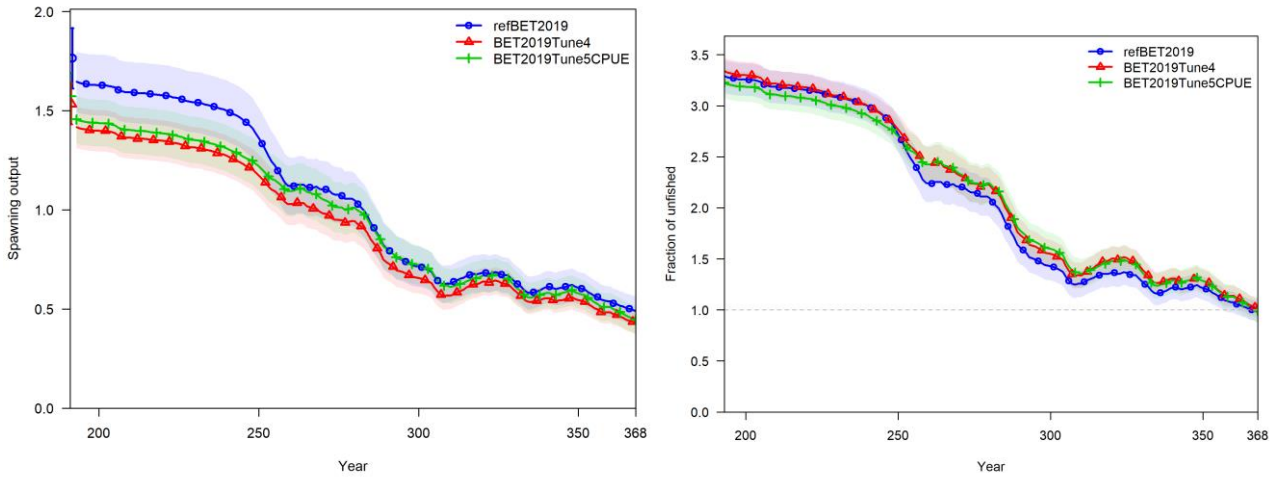


Figure 19. Comparison of the absolute (left) spawning output and SB relative to SBMSY (spawning biomass) (right) from the reference case (refBET2019, blue), the iteratively reweighted model without additional standard error on the abundance indices (BET2019Tune4, red) and estimating additional standard error to that reweighted model (BET2019Tune5CPUE, green).

6 Generating the Operating Model Ensemble

6.1 Fractional Factorial Experimental Design

The concept of fractional factorial design was developed in recognition that it is not practical to run experiments with every possible combination of interactions among a large number of factors, and even if it were possible, appreciable 3 way (and higher) interactions tend to be very rare. As in previous iterations, we used the R package "planor" to propose fractional designs for mixtures of 2 and 3 level factors, opting for a 72 model main effects design. Previous MSE iterations demonstrated negligible MP performance differences between a main effects grid, a much larger 2-way interaction grid and the full factorial grid.

6.2 Parameter estimation sensitivity to initial values

It is recognized that parameter estimation is often sensitive to initial conditions in SS3 assessments (e.g. due to very flat and/or polymodal likelihood surfaces), and the bigeye model is no exception. It would be very difficult to ever conclude that one of these highly-parameterized models has truly reached the global minimum, so we have usually assumed that a model reaching the satisfactory convergence criterion (absolute value of the maximum gradient of the objective function with respect to the parameters < 0.01) should generally be reasonable and informative, even if it is not the best. This may be a risky approach if one is relying on a single (or small number) of models, however, previous iterations of the MSE process demonstrated that MP performance was almost identical regardless whether the OM ensemble consisted of the best or worst fit of the (3 jittered and converged replicate) models. However, we have retained the practice of seeking 3 converged runs from jittered initial conditions, for each model specification, and retaining the one with the lowest objective function value. This should greatly reduce the impact of outlier minimizations.

6.3 Parameters on Bounds

As in previous iterations of OM development, the configuration files for the bigeye OMs had several bounds and prior distributions relaxed relative to the original assessment, to reduce unintended consequences of these somewhat arbitrary constraints. This relaxation presumably could have consequences for the minimization speed and sensitivity to initial conditions in some cases, but this has never been explored.

7 Operating Model OMrefB20.1 characteristics

OMrefB20.1 is derived from the equally-weighted grid OMgridB20.1, i.e. there were no minimization failures in OMgridB20.1, and no models were rejected in the basis of quality of fit diagnostics or the catch likelihood (high F) problem.

Summary diagnostic plots for OMgridB20.1 are shown in Figure 8 - Figure 9 and Figure 21 - Figure 27, from which we note:

- The quality of fit to the CPUE, size composition and tag data are similar to previous iterations.
- The quality of fit (RMSE) between the annualized MP CPUE and OM longline vulnerable numbers is usually better than we would have reason to expect. Note that the RMSE indices presented below are not exactly comparable to those presented in Figure 8. To maintain consistency between the OM and the MP CPUE series, the systematic lack of fit is now described with model-specific CPUE CV and auto-correlation (as described in section 3.1).
- The recruitment variability is in line with recent iterations, and there are no anomalous deviations in the recent period. There were no substantial recruitment residual trends.
- The OM stock status is notably different from both the previous iteration of the OM and the recent stock assessment:
    - OMrefB20.1 current depletion is somewhat more pessimistic than the assessment in terms of SB(T)/SB(MSY) and B(T)/B0, and the variability appears to be of a similar scale. We would have expected the OM to have markedly higher variability due to the larger number of assumptions in the grid. This might indicate that the extra dimensions tend to be of lesser importance (or the OM might actually have a higher CV because the mean is lower and the variance is similar to the assessment).
    - The biggest factor influencing the OM depletion is the catchability trend assumption: 0 or 1% per year. The option of moving the 1% per year catchability trend to a robustness test was briefly discussed at the 2019 WPTT, but the CPUE consultant judged that this was an important option to retain in the reference set OM.
    - OMrefB20.1 MSY is slightly more optimistic than the assessment, and considerably more variable.

The reference set OM projection assumptions were the same as previous iterations, except:

- MP-based management was set to start in 2021, and the bridging catches for the intervening years were updated from the WPTT 2019 figures (2018 “scientific” catch of 81400 t).
- The annual aggregate CV and autocorrelation were derived from the new method discussed in section 3.1.

As a general plausibility check, some standard time series plots are shown in Figure 28 for OMrefB20.1, with a fishing moratorium and constant current catches (starting in 2021, i.e. MPs

CC001 and CC081 respectively). Both scenarios tend to suggest spawning biomass rebuilding over the next 20 years.

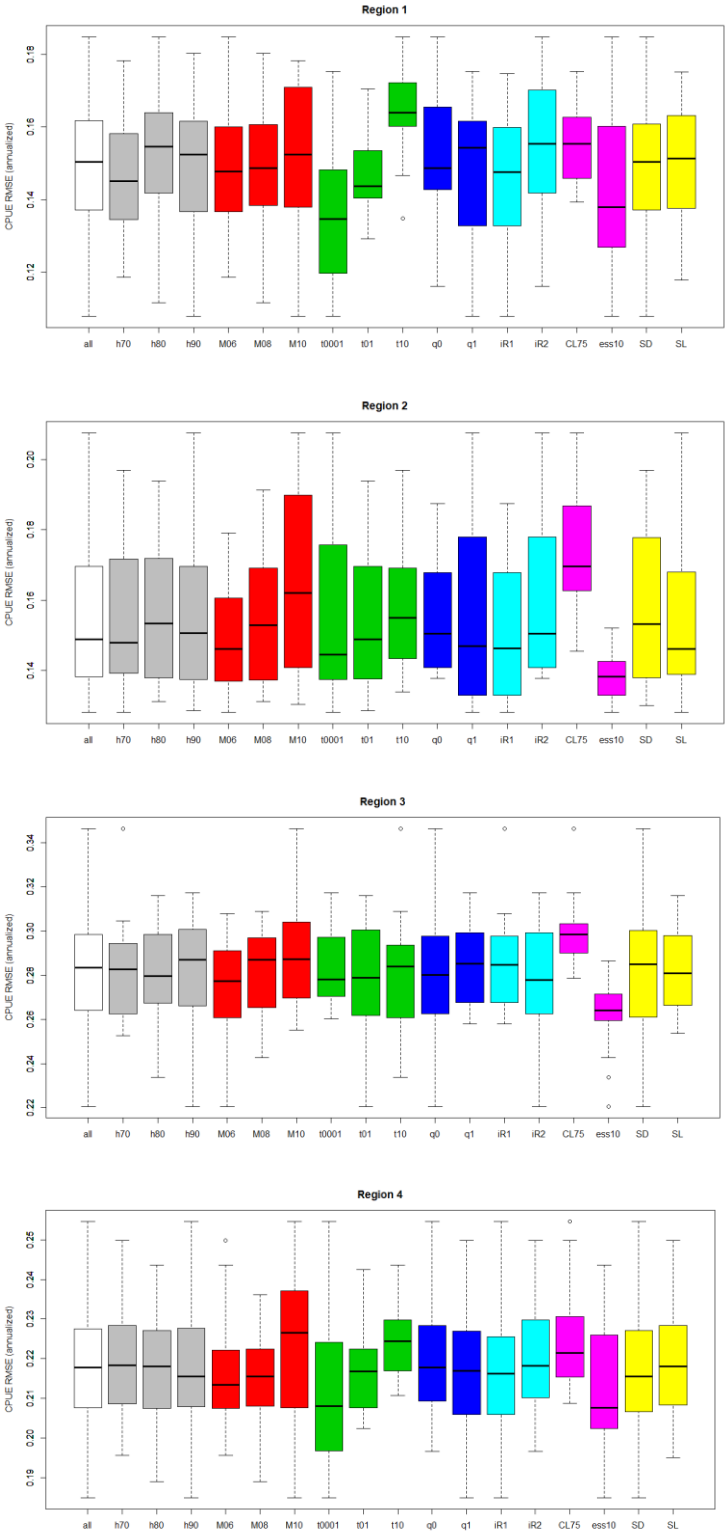


Figure 20. OMgridB20.1 CPUE fit (RMSE) by region for OMrefB20.1.

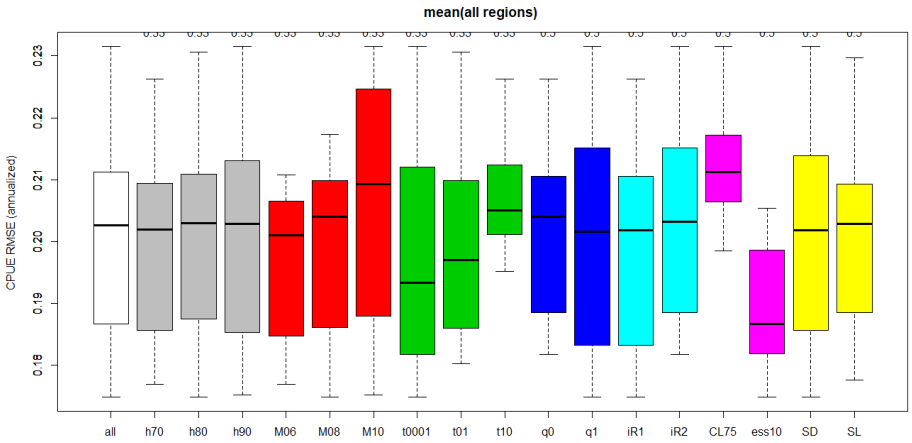


Figure 21. Comparison of annualized CPUE fit among BET OM ensembles, based on the mean among areas.

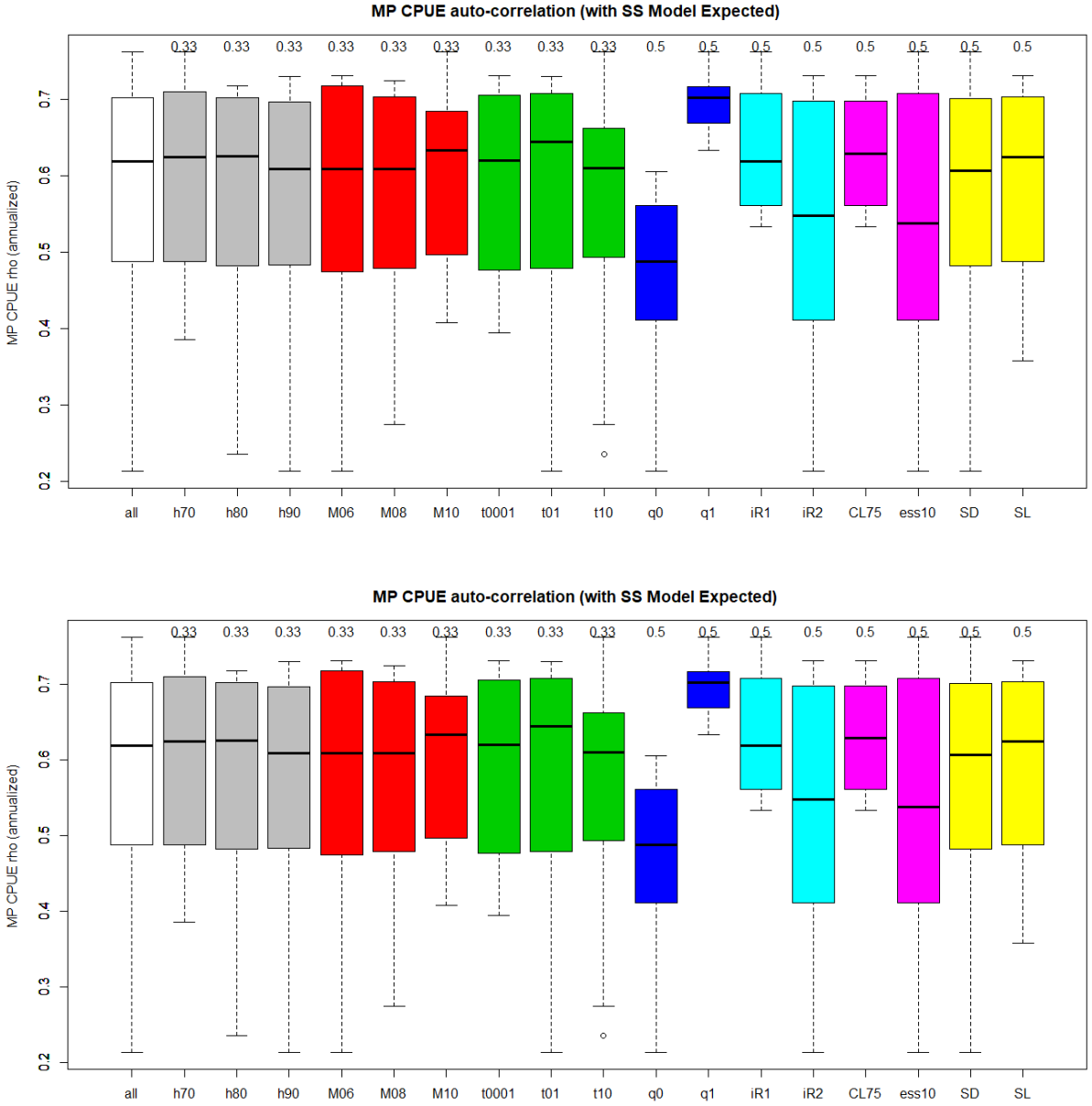


Figure 22. OMgridB20.1 auto-correlation (lag(1)) between the annualized MP CPUE fit to the predicted longline vulnerable biomass. Top panel normalized over the whole of the available time series, bottom panel normalized over the most recent 5 years.

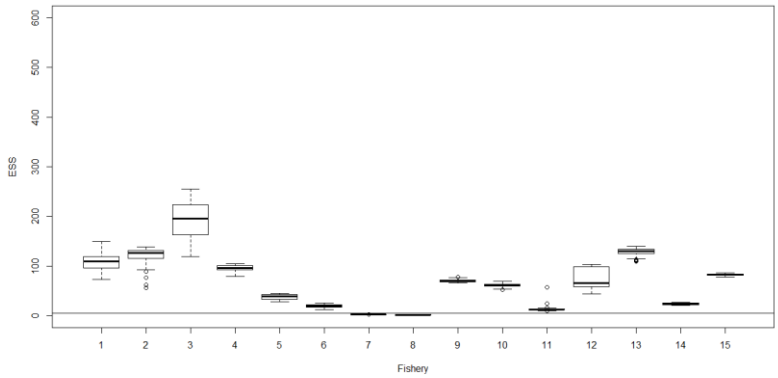
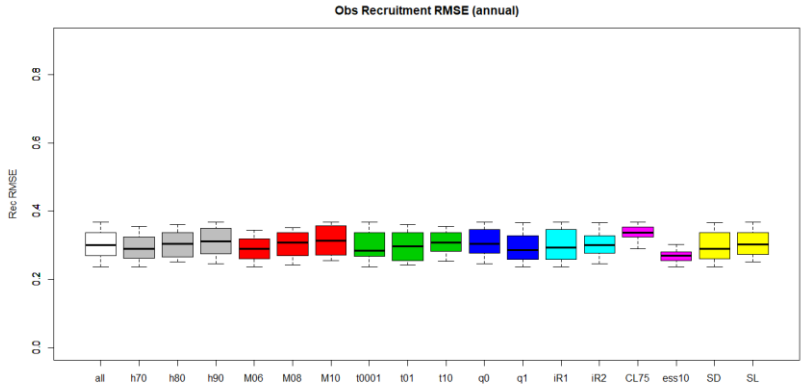


Figure 23. OMgridB20.1 fit to the size composition data by fishery. Each box represents the distribution across all models in the grid.



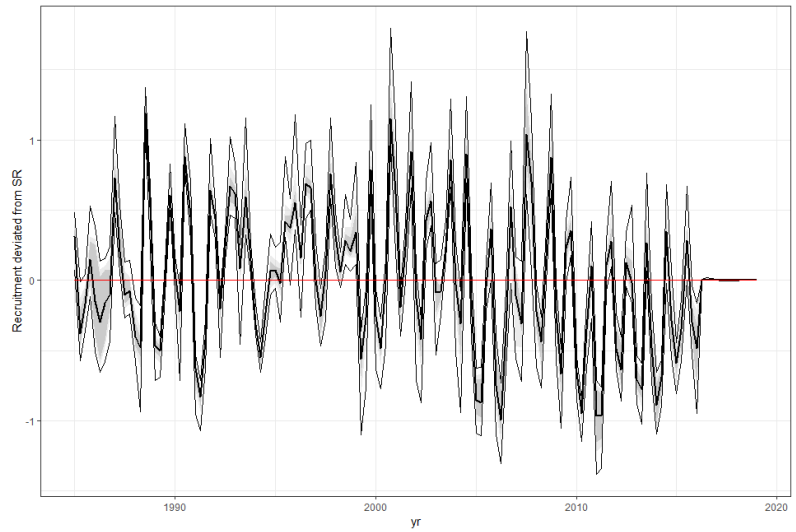


Figure 24. Recruitment deviate CV distribution (top) and time series from OMgridB20.1. Note that constant recruitment from recent assessment and future projection years are included in the lower panel, but initial OM numbers-at-age are subject to stochastic error to account for this.

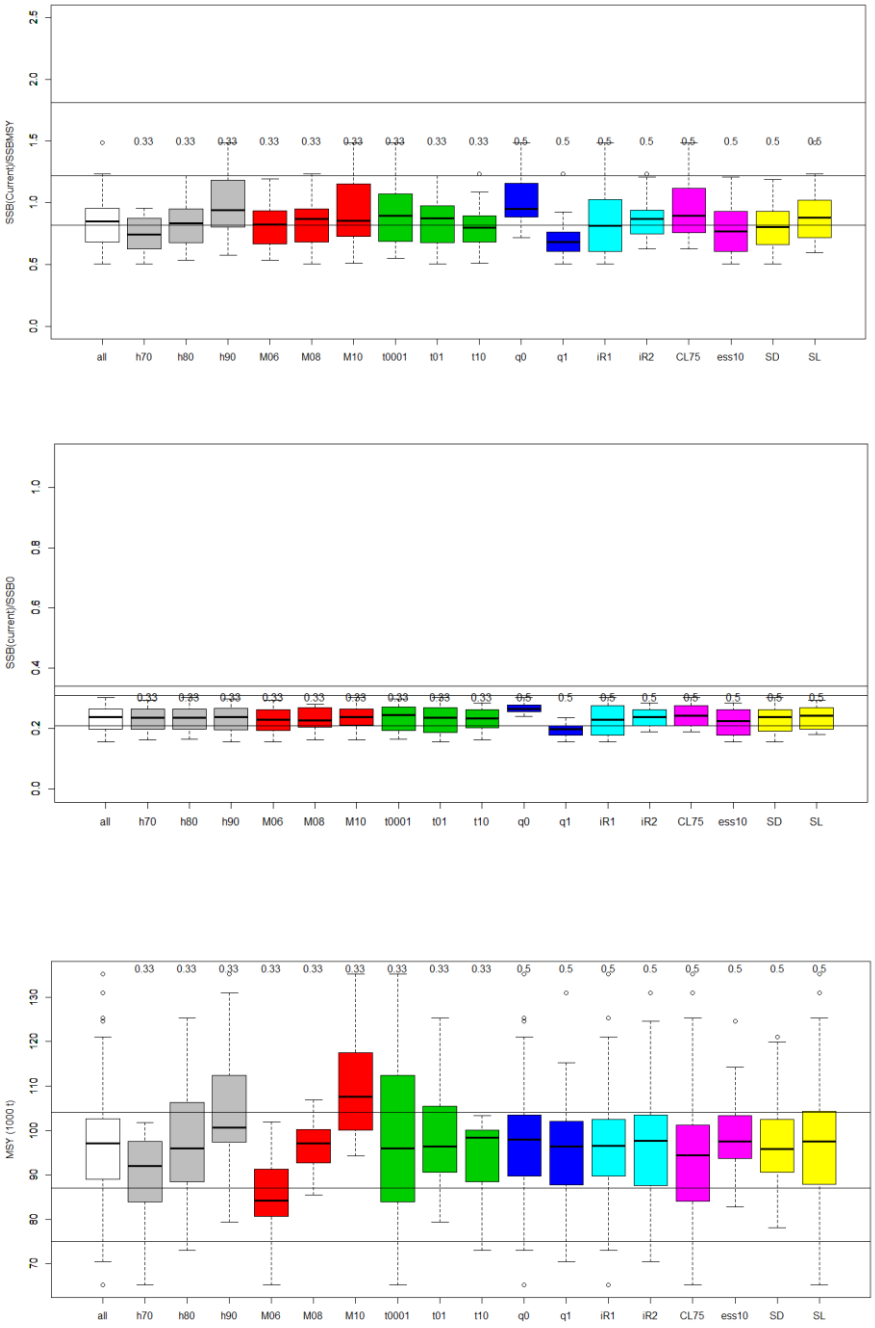


Figure 25. OMgridB20.1 distribution of stock status estimates, partitioned by assumption. Reference lines are the 10, 50 and 90th percentiles reported in the SC stock status reports.

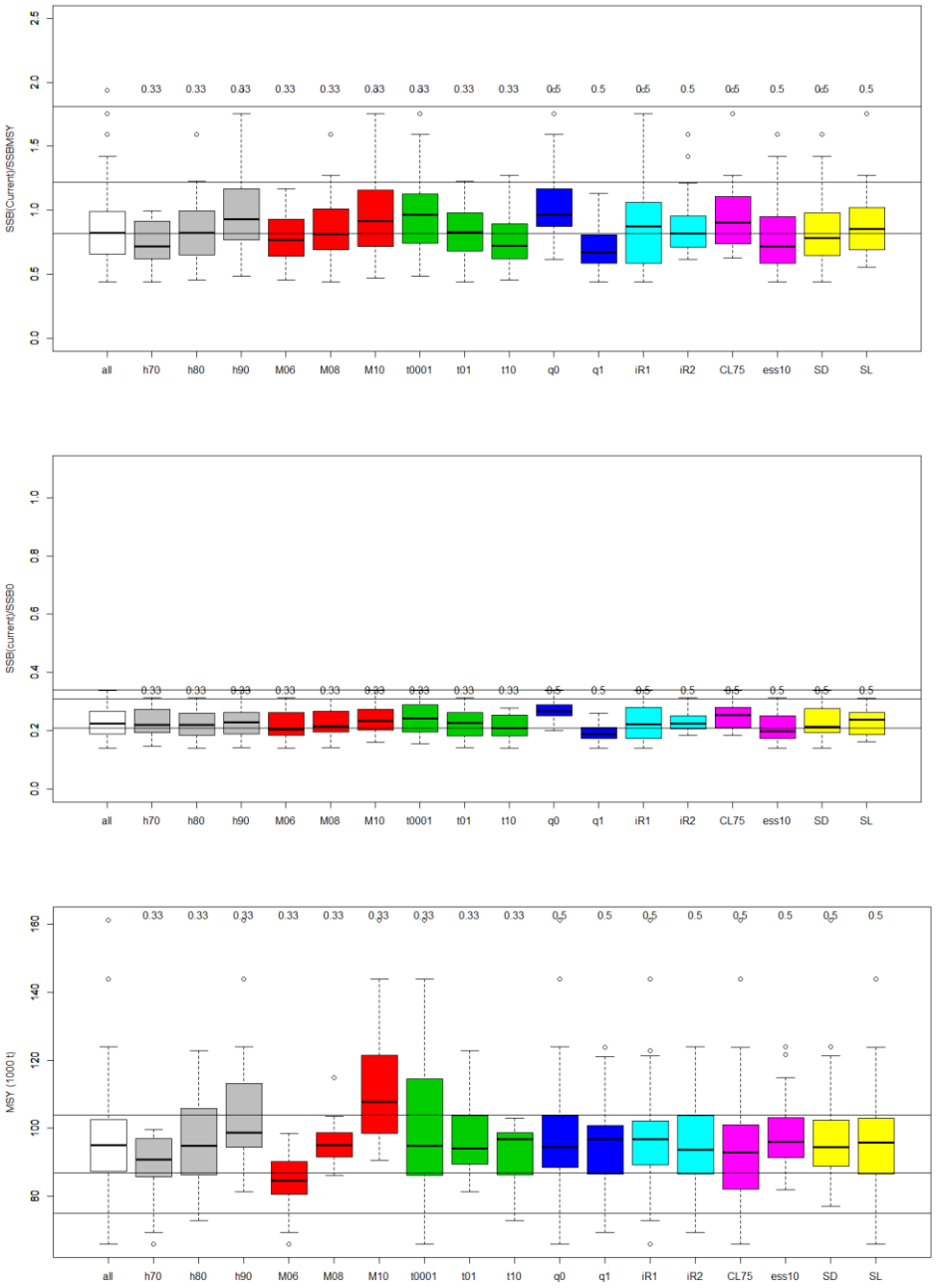
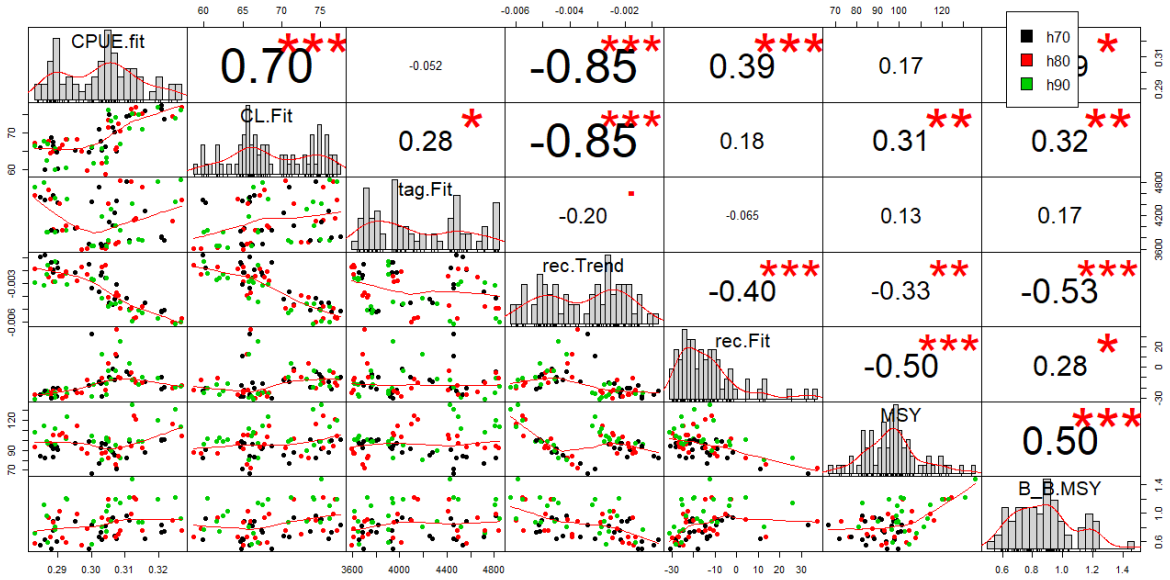
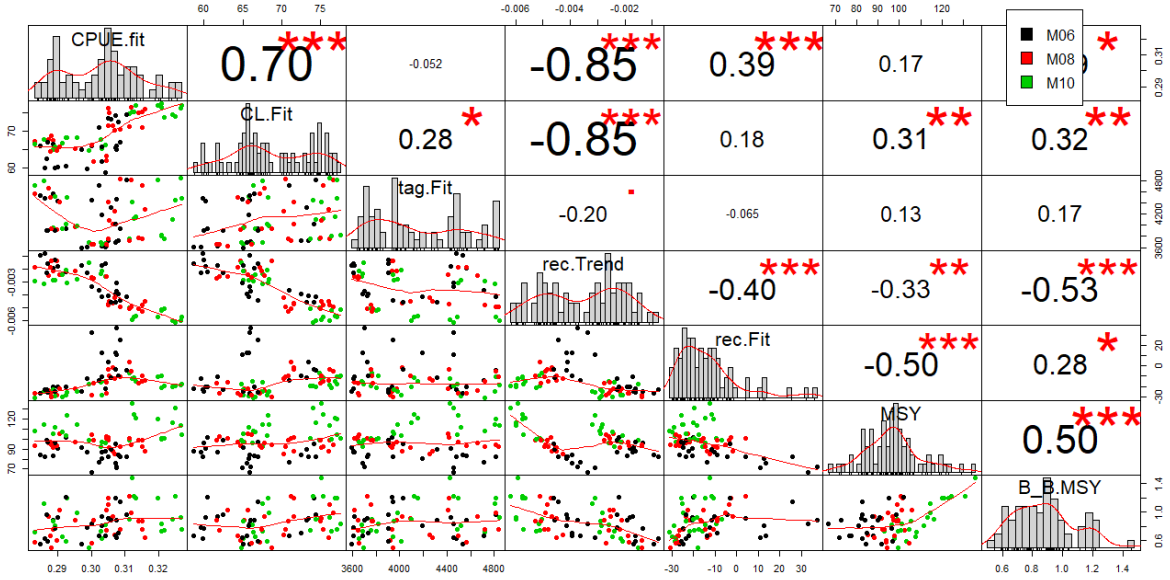
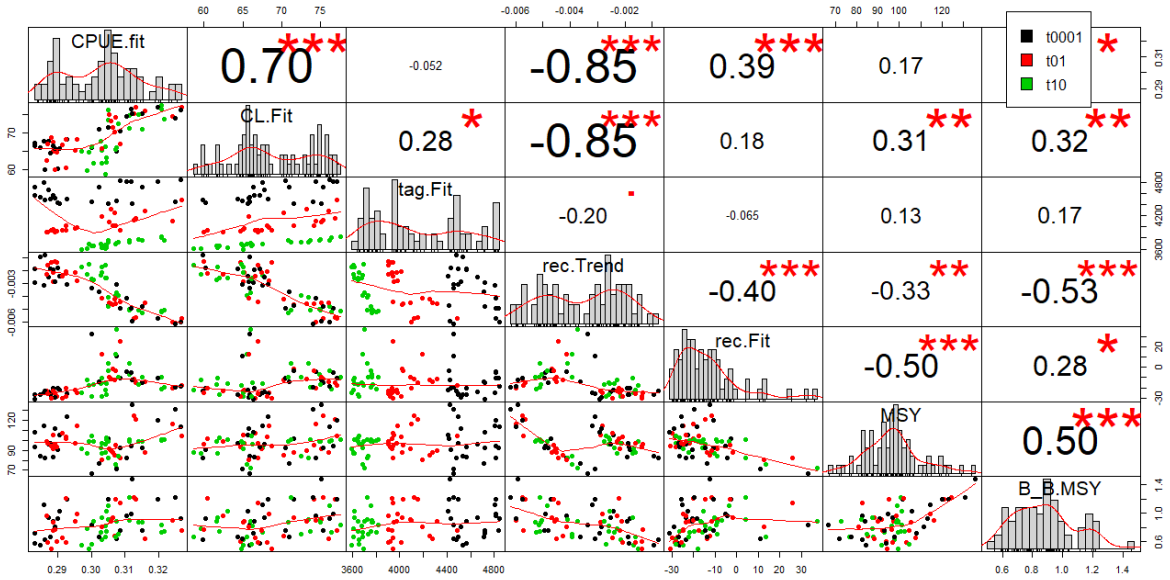
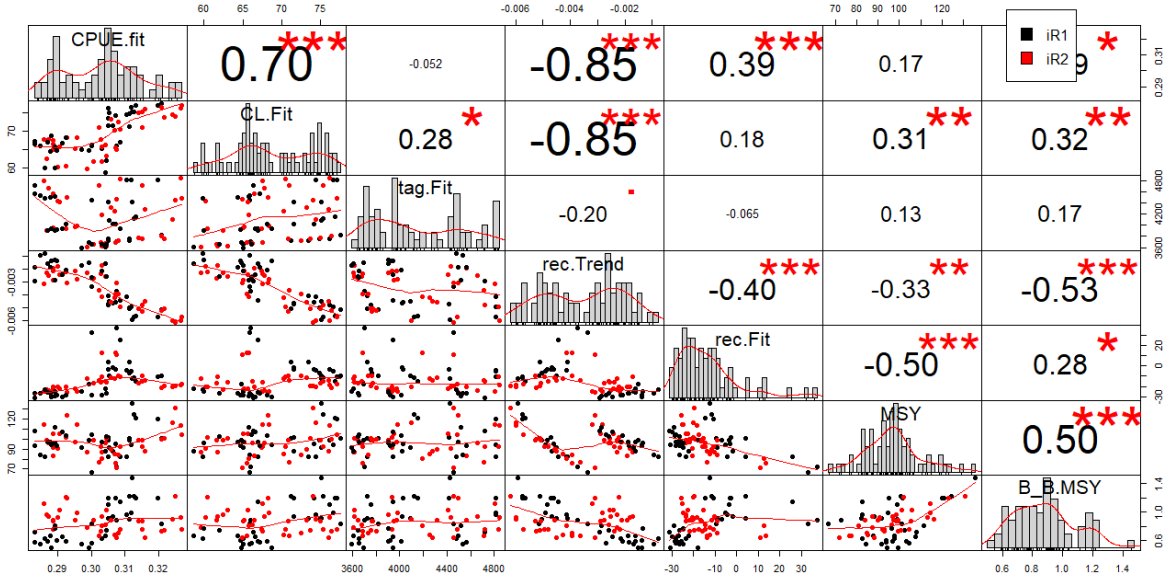


Figure 26. omGridB20.2 distribution of stock status estimates, partitioned by assumption. Reference lines are the 10, 50 and 90th percentiles reported in the SC stock status reports. i.e. Key point is that there is almost no difference relative to the omGridB20.2 without the seasonally-partitioned temperate CPUE


CPUE.fit
CL.Fit
tag.Fit
rec.Trend
rec.Fit
MSY
B_B.MSY
h70
h80
h90
0.70 ***
-0.052
-0.85 ***
0.39 ***
0.17
0.28 *
-0.85 ***
0.18
0.31 **
0.32 **
-0.20 .
-0.065
0.13
0.17
-0.40 ***
-0.33 **
-0.53 ***
-0.50 ***
0.28 *
0.50 ***


CPUE.fit
CL.Fit
tag.Fit
rec.Trend
rec.Fit
MSY
B_B.MSY
M06
M08
M10
0.70 ***
-0.052
-0.85 ***
0.39 ***
0.17
0.28 *
-0.85 ***
0.18
0.31 **
0.32 **
-0.20 .
-0.065
0.13
0.17
-0.40 ***
-0.33 **
-0.53 ***
-0.50 ***
0.28 *
0.50 ***


CPUE.fit
CL.Fit
tag.Fit
rec.Trend
rec.Fit
MSY
B_B.MSY
t0001
t01
t10
0.70 ***
-0.052
-0.85 ***
0.39 ***
0.17
0.28 *
-0.85 ***
0.18
0.31 **
0.32 **
-0.20 .
-0.065
0.13
0.17
-0.40 ***
-0.33 **
-0.53 ***
-0.50 ***
0.28 *
0.50 ***


CPUE.fit
CL.Fit
tag.Fit
rec.Trend
rec.Fit
MSY
B_B.MSY
iR1
iR2
0.70 ***
-0.052
-0.85 ***
0.39 ***
0.17
0.28 *
-0.85 ***
0.18
0.31 **
0.32 **
-0.20 .
-0.065
0.13
0.17
-0.40 ***
-0.33 **
-0.53 ***
-0.50 ***
0.28 *
0.50 ***

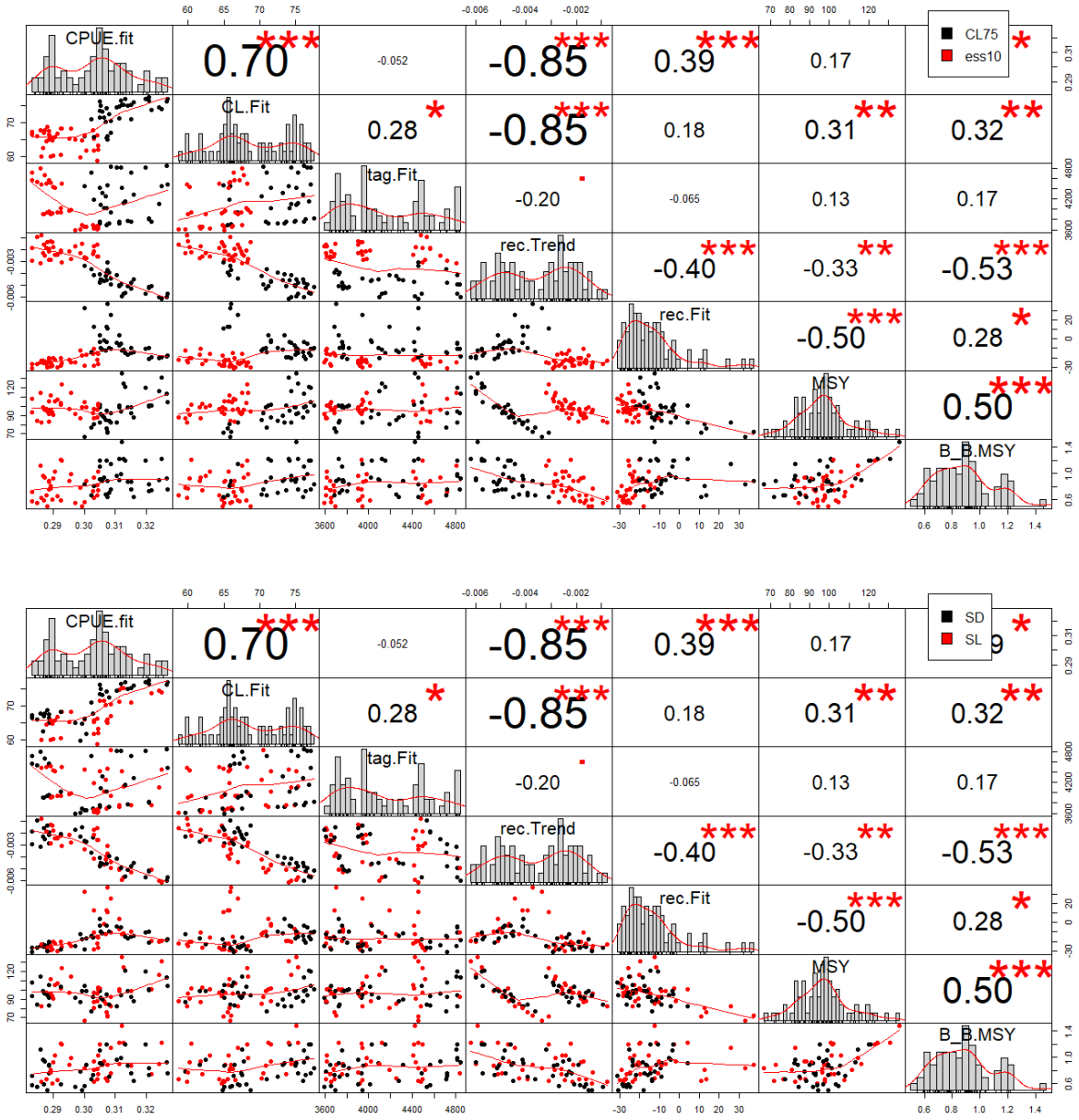


Figure 27. Multiway comparison of OMgridB20.1 model characteristics, partitioned by assumption by colour.

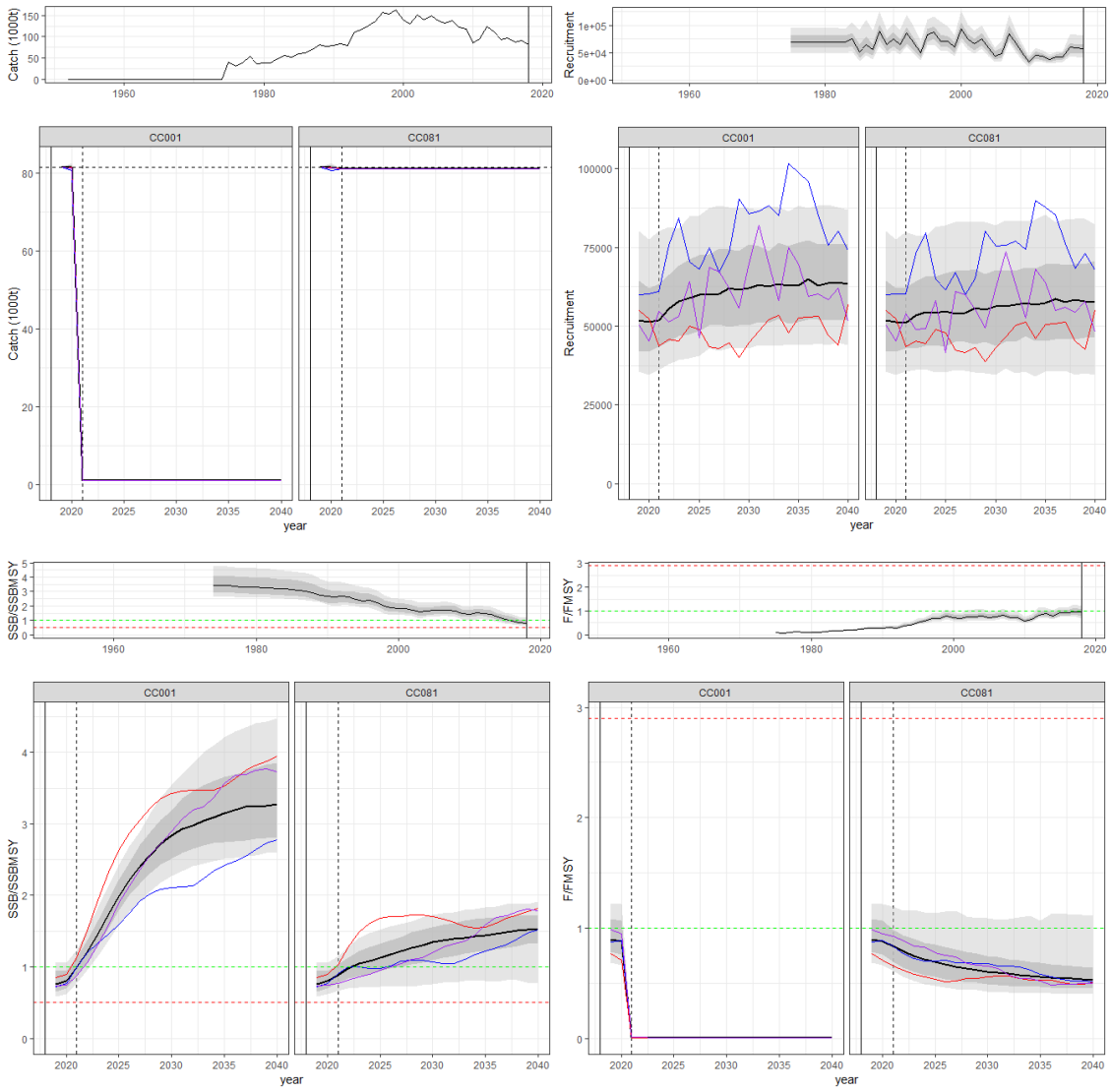


Figure 28. OMrefB20.1 basic time series plots for fishing moratorium (CC001) and constant current catch (CC081) projections.

8 Bigeye Reference set and Robustness test MP evaluation results

A suite of MPs (including constant catch projections, empirical and model-based MPs) were applied with the following conditions:

- Tuning objectives B18.2 and B18.3 (60 and 70% chance of being in the green Kobe zone over 2030-2034) using OMrefB20.1.
- First quota applied in 2021, with intermediate catches set at 2018 levels.
- Quota setting every 3 years, with a maximum 15% change constraint
- 2 year data lag (2021 quota set using data up to 2019).

A subset of 6 MPs (3 per tuning objective) are presented below (Figure 30 - Figure 36), selected to represent contrast and facilitate discussion about the sort of behaviour that is achievable and desirable. The standard plots for the reference set OM are presented first, followed by the corresponding robustness test results (interspersed to facilitate easy comparison).

With respect to the reference set performance, we note:

- MP performance is qualitatively similar to previous iterations. We do not identify any obvious need to provide more performance contrast through alternate tuning objectives. There is some tendency for MPs to increase biomass early on and decrease it toward the end of the projection period, but this tendency is not obviously problematic, particularly when coupled with the expectation that any MP would likely be reviewed within 10 years of adoption and revised as appropriate.
- All MPs are predicted to have higher than current catches on average over the 20 year projection summary period. The feedback-based MPs yield slightly higher catch with a slightly lower probability of breaching SB limits than the constant catch MPs.

Five robustness tests were requested by the WPTT/WPM 2019, all of which can be addressed by changing the OM projection assumptions (as opposed to reconditioning the OM). The six tuned MPs were run against each robustness OM (i.e. without retuning), from which we note

- OMrobB20.1.ICV3 - Annual aggregated CPUE CV = 0.3, auto-correlation = 0.5
    - The elevated CPUE observation error has a negligible effect on the MP performance.
- OMrobB20.1.10overRep - 10% reported over-catch (projections only; reference case conditioning)
    - This level of sustained over-catch increases the overfishing risk, but the probability of breaching biomass limit reference points remains unlikely for the MPs tested.
- OMrobB20.1.10overIUU - 10% unreported over-catch (projections only; reference case conditioning)

- Result is similar to the previous scenario, the effect of 10% over-catch seems similar regardless of whether it is reported.

- OMrobB20.1.qTrend3 - 3% LL catchability trend (projections only; reference case conditioning)
    - A 3% per year catchability trend has an expected adverse effect on feedback-based MP performance (i.e. the MP is increasingly too optimistic about the stock status over time), and no effect on the constant catch scenarios (because CPUE is not used). However, we would question the usefulness of this robustness test. If it was considered realistic it should be applied to the historical conditioning (and would probably result in very pessimistic stock status). If it is considered plausible, it is a strong imperative for finding a better method of monitoring the fish population.
- OMrobB20.1.recShock - Recruitment shock (as for yellowfin - 55% of the otherwise-expected recruitment for 8 quarters, as shown in Figure 29).
    - The weak cohorts reduce the average productivity of the stock and introduce a dip in the biomass time series, but this does not represent a major risk to the population or fishery in most cases.

The constant catch MPs are retained primarily for scientists and MP developers to consider the value of information. They should not be presented to the Commission because they may create unnecessary confusion. However, it is worth considering whether the feedback-based MPs are extracting as much useful information as we would expect. Currently this might not seem to be the case because:

i) The state of the stock is close to reasonable targets, with no urgent need for substantial disruptive management.

ii) The data (particularly CPUE) are not as informative as we would hope (i.e. many CPUE observations may be required to obtain convincing evidence of the trend and/or the observed CPUE might be substantially biased relative to the OM).

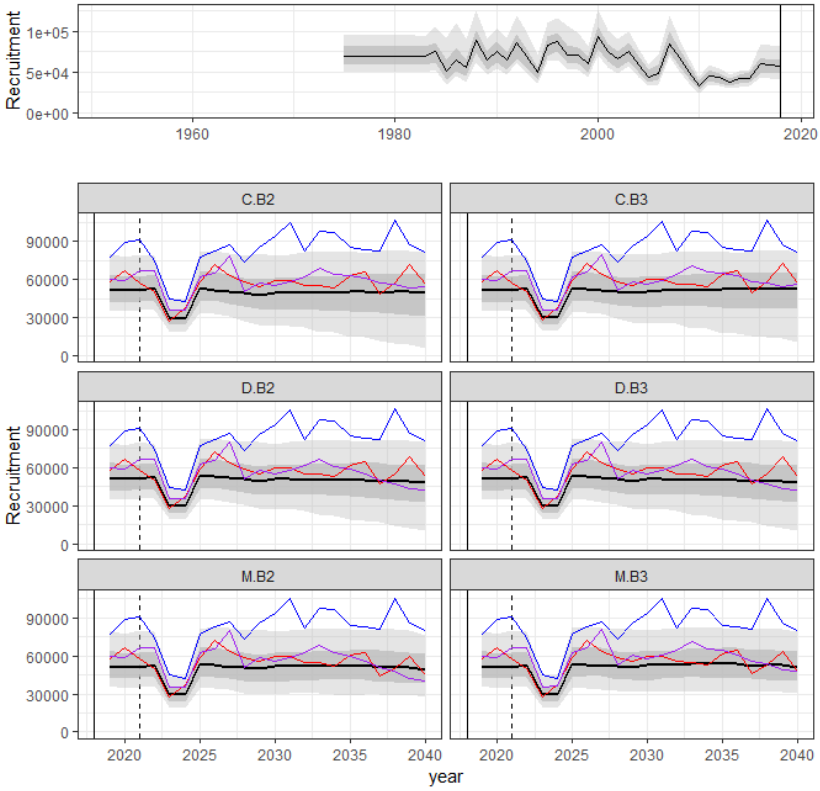
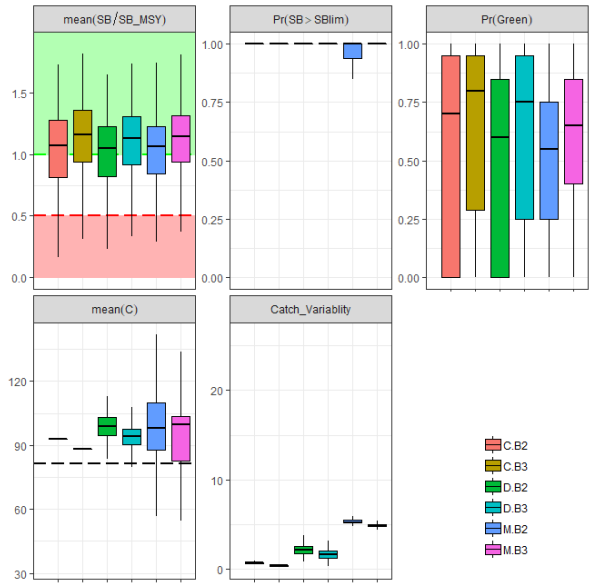
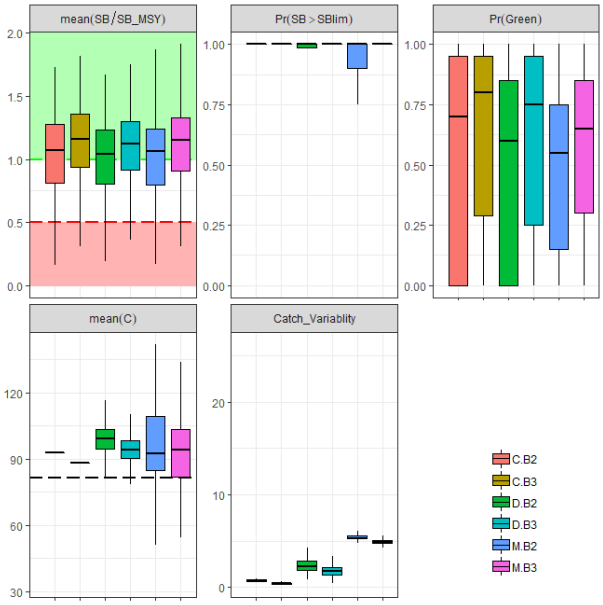


Figure 29. Recruitment time series illustrating the recruitment time series associated with robustness test OMrobB20.1.recShock.

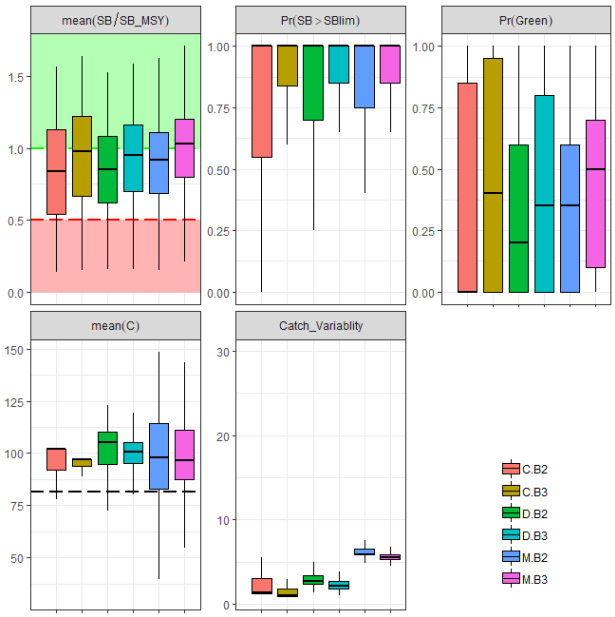


a) OMrefB20.1 – reference set

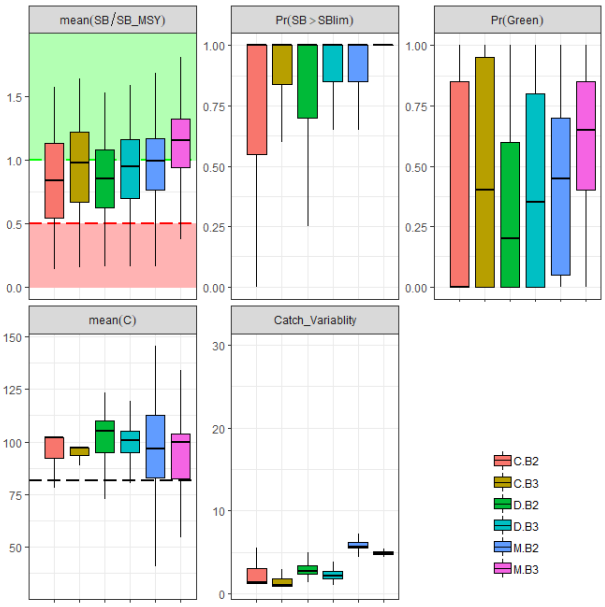


b) OMrobB20.1.ICV3 – CPUE variance = 0.3, all OMs

Figure 30. MP evaluation summaries from the Bigeye reference set OM OMrefB20.1 (a) and robustness tests (b-f). Boxplots compare candidate MPs with respect to key performance measures averaged over the period 2021 - 2040. Horizontal line is the median, boxes represent 25th - 75th percentiles, whiskers represent 10th - 90th percentiles. Red and green horizontal lines represent the interim limit and target reference points for the mean SB/SBMSY performance measure. The horizontal dashed black line is 2018 catch. (Figure 30 continued on following pages)

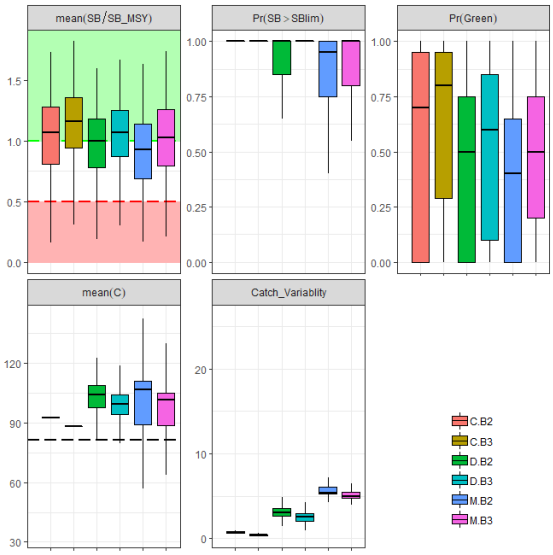


c) OMrobB20.1.10overRep – 10% reported overcatch

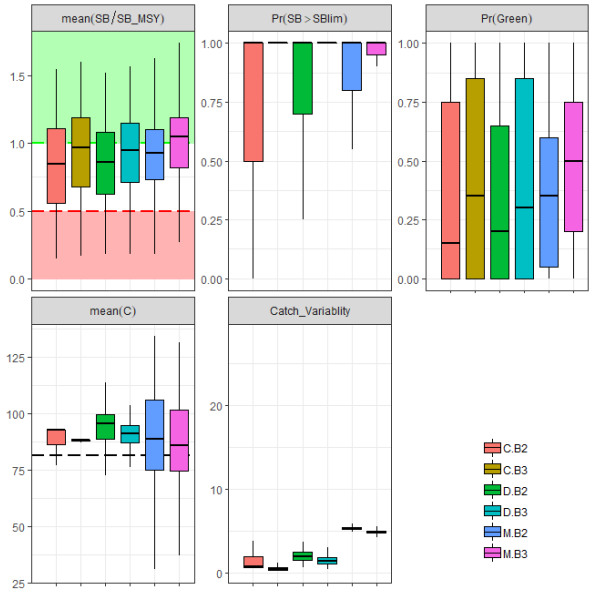


d) OMrobB20.1.10overIUU – 10% unreported overcatch

(Figure 30 cont.)

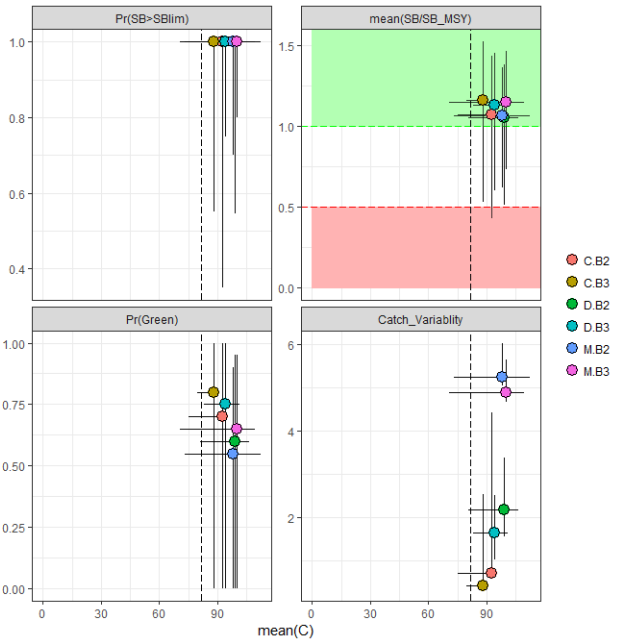


e) OMrobB20.1.qTrend3 – 3% per year longline catchability trend during projection period

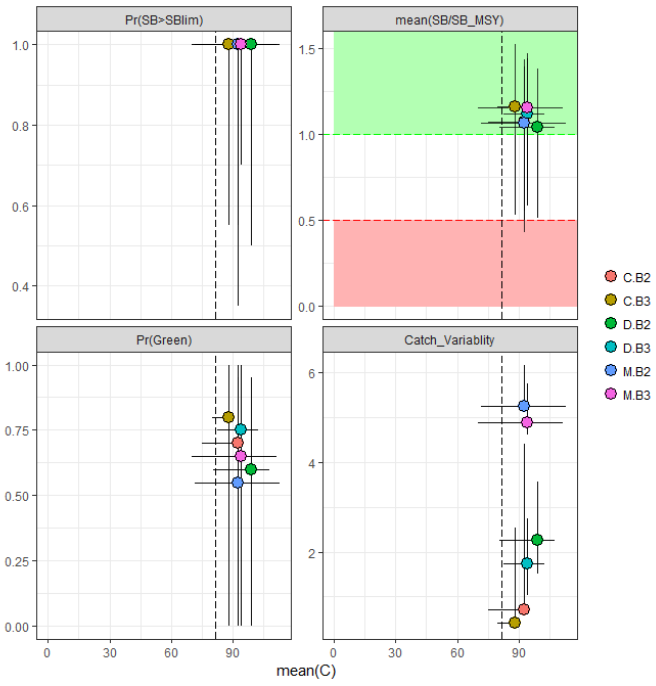


f) OMrobB20.1.recShock – 8 quarters of poor recruitment near start of MP implementation

(Figure 30 cont.)

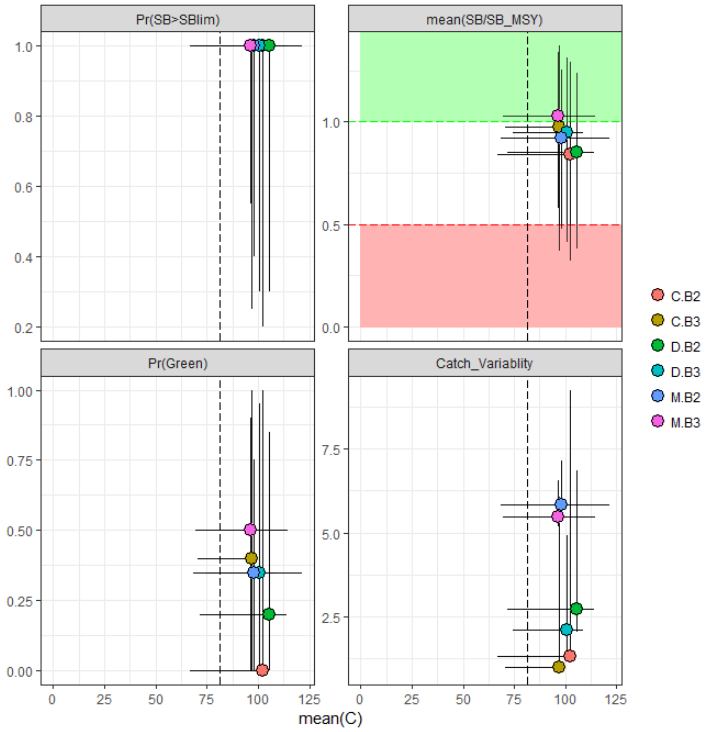


a) OMrefB20.1 – reference set

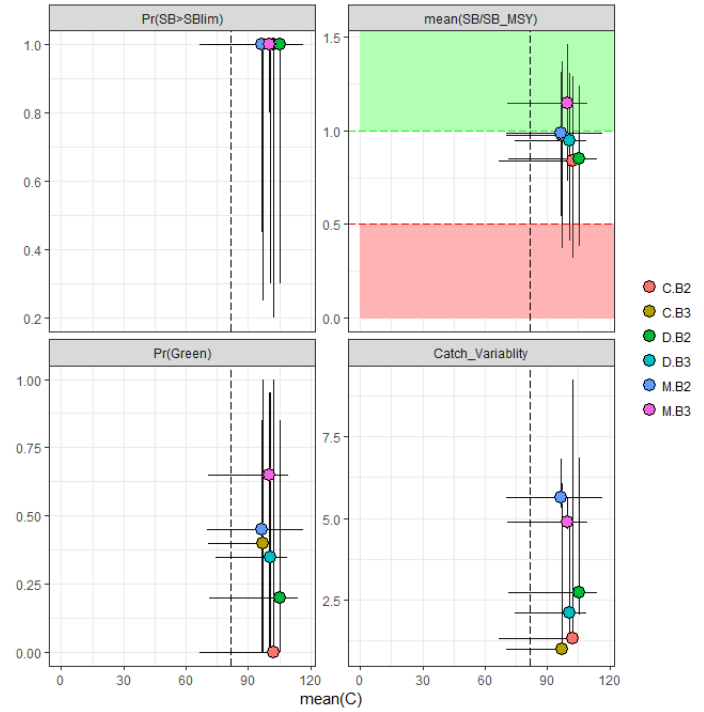


b) OMrobB20.1.ICV3 – CPUE variance = 0.3, all OMs

Figure 31. MP evaluation summaries from the Bigeye reference set OM OMrefB20.1 (a) and robustness tests (b-f). Trade-off plots comparing candidate MPs with respect to catch on the X-axis, and 4 other key performance measures on the Y-axis, each averaged over the period 2019 - 2038. Circle is the median, lines represent 10th-90th percentiles. Red and green horizontal lines represent the interim limit and target reference points for the mean SB/SBMSY performance measure. The dashed vertical black line is 2016 catch. (Figure 31 continued on following pages)

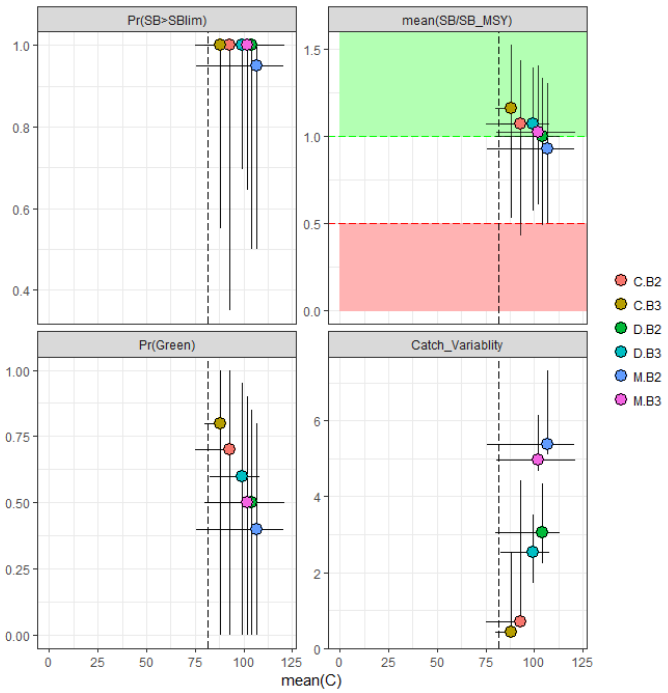


c) OMrobB20.1.10overRep – 10% reported overcatch

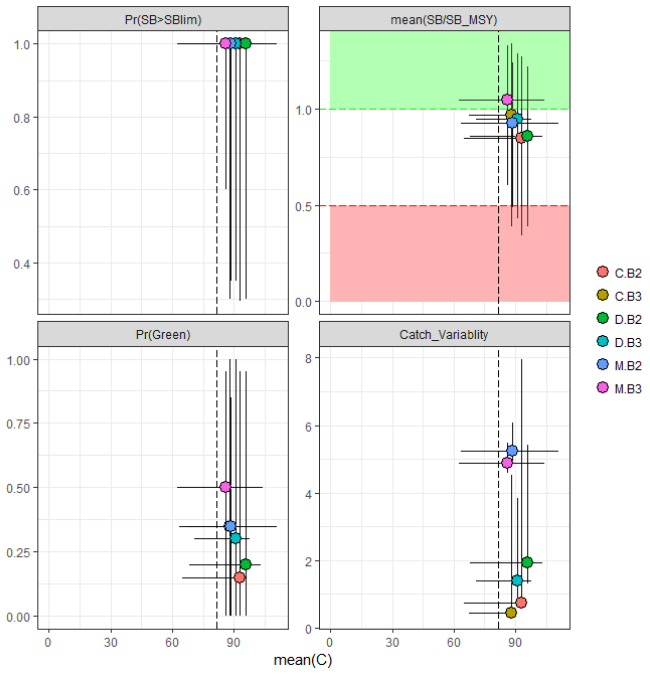


d) OMrobB20.1.10overIUU – 10% unreported overcatch

(Figure 31cont.)

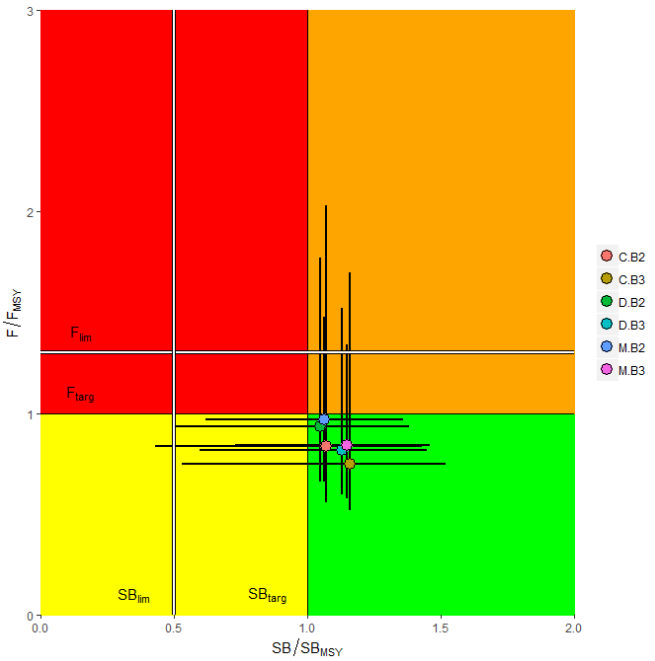


e) OMrobB20.1.qTrend3 – 3% per year longline catchability trend during projection period

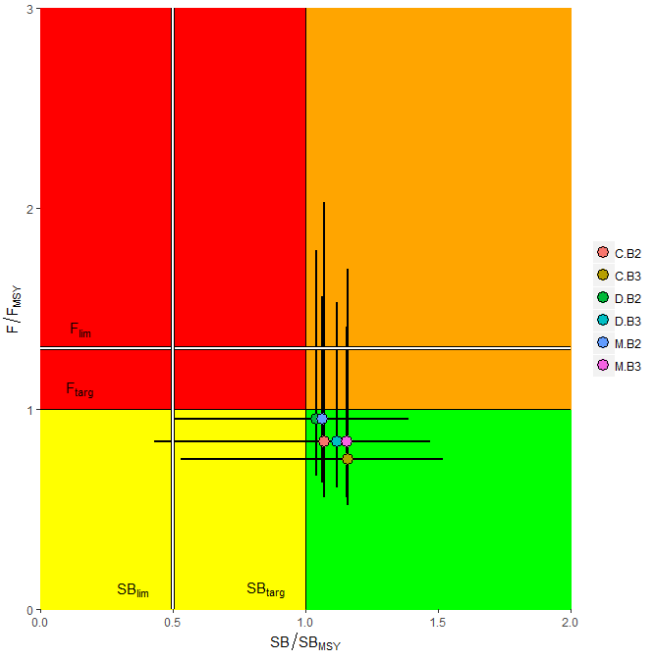


f) OMrobB20.1.recShock – 8 quarters of poor recruitment near start of MP implementation

(Figure 31cont.)

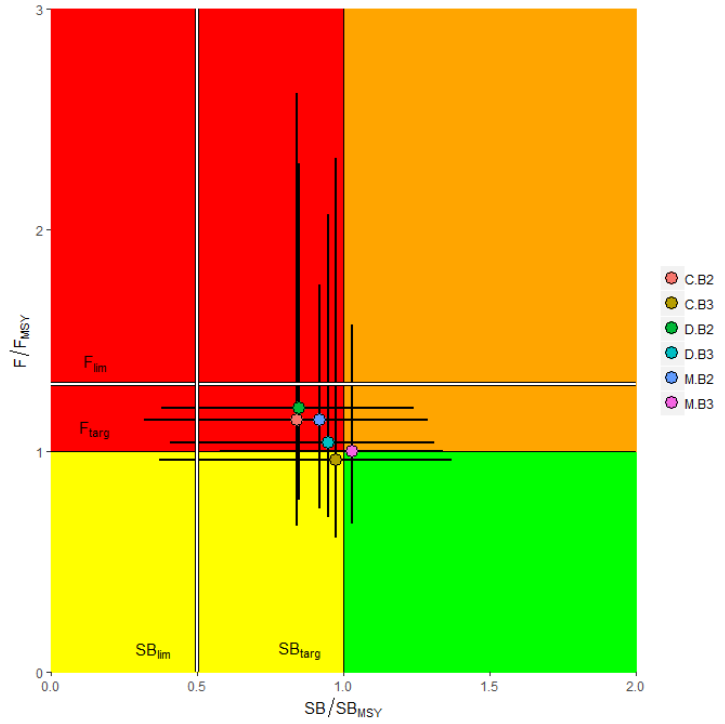


a) OMrefB20.1 – reference set

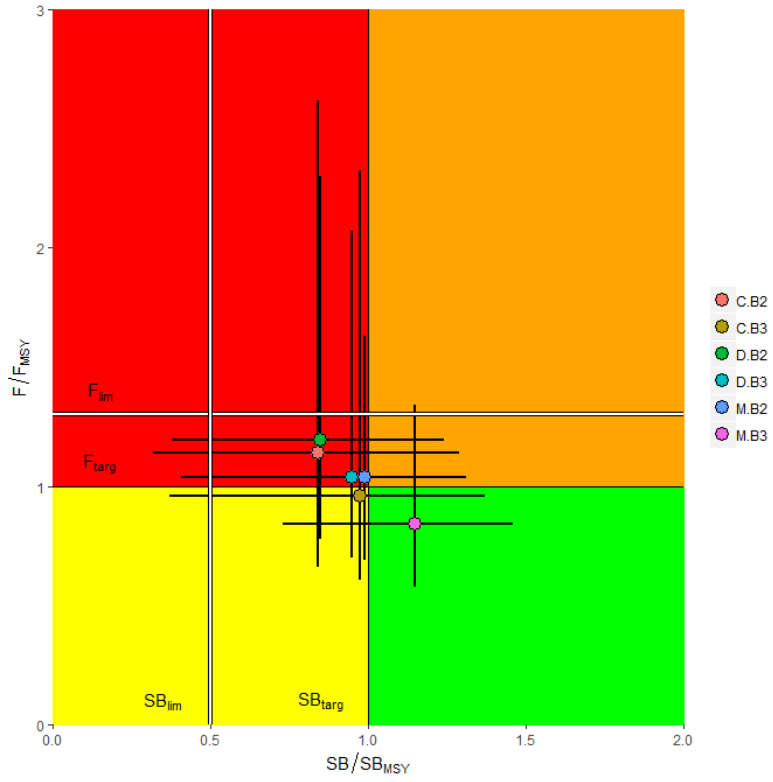


b) OMrobB20.1.ICV3 – CPUE variance = 0.3, all OMs

Figure 32. MP evaluation summaries from the Bigeye reference set OM OMrefB20.1 (a) and robustness tests (b-f). Kobe plot comparing candidate MPs on the basis of the expected 20 year average (2019-2038) performance. Circle is the median, lines represent 10th-90th percentiles. (Figure 32 continued on following pages)

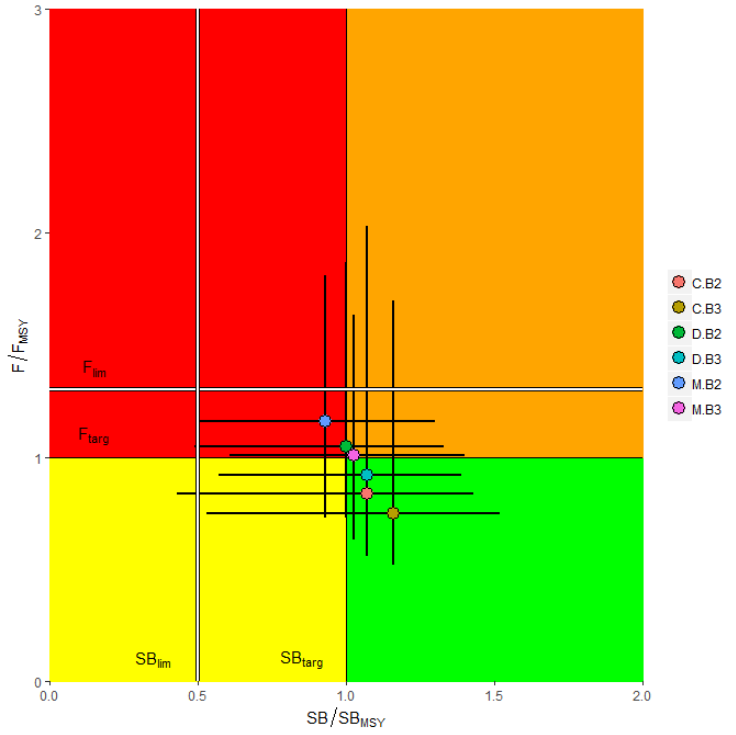


c) OMrobB20.1.10overRep – 10% reported overcatch

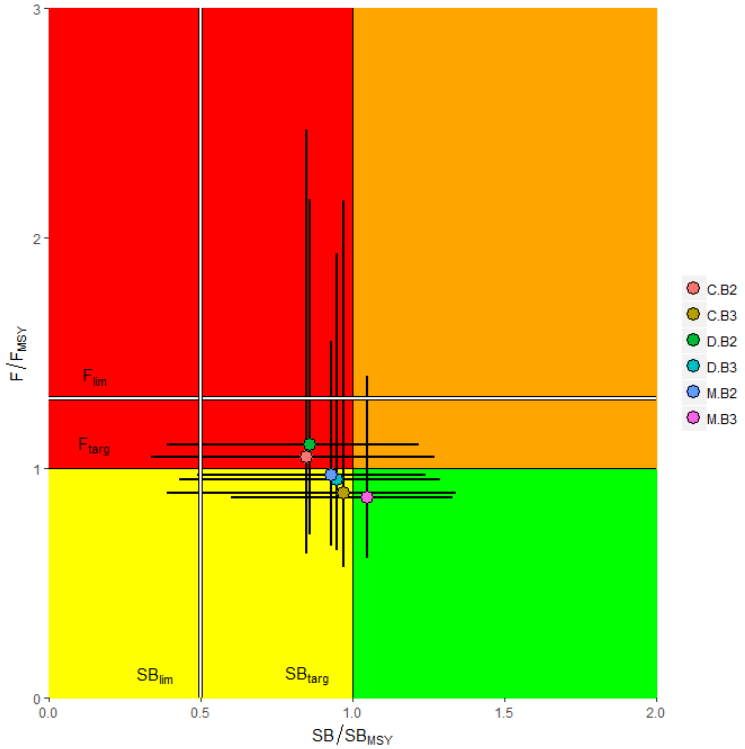


d) OMrobB20.1.10overIUU – 10% unreported overcatch

(Figure 32 cont.)

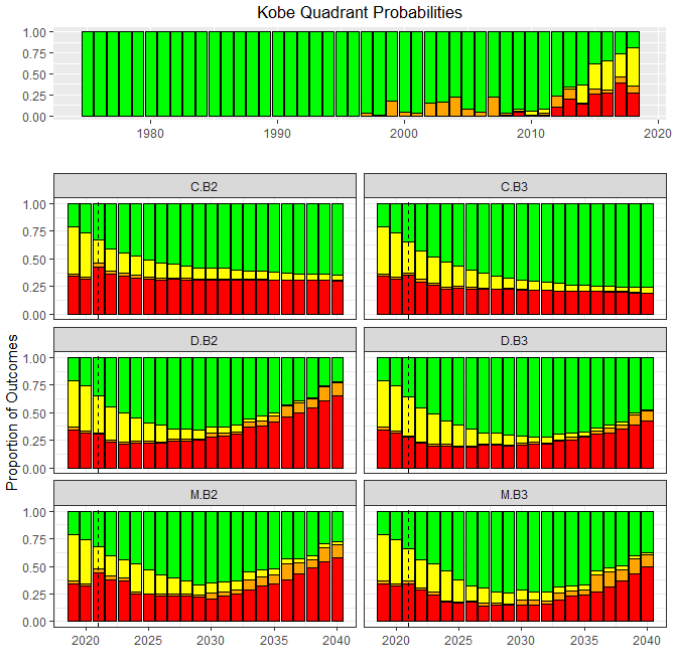


e) OMrobB20.1.qTrend3 – 3% per year longline catchability trend during projection period

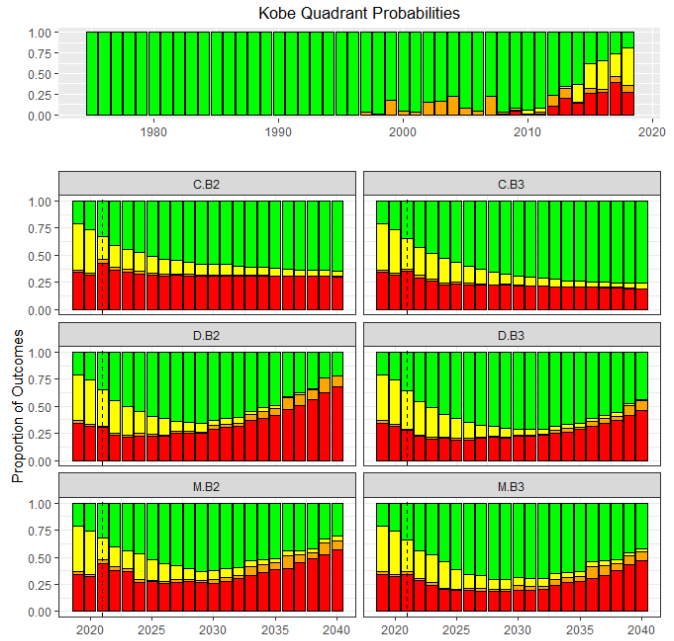


f) OMrobB20.1.recShock – 8 quarters of poor recruitment near start of MP implementation

(Figure 32 cont.)

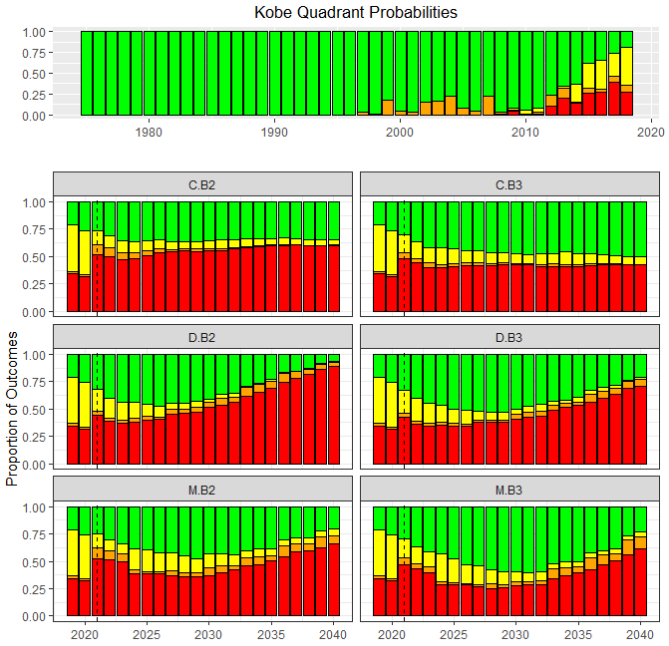


a) OMrefB20.1 – reference set

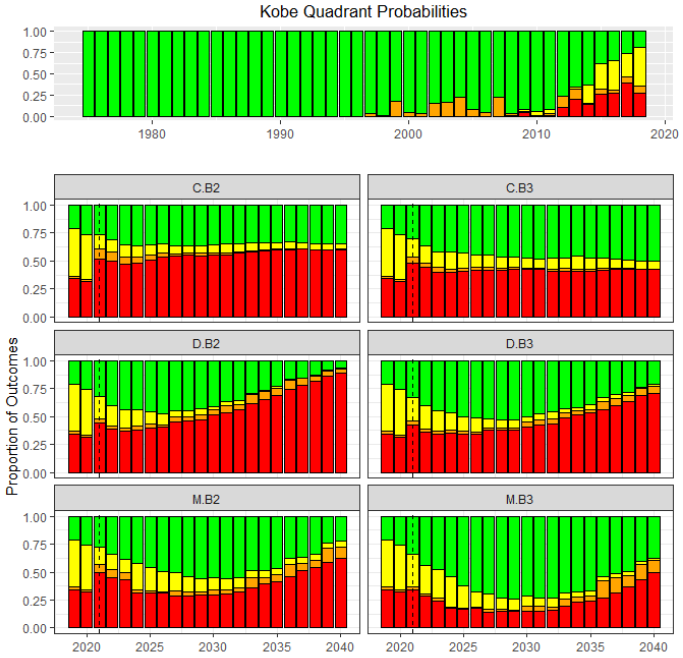


b) OMrobB20.1.ICV3 – CPUE variance = 0.3, all OMs

Figure 33. MP evaluation summaries from the Bigeye reference set OM OMrefB20.1 (a) and robustness tests (b-f). Proportion of simulations in each of the Kobe quadrants over time for each of the candidate MPs. Historical estimates are included in the top panel. The lower panels are projections, with the first MP application indicated by the broken vertical line (2019). (Figure 33 continued on following pages)

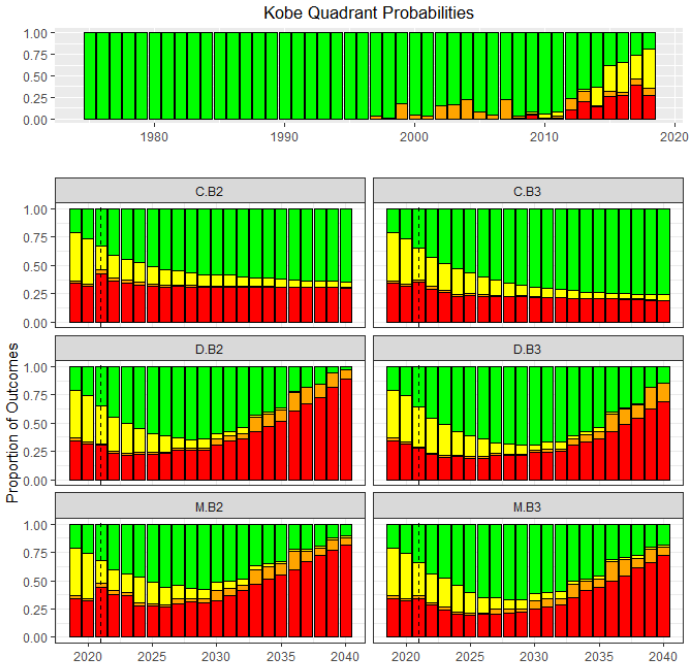


c) OMrobB20.1.10overRep – 10% reported overcatch

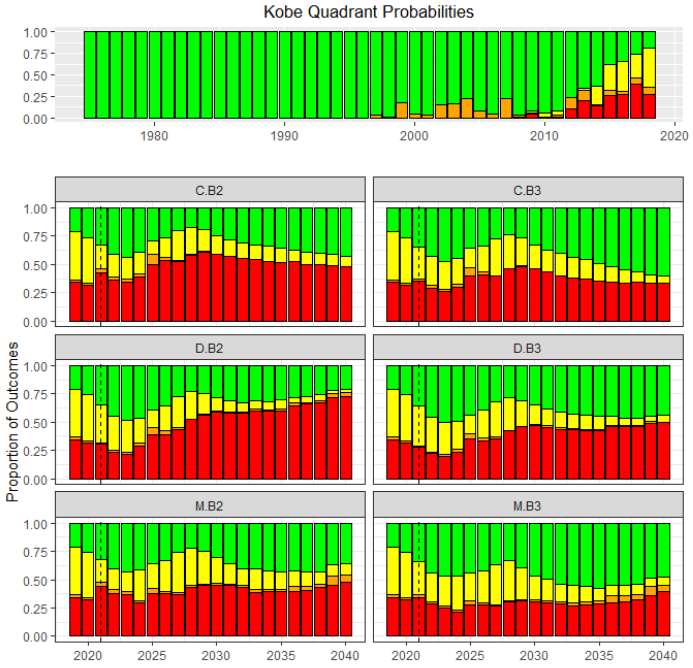


d) OMrobB20.1.10overIUU – 10% unreported overcatch

(Figure 33 cont.)

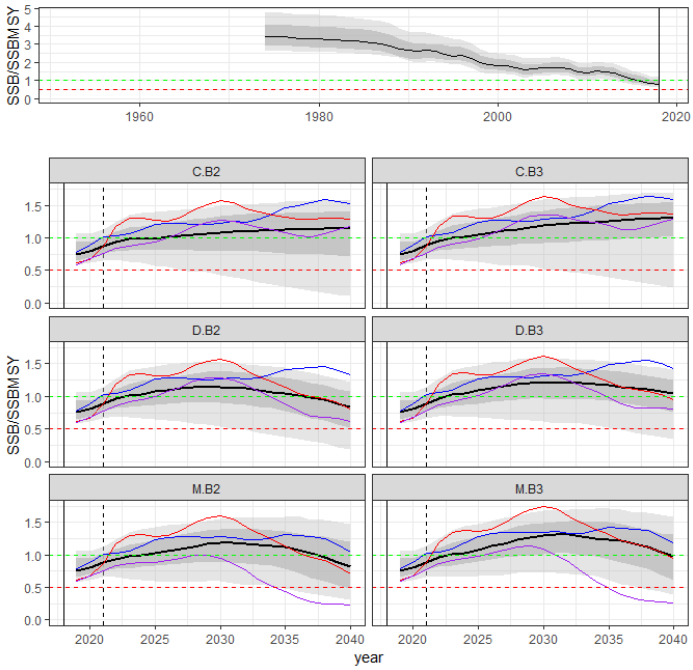


e) OMrobB20.1.qTrend3 – 3% per year longline catchability trend during projection period

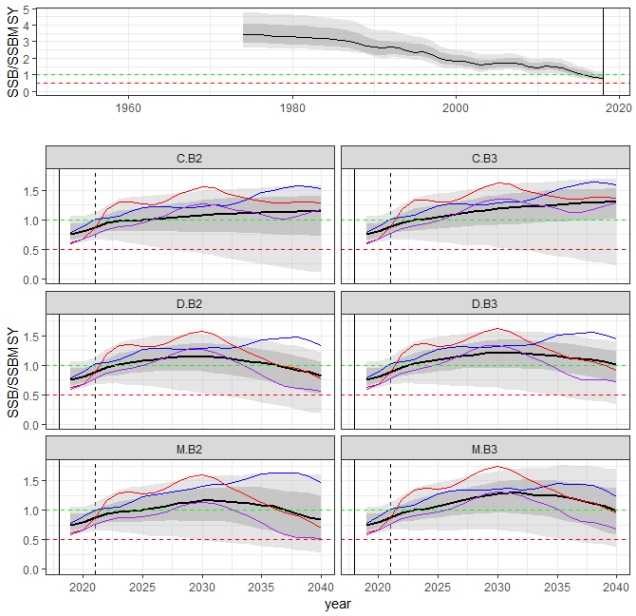


f) OMrobB20.1.recShock – 8 quarters of poor recruitment near start of MP implementation

(Figure 33 cont.)

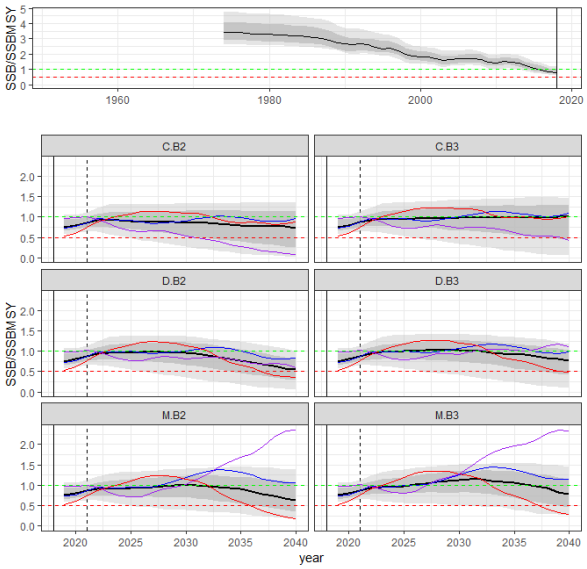


a) OMrefB20.1 – reference set

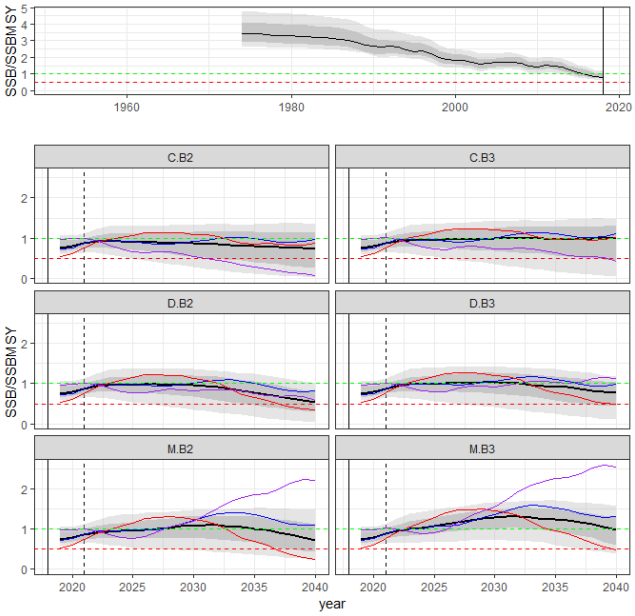


b) OMrobB20.1.ICV3 – CPUE variance = 0.3, all OMs

Figure 34. MP evaluation summaries from the Bigeye reference set OM OMrefB20.1 (a) and robustness tests (b-f). Time series of spawning stock size for the candidate MPs. The top panel represents the historical estimates from the reference case operating model, and lower plots represent the projection period. The solid vertical line represents the last year used in the historical conditioning. The broken vertical line represents the first year that the MP is applied. The median is represented by the bold black line, the dark shaded ribbon represents the 25th-75th percentiles, the light shaded ribbon represents the 10th-90th percentiles. Thick broken lines represent the interim target (green) and limit (red) reference points. The 3 thin coloured lines represent examples of individual realizations (the same OM scenarios across MPs and performance measures), to illustrate that individual variability greatly exceeds the median. (Figure 34 continued on following pages)

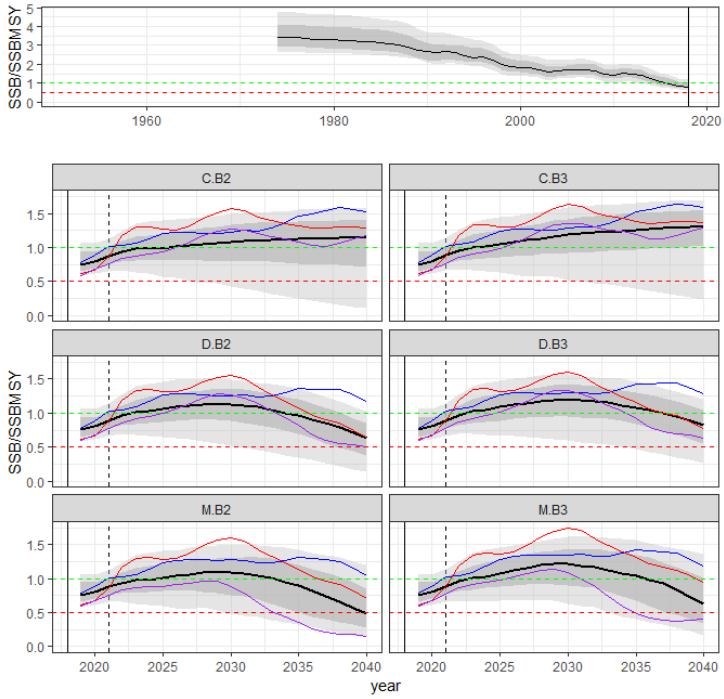


c) OMrobB20.1.10overRep – 10% reported overcatch

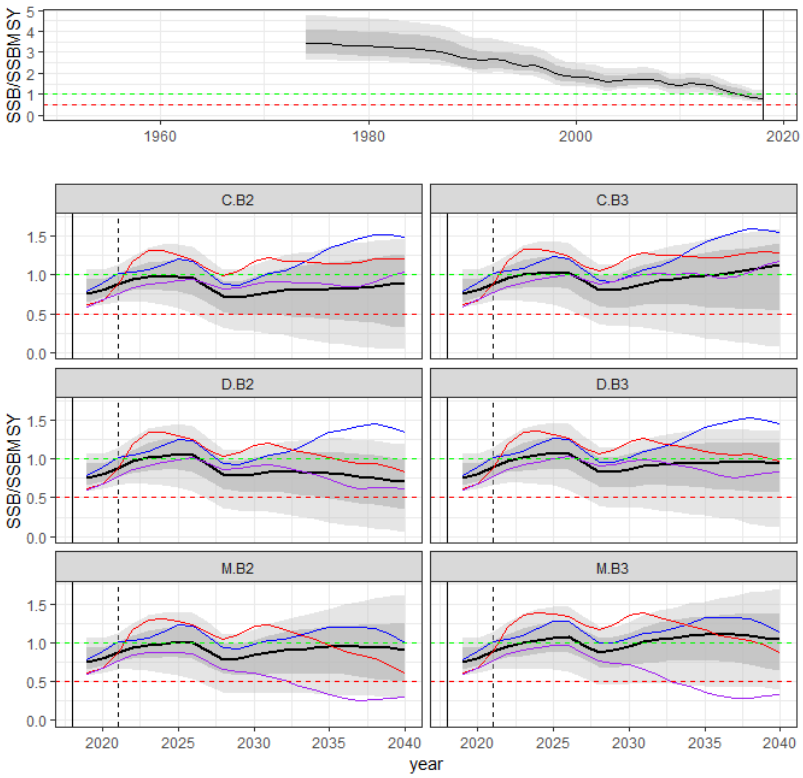


d) OMrobB20.1.10overIUU – 10% unreported overcatch

(Figure 34 cont.)

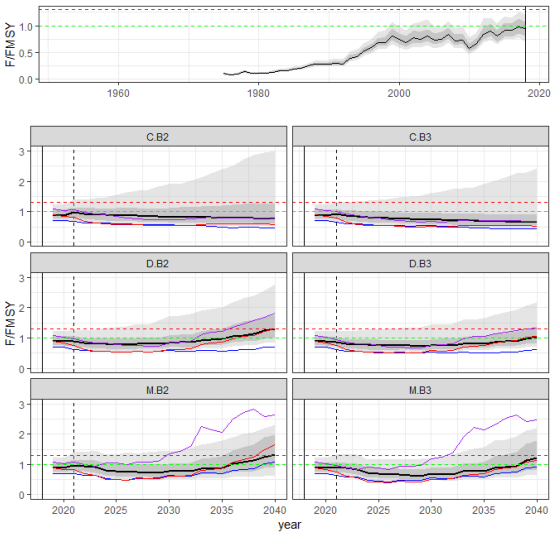


e) OMrobB20.1.qTrend3 – 3% per year longline catchability trend during projection period

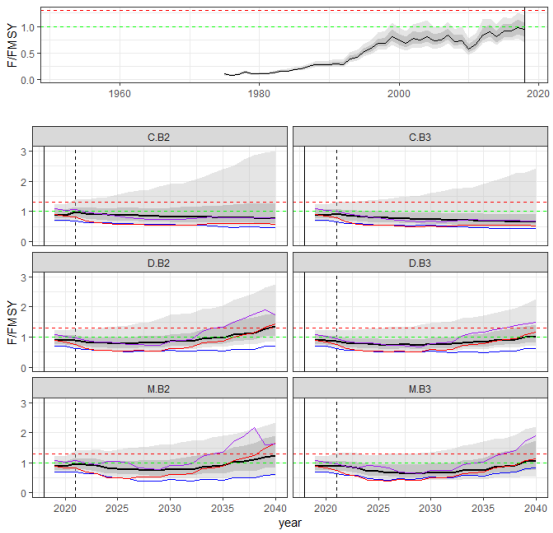


f) OMrobB20.1.recShock – 8 quarters of poor recruitment near start of MP implementation

(Figure 34 cont.)

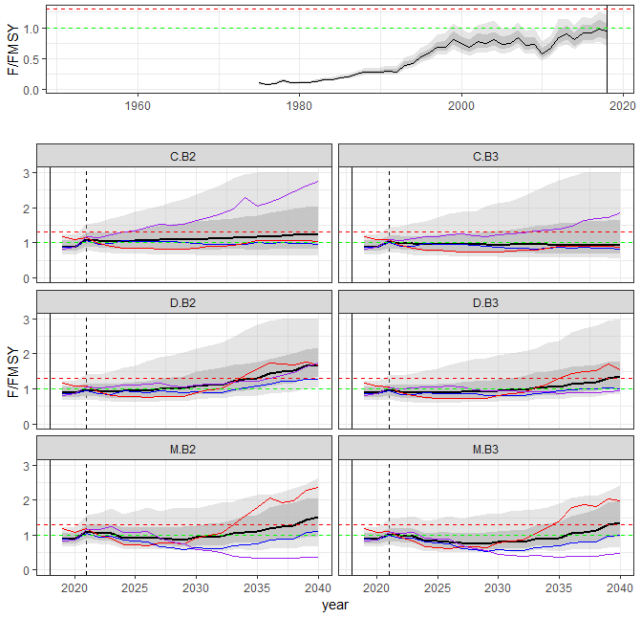


a) OMrefB20.1 – reference set

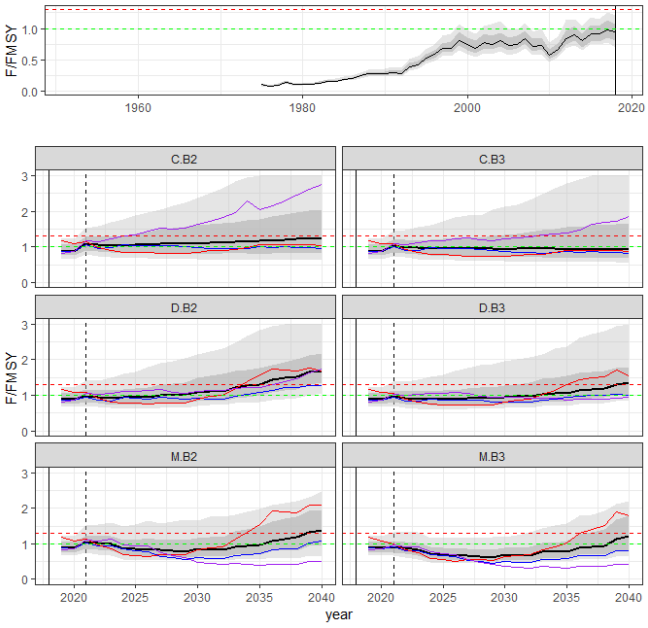


b) OMrobB20.1.ICV3 – CPUE variance = 0.3, all OMs

Figure 35. MP evaluation summaries from the Bigeye reference set OM OMrefB20.1 (a) and robustness tests (b-f). Time series of fishing intensity (Upper bound truncated at F = 3) for the candidate MPs. The top panel represents the historical estimates from the reference case operating model, and lower plots represent the projection period. The solid vertical line represents the last year used in the historical conditioning. The broken vertical line represents the first year that the MP is applied. The median is represented by the bold black line, the dark shaded ribbon represents the 25th-75th percentiles, the light shaded ribbon represents the 10th-90th percentiles. Thick broken lines represent the interim target (green) and limit (red) reference points. The 3 thin coloured lines represent examples of individual realizations (the same OM scenarios across MPs and performance measures), to illustrate that individual variability greatly exceeds the median. (Figure 35 continued on following pages)

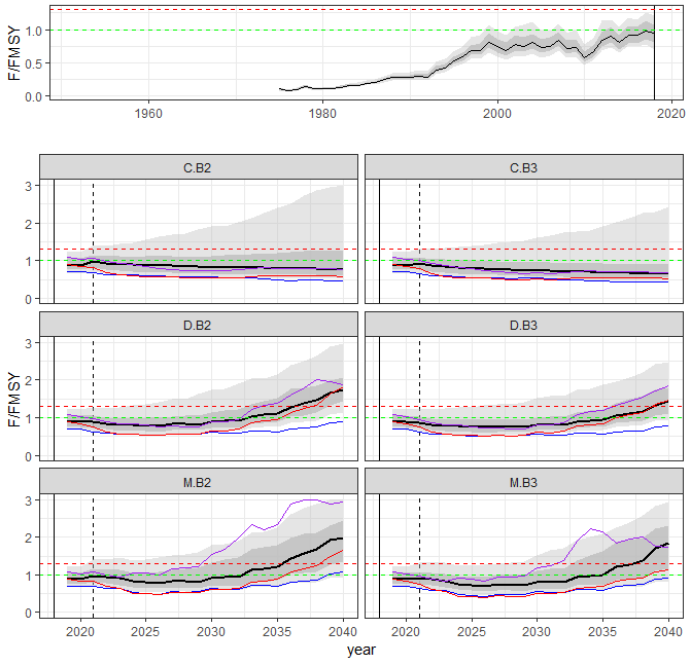


c) OMrobB20.1.10overRep – 10% reported overcatch

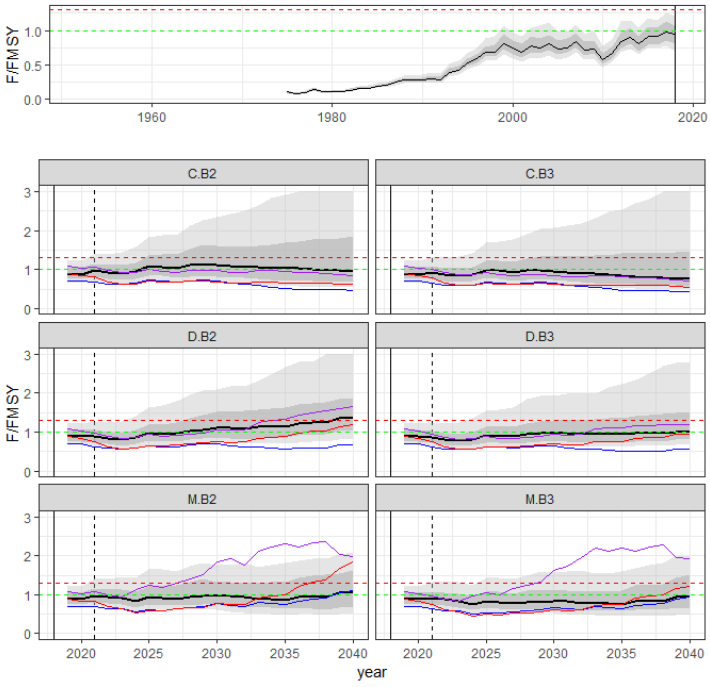


d) OMrobB20.1.10overIUU – 10% unreported overcatch

(Figure 35 cont.)

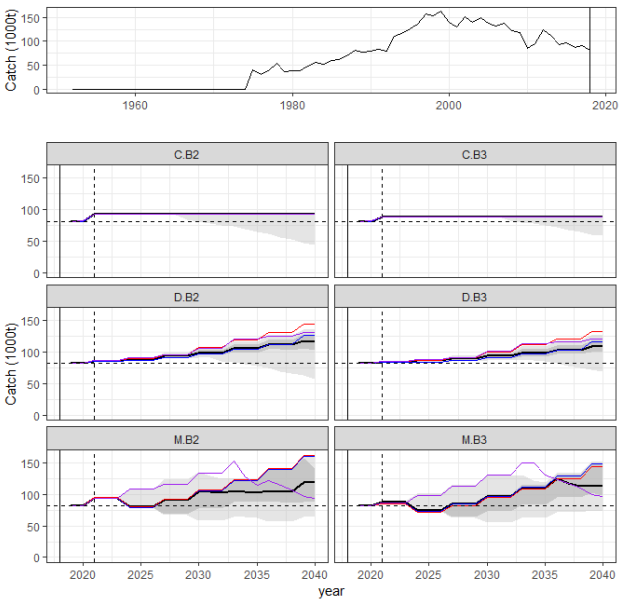


e) OMrobB20.1.qTrend3 – 3% per year longline catchability trend during projection period

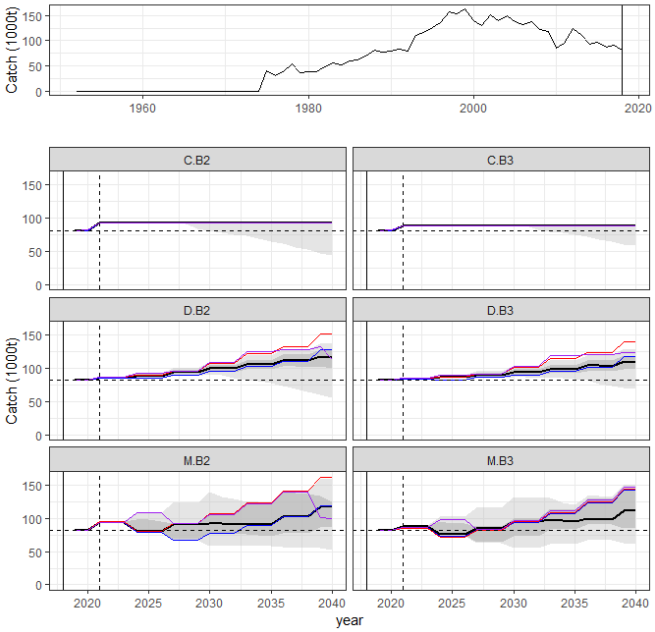


f) OMrobB20.1.recShock – 8 quarters of poor recruitment near start of MP implementation

(Figure 35 cont.)

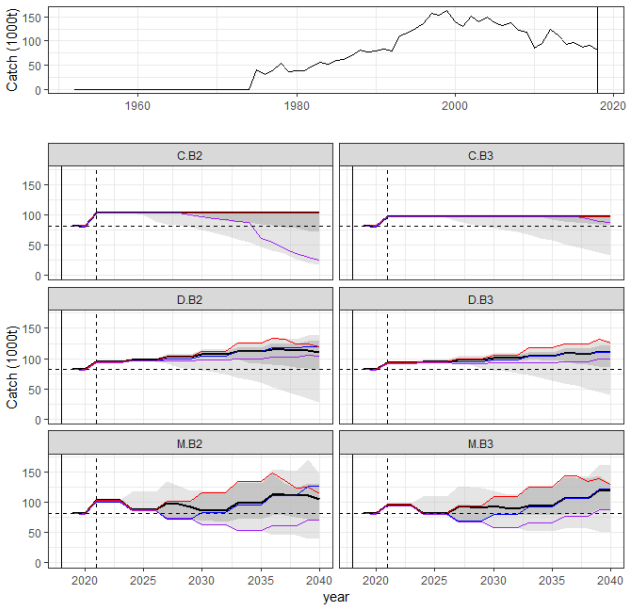


a) OMrefB20.1 – reference set

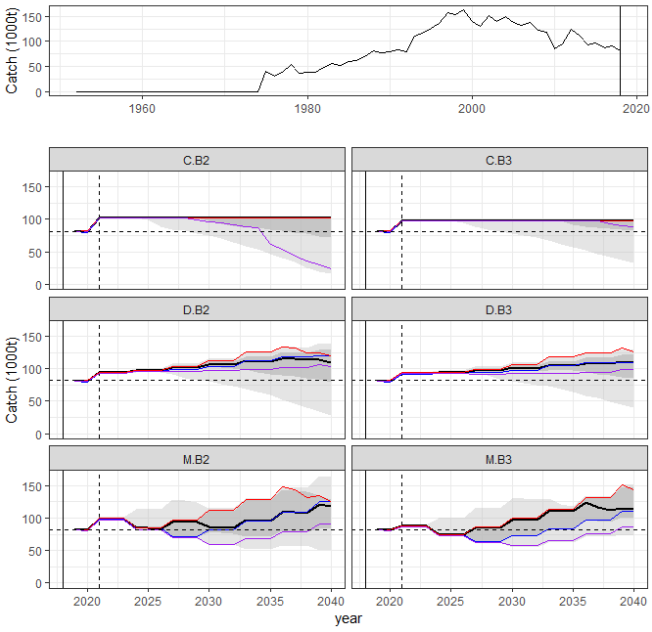


b) OMrobB20.1.ICV3 – CPUE variance = 0.3, all OMs

Figure 36. MP evaluation summaries from the Bigeye reference set OM OMrefB20.1 (a) and robustness tests (b-f). Time series of catch for the candidate MPs. The top panel represents the historical estimates from the reference case operating model, and lower plots represent the projection period. The solid vertical line represents the last year used in the historical conditioning. The broken vertical line represents the first year that the MP is applied. The median is represented by the bold black line, the dark shaded ribbon represents the 25th-75th percentiles, the light shaded ribbon represents the 10th-90th percentiles. The broken black horizontal line represents recent (2016) catch. The 3 thin coloured lines represent examples of individual realizations (the same OM scenarios across MPs and performance measures), to illustrate that individual variability greatly exceeds the median. (Figure 36 continued on following pages)

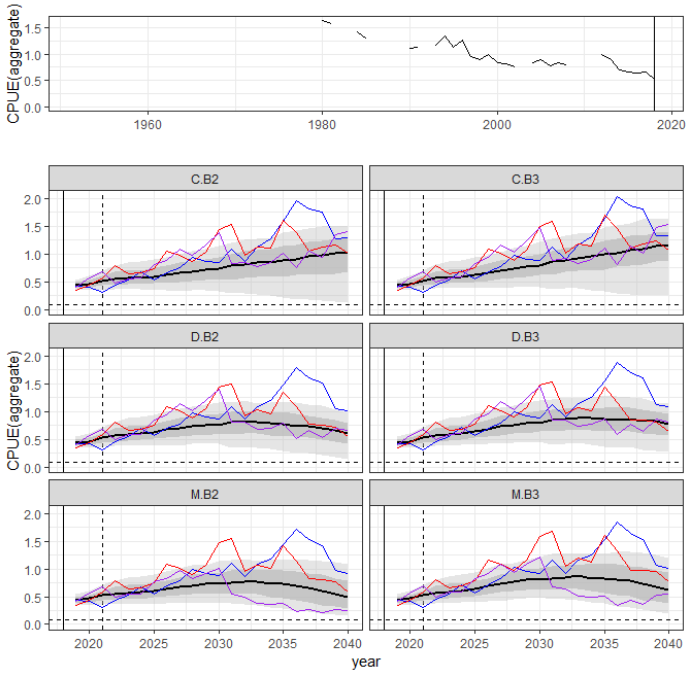


c) OMrobB20.1.10overRep – 10% reported overcatch

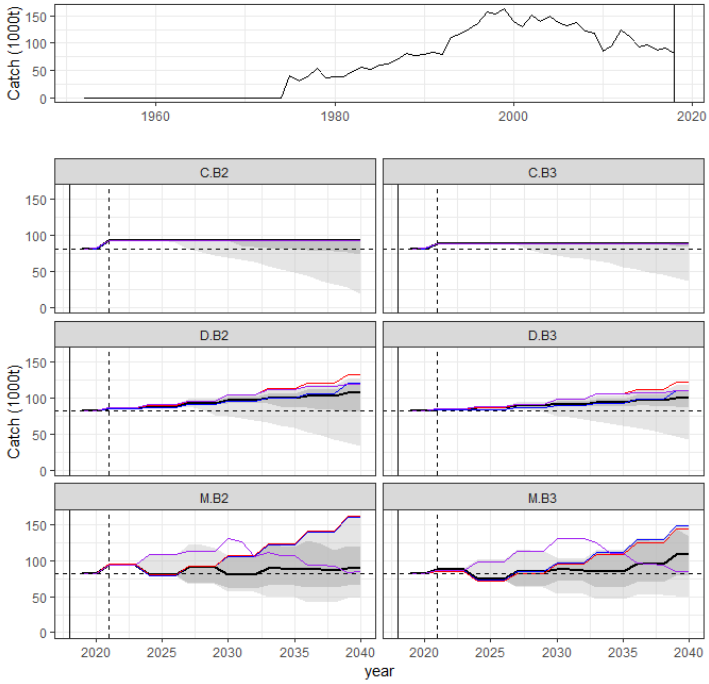


d) OMrobB20.1.10overIUU – 10% unreported overcatch

(Figure 36 cont.)



e) OMrobB20.1.qTrend3 – 3% per year longline catchability trend during projection period



f) OMrobB20.1.recShock – 8 quarters of poor recruitment near start of MP implementation

(Figure 36 cont.)

9 Key Points for the IOTC MSE Task Force Consideration:

We welcome feedback on all elements of the MSE work, and suggest the following priority points for the IOTC MSE Task Force to consider:

1. At the time of writing this document, it remains unclear how the MP development timeline will be altered as a result of Covid-19 disruptions to the IOTC community. While we could produce a 2020 TCMP summary document as in previous years, this does not seem useful because i) the TCMP meeting will likely be cancelled, ii) The results that we have to date appear qualitatively very similar to the previous iteration, and iii) there are remaining issues in the bigeye MSE that require broader input – most notably recommendations and data inputs from the CPUE Working Group.

2. We see no obvious reason to alter the BET OM as proposed by the WPTT/WPM 2020 (Appendix A), except to note that we will require input from the CPUE Working Group to represent the CPUE uncertainty (this could involve alternative CPUE series, catchability trends and/or regional scaling factors).

3. We recommend retaining the fractional factorial design to produce a manageable number of conditioned models (an ensemble of around 50 – 150), which includes all interactions between the 3 level assumptions, allows all main effects of all 2 level options to be estimable. Inclusion of all 2-way interactions is desirable, but did not seem to affect MP evaluation results perceptibly in tests done in the past. We recommend retaining the repeated convergence procedure to minimize the probability of accepting outliers due to extreme numerical convergence problems.

4. The best method for evaluating model plausibility and improving models remains an open topic. There have been a number of recommendations in the context of the stock assessments, but it is not clear that these are either feasible or helpful in the context of an OM. Our explorations to date have included:

- Our efforts at iterative reweighting suggest that i) different analysts will have somewhat different approaches, and ii) the impact on stock status appears to be small relative to the other uncertainty dimensions that are included in the OM (at least this appears to be true as long as some common sense principles of assessment model formulation are adhered to, i.e. ensure that there is a “reasonable” fit to the relative abundance indices, and be aware that sampling design and process errors are not likely to conform to model assumptions).

- Retrospective patterns and high fishing mortality (catch likelihood) seem to suggest that both bigeye assessments and OMs might be somewhat pessimistic.

- We continue to assert that a “substantial” catch likelihood in a Stock Synthesis hybrid F configuration is probably a useful flag for a problematic model. However, it is not clear how to resolve the problem or convert this to a model weighting or filtering criterion. We suspect that the problem might be related to some fundamentally flawed input data or assumption (e.g. total catch, CPUE as a relative abundance index, and/or large-scale seasonal movement). However, we were unable to find a simple and convincing alternative interpretation.

5. We have not yet developed an MP based on a model that admits both process and observation error or age structure considerations. This could be a priority for the months ahead, however, given the level of uncertainty in CPUE that we have to maintain to be internally consistent with the conditioning, it is not clear that more complicated MPs have the capacity to extract more information from the data. We would encourage proponents of alternative MPs to engage directly with the MP development process, to ensure that their ideas are properly represented and tested.

References

Fu, D. 2019. Preliminary Indian Ocean bigeye tuna stock assessment 1950-2018 (Stock Synthesis). IOTC-2019-WPTT21-61

Kolody, D, Day, J, Jumppanen, P. 2020. Indian Ocean Yellowfin Tuna Management Procedure Evaluation Update April 2020. Report prepared for the Indian Ocean Tuna Commission Informal Management Strategy Evaluation workshop 2020 (This document will be assigned a number in the IOTC archive at a future date).

Kolody, D, Jumppanen, P. 2019. Update on IOTC Bigeye Tuna MSE Operating Model Development October 2019. IOTC-2019-WPM10-08.

Matsumoto, T, Yokoi, H, Satoh, K, Kitakado, T. 2018. Diagnoses for stock synthesis model on yellowfin tuna in the Indian Ocean. IOTC–2018–WPTT20–42_Rev1.

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

Pacific Fishery Management Council. 2018. Terms of Reference for the Groundfish and Coastal Pelagic Species Stock Assessment Review Process for 2017-2018 http://www.pcouncil.org/wp-content/uploads/2017/01/Stock_Assessment_ToR_2017-18.pdf.

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

WPTT 2019. Report of the 21st Session of the IOTC Working Party on Tropical Tunas. Donostia-San Sebastian, Spain, 21 - 26 October 2019. IOTC–2019–WPTT21–R[E].

Appendix A. Extracts from the 2019 Methods and Tropical Tuna Working Party reports relevant to bigeye MSE

Working Party on Methods 2019 (draft) report

44. The WPM **NOTED** that the MPs always use the same historical CPUE series, while the OM is conditioned to 8 different series, to represent CPUE standardization uncertainty. Consequently, model predicted CPUE for the projection period may not be consistent with the historical observed CPUE indices. The WPM discussed possible ways to alleviate this discontinuity. To date, the q was re-scaled such that the historical vulnerable biomass and historical CPUE means were equal over the whole time series, which sometimes causes a large discontinuity in the first projection year. An alternative option was proposed in which the rescaling is conducted over the terminal period only. This removes the discontinuity, but may have other consequences. The WPM **AGREED** that the effect of different options for linking the historical CPUE observations used by the MP with the simulated projection CPUE requires further investigation.

45. The WPM **DISCUSSED** whether evaluating the predictive capabilities of models would provide useful criteria for differentially weighting models for inclusion in the OM. The WPM discussed that the hindcasting approach might be one of candidate procedures to address this critical issue, however no decision was reached.

48. The WPM **NOTED** the following points for the next iteration of the BET MSE:

- The WPM did not request any modifications to the reference set OM or robustness tests , but noted that the reference set OM will have to be evaluated in relation to the 2019 assessment at the WPTT to see if reconditioning is required The WPM **REQUESTED** that specific criteria for deciding whether or not reconditioning is required should be developed at the next session of the WPM
- Main effects fractional factorial design appears to be adequate for producing consistent MP evaluation results, with a target of 50-150 models in the reference set OM.
- A standard jitter analysis for every model is probably not necessary, but is likely to reduce the frequency of extreme outliers.
- Retain the 2 year MP implementation data lag unless advised otherwise by the TCMP

Working Party on Tropical Tunas 2019 (draft) report

150. The WPTT **REQUESTED** the addition of a "recruitment shock" robustness test by reducing future recruitment (e.g. by half for 2 years as in YFT) in the Management Procedure (MP) testing, acknowledging that this kind of robustness scenario has commonly been considered in other RFMOs such as CCSBT and IWC.

151. The WPTT **SUGGESTED** modifying an alternative assumption about spatial differences in longline selectivity patterns into the OM conditioning scenarios to be consistent with the 2019 assessment model grid.

152. The WPTT **NOTED** that the relative contribution of fleets to the total catch has been changing over time (e.g. the increasing trend in the proportions of PS catches), and this has implications for MP evaluations. This relates to a Commission request about alternative allocations. The WPTT recognized that allocations are a political decision and therefore **REQUESTED** guidance from the TCMP on specific scenarios to be tested in the MP evaluations.

153. The WPTT **DISCUSSED** the current specification for conditioning, and the WPTT **AGREED** to use tag lambda (a weight to the likelihood from tag recovery data) as 1, 0.1 and 0.001.

154. The WPTT **NOTED** that there is uncertainty about the reported 2018 catch, as was also discussed during bigeye tuna and yellowfin tuna stock assessment sessions. The WPTT **AGREED** that a single "agreed" 2018 catch scenario is used for new OM conditioning, and MP catch allocation assumptions are the 2017-2018 average. The WPTT further **NOTED** that there are also TAC implementation Robustness tests of 10% over-reporting (with and without reporting), and the MP performance was not very sensitive to these errors.

Definition
Stock-recruit function (*h* = steepness) • Beverton-Holt, *h* = 0.7 • Beverton-Holt, *h* = 0.8 • Beverton-Holt, *h* = 0.9
Natural mortality multiplier relative to reference case M vector • 1.0 • 0.8 • 0.6
Tag recapture data weighting (tag composition and negative binomial) • λ = 0.001 • λ = 0.1 • λ = 1.0
Assumed longline CPUE catchability trend (compounded) • 0% per annum • 1% per annum
Tropical longline CPUE standardization method • Hooks Between Floats • Cluster analysis
longline CPUE Regional-scaling factors • reference case • alternate
Longline fishery selectivity • Stationary, logistic, shared among areas • Stationary, logistic in region 1, double-normal (potentially dome-shaped), in other regions
Size composition input Effective Sample Sizes (ESS) • ESS = 10, all fisheries • ESS = One iteration of re-weighting from reference case model, capped at 100.

155. The WPTT **REQUESTED** five bigeye tuna robustness scenarios (all of which assume the reference set OM conditioning):

- What happens if there is a two year recruitment failure (55% of expected + usual stochastic error, as defined for yellowfin tuna)
- What happens if the (annualized aggregate) longline CPUE observation error CV is increased to 30% (auto-correlation 0.5) in projections?
- What happens if there is a consistent 10% future over-catch (accurately reported), equally distributed among fleets?
- What happens if there is a 10% future over-catch (unreported), equally distributed among fleets ?
- What happens if the longline CPUE catchability trend is 2% per year going forward (but remains as in the reference scenario for conditioning)?

232. The WPTT **DISCUSSED** the properties of the production model used in the model-based MP. The WPTT **NOTED** that, in addition to reporting the average performance of a MP across all OM, it is also important to identify when a MP performs and when it fails. For example, in such a model-based MP, the performance may depend on the estimates of shape of the "implicit" production function, the value of r, or the form of process error (variance or frequency) in the OM. Therefore, the WPTT **NOTED** that understanding when an MP fails could help in identifying where resolving uncertainties could improve management performance. The WPTT also **NOTED** that it is worth comparing the true biomass in the OM and estimated biomass by MP in the simulation in addition to the evaluating MP performances.

233. The WPTT **NOTED** that the use of model free cross-validation could potentially identify which data series have good prediction skill and are therefore candidates for use in model free and model base MPs. However, it was further **NOTED** that the MPs (explored to date) pool the regionally-scaled CPUE indices into a single ocean-wide abundance index, and do not use size composition or tag data, so it is not clear how such an analysis would be helpful in this case.

CONTACT US

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

AT CSIRO WE SHAPE THE FUTURE

We do this by using science to solve real issues. Our research makes a difference to industry, people and the planet.

As Australia's national science agency we've been pushing the edge of what's possible for over 85 years. Today we have more than 5,000 talented people working out of 50-plus centres in Australia and internationally. Our people work closely with industry and communities to leave a lasting legacy. Collectively, our innovation and excellence places us in the top ten applied research agencies in the world.

WE ASK, WE SEEK AND WE SOLVE

FOR FURTHER INFORMATION

CSIRO Oceans and Atmosphere
Dale Kolody
t +61 3 6232 5121
e dale.kolody@csiro.au
w https://www.csiro.au/en/Research/OandA