



Indian Ocean Yellowfin Tuna Management Procedure Evaluation Update March 2021

Prepared for the IOTC MSE Task Force Meeting March 2021

Dale Kolody (e: dale.kolody@csiro.au), Paavo Jumppanen

CSIRO Oceans and Atmosphere, Castray Esplanade, Hobart TAS 7000, Australia

Citation

Kolody, D, Jumppanen, P. 2021. Indian Ocean Yellowfin Tuna Management Procedure Evaluation Update March 2021. Working Paper prepared for the Management Strategy Evaluation Task Force of the Indian Ocean Tuna Commission Working Party on Methods Meeting, March 2021. IOTC-2021-WPM12(MSE)-03

© FAO 2021

The designations employed and the presentation of material in this information product do not imply the expression of any opinion whatsoever on the part of the Food and Agriculture Organization of the United Nations (FAO), or of the Commonwealth Scientific and Industrial Research Organisation (CSIRO) concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The mention of specific companies or products of manufacturers, whether or not these have been patented, does not imply that these have been endorsed or recommended by FAO or CSIRO in preference to others of a similar nature that are not mentioned. The views expressed in this information product are those of the author(s) and do not necessarily reflect the views or policies of FAO, or CSIRO.

FAO encourages the use, reproduction and dissemination of material in this information product. Except where otherwise indicated, material may be copied, downloaded and printed for private study, research and teaching purposes, or for use in non-commercial products or services, provided that appropriate acknowledgement of FAO as the source and copyright holder is given and that FAO's endorsement of users' views, products or services is not implied in any way.

All requests for translation and adaptation rights, and for resale and other commercial use rights should be made via www.fao.org/contact-us/licence-request or addressed to copyright@fao.org.

FAO information products are available on the FAO website (www.fao.org/publications) and can be purchased through publications-sales@fao.org

Important disclaimer

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

Contents

Acknowledgments.....	4
Summary	5
1. Introduction	7
2. Spatial structure for the Yellowfin Operating Model – comparison of individual models	8
3. Spatial structure for the YFT Operating Model – comparison of ensembles.....	15
4. Management Procedure Evaluation Results	30
4.1 The relative importance of spatial configurations in relation to the confounded assumptions of including/excluding movement and tags.....	35
4.2 Does the time blocking of recruitment spatial distribution parameters affect MP evaluation results?.....	37
4.3 What effect does the inclusion of 2 extra years of catch data have on the OM retrospective problem?	38
5. How sensitive are the MP evaluation results to “minor” OM projection assumptions related to spatial distributions but not initial status or productivity?	40
5.1 OM diagnostics – the role of catch and effort?.....	50
6. Conclusions and Recommendations.....	55
References	57
Appendix A. Extracts from the 2020 Methods and Tropical Tuna Working Party reports relevant to yellowfin MSE Technical workplan	59
Appendix B. State of the IOTC Yellowfin Tuna MSE Operating Models as of March 2021.	61

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), administered by the Indian Ocean Tuna Commission (UN-FAO). 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 Agurtzane Urtizberea for various files and clarifications related to the 2019 yellowfin tuna stock assessment analyses.

Summary

This working paper describes developments on the Indian Ocean Tuna Commission (IOTC) yellowfin (YFT) Management Procedure Evaluation work, since the 2020 Working Party on Tropical Tunas (WPTT) and Working Party on Methods (WPM). Despite a number of recent investigations, we have not identified a satisfactory Operating Model (OM) upon which we would be confident in providing Management Procedure (MP) advice to the 2021 TCMP. Any attempt to present YFT MP evaluation results at this time might create a misleading perspective that could be counter-productive to the longer term MP adoption process and efforts to provide stock assessment advice. The status of the IOTC YFT stock assessment and management advice provided by the WPTT are in a state of flux, subject to a substantial collaborative review and development process, and with a new assessment scheduled for 2021. Given the close association between the assessment and the OM conditioning, we expect that there should be new perspectives arising from the 2021 assessment, with which to guide the YFT OM revision. The insights provided here are hopefully useful to the assessment as well. Our key concerns about the OM at this time include:

- The MP evaluations conducted here suggest that very large and immediate catch reductions would be required to hit the B_{MSY} rebuilding objective by 2034. However, there are good reasons to believe that the OMs requested by WPTT (2020) are too pessimistic.
- The widening gap between the data used to condition the OMs (2017 end year) and data collected subsequently, provides a very informative hindcast prediction diagnostic, similar to those being encouraged in recent IOTC assessments. When projecting forward from 2017, a large portion of the proposed YFT OM models struggled to remove the catch that was actually reported in 2019. It seems inescapable that any model is too pessimistic (or the space-time distribution of fish and fisheries is not sufficiently realistic), if these observed catches cannot be attained. Among those models that could attain the 2019 catch, the majority failed to remove the remaining bridging catches (assumed equal to 2019) for the 2020-2021 period, i.e. prior to the first MP quota setting.
- In previous development iterations, we have recognized that sometimes, not all fleets are able to remove the TAC recommended by the MP. It was expected that this might cause a disconnect between TACs and catches and how change constraints are applied, that could merit further consideration in the MP definitions. But this was expected to be a minor problem, and it was implicitly assumed that the estimated OM model parameters would still provide an adequate description of the fishery dynamics whether or not this was occurring. However, recent investigation suggests that if the stock is now, or soon to be, near the pessimistic state that the current OM suggests, then how the failure to remove recent catches is modelled could potentially have non-trivial consequences for MP evaluation results and MP selection. i.e. the space-time structure introduces refugia that may prevent some fleets from extracting their allocated quota, and it is not clear that the models adequately represent these refugia and how the fisheries would respond under these conditions.

- The OM has essentially been unrestrained with respect to how effort is allowed to increase to extract the TACs (and bridging catches before the MP is implemented). Due to the way that the Baranov equations are implemented in the C++ projection sub-routine (but not the modified Pope’s approximation originally implemented in R), there is an “effort ceiling” parameter that is set relative to the recent Fs estimated during conditioning. When this value is set to some reasonable number (e.g. if we speculate that the effective fishing mortality could at most double in the period from 2017-2019), the number of minimally plausible realizations (i.e. those that can remove the 2019 catch) is substantially reduced. The role of effort changes may require further consideration in all of the MP evaluations (or at least added as a standard set of robustness tests), to more realistically speculate about how fisheries will respond when quotas cannot be easily reached (i.e. would effort increase, would fleets move or shift targeting?)
- When the MP evaluation results are subset after running, retaining only those realizations that were actually able to extract >95% of the observed 2019 catch, it is notable that:
 - The assumed CPUE CV of 30% in the conditioning was over-represented (97% of realizations vs 3% for the CPUE CV of 10%). The CPUE catchability trend of 0% per year was over-represented relative to 1% per year. This suggests the CPUE series are not very compatible with the recent observed catches in the context of the current model structure.
 - The high M option was retained at a much higher rate than the low M option (despite the fact that the low M option seemed to be more compatible with the tagging studies in the Indian Ocean, and emerging inferences from direct ageing studies in the Atlantic).
 - The down-weighted tag λ option was over-represented relative to the full weighting.

It seems clear that a better mechanism for combining OM characteristics and weighting on the basis of some sort of plausibility diagnostics may be required. However, it is not clear that the OM is in the right structural space that would enable this to be a sufficient solution. The 2 area OMs demonstrated most of the same problems as the 4 area OMs, though to a lesser extent.

A number of additional investigations are described in the main text, but it is not clear how useful the specific inferences are, given the big picture problems with the OM. A full suite of MP evaluation results has not been provided at this time, but it would still be feasible to produce these results before the TCMP 2021, if the MSE Task Force considered this to be useful. We recommend that the OM developers should continue to engage with the YFT assessment team to update everything in 2021, consider broader interpretations of the baseline assumptions to try to resolve the retrospective problem.

1. Introduction

¹No good model ever accounted for all of the facts, since some data were bound to be misleading if not plain wrong

-JD Watson (1988)

This working paper describes developments on the IOTC yellowfin (YFT) reference set and robustness test Operating Models (OM) since the last Working Parties on Tropical Tunas (WPTT 2020) and Methods (WPM 2020). Key development requests are extracted to Appendix A, and a brief self-contained summary of the OMs is included in Appendix B. The OMs are used to simulation test Management Procedures, attain the high priority tuning objectives defined by the Technical Committee on Management Procedures (TCMP), and ultimately quantify the expected performance trade-offs among management objectives. This should eventually allow the Commission to choose a Management Procedure that reflects its medium term goals and risk tolerance. The intended audience for this paper is already familiar with the background of the work, and technical jargon. Other interested parties may need to consult the history of project reports found in the public github repository, <https://github.com/pjumpnanen/niMSE-IO-BET-YFT/>, which also contains the open source MSE software, technical documentation and user manual, and scripts for reproducing key analyses (contacting authors in advance will help ensure that the repository is up to date).

Issues covered in this report include:

- Comparison of the 2 area vs 4 area model structures (as explored in Urtizbera et al, 2020), to determine whether it is the spatial structure per se, or confounded issues of movement and tagging data that lead to apparent differences in model inferences.
- Update the reference set OM grid requested by WPTT (2020, reproduced in Attachment A), and revisit the relative importance of the 2 and 4 area structures in the context of MP evaluations.
- Further investigation of model diagnostics to evaluate the plausibility of conditioned OMs
- Examination of the sensitivity of MP evaluation results to the issue of very high fishing mortality and failure to attain quotas.

The investigation led us to conclude that there are serious problems with the requested OM structure that will require further interaction with the broader IOTC scientific community, and which cannot be resolved in time for the 2021 TCMP. Accordingly, MP evaluation results are only presented for a single reference case MP, to illustrate the OM problems.

¹ This quote was included in author DK's first southern bluefin tuna stock assessment paper in 2001, and proved prophetic to the 2006 revelation of decades of 100-200% unreported catch, with an unknown effect on CPUE series. We do not have any evidence that there is a similar systematic problem for yellowfin tuna at this time, but there are similar frustrations in reconciling different sources of data in the models.

2. Spatial structure for the Yellowfin Operating Model – comparison of individual models

The yellowfin OMs are conditioned to data with Stock Synthesis (SS, e.g. Methot and Wetzel 2013) models, which are derived from IOTC stock assessments (Fu et al. 2018) and associated analyses (Urtizbera et al. 2020). As the stock assessments have moved to a multi-model grid-based approach, the parallels between the OM and stock assessment have increased, though the OMs place a greater emphasis on representing uncertainty, to ensure that Management Procedures (MPs) are robust to plausible alternative interpretations of the data. Urtizbera et al (2020) proposed a grid of YFT assessment models for the provision of revised management advice, that compared 2 area and 4 area configurations (plus 1 area models that were rejected). As far as we understand, the original justification for the 2 area configuration was parsimony (reducing overparameterization, and increasing numerical stability). WPTT (2020) initially requested the OM to also include 2 area and 4 area models, but a number of disadvantages to this were recognized, including:

- There was no movement in the 2 area configuration, which implies two independent populations (except for the aggregate stock-recruit relationship). This potentially might lead to troubling interpretation of management advice (e.g. why reduce fishing effort here, if the problem is in a different population on the other side of the ocean?)
- Tag mixing assumptions are less likely to be met across larger areas, and hence the tagging data were not included in the 2 area configurations. This confounding of tagging assumptions and spatial structure potentially unbalances the overall grid in a manner that was not intended (i.e. tagging data only given full weight in a small portion of the overall ensemble).
- Urtizbera et al (2020) ranked the ensemble of assessment models on the basis of diagnostics of internal consistency, with a tentative recommendation to retain 18 models for the management advice, 83% of which were based on the 4 area structure.
- There was no specific demonstration that the inferences from the 2 area and 4 area models were substantively different. And if there were important differences, there was no way to distinguish whether this was caused by the differing spatial structure per se, or confounded assumptions (i.e. inclusion or exclusion of tags, or the restriction of West-East movement).

Given that the MSE source code would require substantial modification to mix OMs of multiple spatial structures within an ensemble, further investigation was suggested before adopting multiple spatial structures within the YFT OM.

Seven models are defined in Table 1. Six were originally intended to compare the 2 and 4 area structures in relation to the confounding assumptions, and one additional model was added to revisit the influence of environmental links to movement (investigation of OM ensembles is described in the following section). The (spatially-aggregated) behaviour of these models is very similar with respect to spawning biomass, fishing mortality and recruitment patterns (Figure 1).

The 2 area structure is slightly more optimistic than the 4 area configurations. Among the 4 area models, there is a trivial difference between tag $\lambda = 0$ vs 0.1, and open vs: blocked West-East migration. The biggest difference is observed between the 4 area model with $\lambda = 1.0$ and the others ($\lambda = 0$ or 0.1). Removing the environmental link to movement had a modest effect, qualitatively similar, but less extreme than the tag weighting.

Figure 2 - Figure 4 compare some regional inferences among the 2 area configuration and four of the 4 area models that differ only in the value of tag λ and the removal of the environmental link to movement, from which we note:

- There is no appreciable visual difference between the 4 area models with tag $\lambda = 0$ or 0.1.
- The 2 area model does not appear to suffer from the very high F problem, with a maximum among individual fisheries of 0.84 (fishery 25 - fresh tuna LL region 4). In contrast, the 4 area models with tag $\lambda = 0$ and 0.1 appear to hit the default maximum $F = 2.9$ (fishery 15 - troll region 4), while the models with tag $\lambda = 1.0$, and removal of the environmental link to movement do not reach the limit, with a peak $F \sim 1.8$ and 2.0 (also fresh tuna LL region 4).
- All of the models are qualitatively very similar with respect to the West – East biomass distribution. The models with environmental-linked movement show some low amplitude seasonal patterns, but similar annual trends to the other configurations.
- All of the models appear to have very similar temporal trends in the estimated recruitment spatial distribution, with a gradual increase in the area 1 proportion, until the most recent (data-limited) period which reverts to the initial pattern (i.e. identical to the period before deviates are estimated).

We cannot be certain that these conclusions would remain valid across other OM assumptions (and the yellowfin models are somewhat prone to identifying local minima). The similarity among models suggests that the spatial structure per se is not introducing much variability to the overall estimates of stock status and productivity, but the minor region-specific fishing mortality estimates are somewhat sensitive. We would hope that these sensitivities do not have much impact on MP evaluations, but their importance may be magnified under circumstances that differentially prevent some fisheries from attaining their allocated quotas.

Tentative Conclusions:

- The biomass and recruitment trends are similar between 2 and 4 area models, including the West-East spatial distribution. The most influential assumption appears to be the inclusion of the tagging data (which supports more pessimistic outcomes).
- The suspiciously high F values for the 4 area models, suggest that this spatial configuration might be too disaggregated, and simply incapable of allocating the fish to the right locations (e.g. perhaps mean movement rates are inadequate if there is significant seasonal and interannual variability).
- There does not appear to be much value in retaining both the tag $\lambda = 0$ and 0.1 options in the 4 area grid (with environmental links). However, the opposite view was evident in subsequent sections (Figure 30).

Table 1. Individual SS models defined to explore spatial effects (Other model assumptions and R1, R13 and R14 labels are adopted from Urtizbera et al 2020). ENV indicates a link between movement and environmental indices as in the Fu et al (2018) assessment.

Model	Areas	Tag λ	Movement
2A_lbda0_MB_GF_h08_nEW (R1)	2	0.0	none
4A_lbda0_MB_GF_h08_mv	4	0.0	NS & WE (ENV)
4A_lbda0_MB_GF_h08_nEW	4	0.0	NS only (ENV)
4A_lbda01_MB_GF_h08_nEW	4	0.1	NS only (ENV)
4A_lbda01_MB_GF_h08_mv (R13)	4	0.1	NS & WE (ENV)
4A_lbda1_MB_GF_h08_mv (R14)	4	1.0	NS & WE (ENV)
4A_lbda01_MB_GF_h08_mv_noENV	4	0.1	NS & WE (<u>No</u> ENV)

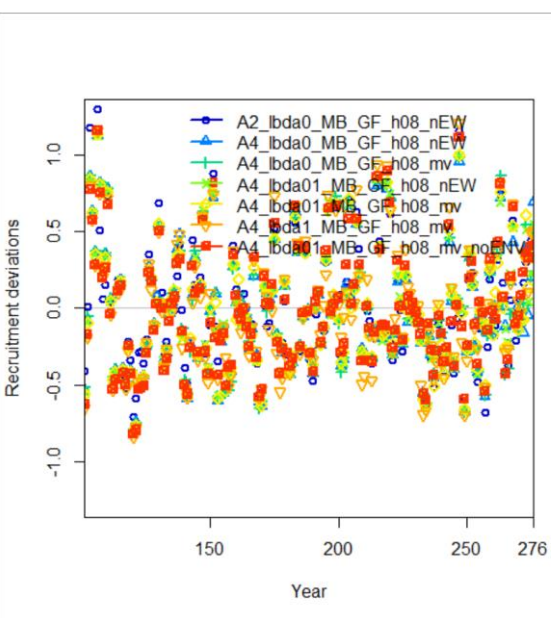
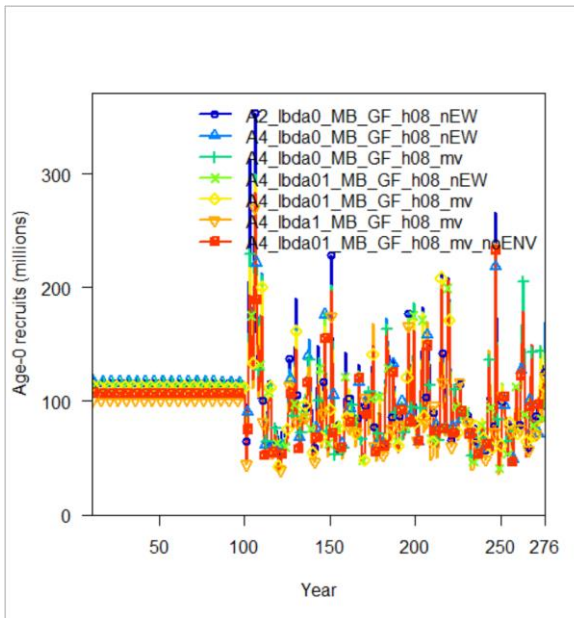
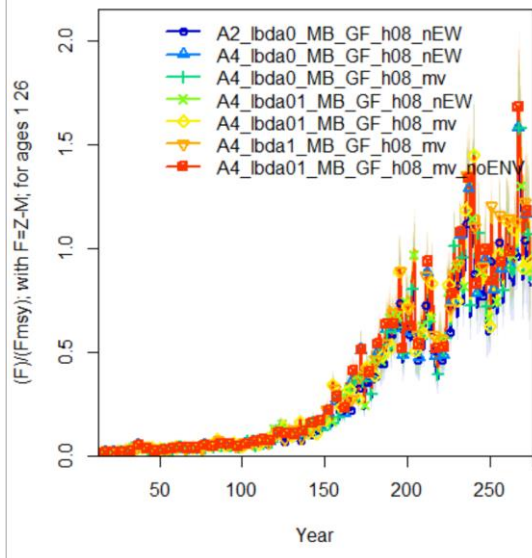
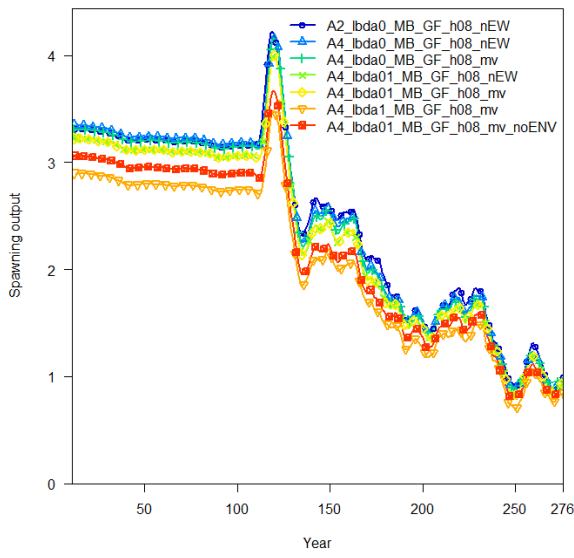
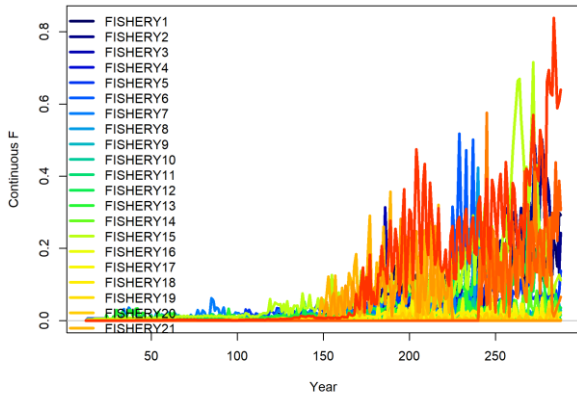
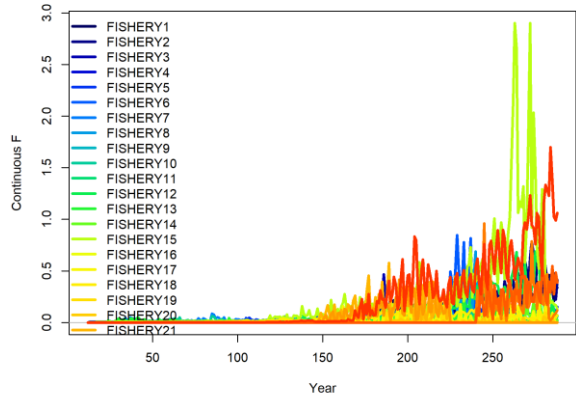


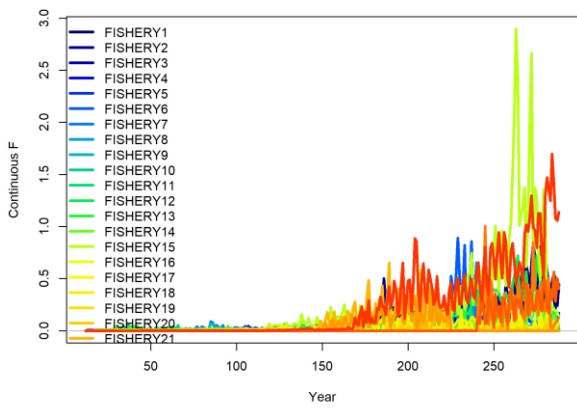
Figure 1. Comparison of aggregate dynamics from models defined in Table 1.



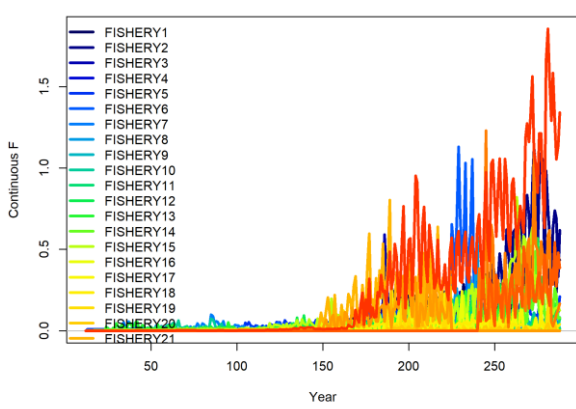
A) 2 Areas, no tags, no movement



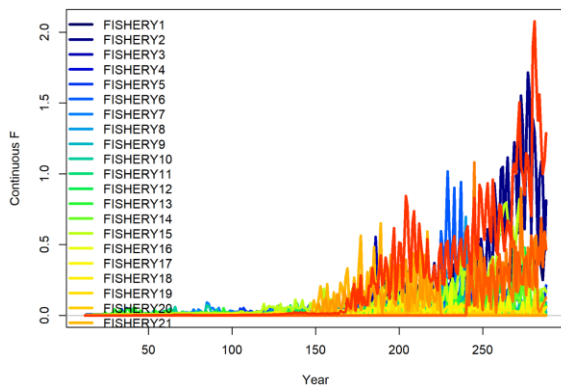
B) 4 Areas, tag $\lambda = 0$, full movement



C) 4 Areas, tag $\lambda = 0.1$, full movement

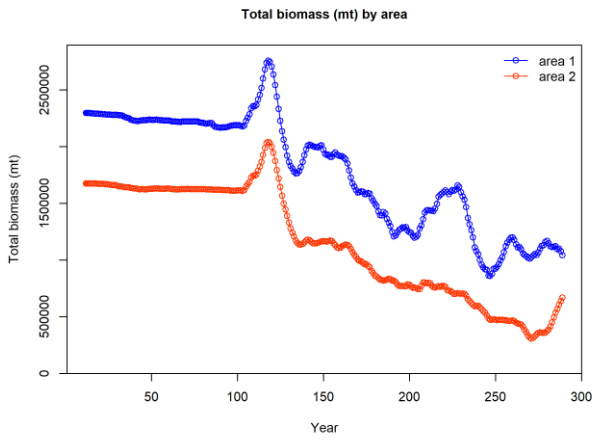


D) 4 Areas, tag $\lambda = 1.0$, full movement

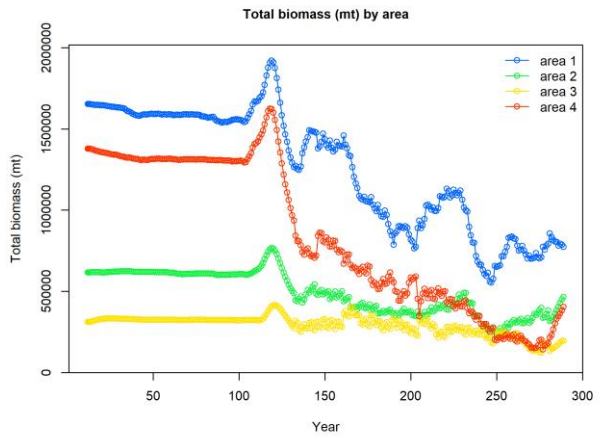


E) 4 Areas, tag $\lambda = 0.1$, full movement not linked to environmental indices

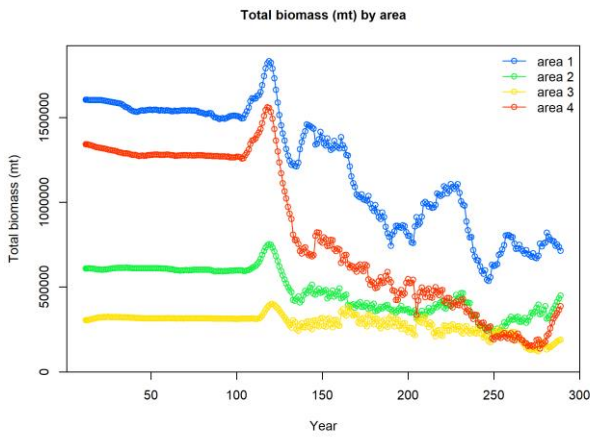
Figure 2. Fishing mortality by fishery for select models defined in Table 1.



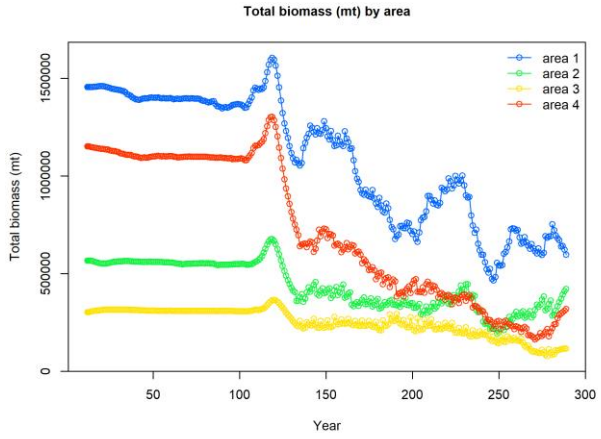
A) 2 Areas, no tags, no movement



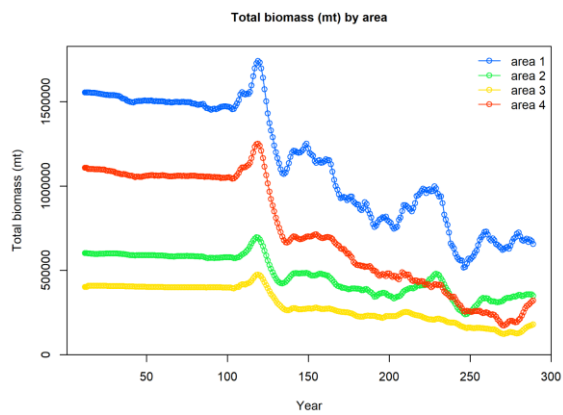
B) 4 Areas, tag $\lambda = 0$, full movement



C) 4 Areas, tag $\lambda = 0.1$, full movement

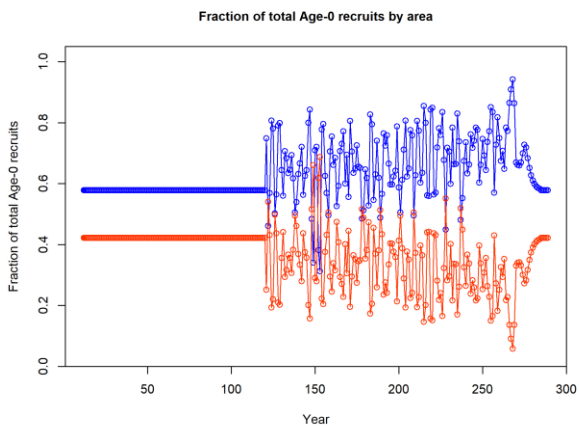


D) 4 Areas, tag $\lambda = 1.0$, full movement

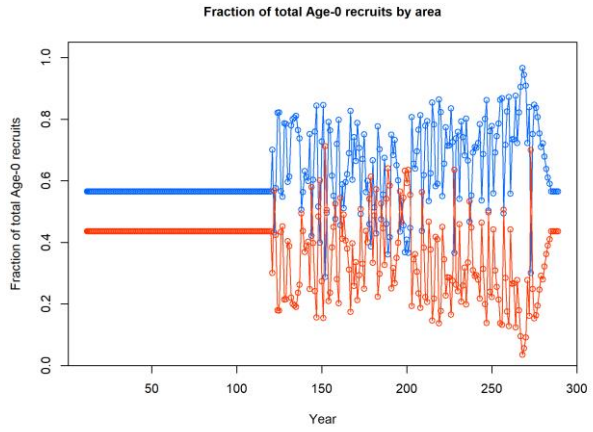


E) 4 Areas, tag $\lambda = 0.1$, full movement not linked to environmental indices

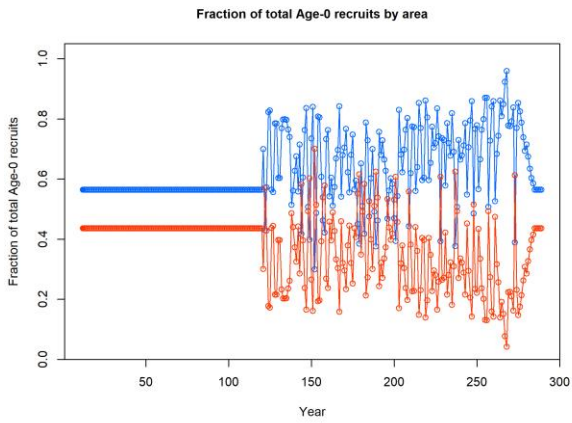
Figure 3. Total biomass distribution by region for select models defined in Table 1.



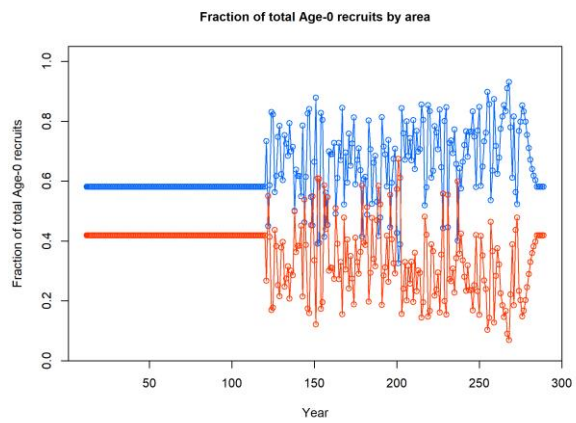
A) 2 Areas, no tags, no movement



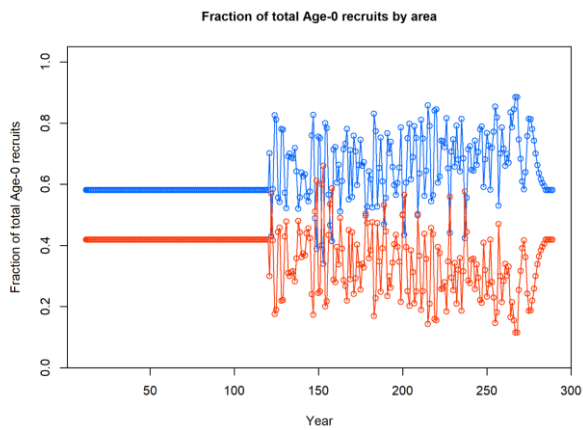
B) 4 Areas, tag $\lambda = 0$, full movement



C) 4 Areas, tag $\lambda = 0.1$, full movement



D) 4 Areas, tag $\lambda = 1.0$, full movement



E) 4 Areas, tag $\lambda = 0.1$, full movement not linked to environmental indices

Figure 4. Recruitment distribution (west in blue, east in red) of select models defined in Table 1.

3. Spatial structure for the YFT Operating Model – comparison of ensembles

The reference set OM grid structure represented a mix of features from the 2018 YFT OM, and new elements derived from the latest efforts of the YFT assessment team (Urtizbera et al 2020) as defined by WPTT (2020). Table 2 explains the option abbreviations used in this paper (and corresponds to the reference set of options for the requested 4 area configuration). The 4 area reference set OM grid and potential 2 area alternative (or addition) is defined in appendix B. Seven OM grids were fit (and several other OM ensembles were adapted on the basis of these existing grids, Table 3) to further consider the importance of spatial issues. The comparison of these grids was intended to:

- Examine the general plausibility of the different OM ensembles.
- Determine the relative importance of spatial configurations in relation to the confounded assumptions of including/excluding movement and tags.
- Determine the extent to which time blocking of recruitment spatial distribution estimates affects MP evaluation results.
- To confirm that the inclusion/exclusion of tags is the biggest cause of stock status differences between the requested 2 area and 4 area configurations
- Test the implications of including 2 additional years of catch data in the conditioning.

Time blocking refers to the option of estimating temporal variability in SS parameters. The default assumption for the assessment and OM to date has been a stationary distribution of recruitment spatial deviates (unchanging over time). In 2020, the YFT assessment team identified a concern about this assumption in relation to problems with the SS assessment projections (Max Cardinale, pers. comm.), which manifests as shown in Figure 4. It is not known whether this is a real phenomenon, or an artefact of some other model problem. But the issue can be modelled in a manner that is more internally consistent, and should yield projections that reflect recent recruitment patterns. The method that we have adopted to explore this problem involves fitting the recruitment spatial distribution parameters in two time blocks 1950-2007 and 2008-2017. This essentially assumes a step function regime shift in 2008. This is probably not realistic, but it does allow the projections to start with a recruitment distribution that is consistent with recent estimates. The intent was not to determine whether the shift in parameters was real, or to decide how best to model it. Rather the intent was to find out whether it would make any difference to MP performance. The expectation was that it would only matter under dire circumstances, in which fisheries are struggling to remove their quotas from the region in which they are operating.

The models in the OM grids all differed from those run in the previous section and Urtizbera et al (2020), in that the environmental links to movement were removed. This is consistent with all previous YFT OMs except for a brief exploration of using multiyear mean environmental links in 2020, which did not appear to make a substantive difference to the model inferences (i.e. the main expectation was an improvement to the fit of the seasonality in the CPUE series, and tags to a lesser extent, but this was not very successful).

The OM grid calculation protocols were similar to, but streamlined relative to the reference set OMs adopted in previous iterations:

- 54 model fractional factorial design in all cases (the full factorial cross would be 432 models for the 4 area structure).
- Convergence was deemed successful if the maximum gradient was < 0.01 (and varied among OMs as summarized in Table 3).
- Each individual model was run with a repeated jittered minimization, with a goal of 3 successful convergences, in a maximum of 10 attempts. The fitting with the lowest objective function (that also met the convergence criterion) was retained for the OM.
- Parameters on bounds issues were assumed to be minimal, because the template files were adopted from previous iterations in which important bounds were progressively relaxed (and seemed okay for the grids that were checked).
- Consideration of model diagnostics was minimal – cursory inspection for outlier behaviour, of which there did not seem to be much evidence for concern. Models with a “substantial” catch likelihood (indicating failure to remove the catch from at least one time/age/area strata) were retained.

Key summary results are shown for a subset of the OM grids in (Figure 5 - Figure 13), from which we note:

- The 2 area ensemble was not affected by the very high F problems that affected the 4 area grids (Figure 5). The catch (negative log-) likelihood tends to resolve into two modes, with very few models in the intermediate region of 10^{-7} to 10^{-3} . Our understanding is that the higher values occur when the SS hybrid F configuration cannot remove the observed catch because $F > 2.9$ (the default setting in which the highest exploitation rate for an individual age/region strata $\sim 95\%$). Alternatively, this can occur because there are not enough iterations specified for the catch equation algorithm, but we have not found this to be a problem (with the default setting of 4 iterations raised to 7). About two thirds of models were affected in the 4 region ensemble, compared with 0 in the 2 region ensemble. The consequences of ignoring the catch likelihood are not fully understood, such that we have sometimes used it as a criterion for rejecting models as implausible, while it has also been knowingly ignored in assessments. We expect there are circumstances where the effect would be similar to hitting a hard parameter bound and might have a large influence on model dynamics. In other cases, it might simply indicate that the distribution of fish was slightly wrong in one historical age/region/quarter strata, and could have a trivial effect on the overall dynamics.
- The stock status distributions and summary diagnostics were almost identical regardless whether the recruitment spatial distribution was estimated with one or two time blocks (only a few representative comparisons are included, e.g. Figure 6 - Figure 8, Figure 9). This was expected, and the real question is whether it matters for the projections (next section).
- Both the 2 and 4 area OM ensembles estimate the stock status to be in a state that is somewhat more pessimistic than the 2018 assessment (Figure 6 - Figure 8). This would be

expected a priori, simply on the basis that the assessment does not have the 1% per year CPUE catchability trend assumption that represents 50% of the OM grid elements. The distribution of reference points between the two ensembles appears to be very similar, with the 4 area OM somewhat more pessimistic. This latter discrepancy might be largely a result of including the tag data in the 4 area OMs.

- Both spatial ensembles fit the MP CPUE (i.e. annual index aggregated over seasons and regions) better than we would have reason to expect (RMSE < 0.17). There is substantial systematic lack of fit to the CPUE for some models in both ensembles maximum lag(1 y) auto-correlation ~ 0.8 , but we do not consider this to be a serious problem, given that the RMSE is low.
- Both spatial ensembles fit the size composition better than the (very low) input effective sample sizes, with very similar characteristics and little evidence for outlier behaviour as indexed by the mean post-fit Effective Sample Sizes. The only exception was 1-2 models (for both ensembles) and fisheries 4 and 14 (which are both miscellaneous aggregations, probably heterogenous and poorly sampled).
- The estimated (annual aggregate) recruitment variability is modestly higher for some of the 4 area models than the 2 area models (max $\sigma_R \sim 0.5$ vs 0.4, Figure 12). The patterns of recruitment deviates were always similar (Figure 13), with multiple short (< ~ 5 y) positive or negative stanzas, but no substantial long term trends evident.

These results indicate that, at the basin scale, the inferences from the 2 area and 4 area OM grids have a large degree of overlap. The most substantial difference between the two appears to be related to the exclusion or inclusion of the tagging data. Higher weight to the tagging data has a more pessimistic influence on the stock status.

Figure 14 illustrates the recruitment spatial distribution from some contrasting 2 area and 4 area OM models. The trend in the recruitment spatial pattern evident in Figure 4 is not as pronounced in these example OM models (i.e. it does not appear to exist in one case).

We have not repeated some of the conditioning diagnostics that have been presented in previous iterations, because there has not been a substantial data update, and there has been very little fundamental change to the general characteristics of the OM.

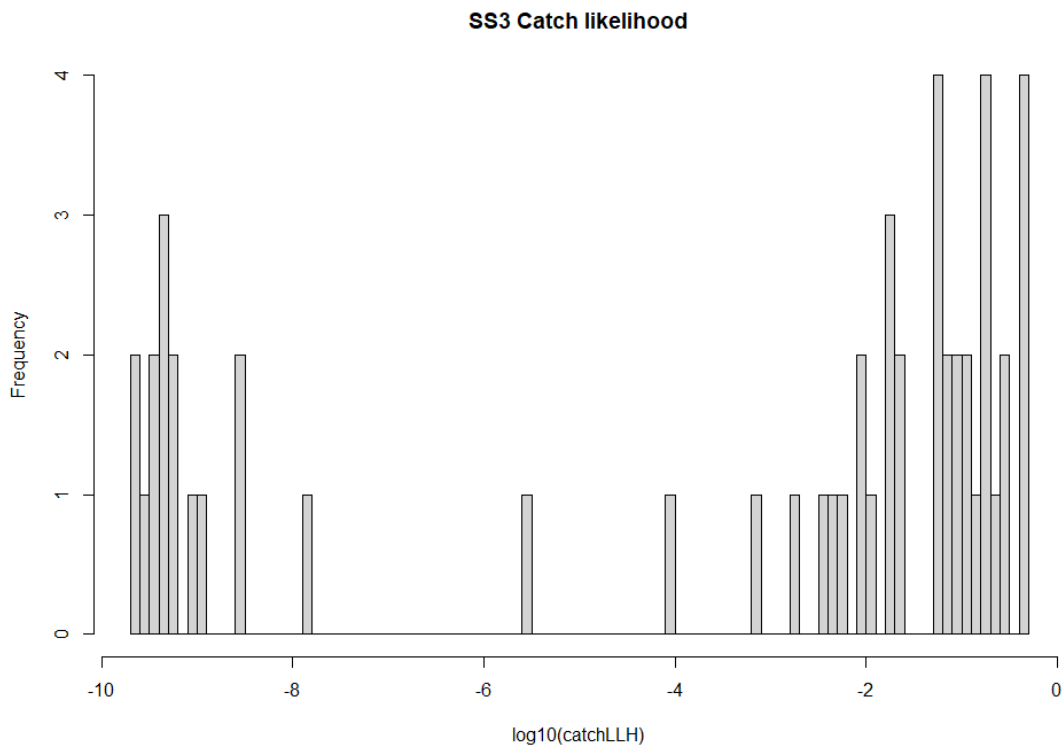
Section 3 Tentative Conclusions:

- 1) The problematic temporal trend in the spatial distribution of recruitment is not evident in all models within the requested 2 and 4 area OM ensembles, and estimating the recruitment distribution in either one or two time blocks does not appear to make a noticeable difference to the large scale stock status summaries and diagnostics considered here. This issue is revisited in the following section to see if it is important in the MP evaluations. However, a potential unresolved issue was identified late in the preparation of this paper – the recruitment deviation parameters for the second time block appeared to be estimated with a value very close to 0 in the models checked, despite a relaxed prior (sd = 0.5). This may indicate a problem in the SS control file set up.

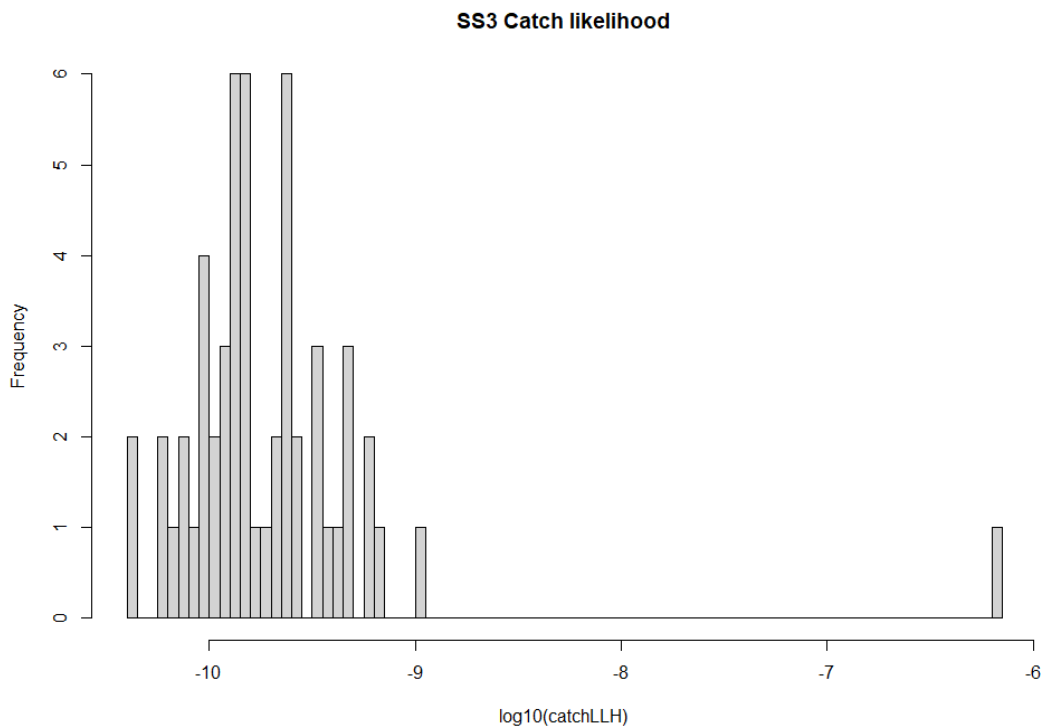
- 2) From the perspective of stock status uncertainty within the OM, it appears that choosing between the 2 and 4 area spatial structure makes very little difference, and the more important issue is whether the tags are included.
- Retaining the 4 region structure is more defensible from the perspective of retaining the tags, and disaggregating tropical and temperate fisheries.
 - Adopting the 2 area structure is more attractive from the perspective of distributing the fish in such a way that all fisheries can attain the reported catches. This issue is revisited in the following section to see if it is important in the MP evaluations.

Table 2. Abbreviations used in this paper to refer to the YFT (4 area) reference set OM uncertainty dimensions defined by the WPTT (2020).

Abbreviation	Definition
	<u>Stock-recruit function ($h =$ steepness)</u>
h70	Beverton-Holt, $h = 0.7$
h80	Beverton-Holt, $h = 0.8$
h90	Beverton-Holt, $h = 0.9$
	<u>Natural mortality (multiplier relative to reference case M vector M10)</u>
M10	1.0 - Base case (1.0)
M08	0.8 – intermediate M (to smooth bimodal OM results)
M06	0.6 – low M
	<u>Tag recapture data weighting (tag composition and negative binomial)</u>
t0001	$\lambda = 0.001$
t01	$\lambda = 0.1$
t10	$\lambda = 1.0$
	<u>Growth curve</u>
gr2	Fonteneau (c. 2012)
gr3	Dortel et al. (2014) model 3 lognormal (with compromised variance)
	<u>Assumed longline CPUE catchability trend (compounded)</u>
q0	0% per annum
q1	1% per annum
	<u>Longline CPUE error assumption (quarterly observations)</u>
i3	$\sigma_{\text{CPUE}} = 0.3$
i1	$\sigma_{\text{CPUE}} = 0.1$
	<u>Tag mixing period</u>
x4	4 quarters
x8	8 quarters

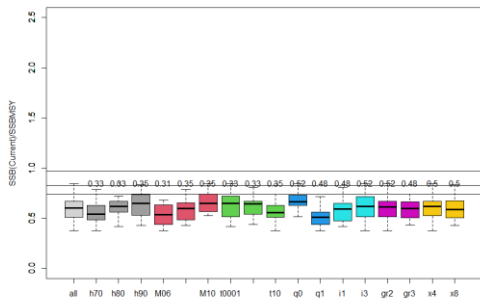


4 Areas (OMgridY21.1)

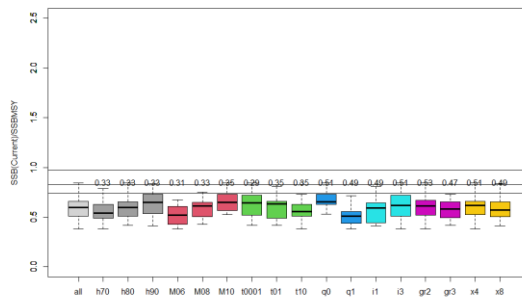


2 Areas (OMgridY21.2)

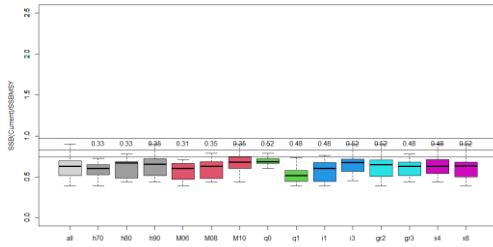
Figure 5. Catch likelihood distribution (indicator of dubiously high fishing mortality and failure to completely extract catch for at least one model strata), for the reference set 4 area and 2 area OMs. Note that X-axes differ.



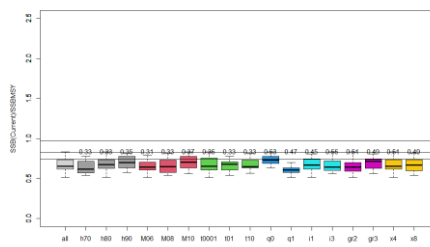
4 Areas (OMgridY21.1)



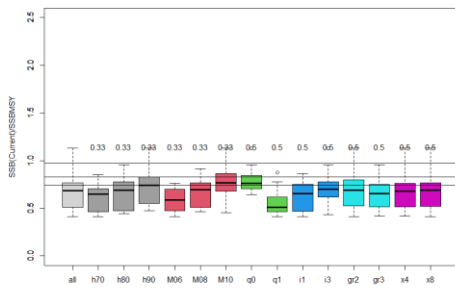
4 Areas (OMGridY21.3)



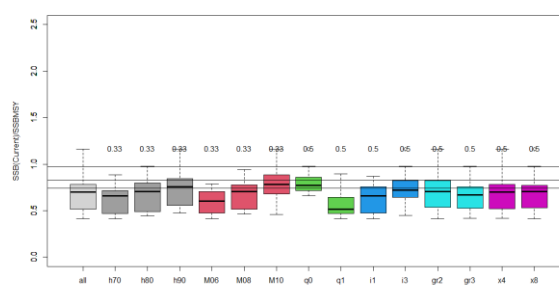
4 Areas (OMgridY21.5)



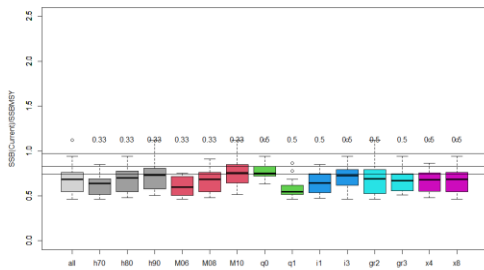
4 Areas (OMGridY21.6)



2 Areas (OMgridY21.2)

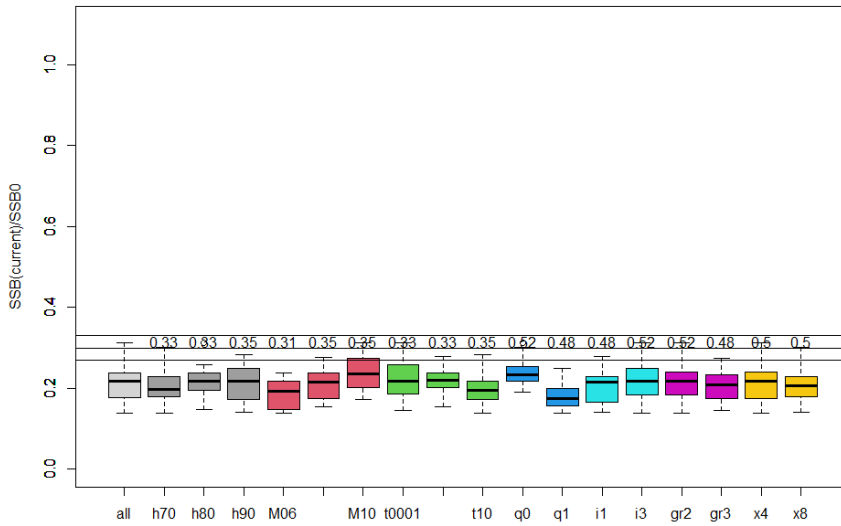


2 Areas (OMGridY21.4)

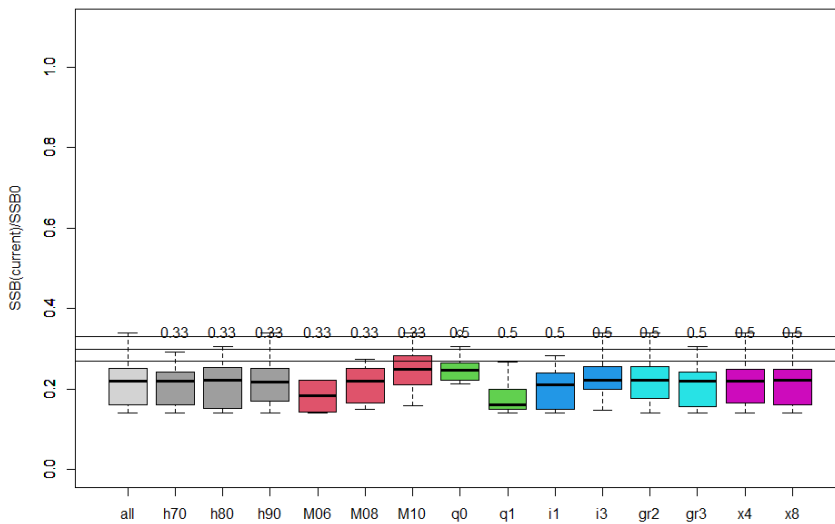


2 Areas (OMgridY21.8)

Figure 6. Comparison of OM ensemble B(T)/BSMSY, for a range of OM grids (panels). Each box is a summary for a particular assessment option, marginalized over all other option dimensions in the grid. Grid option abbreviations defined in Table 2.

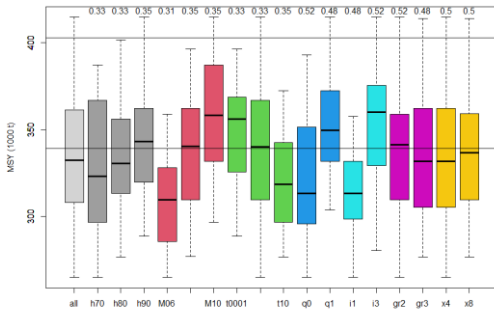


4 Areas (omRefY21.1)

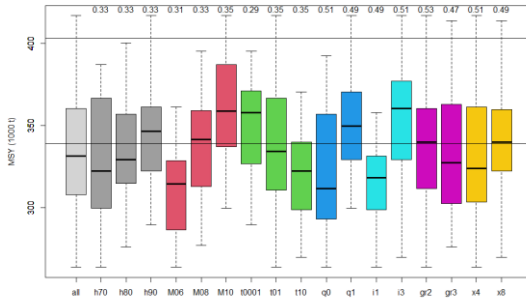


2 Areas (omRefY21.2)

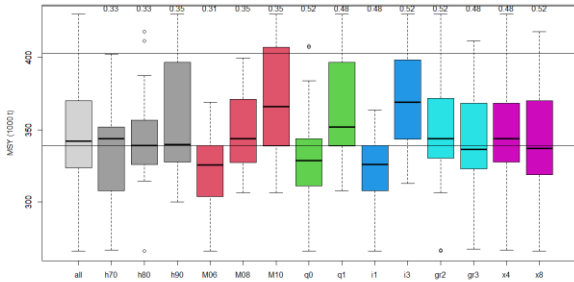
Figure 7. Comparison of OM ensemble depletion $B(T)/B_0$ for a range of OM grids (panels). Each box is a summary for a particular assessment option, marginalized over all other option dimensions in the grid. Grid option abbreviations defined in Table 2.



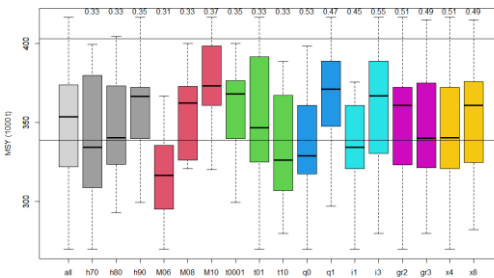
4 Areas (OMgridY21.1)



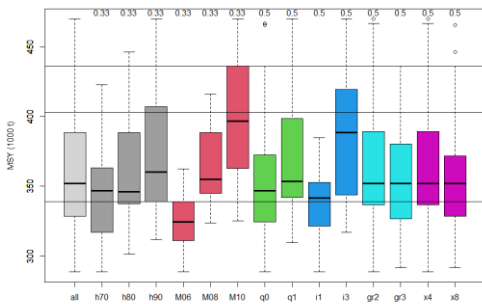
4 Areas (OMGridY21.3)



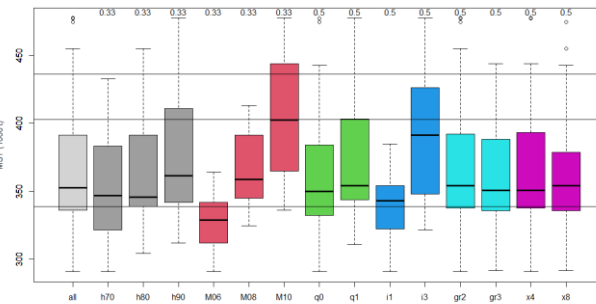
4 Areas (OMgridY21.5)



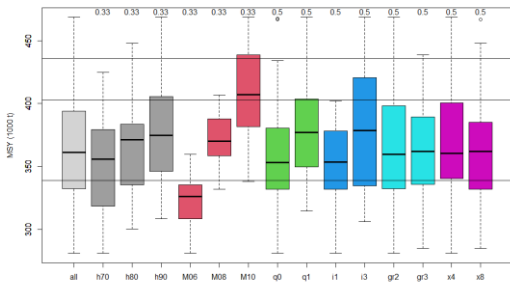
4 Areas (OMGridY21.6)



2 Areas (OMgridY21.2)

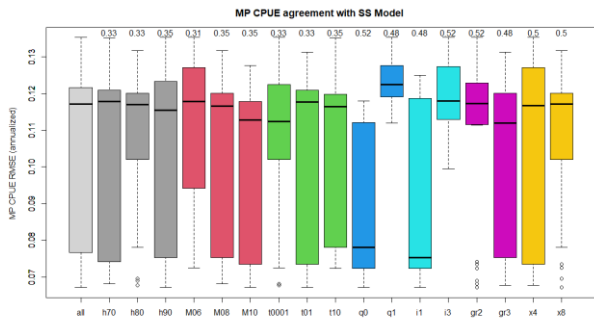


2 Areas (OMGridY21.4)

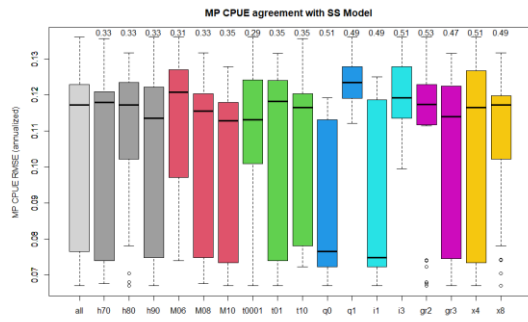


2 Areas (OMGridY21.8)

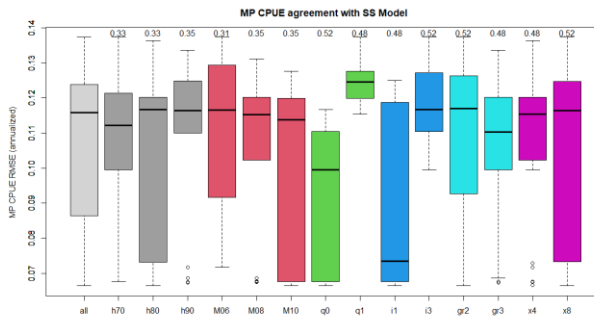
Figure 8. Comparison of OM ensemble MSY for a range of OM grids (panels). Each box is a summary for a particular assessment option, marginalized over all other option dimensions in the grid. Grid option abbreviations defined in Table 2.



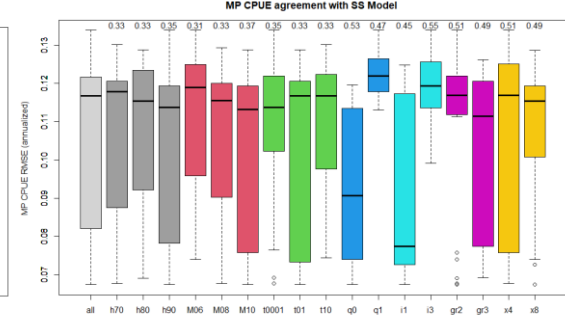
4 Areas (OMgridY21.1)



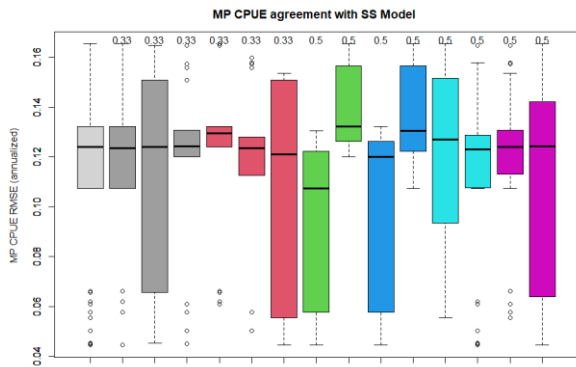
4 Areas (OMGridY21.3)



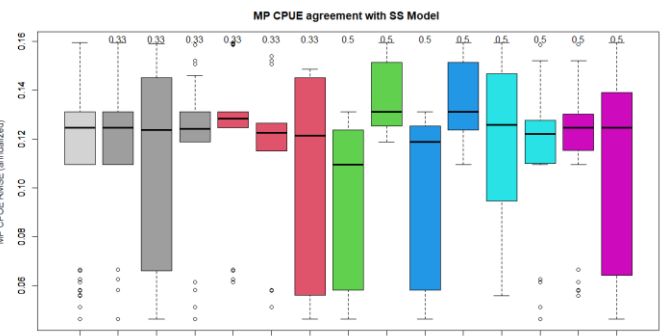
4 Areas (OMgridY21.5)



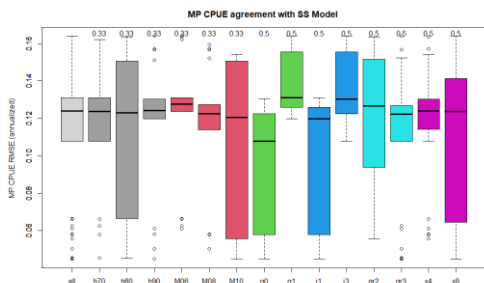
4 Areas (OMGridY21.6)



2 Areas (OMgridY21.2)

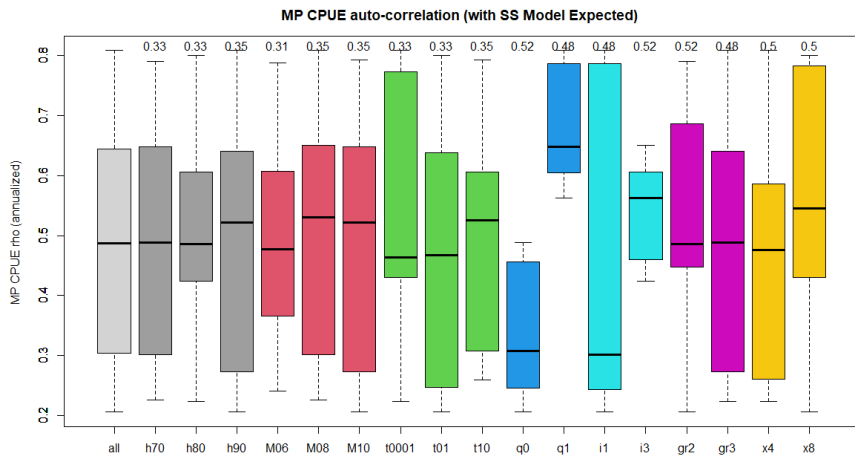


2 Areas (OMGridY21.4)

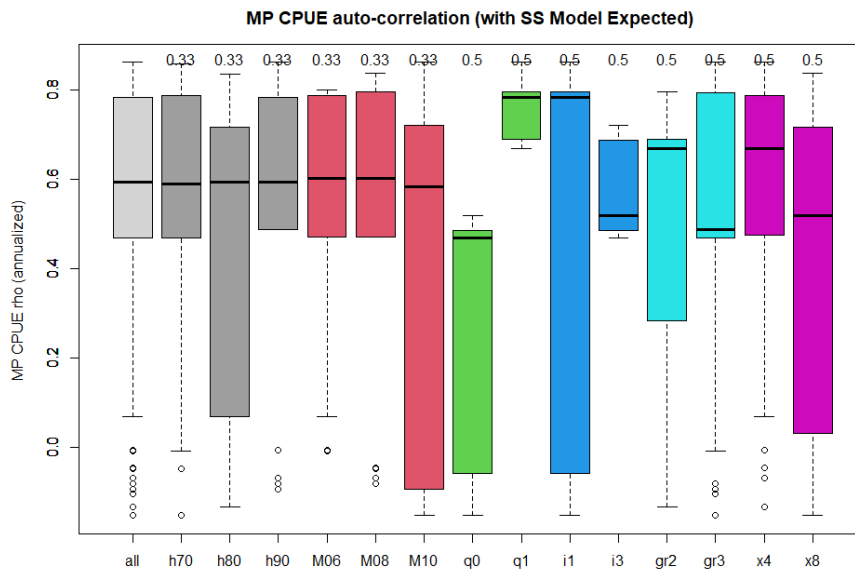


2 Areas (OMGridY21.8)

Figure 9. Quality of fit (CPUE RMSE) between observed MP CPUE and model LL vulnerable biomass (aggregated over seasons and regions as used in the MPs). Grid option abbreviations defined in Table 2.

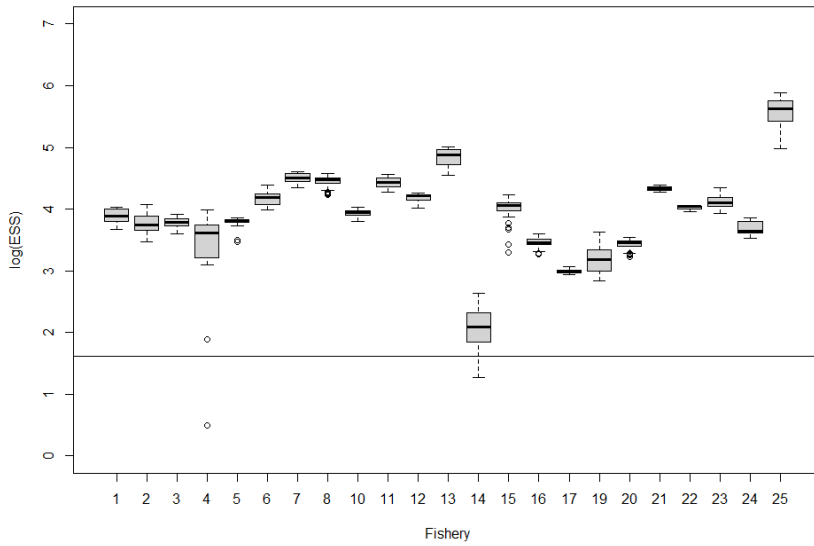


A) 4 Areas (omRefY21.1)

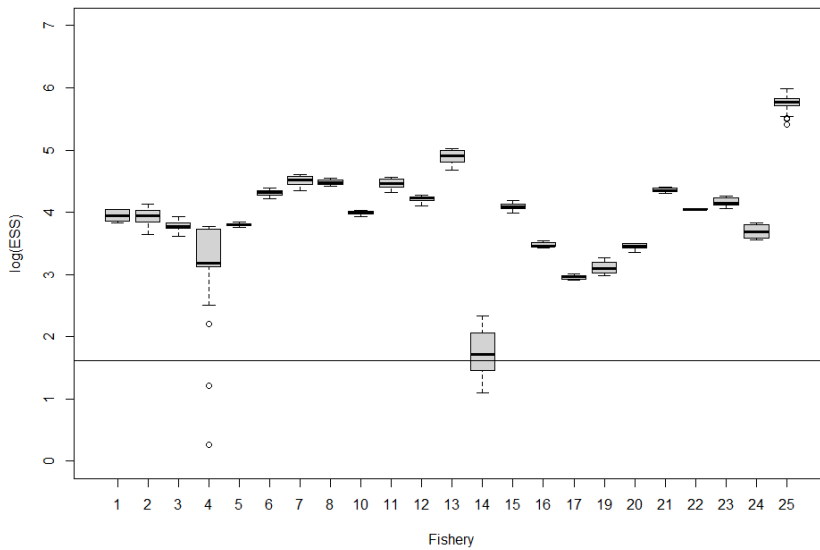


B) 2 Areas (omRefY21.2)

Figure 10. Systematic lack-of-fit (lag (1y) auto-correlation) between observed MP CPUE and model predictions. Grid option abbreviations defined in Table 2.

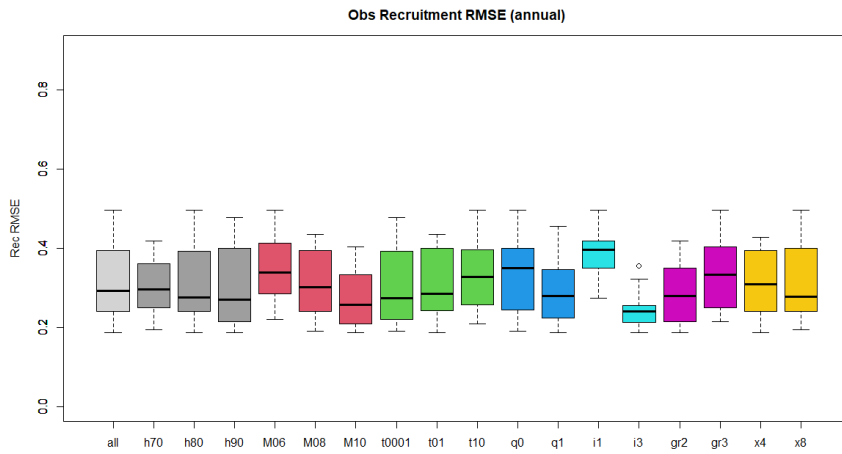


A) 4 Areas (omRefY21.1)

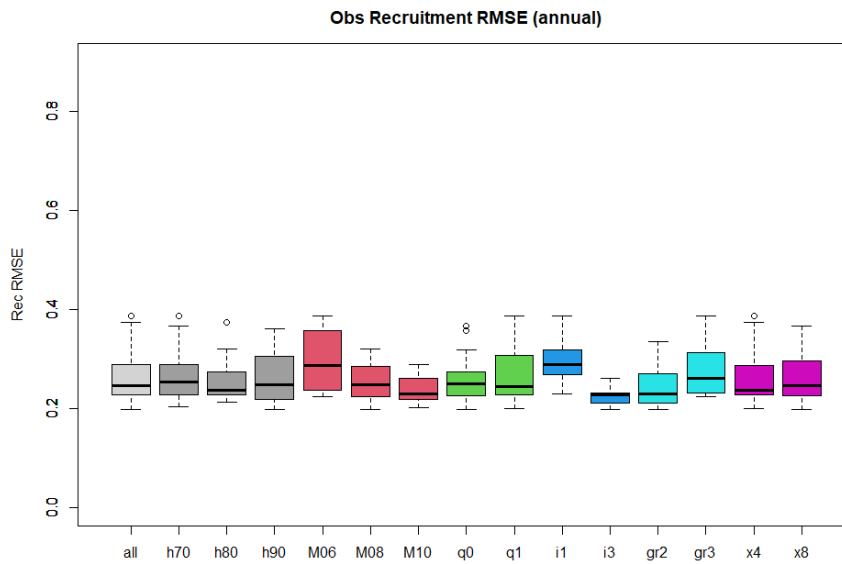


B) 2 Areas (omRefY21.2)

Figure 11. Quality of fit to catch-at-length distributions as indexed by the post-fit Effective Sample Size. All elements of the OM ensemble combined

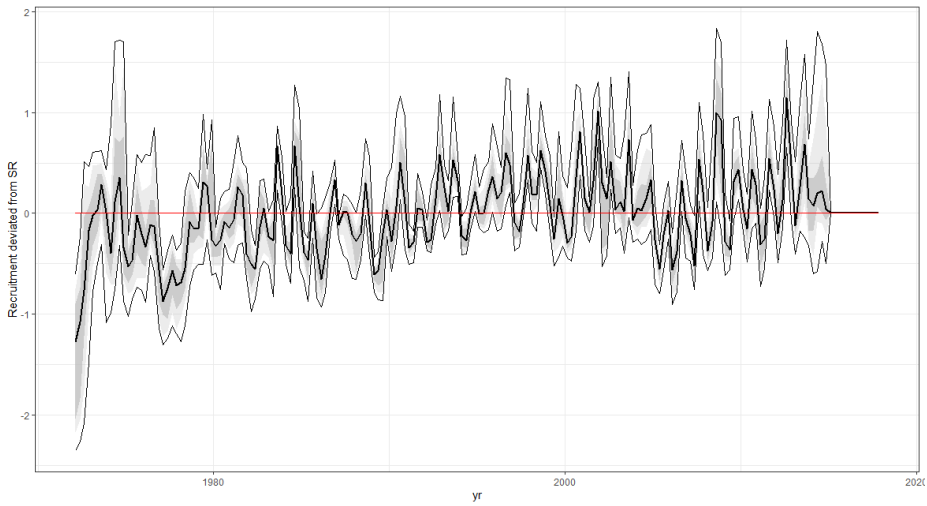


A) 4 Areas (omRefY21.1)

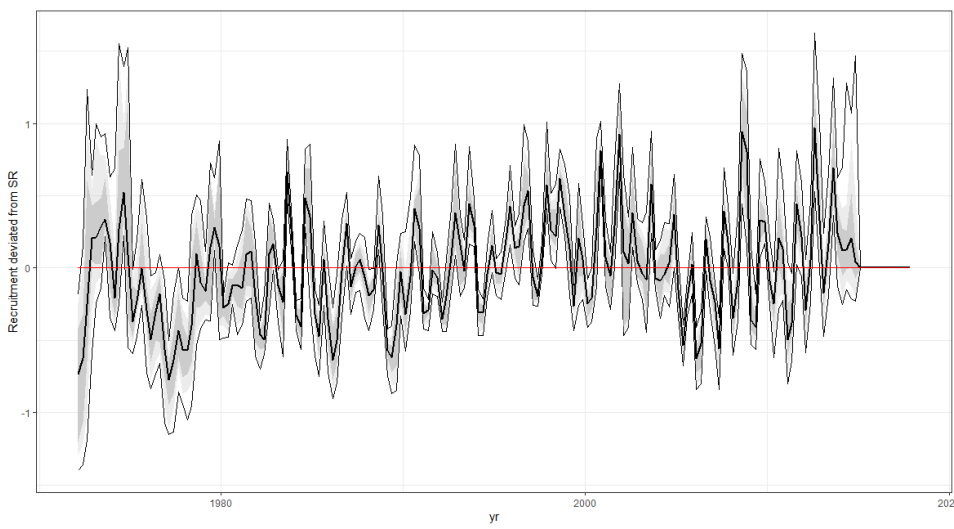


B) 2 Areas (omRefY21.2)

Figure 12. Annualized recruitment variability of the requested 2 Area and 4 Area OM ensembles. Grid option abbreviations defined in Table 2.



A) 4 Areas (omRefY21.1)

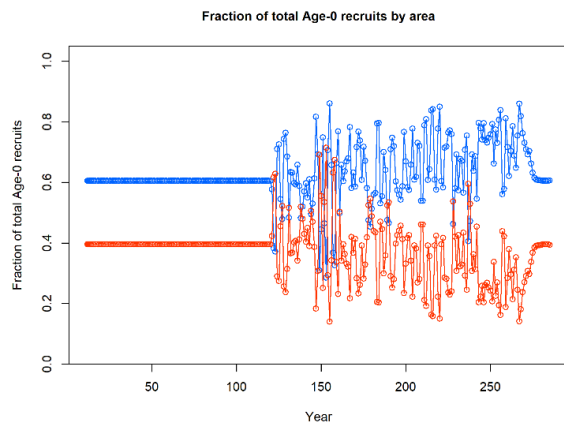


B) 2 Areas (omRefY21.2)

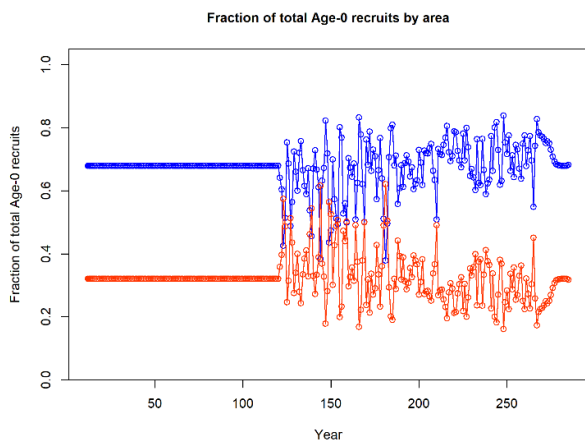
Figure 13. Distribution of (spatially-aggregated) quarterly recruitment deviation times series for the WPTT requested 2 area and 4 area OMs (all SS models).



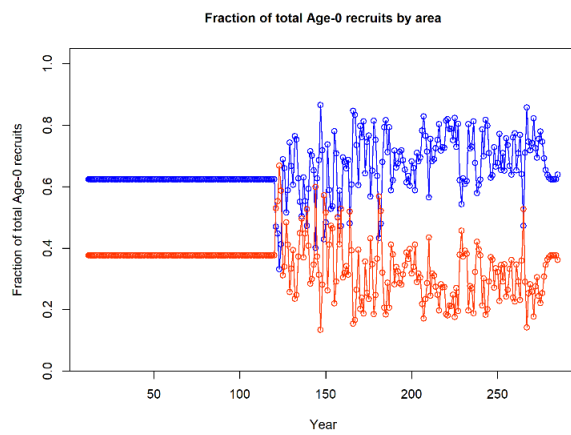
A) 4 Areas - h70_M06_t0001_q1_i1_gr3_x8



B) 4 Areas - h90_M10_t10_q0_i3_gr2_x8



C) 2 Areas - h70_M10_q1_i3_gr3_x8



D) 2 Areas - h90_M06_q0_i1_gr3_x4

Figure 14. Some example recruitment deviation spatial distributions from contrasting models from OMgridY21.1 (A - 4 areas, one time block of mean recruitment parameters), OMgridY21.3 (B - 4 areas, two time blocks of mean recruitment parameters), OMgridY21.2 (C - 2 areas, one time block of mean recruitment parameters), OMgridY21.4 (D - 2 areas, two time blocks of mean recruitment parameters).

4. Management Procedure Evaluation Results

In this section, we would ordinarily contrast the performance of a suite of different MPs, evaluated against a reference set OM, with some robustness tests. For reasons evident in the following, we did not consider it appropriate to compare MPs at this time. Instead we delve further into some problems with the OMs. The figures are mostly the same as those agreed as part of the TCMP presentation standard, but the intent is reversed. Rather than showing multiple MPs evaluated with a single OM, we are using the same MP, and testing how it performs with different OMs. This is more in the spirit of new robustness tests that attempt to explore key issues that have been identified in the proposed reference set OMs (and stock assessment).

The results below are all based on the single MP:

- TMB-based Pella-Tomlinson model with internal 10 year projections (described in companion paper Kolody and Jumppanen 2021). i.e. Similar to how the Commission might interpret the Kobe 2 Strategy Matrix, in each MP application, a model is fit, then solves for the constant TAC that will hit a particular depletion level 10 years in the future. The depletion level is the MP tuning parameter (i.e. chosen across the aggregate OM to hit the TCMP tuning objective).
- Tuning was done only once, for one of the rebuilding targets from the 2018 TCMP (50% probability of rebuilding to BMSY by 2034) using the OMgridY21.5 ensemble. OMgridY21.5 was selected for tuning because it was expected to represent an intermediate specification between the 2 area and 4 area reference set proposals (i.e. it includes 4 areas, but excludes the tags).
- TAC setting was subject to a 50% TAC change constraint, (change constraint of 35% or less not sufficient to allow successful tuning) updated every 3 years.

There is nothing particularly special about the MP and tuning objectives chosen – they were selected purely for the purposes of illustrating some of the implications of the different OMs, and much of the interesting behaviour occurs in the bridging period, between the first year of projections and before the first active quota setting. Time series results for this tuning are shown in Figure 15, and illustrate some general features of most or all of the OM grids explored:

- Very large quota reductions are required to hit the 2034 rebuilding objective.
- In > 90 % of realizations, the fisheries are unable to extract their quotas in the first few years of the projections, before the first MP-based TAC setting.
- Furthermore, more than 50% of the MSE realizations did not extract the catch that was actually reported in 2019 (2018 and 2019 were not included in the OM conditioning). And hence, this key feature of interest is independent of the specific MP tested.

Failure to extract the observed recent catch is consistent with the problematic retrospective pattern that has been recognized in recent years (i.e. when the model is fit with new data in year T+1, the stock status for year T tends to be slightly more optimistic when fit with data only up to year T). Addressing this problem is obviously desirable, but it remains unclear which structural

assumptions or data biases are causing the problem. The brief attempt to explore CPUE hyperdepletion in 2020 was not conclusive.

The retrospective problem manifests in all of the following results to some degree, and conceivably represents a higher priority for future iterations of the assessment and OM development than those explored below.

The issue of failing to extract the TAC has always been recognised as a potential issue that may be sensitive to nuisance parameters like seasonal movement and recruitment distributions. These may be difficult to estimate (and possibly non-stationary, or at least dependent on unpredictable environmental variation). It would be preferable if this problem could be avoided, but it is probably unavoidable if the population becomes highly-depleted as the current YFT OMs suggest. It is beyond the scope of the current iteration of the project to attempt to represent the fleet dynamics that would likely occur under these conditions. This would represent a non-trivial and inconclusive study in its own right (i.e. when will each fleet stop adding effort, change targeting and/or move to a different region?). It is possible that these complications might justify a reduction in spatial complexity in the OM context.

Time-integrated performance plots (of the first 15 years of projections) for the tuned MP for the relevant exploratory OM grids defined in Table 3 are shown together in Figure 17 and Figure 18. It appears that the MP performance for all of these OMs is reasonably similar, with the exception of OMgridY21.6, which is somewhat more optimistic. Subsets of these OMS are discussed in the following sections.

Table 3. Suite of YFT test grid OMs defined to explore sensitivity to spatial issues. In all cases the original grid specification included 54 models, though the grids differed depending on the spatial assumptions.

Grid (N Regions)	N models converged (Catch LLH < 10⁻⁵)	Grid features
gridY21.1 (4)	52 (16)	4 areas, no time series structure in recruitment spatial distribution. This conforms to the specification requested by the WPTT 2020, Appendix A.
gridY21.2 (2)	54 (54)	2 areas, no time series structure in recruitment spatial distribution (plus no tags and no movement). This conforms to the alternative specification requested by the WPTT 2020, Appendix A.
gridY21.3 (4)	51 (14)	4 areas, recruitment spatial distribution parameters split into 2 blocks with a breakpoint 10 years before the end of the model data
gridY21.4 (2)	54 (54)	2 areas, recruitment spatial distribution parameters split into 2 blocks with a breakpoint 10 years before the end of the model data (plus no tags and no movement)
gridY21.5 (4)	52 (18)	4 areas, with tags extremely downweighted, to confirm that the inclusion/exclusion of tags is the biggest cause of stock status differences between the requested 2 area and 4 area configurations
gridY21.5cpp (4)	52 (18)	As 21.5, except the OM adopted the C++ (Baranov) solution to the catch equations.
gridY21.5MU (4)	52 (18)	As 21.5, except the migration parameters for all ages were altered to represent a uniform distribution for every age each time-step. This resulted in inconsistencies with the CPUE series for reasons that remain unclear.
gridY21.6 (4)	49 (5)	as gridY21.1, except two extra years of total catch data were added (duplication of 2017, recognizing that this is a bit lower than reported in 2018 and 2019). The intent was to see if the extra years change the dynamics in a way that partially counteracts the retrospective pattern, and how this influences the quality of fit.
gridY21.7 (4)	49 (0)	as gridY21.6 except the 2017 catch data was repeated from 2018-2021. Not reported in detail, because all models failed the catch penalty criterion and there was no basis for justifying the catch for 2020-21.
gridY21.8 (2)	54 (20)	as gridY21.2 except the 2017 catch data was repeated from 2018-2019.

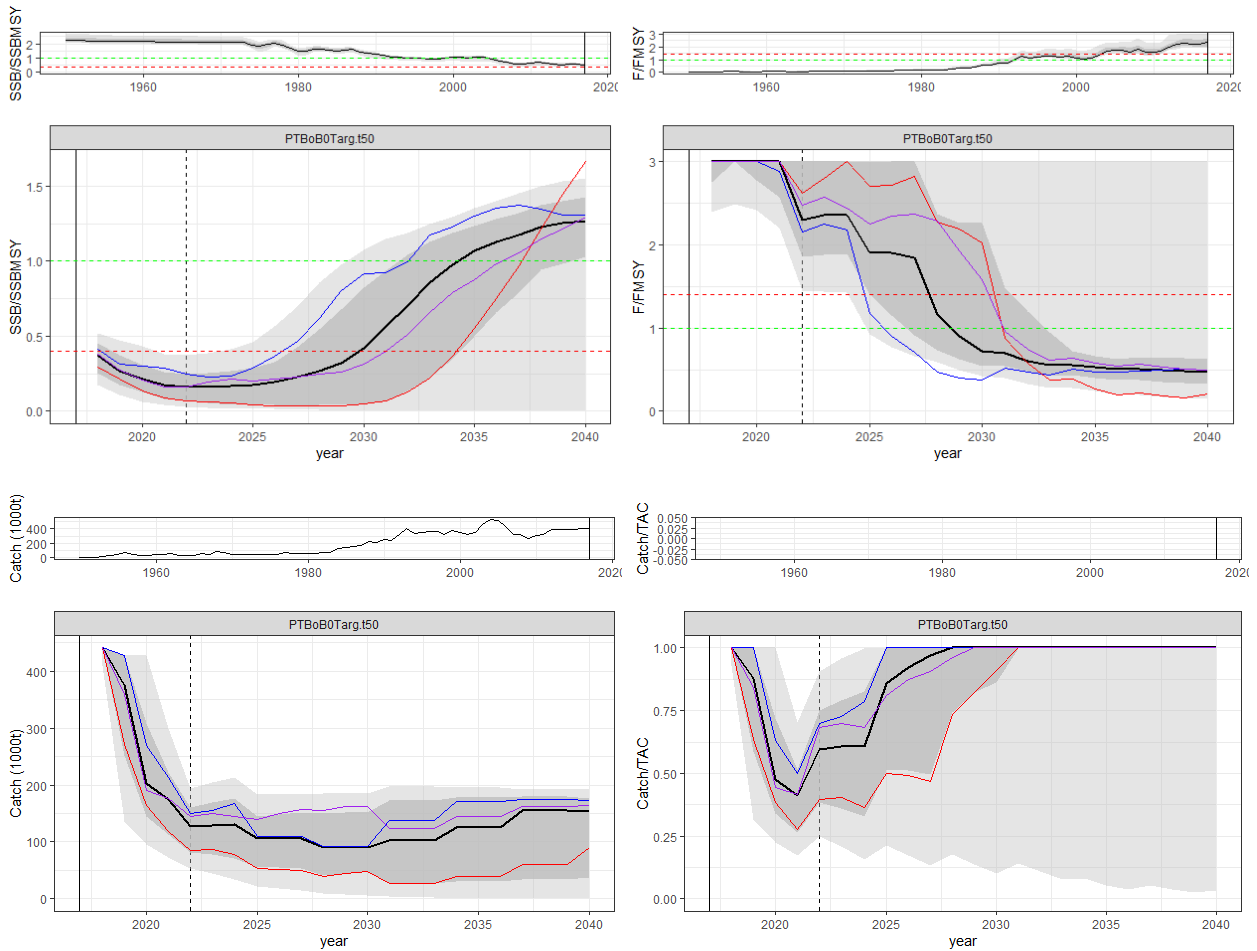


Figure 15. MP behaviour for the intermediate case OM OMgridY21.5, with the PTRE projection MP tuned for 50% recovery in 2034. Time series of biomass, fishing mortality, catch, and catch/TAC. The top section of each panel represents the historical estimates from the OM, and lower plots represent the projection period. The solid vertical line represents the last year of data 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 3 thin coloured lines represent examples of individual realizations to illustrate that individual variability greatly exceeds the median.

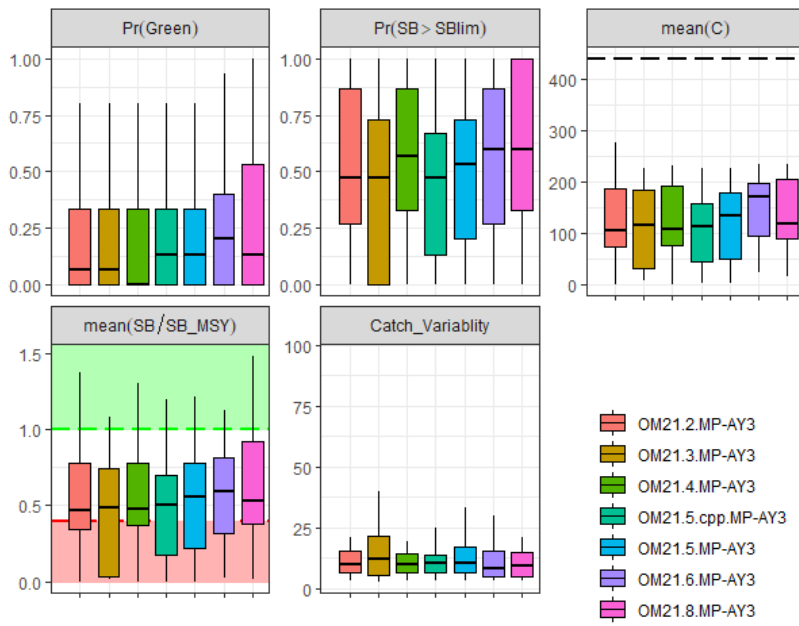


Figure 16. Fifteen year performance summary plots for the same MP when applied to the suite of candidate OMs in Table 3.

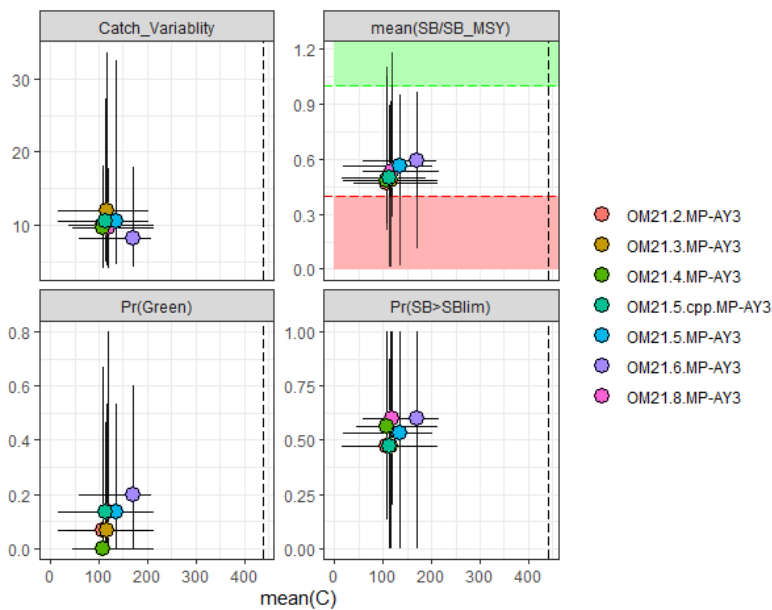


Figure 17. Fifteen year (2021-2035) performance summary plots for the same MP when applied to the suite of candidate OMs in Table 3. MP is the PTRE projection-based MP tuned for 50% recovery to BMSY in 2034 for OMgridY21.5.

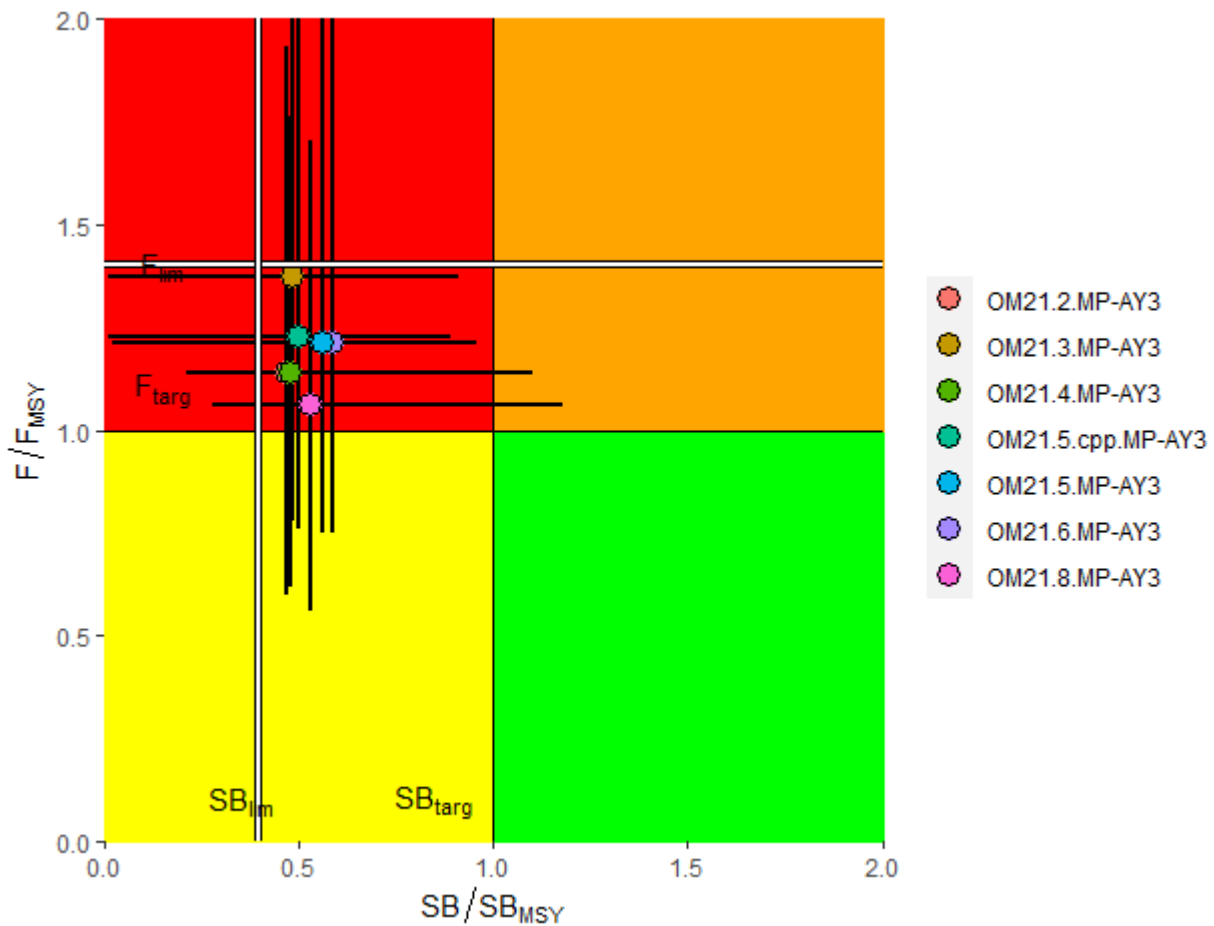


Figure 18. Fifteen year (2021-2035) MP performance summary plots for the same MP when applied to the suite of candidate OMs in Table 3. MP is the PTRE projection-based MP tuned for 50% recovery to BMSY in 2034 for OMgridY21.5.

4.1 The relative importance of spatial configurations in relation to the confounded assumptions of including/excluding movement and tags.

Figure 19 compares the example MP performance for the reference set proposed by WPTT (2020) - OMgridY21.1 (4 areas with tags and movement) and the 2 area alternative OMgridY21.2 (no tags or movement), and an intermediate configuration OMgridY21.5 (4 areas, no tags),:

- Most or all models converged in these configurations (52-54 out of 54, Table 3)
- No 2 area models were identified by the catch LLH flag, compared to a third of 4 area models (16-18), i.e. as might be expected greater disaggregation appears to be associated with more problems of locating the fish where the fisheries are trying to extract them.
- The MP evaluation results are fairly similar, though OMgridY21.5 seems to be the most different and, a priori, we had expected it to be intermediate.

- The time series dynamics of the MP performance are similar to Figure 15 for all three OMs (not shown). More of the 4 area OM realizations have trouble removing the TACs and bridging catches than the 2 area OM (Figure 20), but the issue is substantial in all cases.

This was intended to address the question of whether the 2 or 4 area spatial option (or both) should be included in the OM grid. It is perhaps reassuring that MP performance does not appear to be very sensitive to the spatial assumption and tag options among these grids, but neither option is very compelling.

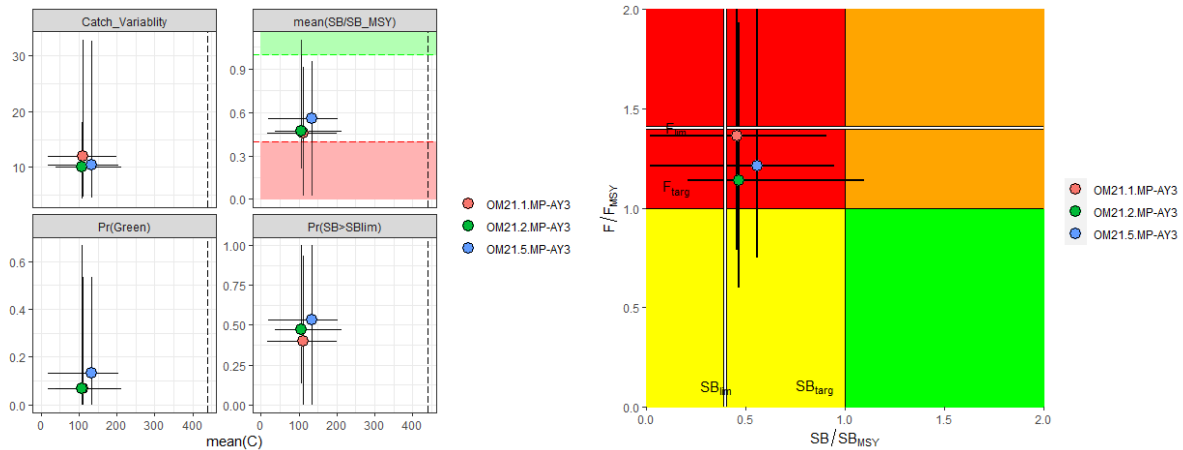


Figure 19. Comparison of MP results for 3 different OMs with contrasting spatial and tagging options. Fifteen year (2021-2035) MP performance summary plots for the same MP when applied to the suite of candidate OMs in Table 3. MP is the PTRE projection-based MP tuned for 50% recovery to BMSY in 2034 for OMgridY21.5.

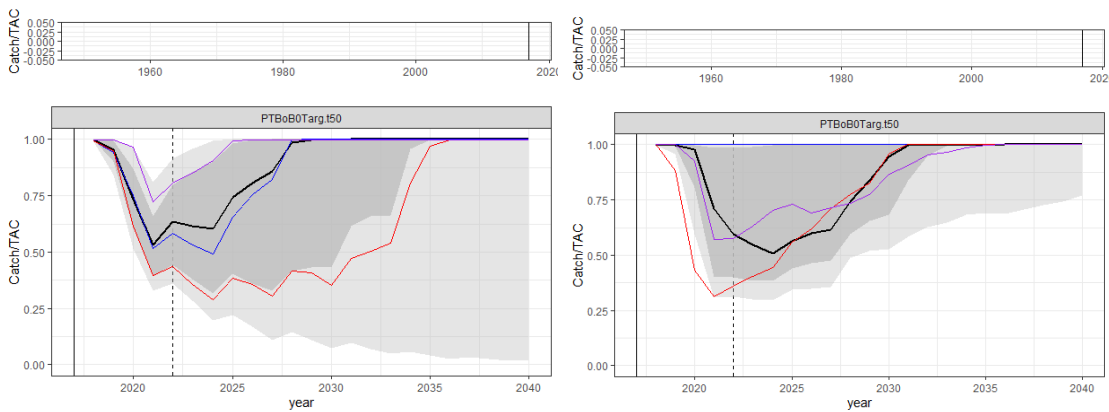


Figure 20. Time series of the ratio of observed catch over TAC, i.e. indicates the proportion of simulations in which the TAC cannot be removed due to a shortage of fish in the right place at the right time. Left panel = 4 Area (OMgridY21.1), right panel = 2 areas (OMgridY21.2). The broken vertical line 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 3 thin coloured lines represent examples of individual realizations to illustrate that individual variability greatly exceeds the median.

4.2 Does the time blocking of recruitment spatial distribution parameters affect MP evaluation results?

Figure 21 compares the example MP performance for 4 OMs, with 2 different options for treating the spatial distribution of recruitment, across 2 spatial options. The results suggest that estimating an additional spatial recruitment deviation parameter (for the most recent 10 years) to account for non-stationarity in this process does not have an effect that would likely make any difference for MP performance and selection, at least in the current context (and certainly less than the spatial structure option). The reported significance of this factor in the recent stock assessment investigations might indicate an important difference between SS projections and the OM, perhaps related to an interaction with environmental-linked movement in the assessment, or the SS projection equations.

Unfortunately, a possible problem was identified in the 2 time-block recruitment spatial distribution SS configuration, too late to investigate thoroughly. The individual SS grid models examined indicated that the time block deviation parameters were estimated suspiciously close to 0, despite relaxed bounds and a prior $\sigma = 0.5$. We would have expected larger deviations (given the apparent trends in Figure 14). Thus we are reluctant to conclude that the non-stationarity of spatial recruitment deviations can be dismissed at this time, but we are confident that it is not the highest priority OM problem.

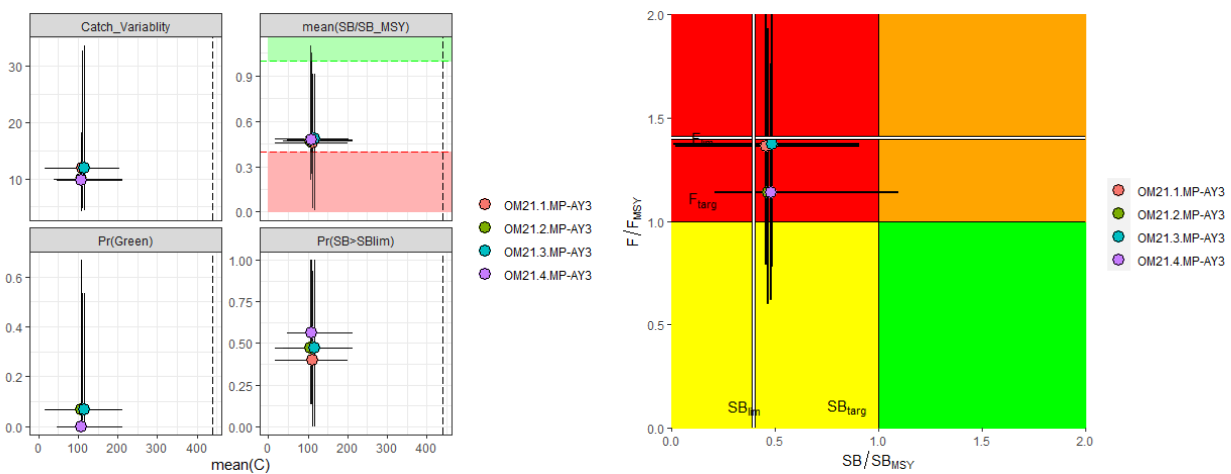


Figure 21. Comparison of MP results for 4 different OMs with contrasting recruitment spatial deviation time series options (crossed with the 2 and 4 area spatial options). Fifteen year (2021-2035) MP performance summary plots for the same MP when applied to the suite of candidate OMs in Table 3. MP is the PTRE projection-based MP tuned for 50% recovery to BMSY in 2034 for OMgridY21.5.

4.3 What effect does the inclusion of 2 extra years of catch data have on the OM retrospective problem?

The current iteration of the OM was conditioned using data that is now 2+ years out of date. Most projections are unable to extract the TAC in the first few years, which include catches that have actually been reported subsequent to the data used in conditioning. This clearly demonstrates a problem, presumably linked to the general retrospective issue, and probably a useful diagnostic for filtering out implausible models. When tested against an OM that was conditioned with two additional years of recent catch data (OMgridY21.6), the result was slightly more optimistic, as would be expected (Figure 16 - Figure 18). The proportion of simulations that fail to extract the TAC and bridging catches are appreciably reduced (Figure 22), but the issue still affects >90% of realizations.

The new data might reduce the recent pessimistic bias in the conditioning, as the model is forced to try to sustain higher recent catches. But it is not completely successful, and unfortunately, there is no reason to expect that this would solve a long term retrospective problem caused by systematic structural problems.

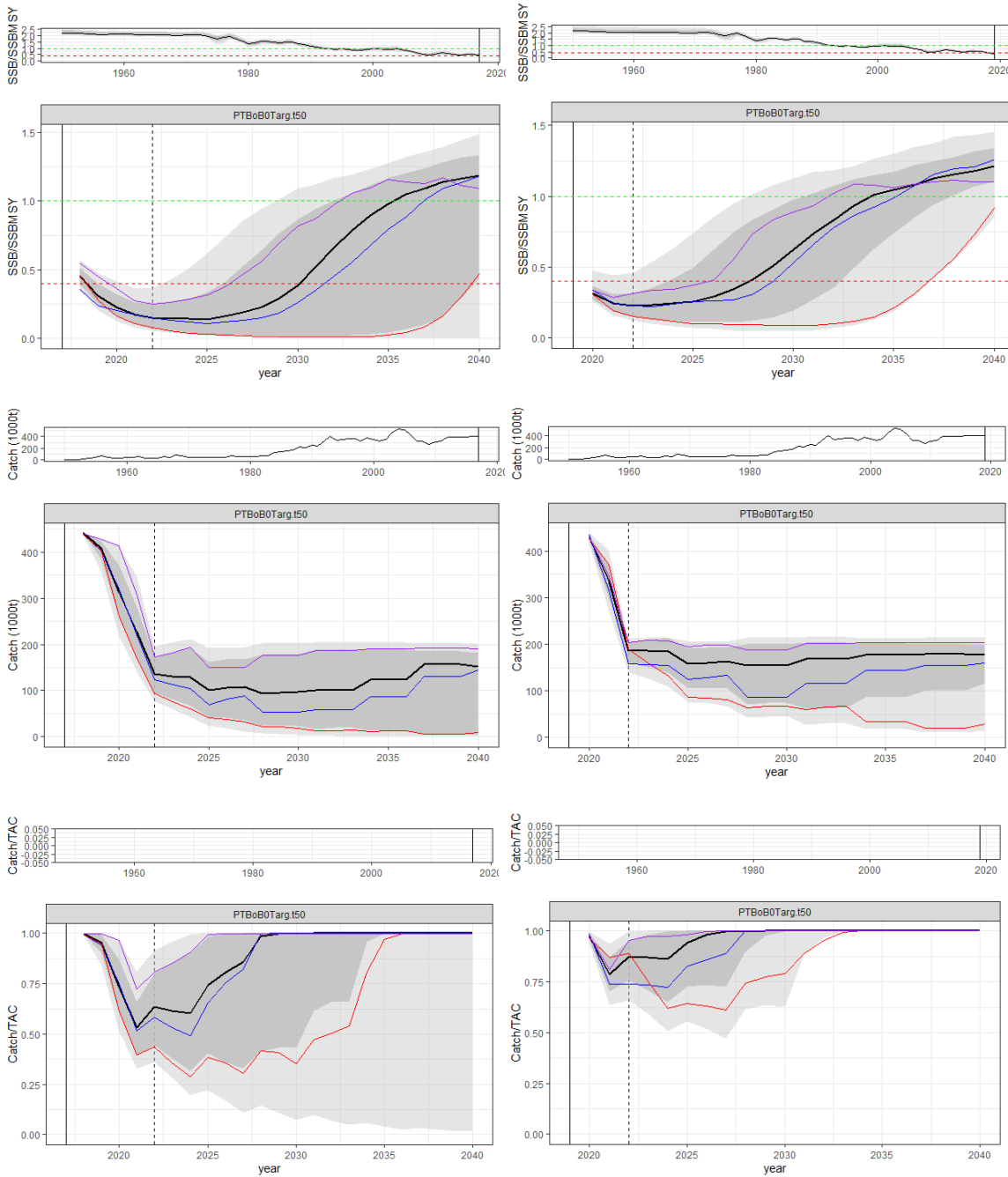


Figure 22. Comparison of MP evaluation time series for the 4 area grid requested by WPTT 2020 (left column), and the equivalent grid with two additional years of recent catch (approximately) included in the conditioning (right column). The top section of each panel represents the historical estimates from the OM, and lower plots represent the projection period. The solid vertical line represents the last year of data 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 3 thin coloured lines represent examples of individual realizations to illustrate that individual variability greatly exceeds the median.

5. How sensitive are the MP evaluation results to “minor” OM projection assumptions related to spatial distributions but not initial status or productivity?

At several points in the MSE process, we have noted that there are some spatial assumptions which we are not particularly satisfied with, and which might have consequences for the MP evaluation and selection. Our interpretation, seemingly with the tacit endorsement of the MSE Task Force, has been that i) the best available assumptions and parameterizations for spatial processes will be derived from the stock synthesis assessment models, and ii) Issues related to high fishing mortality (e.g. sensitivity to whether the Baranov equations or Pope’s approximation are adopted, and what happens if there are not enough fish to remove the TAC, etc) should not be a real problem for any sensible MP that can maintain the stock near appropriate biomass targets. We expect this remains the case, unfortunately, the current state of the OM makes it important to revisit the issue.

Table 4 defines 6 OMs derived from OMgridY21.5 (plus the 2 area and 4 area reference set OMs proposed by the WPTT, subject to alternative projection assumptions). The first 6 OMs all used the same conditioning, hence the initial stock status and productivity at the start of the projections are identical. They differ only in terms of the fishery equations and fish movement assumptions:

- “Pope” indicates that a modified form of Pope’s approximation to the catch equation was used as implemented in the R-based projection code. Additionally, 4 equal quarterly TACs are removed, with each season evaluated independently.
- “ECx” indicates that the Baranov catch equations were used as implemented in the C++ projection code, with an “Effort Ceiling” of x. This C++ code should be essentially the same as the SS assumptions used in OM conditioning. In the projections, the TAC extraction is solved for all 4 quarters simultaneously, such that if there are not enough fish in one season, this can be counter-balanced by extracting more fish in other quarters. Solving the catch equation involves solving for an annual effective effort multiplier for each fleet. This co-efficient re-scales the seasonal pattern of F_s estimated in the most recent period (2016-2017 in this case) in each SS model, until the allocated quota for each fleet can be removed. The effort ceiling is a cap on the co-efficient, i.e. where 20 means that the effective F cannot exceed 20X the value estimated from the recent historical period.
- “MU” indicates that movement parameters were altered in the projections, such that all ages redistribute uniformly among regions every quarter. This is intended to approximate a spatially-aggregated context, i.e. the effect of potential spatial refugia is minimized. OMs without the “MU” designation simply retained the movement parameter estimates from the individual SS models.

As in the previous section, the same tuned MP was applied to each of these OMs. A comparison of the MP performance for this set of OMs shows a large range of behaviour (Figure 23). As shown in

Figure 24 - Figure 26, the MP behaviour is qualitatively similar in most cases (only 4 shown), but substantially different for OMgridY21.5MU.EC20. The difference in performance is initiated at the very beginning of the projections, in which the fisheries tend to struggle to remove the bridging catches, before the MP is active. OMgridY21.5MU.EC20 (which approximates a spatially-aggregated model with unconstrained effort) initially extracts a slightly higher proportion of the bridging catches, and more than 50% of the realizations are driven into a collapse which the MP cannot prevent. Figure 27 illustrates that the problem of removing the TAC is shared among most fisheries (for OMgridY21.5EC2).

It appears that seemingly minor assumptions about how the catches are extracted can have big implications for the MP evaluation performance, and presumably how the MP would be selected. These differences appear greater than the differences among the structurally diverse OMs shown in Figure 17 - Figure 19. If the OMs provide a realistic representation of the stock status, further consideration may be required as to what would really happen to the fisheries as the stock declines, i.e. at what level would fleets simply stop fishing, change targeting or move to different areas? However, we expect that the real problem is the inherent pessimism in the current suite of OMs, in that they struggle to remove catches that the fleets are reported to have already extracted, without any dramatic increases of effort (so far as we are aware). When similarly tested, the (more optimistic) bigeye results are much more robust to these assumptions (see companion paper Kolody and Jumppanen 2021).

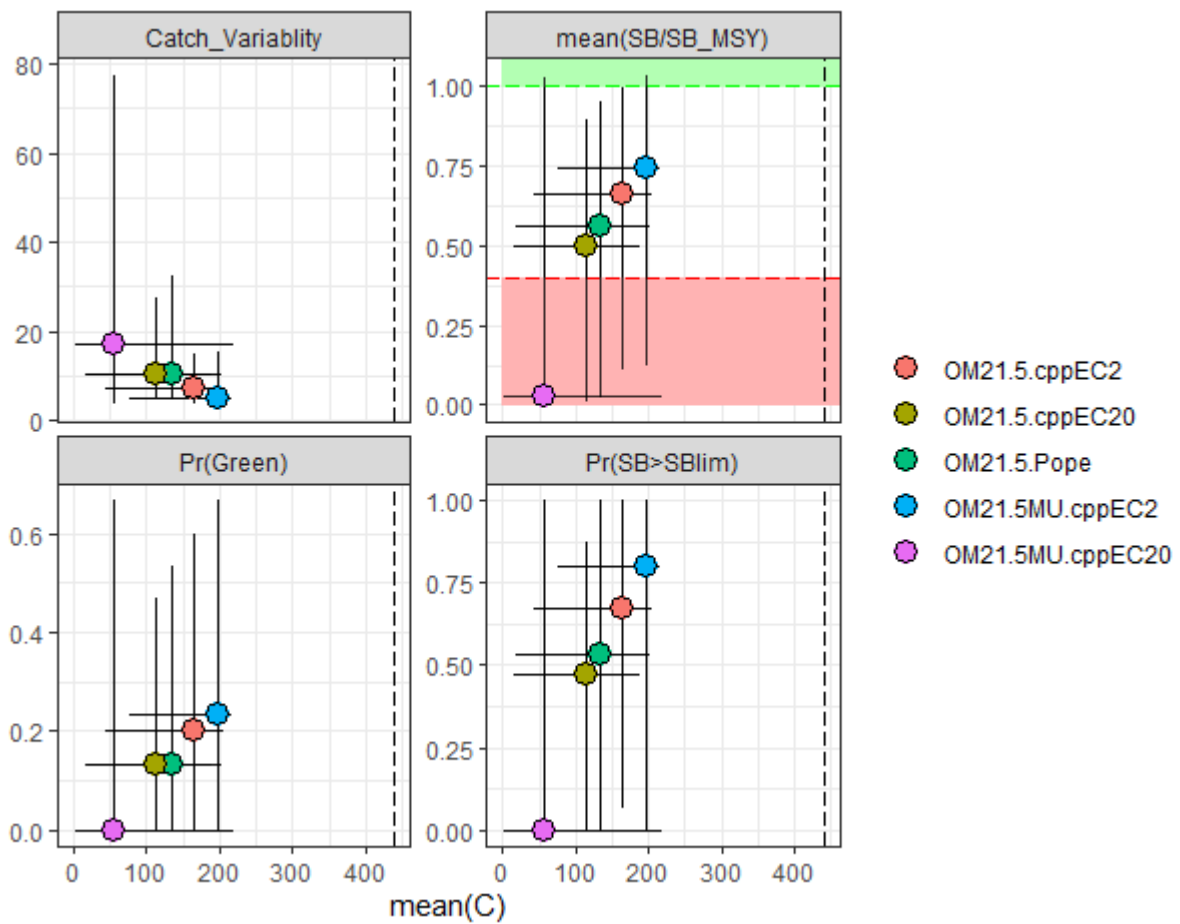
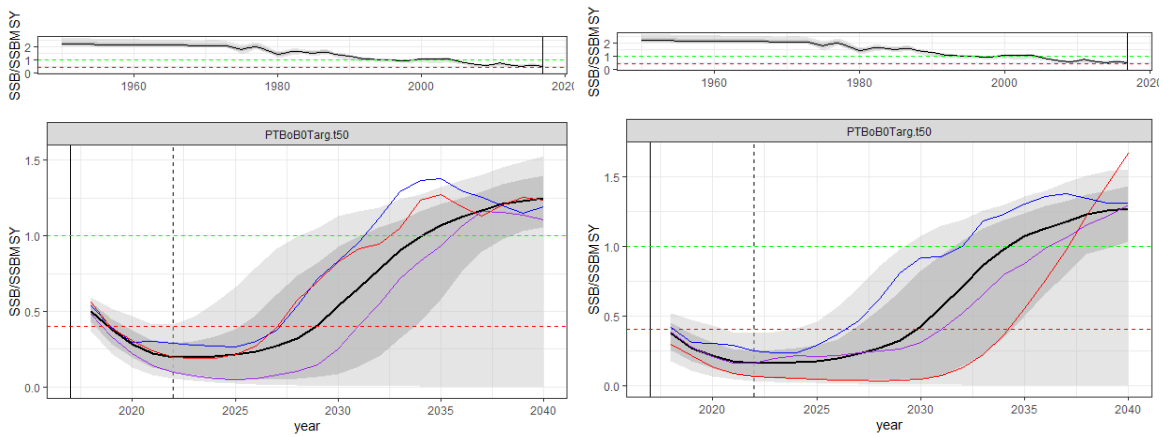


Figure 23. Comparison of MP performance, for the same tuned MP, for a series of OMs which differ only in the catch extraction assumptions and movement rates.

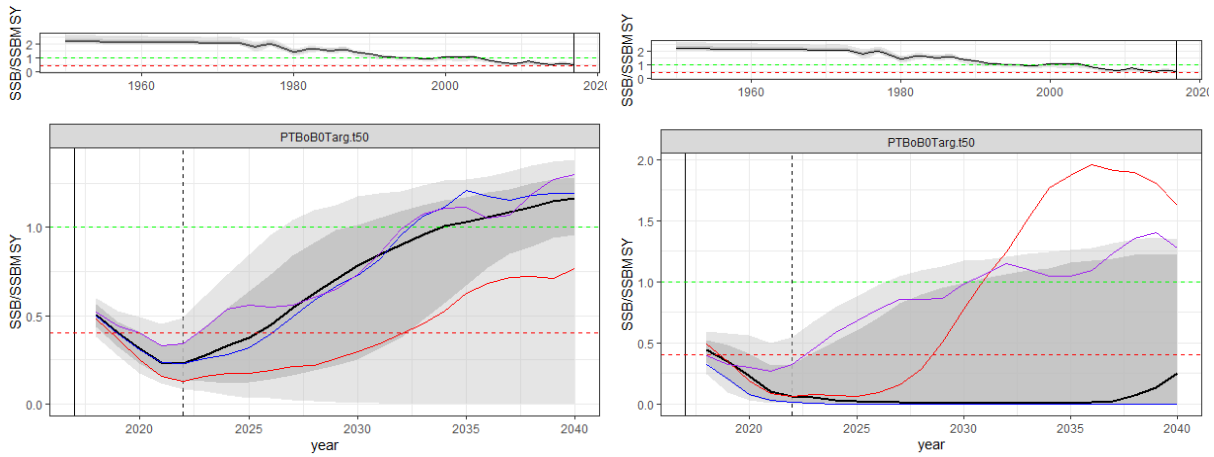
Table 4. Proportion of OM realizations (conditioned with data up to 2017) that can extract the observed 2019 catches, and the assumed bridging catch in 2021, before the first MP implementation. The effort ceiling is the maximum F scalar that can be applied to each fishery, relative to the terminal (seasonal pattern of) Fs estimated in the SS conditioning. SS movement is the estimate from the SS conditioning. Uniform movement means that fish of all ages are uniformly redistributed every quarter. The first 6 OMs are derived from the same set of conditioned models (OMgridY21.5), and differ only in the projection assumptions. The final 2 OMs correspond to the 2 area and 4 area reference set OMs proposed by the WPTT (Appendix A), except with the effort ceiling of 2 applied in projections.

OM	F Method (effort ceiling)	Movement	Percent realizations remove >95% observed catch 2019	Percent realizations remove >95% bridging catch in 2021
OMgridY21.5.Pope	Pope	SS	65	5
OMgridY21.5.EC20	Baranov (20)	SS	41	5
OMgridY21.5MU.Pope	Pope	Uniform	57	25
OMgridY21.5MU.EC20	Baranov (20)	Uniform	86	53
OMgridY21.5.EC2	Baranov (2)	SS	20	2
OMgridY21.5MU.EC2	Baranov (2)	Uniform	44	24
OMgridY21.1.EC2 <i>(reference set 4 area)</i>	Baranov (2)	SS	10	1
OMgridY21.2.EC2 <i>(reference set 2 area)</i>	Baranov (2)	SS	59	20



OMgridY21.5.Pope

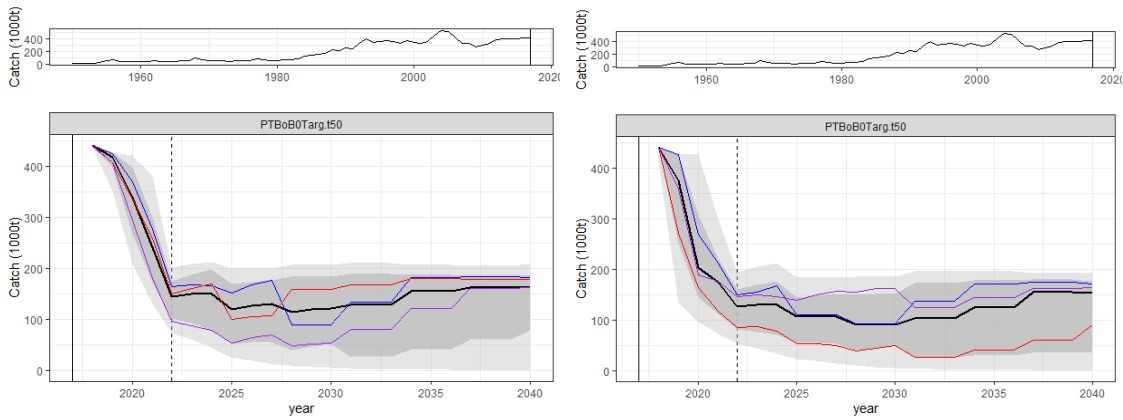
OMgridY21.5.EC20



OMgridY21.5MU.Pope

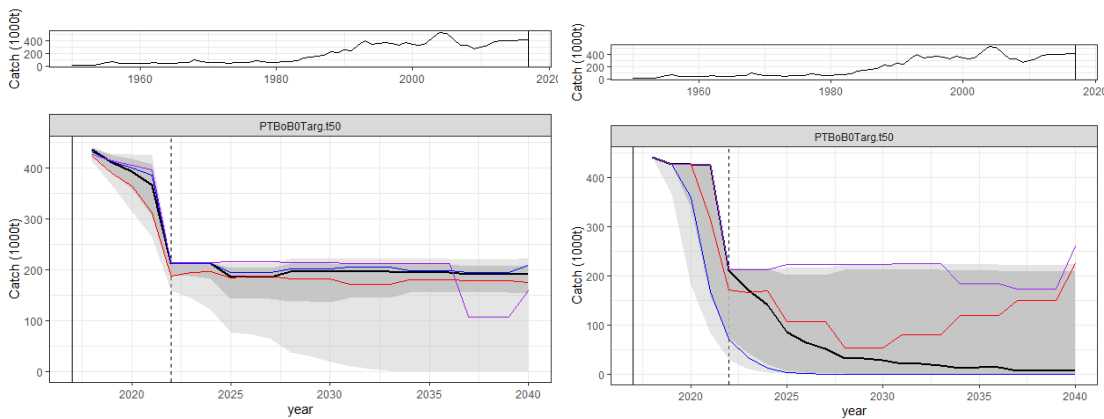
OMgridY21.5MU.EC20

Figure 24. Biomass trajectories for the tuned baseline MP, from 4 OMs representing the same production dynamics, but different movement and fishing mortality assumptions. . The top section of each panel represents the historical estimates from the OM, and lower plots represent the projection period. The solid vertical line represents the last year of data 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 3 thin coloured lines represent examples of individual realizations to illustrate that individual variability greatly exceeds the median.



OMgridY21.1

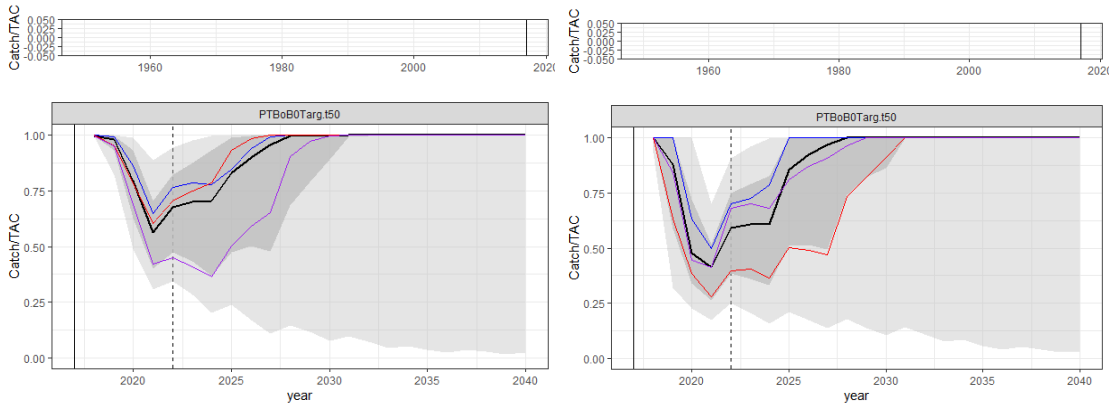
OMgridY21.1cpp



OMgridY21.1MU

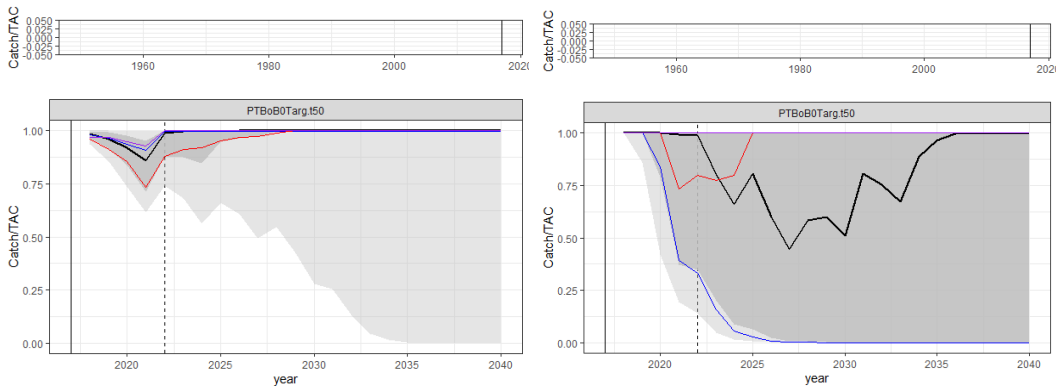
OMgridY21.1MUcpp

Figure 25. Catch trajectories for the tuned baseline MP, from 4 OMs representing the same production dynamics, but different movement and fishing mortality assumptions. . The top section of each panel represents the historical estimates from the OM, and lower plots represent the projection period. The solid vertical line represents the last year of data 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 3 thin coloured lines represent examples of individual realizations to illustrate that individual variability greatly exceeds the median.



OMgridY21.1

OMgridY21.1cpp

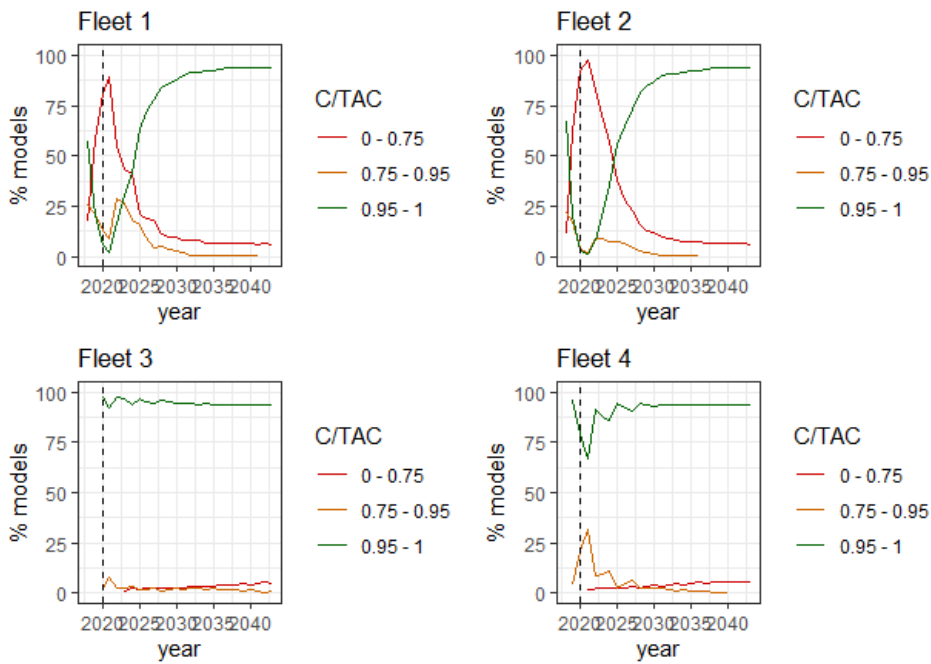


OMgridY21.1MU

OMgridY21.1MUcpp

Figure 26. Catch/TAC ratio time series for the tuned baseline MP, from 4 OMs representing the same production dynamics, but different movement and fishing mortality assumptions. . The top section of each panel represents the historical estimates from the OM, and lower plots represent the projection period. The solid vertical line represents the last year of data 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 3 thin coloured lines represent examples of individual realizations to illustrate that individual variability greatly exceeds the median.

C / TAC for PTBoB0Targ.t50



C / TAC for PTBoB0Targ.t50

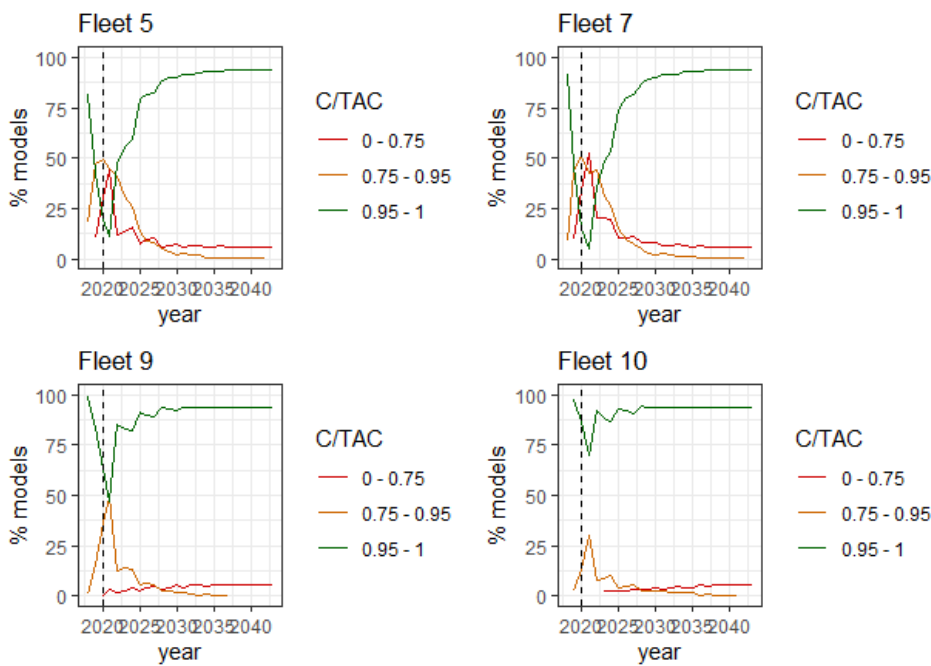
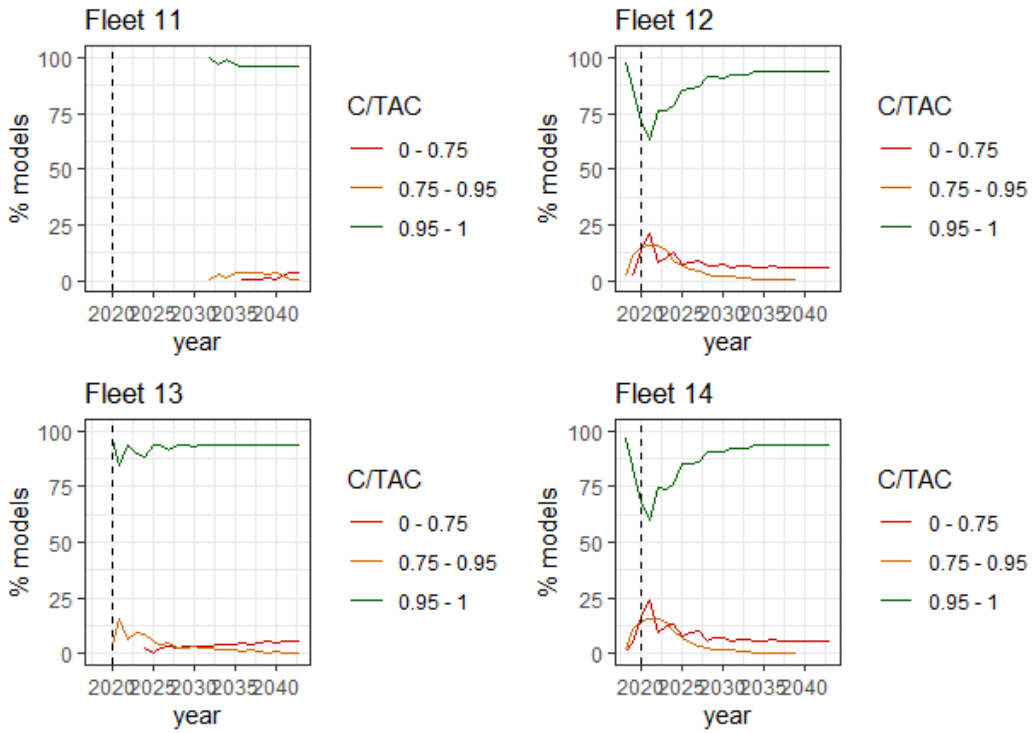


Figure 27. Catch/TAC time series by fishery for OMgridY21.5.EC2. (continued on next page)

C / TAC for PTBoB0Targ.t50



C / TAC for PTBoB0Targ.t50

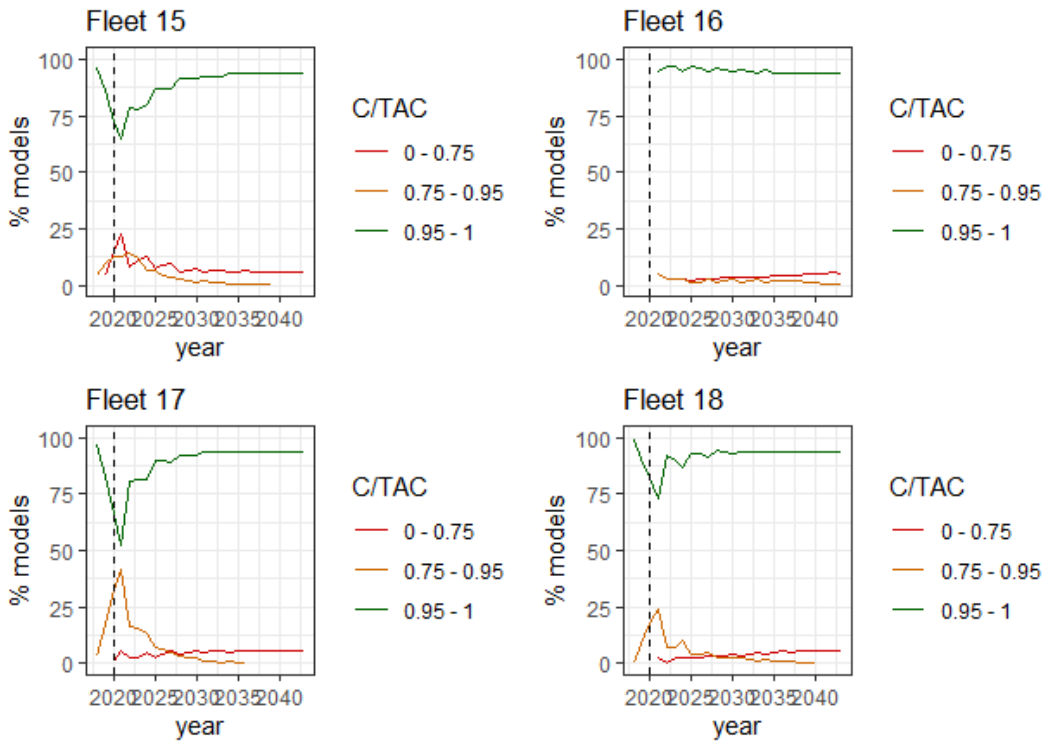


Figure 27 (continued)

C / TAC for PTBoB0Targ.t50

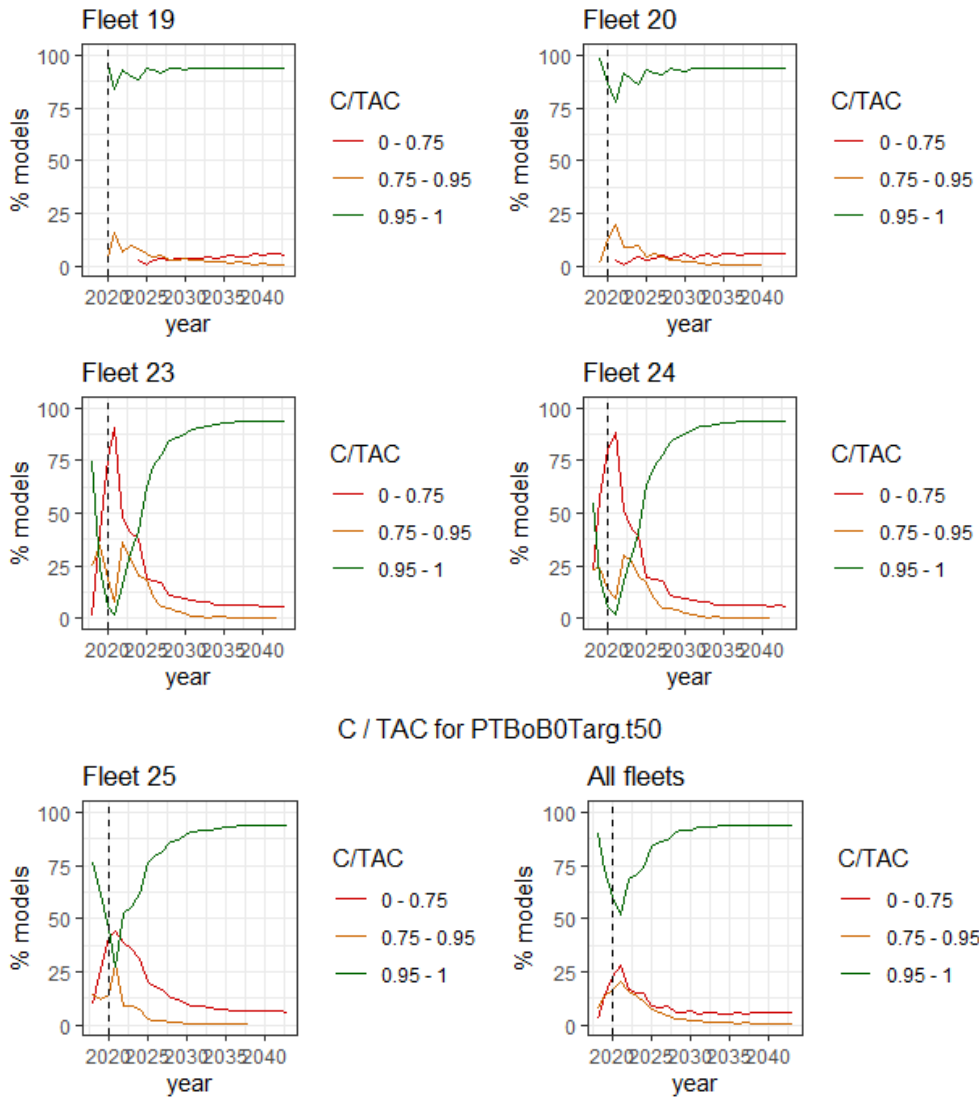


Figure 27 (continued)

5.1 OM diagnostics – the role of catch and effort?

There has been a movement within the IOTC to adopt a standard suite of model diagnostics (e.g. Matsumoto et al. 2018) for the stock assessments to ensure that implausible models are not influencing the management advice inappropriately. Those proposed include:

- runs tests – assessment models tend to assume that data errors are independent across time. If there are systematic patterns in the deviations between model predictions and observations, this suggests that there is either a systematic structural problem in the model, or a more complicated time series error structure should be assumed in the likelihood functions.
- retrospective analyses identify whether the addition of new data tends to identify systematic biases in the historical estimates. i.e. data to year T should be more informative about stock status in year T , than data to year $T + x$. If inferences at time T given data to time T are more pessimistic than inferences at time T given data to time $T + x$, for a long period of different T , then there is reasons to expect that this bias will continue to be a problem.
- hindcast prediction – related to retrospectives, if you truncate the data at T , fit the model and predict the data for $T + x$, how does this compare with the observed data at $T + x$? A substantial deviation might be indicative of a biased or overparameterized model.

While these diagnostics are useful in principle, the best way to use them remains a work in progress for a few reasons: i) at a technical level, the available software is not easily applicable to models with the SS real-seasons-defined-as-model-years configuration currently used for YFT (and BET), ii) the retrospective calculations are very slow for a large grid of complicated models), and iii) there is a lot of subjectivity and arbitrariness about how to evaluate the importance of these diagnostics, how to combine them, and how to apply them to the elements of a grid (e.g. retention/rejection, weighting, etc).

It is also not clear how the role of diagnostics should differ in stock assessment grids and OMs, given that OMs are intended to encompass a broader range of uncertainty and ensure that MPs are robust to more diverse circumstances than a stock assessment would normally describe. e.g. If the CPUE working group is convinced that an effort creep trend of 1% per year is plausible and important to entertain in the OM, it is not clear how to address potential conflicts, e.g. i) Is it appropriate to simply dismiss all models with a 1% effort creep on the basis of diagnostics, ii) is it acceptable to keep the models despite the diagnostic failure, or iii) should further effort be invested in model development to ensure that the OM is re-structured in such a way that the 1% effort creep scenario can be retained and pass the diagnostic test?

In the OM development to date, we have primarily looked at:

- indices of quality of fit between model predictions and observations (e.g. CPUE RMSE, auto-correlation, size composition post-fit Effective Sample Size). In the past, a small number of models were removed from a grid if there was clear evidence of outliers. This was not relevant in the OMs discussed this iteration.

- Systematic deviations from the stock recruitment relationship (notably trends that appear to explain biomass decline as a function of recruitment decline rather than fishery impact). This has been problematic in the past, but not an obvious concern in this iteration.
- Catch-likelihood indicating that the observed catch cannot be removed in at least one age/region/time strata. This has been used in the past, but not consistently. In some cases it was evident that there was a shortage of fish in a particular historical quarter, which might be irrelevant for the overall model dynamics. However, if the observed catch cannot be removed near the end of the time series, this may be an indication of something more serious in the model, and the model is likely going to start the projections from a state that is too pessimistic.

Given that the data for the current OM end in 2017, but there are actually new catch data reported for 2018 and 2019, (that are included in the bridging catches of the OM projections. i.e. spanning the years between the last assessment year and the first MP implementation), this provides a concrete diagnostic in the spirit of hindcast projection. If the OMs cannot remove the catches observed in 2018 and 2019, there is clearly a problem. Furthermore, if the models cannot remove the subsequent bridging catches (that have been assumed but not yet observed), either there is a problem with the conditioning or the bridging catch assumption.

Additionally, the effort required to remove the catches might be indicative of a problem. i.e. by default, in removing the catch, the OM C++ projection sub-routine increases the fishing mortality in proportion to the recent historical pattern estimated to have occurred seasonally, with a default scaling limit of 20. i.e. If the effort ceiling has been hit, this implies that the effective fishing effort of the relevant fishery is 20X higher than it was during the recent past. This does not seem very credible, at least for any of the major fisheries (though the link between F and effort is unclear particularly when fish aggregate). If this was to be used as a diagnostic criteria, a more appropriate cap for the effort increase in two years is presumably <2 for any substantive fleet.

Table 4 also describes the percentage of MSE realizations that can attain $>95\%$ of the catch observed in 2019 (and bridging catch assumed for 2021) for the reference MP for a series of OMs. If we adopted the attainment of 2019 catch (from Table 4) as some sort of post-MSE diagnostic filter, the number of retained realizations ranges from 10% (for an effort ceiling of 2, with the SS estimates of fish movement) to 86% (for an effort ceiling of 20, and a uniform redistribution of fish every quarter). The most restrictive filtering would reject 90% of the (most pessimistic) realizations from the original grid, and is certainly more optimistic than the base set. But, as also indicated in Table 4, $>90\%$ of those remaining runs are still not able to extract the current catches assumed in the bridging period to 2021. Effectively the MSE results would be reduced to $\sim 1\%$ of the original OM grid (and even those retained realizations might be the result of fortuitous stochastic recruitment rather than configuration assumptions).

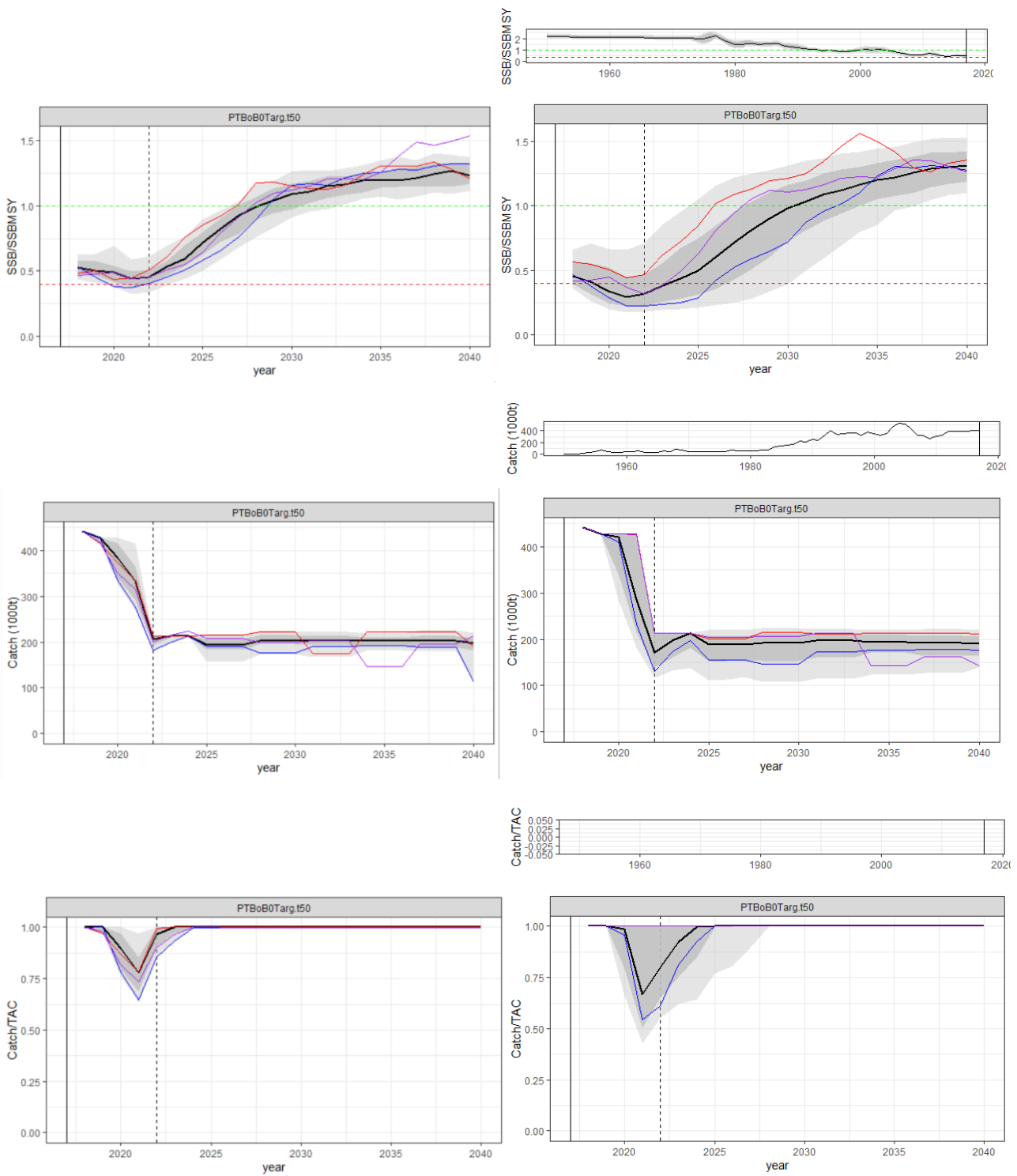
If we apply this same diagnostic filter (Catch in 2019 and effort ceiling of 2) to the original 4 area (and 2 area) reference set OMs proposed by the WPTT (Appendix A), the dynamics are shown in Figure 28. The MP would prescribe similar large initial quota reductions as in the default tuning case, but rebuilding to BMSY would be brought forward from 2034 to 2028 in the 4 area case (2030 in the 2 area case). But the majority of OMs (90% for the 4 area OM and 67% for the 2 area OM) would still struggle to remove the bridging catches assumed in 2021.

This sort of filtering diagnostic seems to be very useful for helping to identify the plausible OM space in this case. The diagnostics that have been proposed for the assessment have a similar intent, but there is perhaps a key difference here worth considering. The assessment diagnostics are all geared toward improving internal model consistency, but will not necessarily help if major assumptions are wrong (e.g. identifying a model that fits the CPUE series without any systematic lack of fit is not very helpful if the assumption of a proportional relationship between CPUE and abundance is fundamentally wrong to begin with). In this case, we were able to use a diagnostic based on data that were external to the model (2019 catch), and external arguments about how effort is likely to be constrained, to conclude that we have good reason to be sceptical of the whole ensemble. Unless there is evidence of fisheries failing to attain recent catches in 2020 and 2021, or there are massive recent effort increases, these OMs just do not seem credible.

If we were to employ the diagnostic criterion above (max. effort multiplier = 2, $C(2019 \text{ predicted}) > 0.95 C(2019 \text{ observed})$), the retained OM grid elements shown in Figure 29 would result. 97% of the models included the down-weighted CPUE option and 77% of the models included the 0% per year catchability trend option, which both support the notion that there is some inherent pessimism in the CPUE series that is not compatible with the recently observed catches, at least in the context of the other structural assumptions. Only 6% of these models included the lowest (M06) option (which seems to have the most support from tagging and recent ageing studies in the Atlantic), while 64% of the options were associated with the highest M option (M10). The pattern is similar (Figure 30) when the candidate 4 area reference set OM (OMgridY21.1.EC2) is constrained with an effort ceiling of 2. With the default 2 area model effort ceiling 2 (OMgridY21.2.EC2), the pattern is qualitatively similar, but the rejection rate is less extreme (Figure 31).

It is perhaps worth highlighting that these complications make it difficult to reach a simple conclusion about which of the 4 area or 2 area OMs is more optimistic. The 4 area OM might appear to be in a more pessimistic state initially, but not necessarily if the diagnostic filtering is applied. Furthermore, the 4 area structure might create spatial refuges that might lead to more optimistic MP evaluation outcomes. We may need to be careful in considering whether this is realistic.

At this time, we do not think it makes sense to proceed with any of the current YFT OMs under consideration for the purpose of providing MP advice to the 2021 TCMP. We propose the next OM update should be undertaken in conjunction with the 2021 YFT assessment update, which will include the latest data and considerable collaborative development on multiple fronts. We would encourage that group to consider carefully whether there is any avenue of investigation that might cast a very different perspective on the assessment. If the MSE Task Force considers it essential to produce interim YFT results for the TCMP 2021, we would propose something along the lines of Appendix B.



4 Area reference set OM

2 Area reference set OM

Figure 28. MP performance time series for OMgridY21.1, with all realizations removed that could not extract >95% of the observed 2019 catch with an Effort Ceiling of double the 2017 level. . The top section of each panel represents the historical estimates from the OM, and lower plots represent the projection period. The solid vertical line represents the last year of data 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 3 thin coloured lines represent examples of individual realizations to illustrate that individual variability greatly exceeds the median.

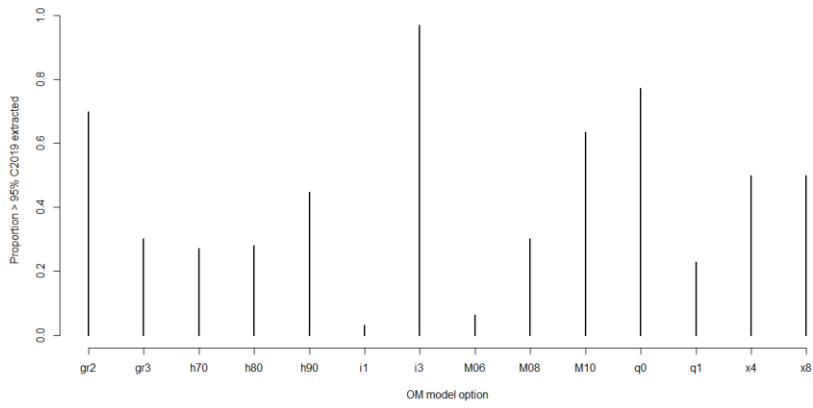


Figure 29. Distribution of model options retained in OMgridY21.5cppEC2, if a criterion was applied that required 95% of the observed catch in 2019 to be extracted with an effort increase of no more than double for each fishery in the period from 2017-2019.

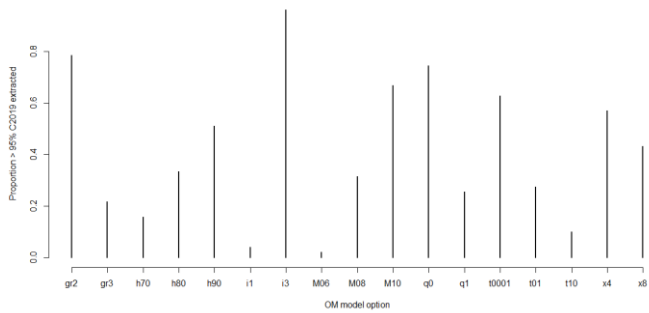


Figure 30. Distribution of model options retained in OMgridY21.1cppEC2, if a criterion was applied that required 95% of the observed catch in 2019 to be extracted with an effort increase of no more than double for each fishery in the period from 2017-2019. (Similar to figure above, except this is the default 4 Area OM grid configuration requested by the WPTT 2020)

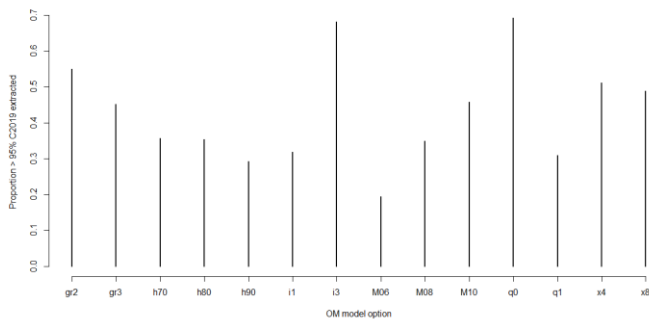


Figure 31. Distribution of model options retained in OMgridY21.2cppEC2, if a criterion was applied that required 95% of the observed catch in 2019 to be extracted with an effort increase of no more than double for each fishery in the period from 2017-2019. (Similar to figures above, except this is the default 2 Area grid configuration requested by the WPTT 2020)

6. Conclusions and Recommendations

- 1) At this time, we find that all of the YFT OM ensembles entertained appear to be unrealistically pessimistic, as indicated by the difficulty in extracting recently observed catches. In some cases (for many of the 4 area models), this may also be evident in the catch likelihood, that (usually) indicates some element of the observed catch could not be removed during conditioning (and which we have used as a model rejection criterion in the past). But more importantly, most models were unable to extract the recently observed catches that were reported subsequent to the data adopted for conditioning. This is consistent with earlier observations of the persistent retrospective pattern, in which newer data tends to suggest that the preceding assessment would have been too pessimistic to explain subsequent observations.
- 2) The pervasiveness of the problem across model assumptions leads us to suspect that there may be something fundamentally misleading across all OMs considered (e.g. perhaps a fundamental bias in the reported catches, or misinterpretation of the relationship between longline CPUE and abundance). The problem can be reduced by conditioning with updated data, but we suspect that this might be simply kicking the problem down the road a couple years.
- 3) Given the high fishing pressure that the OMs must exert to remove initial catches, the importance of spatial assumptions (movement, and how fleets would react) has the potential to be very influential to MP performance. We are not confident that any of the OMs are currently capable of representing these processes reliably. However, we would also not expect these sensitivities to be nearly as important for situations in which the stock was less depleted.
- 4) It is somewhat reassuring that the MP performance seemed to be relatively robust to most of the OM specification options that we investigated, i.e. 2 vs 4 areas, time-blocking of spatial recruitment deviations, but since all of the OM ensembles under consideration seem to be badly flawed, we are not convinced that this is a general conclusion.
- 5) The role of diagnostics in weighting the elements of the OM ensembles should be given further consideration going forward. The suite of diagnostics proposed for the assessments seems like a reasonable starting point, but we are concerned that they are primarily focused on issues of internal consistency (i.e. internal consistency is certainly a desirable model feature, but not sufficient if the model assumptions are fundamentally wrong). The diagnostics that we considered here (i.e. Can we explain the observed catches subsequent to the data used in conditioning, and does it make sense to allow effort to increase without limit?) were considerations external to the model.
- 6) It is perhaps notable that the 2 area OM configuration tends to be slightly more optimistic than the 4 area configuration. However, if the OM is subset according to the requirement of removing 95% of the 2019 catch, the 4 area OM is more optimistic. This presumably arises for two reasons: i) the 4 area OM is more heavily filtered (because it is initially more

pessimistic due to the underlying production dynamics), and ii) the 4 area models have larger refugia (space/time structure where fish might not be vulnerable to the fleet that has unfilled quota).

- 7) We recommend against reporting YFT MP evaluation results to the TCMP for management objective feedback, until after the known problems in the stock assessment can be further considered by WPTT (2021), and reconciled with the operating models.
- 8) If the MSE Task Force concludes that it would be productive to provide MP evaluation advice to the TCMP 2021, a proposal for a stopgap OM is included in Appendix B.
- 9) We would strongly encourage strategic investigations by the broader IOTC community, that might help to identify fundamentally new insights into the YFT population dynamics. Close-Kin Mark Recapture is the most promising tool that we can think of. The role of acoustic FADs needs serious consideration. The practical relevance (or not) of the recent stock structure study should be formally established.

References

- Dortel, E, Sardenne, F, Bousquet, N, Rivot, E, Million, J, Le Croizier, G, Chassot, E. 2014. An integrated Bayesian modeling approach for the growth of Indian Ocean yellowfin tuna. *Fisheries Research*. 163. 10.1016/j.fishres.2014.07.006.
- Fu, D, Langley, A, Merino, G, Ijurco, AU. 2018. Preliminary Indian Ocean Yellowfin Tuna Stock Assessment 1950-2017 (Stock Synthesis). IOTC–2018–WPTT20–33.
- Kolody, D, Jumppanen, P. 2021. Indian Ocean Bigeye Tuna Management Procedure Evaluation Update March 2021. Working Paper prepared for the Management Strategy Evaluation Task Force of the Indian Ocean Tuna Commission Working Party on Methods Meeting, March 2021. IOTC-2021-WPM12(MSE)-04.
- 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, RD, Wetzel, CR, 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. *Fish. Res.* 142: 86– 99.
- Urtizbera, A, Cardinale, M, Winker, H, Methot, R, Fu, D, Kitakado T, Fernández, C, Merino, G. 2020. Towards providing scientific advice for Indian Ocean yellowfin in 2020. IOTC-2020-WPTT22(AS)-21.
- WPM 2020. Report of the 11th Session of the IOTC Working Party on Methods. Microsoft Teams Online, 14 -15 October 2020. IOTC–2020–WPM11–R[E].
- WPTT 2020. Report of the 21st Session of the IOTC Working Party on Tropical Tunas. Virtual Meeting, 19-23October2020. IOTC–2020–WPTT22(AS)–R[E]_Rev1.

CONTACT US

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

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>

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

Appendix A. Extracts from the 2020 Methods and Tropical Tuna Working Party reports relevant to yellowfin MSE Technical workplan

Working Party on Methods

WPM 2020 deferred explicit workplan requests to the WPTT, while the following points offer some guidance for problematic and subjective decisions:

37. The WPM NOTED the progress made in 2020. Also, the WPM DISCUSSED the use of some of the diagnostic tools provided in the SSdisag R package (Winker et al. 2019) for the screening of OMs. In particular, the WPM NOTED that the 'run' tests are non-parametric tests to test for non-random distribution of residuals (sequence positive/negative), but they do not provide a measure of the scale of errors. Therefore, the runs tests are useful to detect misspecifications in the observation process of models (Carvalho et al. 2016), but may not be suitable for examining how well the model "fits" are. The WPM further NOTED the metrics such as Mean Squared Errors can measure the scale of the errors, but they are influenced by the fixed sample size or assumed observation error and, therefore, may not necessarily work well.
38. With regards to other diagnostics, the authors clarified that "jittering" can also be added to the list of diagnostics in future model configurations. In this regard, the WPM NOTED that the Jittering analysis is important diagnostics to ensure model stability and convergency, but it would not be feasible to perform it all the models in the OM. Jittering is probably best placed for a representative subset of the final candidate models.
39. The WPM NOTED that there is a need for a common definition of "plausibility" to separate models that are "plausible" for the grid of OMs and those that are "not plausible", that could be discarded or moved to a robustness grid.
40. The WPM NOTED that autocorrection is not accounted for when fitting the CPUE indices in the OM, but has been included when generating simulate future abundance observations.
42. The WPM NOTED the problematic retrospective pattern observed in the reference model and discussed some of the possible causes. The WPM NOTED that one mechanism for introducing this retrospective pattern might be a non-linearity between CPUE and abundance, but the investigation conducted so far has not been very conclusive. The WPM further noted the retrospective pattern are mainly in the scale of the biomass, but not in the trend, and thus queried whether this might be related to changes in dome-shaped fishery selectivity.
43. The WPM NOTED that the Stock Synthesis (SS) maximum fishing mortality setting used in the assessment ($F_{max} = 2.9$) results in a "non-trivial" catch likelihood term for the majority of OM specifications. Raising the max F constraint to 6.0 allows the majority of models to avoid this problem. However, in either cases, a large number of models have produced a very high exploitation rate (>95%), which seems not realistic. The WPM NOTED additional approaches to further improve model stability and behaviour include exploring the option of estimating fishing mortality F as a random effects vector which are constrained by AR1-type penalty on the annual deviation of F, with a vague, lognormal prior on the unfished biomass.

Working Party on Tropical Tunas

WPTT (2020) made the following comments and requests with respect to the YFT MSE workplan (with other elements of the work to remain unchanged from previous iterations):

141. The WPTT **NOTED** that a very high fishing mortality was estimated for some age/region/quarter strata in the previous OM iteration, and this is expected to recur. The problem manifests on a continuum, such that there is no obvious criterion for model retention/rejection. It remains unclear the extent to which this represents i) a genuine problem that has serious effects on inferences, ii) a genuine situation with a trivial effect on inferences, or iii) an artefact of misleading labels in Stock Synthesis or r4ss outputs. Additional plausibility constraints will likely need to be applied, perhaps in parallel with new insights from the 2021 YFT assessment
142. The WPTT **NOTED** that different spatial structures (2-area and 1-area models) have been explored as alternatives to the 4-area model in the yellowfin tuna stock assessment (the 1 area model was not retained for management advice). However, the OM for the yellowfin MP development currently retains the 4-area model structure only. The WPTT **AGREED** to retain the 4-area model structure in the OM, noting that the current MSE software would require modification to support multiple spatial structures. The WPTT **AGREED** to explore the potential to include the 2-area model in future OMs, pending the outcome of further comparison of the 2 area and 4 area models to determine whether important inferences and challenges for candidate MPs actually exist, and if so, whether they depend on the spatial structure or other confounding factors (e.g. differing use of tags, restriction of east-west movement and/or interpretation of CPUE regional scaling factors).
144. The WPTT **NOTED** the difficulty in running diagnostics on all models in the OM and the redundancy that arises from having a large number of models in the OM in the centre of the model distribution. The WPTT **NOTED** the value in running diagnostics on the models at the 'corners' of the OM domain, rather than across all models in the OM, to identify the plausibility of different models. This can assist with objective elimination of the least plausible Operating Models based on three quantifiable criteria: (1) fit to the data, (2) model internal consistency and (3) prediction skill.
150. The WPTT **NOTED** that the two robustness scenarios for catch implementation error (10% overcatch reported; 10% overcatch not reported) may not represent likely scenarios and **SUGGESTED** using an additional robustness scenario that includes both 5% overcatch reported and 5% overcatch not reported.

The table of OM reference set options has been included in the main text and Attachment B.

There was an additional comment about the recent yellowfin stock assessment modelling efforts, that may have important implications for the MSE Operating Model, but for which there were no specific requests:

110. The WPTT **NOTED** that following the work with the yellowfin stock assessment model, the analysts encountered a potential problem with the projections that were run in 2018 to build the K2SM that is currently in the YFT Executive Summary. This preliminary presentation indicated that the way in which total **recruitment** is allocated between two of the four areas of the model may be causing the models to crash and therefore potentially producing bias in the probabilities estimated for the K2SM. The WPTT **NOTED** that the group of analysts will continue looking at this issue more carefully and report to the group when they have more definitive conclusions.

Appendix B. State of the IOTC Yellowfin Tuna MSE Operating Models as of March 2021.

State of the IOTC Yellowfin Reference Set Operating Model for Management Procedure evaluation March 2021

Dale Kolody (dale.kolody@csiro.au)

Paavo Jumppanen

CSIRO, Australia

Introduction

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

As discussed in the main text, at this time, we do not consider that there is a satisfactory yellowfin Operating Model, with which to present MP evaluation results to the TCMP 2021. We argue that presenting the next iteration of results should be delayed until the full deliberations of the WPTT are conducted in 2021, in relation to the updated data, and rapidly evolving stock assessment process. This appendix describes what we might propose as the best option, if the MSE Task Force judged that something was urgently required:

- The base 4 area grid (to retain the potential for describing the influence of tags and CPUE disaggregation into tropical and temperate regions)
- Fractional factorial design much larger than 54 models (i.e. to broaden the diversity of the OM by representing more of the interactions among model options)
- Diagnostic filtering of models, including:
 - i. substantial catch likelihood indicating catch removal failure in conditioning
 - ii. A prior weighting on model assumption options similar to that described in the main text, to reduce the chance of failing to remove observed 2018 and 2019 catches with an effort ceiling of relative to the recent period (a level to be determined by the MSE Task Force (e.g. ~1-3))
- Bridging catches for 2020 and 2021 determined by projecting the constant effective effort from 2019. As soon as the reported catches from 2020 become available, this should be incorporated in the conditioning or diagnostics.
- MP implementation with an effort ceiling relative to the recent historical period at a level to be determined by the MSE Task Force (e.g. ~1-3)

This might provide an OM that cannot be immediately dismissed as implausible, but it seems like a superficial plan that does not attempt to address the fundamental retrospective problem that afflicts the stock assessment and OM.

The details below are general characteristics of the YFT OMs under consideration at this time.

Conditioning Software

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

Projection Software

The projection software is available from <https://github.com/pjumpnanen/niMSE-IO-BET-YFT/>. The population dynamics equations conform to fairly standard fisheries stock assessment model assumptions, and are fully documented in the technical reference (also on github). In most respects, the OMs attempt to mirror the equations used in fitting a Stock Synthesis model, and the same fixed and estimated parameters are used for the MSE projections. Some deviations are noted below.

Reference Set OM

The various models considered in the OM ensemble are mostly derived from the reference case stock assessment (supplied by Dan Fu, IOTC secretariat, and defined in Fu et al (2018)), with additional modifications from the interim assessment development work described by Urtizberea et al. 2020). Key assumptions include:

- 4 regions (Figure 1) with age-dependent movement. *The option of 2 areas remains under consideration, in which the population is split into Western and Eastern Areas, with no movement (this is discussed in the main text).*
- Quarterly dynamics (implemented with calendar quarters as SS model-years)
- 25 fisheries (Table 1) - 21 with some temporal variation represented as independent fisheries
- Parameter estimation objective function includes
 - Total catch penalty (active if some component of catch cannot be removed)
 - Standardized longline CPUE (one series per region – constant catchability is assumed following the application of regional scaling factors to establish a relative abundance linked among regions). *Regionally-scaled indices from north and south are merged in the 2 area models*
 - Size composition data
 - Tags (down-weighted to be essentially excluded in some OM scenarios, *including all 2 area models*)
 - Recruitment penalties on deviations from stock recruit relationship and mean spatial distribution
 - Diffuse priors on all estimated parameters
- Estimated parameters:
 - Fishery selectivity (various functional forms, parameters shared among some fleets)
 - Longline catchability (in aggregate - regional scaling factors are used to scale relative density to relative abundance among regions)
 - Virgin recruitment

- Recruitment deviations from the Beverton-Holt stock-recruit relationship, recruitment spatial partitioning among tropical regions (1 and 4) and deviations from the mean spatial distribution.
- Juvenile and adult movement rates
- Initial fishing mortality
- Modifications to the reference case assessment for the OM included:
 - Removing the movement-environment link
 - Constraining 12 quarters of recruitment deviations to the stock-recruit function (instead of 8)
 - Adding additional age-dependent error to the terminal SS age structure before the MSE projections begin
 - Relaxing some parameter bounds (i.e. so that an arbitrary hard bound does not constrain parameter estimates)

OM Reference Set Grid

- Model structural and parameter uncertainty would be introduced to the OM through the alternative assumptions listed in Table 2 (4 area grid only). Only the point estimates (maximum posterior density) of parameters and initial states from each model specification are retained for the OM.
- A fractional-factorial experimental design of 50-150 models is used to reduce the dimensionality of the full factorial cross. In an experimental design context, all main effects should be estimable at a minimum.
- In recognition that the IOTC yellowfin assessment model parameter estimates can be sensitive to initial starting conditions, minimization was repeated from randomly jittered starting conditions until either (i) successful minimization was achieved 3 times (maximum gradient of the objective function with respect to the estimated parameters <0.01) or (ii) 10 attempts were made without reaching 3 successful minimizations.
- Projection assumptions are defined in Table 3.

References

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

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

Urtizbera, A, Cardinale, M, Winker, H, Methot, R, Fu, D, Kitakado T, Fernández, C, Merino, G. 2020. Towards providing scientific advice for Indian Ocean yellowfin in 2020. IOTC-2020-WPTT22(AS)-21.

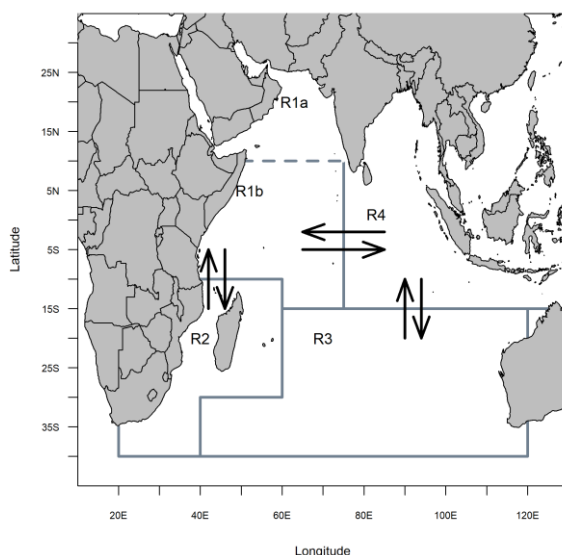


Figure 1. Spatial structure for the yellowfin tuna OM (figure from Fu et al. 2018).

Table 1. IOTC Yellowfin assessment fishery definitions.

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

Table 2. Candidate configurations and assumptions for the reference set OM as defined by WPTT (2020, Table 4)

Definition
<u>Spatial Structure – 4 regions</u> (2 region option to be further investigated for potential inclusion)
<u>Stock-recruit function (h = steepness)</u> Beverton-Holt, $h = 0.7, 0.8, 0.9$
<u>Natural mortality (multiplier relative to reference case M vector M10)</u> 1.0, 0.8, 0.6
<u>Tag recapture data weighting (tag composition and negative binomial)</u> $\lambda = 0.001, \lambda = 0.1, \lambda = 1.0$ if 2 area model is added, tag $\lambda = 0$, and 4 Area $\lambda = 0.001$ will be dropped
<u>Growth curve</u> Fonteneau (2008) Dortel et al. (2014) model 3 (as approximated in <u>2020 YFT Management advice paper</u>)
<u>Assumed longline CPUE catchability trend (compounded)</u> 0% per annum 1% per annum
<u>Tropical longline CPUE standardization method</u> Hooks Between Floats only
<u>Longline CPUE error assumption (quarterly observations)</u> $\sigma_{CPUE} = 0.1, 0.3$
<u>Tag mixing period</u> 4 quarters 8 quarters

Table 3. OM Projection assumptions in the yellowfin reference set and robustness sets. Reference set values not listed are identical to the model-specific conditioning assumptions/estimates. Robustness case values not listed are identical to the reference set except as noted.

OM	Projection assumption	Value
(TBD)	Reference set OM	
	Initial population error CV (a = age in quarters)	$0.6\exp(-0.1a)$
	Recruitment deviation penalty	$\max(\sigma_R = 0.42, \text{SS estimate})$
	Recruitment deviation lag(1) auto-correlation (these are annual values, but they are parameterized by the quarterly quarterly equivalents)	$\max(\rho_R = 0.21, \text{SS estimate})$
	CPUE observation error	$\max(\sigma_I = 0.2, \text{SS estimate})$
	CPUE observation error lag(1) auto-correlation (implemented annually)	$\max(\rho_I = 0.5, \text{SS estimate})$
	Multinomial Catch-at-length sample size (all fisheries, but not used in MPs to date)	100
	Selectivity stationary for all fisheries	
	Quota Implementation error	CV = 0
	First MP quota year	2022
	Bridging catches 2018, 2019-2021 (1000 t)	440, 417
	MP data lag (i.e. data from 2018 informs 2021 quota)	2 years
	Quota allocation (average observed over period)	2016-2017
	Baranov Catch Equation	
	Seasonal F averaged over SS period	2016-2017
	Effort Ceiling (relative to seasonal F above)	20
	Robustness tests (other features as reference set)	
	1) Increased Longline CPUE error variance $\sigma_I = 0.3, \rho_I = 0.5$	
	2) 10% overcatch, accurately reported	
	3) 10% overcatch, unreported	
	4) 10% overcatch, 5% reported, 5% not reported	
	5) 8 consecutive quarter recruitment shock (55% of average, near start of projections)	
	6) 3% per year LL catchability trend (not in SS conditioning; projections only)	